IDENTIFYING REGULATED GENES THROUGH THE CORRELATION STRUCTURE OF TIME DEPENDENT MICROARRAY DATA

A Thesis
Submitted to the Faculty
of
Purdue University
by
Martina Muehlbach Bremer

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

December 2006

Purdue University
West Lafayette, Indiana
To my husband Timo, for his strength and support,
and to my parents, who taught their daughters to think
with the heart as well as the head.
ACKNOWLEDGMENTS

I sincerely thank my advisor, Rebecca W. Doerge, for her support, advice, and friendship over the last four years. Her grace, generosity, and dedication to her profession are an inspiration for her students. I also thank my advisory committee members Professors Mary Ellen Bock, Jayanta Ghosh, and Min Zhang for their insights and helpful comments.

The Department of Statistics at Purdue University, headed by Professor Mary Ellen Bock, provides an environment in which teaching and learning are valued and encouraged. The atmosphere of respect and friendship among students and professors has made the time here rewarding as well as enjoyable. Productive discussions with fellow students were much appreciated, especially with the former and present members of the Doerge research group: Lingling An, Riyang Chen, Hongmei Jiang, Kyunga Kim, Alex Lipka, Gayla Olbricht, John Stevens, Suk-Young Yoo, and Lianbo Yu. Without the computational resources provided by Doug Crabill and My Truong the simulations conducted for this dissertation would not have been possible.

For the past year I consider myself fortunate to have collaborated with a truly special group of scientists working on a National Science Foundation Plant Genome project that studies gene expression in polyploids (NSF grant 0501712-DBI). My gratitude goes especially to Ed Himelblau and Andreas Madlung who have shown me that dedication to teaching is compatible with continued involvement in research.

I thank my husband Timo for his unwaivering love and support. His knowledge and patience for computer related questions were invaluable. The curiosity for science I owe to my parents. You have taught me the pleasure of learning and have encouraged and challenged me to pursue my dreams. Thank you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>SYMBOLS</td>
<td>xii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Genetics</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Measuring Gene Expression</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Gene Regulation</td>
<td>6</td>
</tr>
<tr>
<td>2 Microarray Technology</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Fabrication of Arrays</td>
<td>8</td>
</tr>
<tr>
<td>2.1.1 Spotted cDNA Microarrays</td>
<td>9</td>
</tr>
<tr>
<td>2.1.2 Spotted Oligonucleotide Microarrays</td>
<td>10</td>
</tr>
<tr>
<td>2.1.3 Affymetrix® Technology</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Normalization of Microarray Data</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Protein Arrays</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Summary</td>
<td>19</td>
</tr>
<tr>
<td>3 Time Series Analysis</td>
<td>20</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>20</td>
</tr>
<tr>
<td>3.2 ARMA Processes</td>
<td>21</td>
</tr>
<tr>
<td>3.3 Multivariate Time Series Analysis</td>
<td>23</td>
</tr>
<tr>
<td>3.3.1 Multivariate ARMA Processes</td>
<td>25</td>
</tr>
<tr>
<td>3.4 State Space Models</td>
<td>26</td>
</tr>
<tr>
<td>3.5 The Kalman Recursions</td>
<td>29</td>
</tr>
<tr>
<td>3.6 Application to Biology</td>
<td>32</td>
</tr>
<tr>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>The Current State of Microarray Time Series Data Analysis</td>
</tr>
<tr>
<td>4.1</td>
<td>Mathematical Background</td>
</tr>
<tr>
<td>4.2</td>
<td>Time Series Microarray Applications</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Dynamic Bayesian Networks</td>
</tr>
<tr>
<td>4.3</td>
<td>Motivation of the KM-Algorithm</td>
</tr>
<tr>
<td>5</td>
<td>The KM-Algorithm</td>
</tr>
<tr>
<td>5.1</td>
<td>Overview</td>
</tr>
<tr>
<td>5.2</td>
<td>Initial Values</td>
</tr>
<tr>
<td>5.3</td>
<td>The KM-Algorithm Iterations</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Kalman Smoothing Step</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Model Fitting Step</td>
</tr>
<tr>
<td>5.4</td>
<td>Cholesky Decomposition</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Gradient Ascent</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Partial Update</td>
</tr>
<tr>
<td>5.5</td>
<td>Termination Criterion</td>
</tr>
<tr>
<td>5.6</td>
<td>Symmetries in the Likelihood Function</td>
</tr>
<tr>
<td>5.7</td>
<td>Regulation Criterion</td>
</tr>
<tr>
<td>5.8</td>
<td>Numerical Computation of the KM-Algorithm</td>
</tr>
<tr>
<td>5.9</td>
<td>Unequally Spaced Observations</td>
</tr>
<tr>
<td>5.10</td>
<td>Missing Data</td>
</tr>
<tr>
<td>5.11</td>
<td>Model Selection</td>
</tr>
<tr>
<td>6</td>
<td>Simulated Data and KM-Algorithm Results</td>
</tr>
<tr>
<td>6.1</td>
<td>Data Sets used in the Simulation</td>
</tr>
<tr>
<td>6.2</td>
<td>Model Likelihood</td>
</tr>
<tr>
<td>6.3</td>
<td>Regulation Criterion</td>
</tr>
<tr>
<td>6.4</td>
<td>Goodness of Ranking</td>
</tr>
<tr>
<td>6.5</td>
<td>Power of Regulation Criterion</td>
</tr>
<tr>
<td>6.6</td>
<td>Gene Variance and Regulation</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>6.7</td>
<td>Model Selection</td>
</tr>
<tr>
<td>6.7.1</td>
<td>Stability of Regulation Ranking with Respect to Model Selection</td>
</tr>
<tr>
<td>6.7.2</td>
<td>Stability of GR with Respect to Model Selection</td>
</tr>
<tr>
<td>6.7.3</td>
<td>Stability of GR with Respect to Initial Value Choice</td>
</tr>
<tr>
<td>6.8</td>
<td>Run-Times</td>
</tr>
<tr>
<td>6.9</td>
<td>Partitioning of Large Data Sets</td>
</tr>
<tr>
<td>7</td>
<td>Application to Real Microarray Data</td>
</tr>
<tr>
<td>7.1</td>
<td>Caulobacter Data Set</td>
</tr>
<tr>
<td>7.2</td>
<td>Synechocystis Data Set</td>
</tr>
<tr>
<td>7.3</td>
<td>Spellman CDC15 Yeast Data Set</td>
</tr>
<tr>
<td>8</td>
<td>Summary and Future Work</td>
</tr>
<tr>
<td>8.1</td>
<td>Future Directions</td>
</tr>
<tr>
<td>8.1.1</td>
<td>Regulator Profiles</td>
</tr>
<tr>
<td>8.1.2</td>
<td>Partitioning Method</td>
</tr>
<tr>
<td>8.1.3</td>
<td>Model Selection Method</td>
</tr>
<tr>
<td>8.1.4</td>
<td>Biological Validation</td>
</tr>
<tr>
<td>8.2</td>
<td>Conclusions</td>
</tr>
</tbody>
</table>

LIST OF REFERENCES | 117 |

VITA | 125 |
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Selected model dimensions for nine simulated data sets using different maximum relevant biological time lags $p$. Top $p = 1$, middle $p = 2$, bottom $p = 3$. The true model dimension in each case is $m = 10$.</td>
</tr>
<tr>
<td>6.2</td>
<td>Run-times for the KM-algorithm with a termination criterion value of $c = 0.0005$.</td>
</tr>
<tr>
<td>6.3</td>
<td>Comparison of regulation results for three data sets with $n = 2000$ genes and $T = 20, 40$, and $100$ observations, respectively. The regulation results obtained by applying the KM-algorithm to the large data set as a whole are compared with the results from partitioning the data set into smaller parts and applying the KM-algorithm to the smaller parts separately.</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The structure of a DNA molecule. Image courtesy: National Human Genome Research Institute.</td>
</tr>
<tr>
<td>1.2</td>
<td>Central dogma of molecular genetics</td>
</tr>
<tr>
<td>2.1</td>
<td>Laser scanned image of a microarray slide with one block enlarged. Image courtesy of <a href="http://www.wikipedia.org">http://www.wikipedia.org</a>.</td>
</tr>
<tr>
<td>2.2</td>
<td>Probe set representing a gene: Here eight short nucleotide sequences of different length represent the gene together as a set.</td>
</tr>
<tr>
<td>2.3</td>
<td>Spotted oligo microarray</td>
</tr>
<tr>
<td>2.4</td>
<td>Probe set in Affymetrix technology</td>
</tr>
<tr>
<td>2.5</td>
<td>Affymetrix microarray image</td>
</tr>
<tr>
<td>2.6</td>
<td>Schematic of a dye swap experiment</td>
</tr>
<tr>
<td>4.1</td>
<td>Temporal profiles of three genes in a microarray time series experiment.</td>
</tr>
<tr>
<td>4.2</td>
<td>Determination of differential expression for two temporal log-ratio gene expression profiles. (a) The observed genes are assumed to have the same function as their temporal profiles are similar. (b) The temporal profiles are not similar and the genes are declared differentially expressed over time.</td>
</tr>
<tr>
<td>4.3</td>
<td>Clustering of similar temporal gene expression profiles.</td>
</tr>
<tr>
<td>4.4</td>
<td>(a) Smoothing of averaged gene expression profiles. (b) The red cluster is assumed to activate the blue.</td>
</tr>
<tr>
<td>4.5</td>
<td>Bayesian network: The circles represent observable genes and unobservable regulators and the directed arcs describe regulation between them.</td>
</tr>
<tr>
<td>5.1</td>
<td>Flowchart of the proposed KM-algorithm. Here, $Z_t$ are the $n$ dimensional observations and $Y_t$ are the $m$ dimensional unobservable regulators. $\mathbf{F}, \mathbf{G}, \mathbf{\Sigma}_\delta, \mathbf{\Sigma}_e$ are the parameters of the state space model in Equations 3.1 and 3.2.</td>
</tr>
<tr>
<td>5.2</td>
<td>$\Delta t$ is the highest common factor between measured time points.</td>
</tr>
</tbody>
</table>
6.1 Regulators from Example 6.0.1 in a simulation for $T = 40$ time points. (a) Regulators $Y_1$ and $Y_8$ with different autocorrelation; (b) Regulators $Y_2$ and $Y_9$ with different variance. ............................. 73

6.2 Increase in the model log-likelihood for different data sets: (a) Real microarray data sets. The *Caulobacter crescentus* data set from Laub et al. [33] as well as the *Synechocystis* data set from Gill et al. [34]. (b) Simulated data with $n=500$ genes. ................................. 76

6.3 Increase in the model log-likelihood for simulated data with $n=1000$ genes. ......................................................... 77

6.4 Observations on $n = 1000$ genes at $T = 100$ time points are simulated according to the model in Example 6.0.1. The KM-algorithm is implemented five times with different initial values. This plot shows the averaged sum of squares of the row entries in the gene regulation matrices $G$ against the temporal variance of each gene. The simulated regulated genes are plotted as red stars and the unregulated genes as black dots. ............................................................. 79

6.5 Observations on $n = 1000$ genes at $T = 100$ time points are simulated according to the model in Example 6.0.1. The KM-algorithm is implemented five times. The averaged proposed regulation criterion is plotted against the temporal variance of each gene. The simulated regulated genes are plotted as red stars and the unregulated genes as black dots. ............................................................. 80

6.6 Principle behind the proposed goodness of ranking measure to evaluate the performance of the regulation ranking. ................................. 81

6.7 Goodness of ranking obtained from averaging $k$ independent regulation criterion results to assess the power of the KM-algorithm. Implementing the KM-algorithm more often and averaging the results provides more consistent and improved ranking results. ................................. 82

6.8 For nine simulated data sets of different size $n(500, 1000, 2000)$ and length $T(20, 40, 100)$, the KM-algorithm is implemented five times each. The regulation results are averaged and the combined regulation criterion value for each gene is plotted on the $y$-axis. On the $x$-axis, the maximum absolute expression value of the gene over the time course of observations is plotted. Regulated genes are shown as red stars and unregulated genes are shown as black dots. ............................. 85
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.9 Magnitude of singular values for nine data sets of different size $n$ and length $T$ for lag-values $p = 1$ (top), $p = 2$ (bottom). The solid line represents 80% of the largest singular value. The dotted lines indicate the selected model dimension for each data set.</td>
<td>88</td>
</tr>
<tr>
<td>6.10 Magnitude of singular values for nine data sets of different size $n$ and length $T$ for lag-values $p = 3$. The solid line represents 80% of the largest singular value. The dotted lines indicate the selected model dimension for each data set.</td>
<td>89</td>
</tr>
<tr>
<td>6.11 Quantile plots comparing the ranking results obtained with the true state space dimension to the results obtained from implementing the KM-algorithm with the state space dimension obtained through model selection. Top: Comparing selected state space dimension $m = 4$ to true state space dimension $m = 10$ for $n = 500$ genes simulated at $T = 100$ time points. Bottom: Comparing selected state space dimension $m = 9$ to true state space dimension $m = 10$ for $n = 1000$ genes simulated at $T = 100$ time points.</td>
<td>91</td>
</tr>
<tr>
<td>6.12 Quantile plots comparing the ranking results obtained with the true state space dimension to the results obtained from implementing the KM-algorithm with the state space dimension obtained through model selection. Comparing selected state space dimension $m = 18$ to true state space dimension $m = 10$ for $n = 2000$ genes simulated at $T = 100$ time points.</td>
<td>92</td>
</tr>
<tr>
<td>6.13 Goodness of ranking (GR) obtained from running the KM-algorithm five times at state space dimension $m$ and averaging the results. Shown here are GR results for the true dimension ($m = 10$) and the dimensions obtained by model selection with maximum relevant biological time lag $p = 1, 2, 3$.</td>
<td>94</td>
</tr>
<tr>
<td>6.14 Goodness of ranking (GR) obtained from averaging the regulation criteria resulting from five implementations of the KM-algorithm. The same five initial gene regulation matrices $G$ are used on a data set with $T = 100$ observations on $n = 500$ genes in which the order of genes is permuted randomly one hundred times.</td>
<td>95</td>
</tr>
<tr>
<td>6.15 A large microarray data set is partitioned into smaller sets which are analyzed in parallel using the KM-algorithm. Results are combined to yield one final regulation estimate for each gene in the data set.</td>
<td>97</td>
</tr>
<tr>
<td>7.1 Magnitude of singular values for the Caulobacter data set obtained by singular value decomposition of the Hankel matrix $H$ for different maximum relevant biological time-lags $p = 1, 2, 3$.</td>
<td>101</td>
</tr>
</tbody>
</table>
7.2 For the genes in the Caulobacter data set [33], the regulation criterion value is plotted against the maximum absolute log-fold expression value over the time course of observations. Genes that had been determined to be cell-cycle regulated previous to the analysis of Laub et al. are shown in green. Genes that were found to be regulated by Laub et al. are shown in red. .......................................................... 102

7.3 For the genes in the Caulobacter data set [33], the regulation criterion value is plotted against the temporal variance of the gene. Genes that had been determined to be cell-cycle regulated previous to the analysis of Laub et al. are shown in green. Genes that were found to be regulated by Laub et al. are shown in red. .......................................................... 103

7.4 Magnitude of singular values of the Bacteria data set obtained by singular value decomposition of the Hankel matrix $H$ for maximum relevant biological time-lags $p = 1, 2$. .......................................................... 105

7.5 For each gene in the Bacteria *Synechocystis* data set the regulation criterion is plotted against the maximum absolute expression value. Genes classified into response classes by Gill et al. [34] are shown in green, red, blue, and orange, respectively. Genes that were not found to be involved in the light-dark reaction of the cell by Gill et al. are plotted in black. .......................................................... 107

7.6 Regulation criterion value obtained by the KM-algorithm plotted against the temporal variance for each ORF representing a gene in Spellman’s CDC15 yeast data set. The genes found to be regulated in a cell cycle-dependent manner by Spellman are plotted in red, with the genes found to be directly involved in cell cycle processes plotted in green. . 110

7.7 Regulation criterion value obtained by the KM-algorithm plotted against the maximum absolute expression value for each ORF representing a gene in Spellman’s CDC15 yeast data set. The genes found to be regulated in a cell cycle-dependent manner by Spellman are plotted in red, with the genes found to be directly involved in cell cycle processes plotted in green. . . 111
SYMBOLS

DBN dynamic Bayesian network
DNA deoxyribonucleic acid
cDNA complementary DNA
RNA ribonucleic acid
mRNA messenger RNA
rRNA ribosomal RNA
tRNA transfer RNA
EST expressed sequence tag
PCR polymerase chain reaction
MA\( (q) \) moving average process of order \( q \)
AR\( (p) \) autoregressive process of order \( p \)
ARMA\( (p,q) \) autoregressive moving average process
GR goodness of ranking
T length of observed time series
n number of observations at each time point
m number of state variables
\( Y_t^{t-1} \) Kalman prediction estimate
\( Y_t^t \) Kalman filtering estimate
\( \hat{Y}_t \) Kalman smoothing estimate
ABSTRACT

Muehlbach Bremer, Martina. Ph.D., Purdue University, December, 2006. Identifying Regulated Genes through the Correlation Structure of Time Dependent Microarray Data. Major Professor: Rebecca W. Doerge.

Since microarray technology has become widely available, it is possible to study the transcription of thousands of genes simultaneously. Experiments can be conducted in which measurements of transcription levels on the same set of genes are taken repeatedly over time. Often these time course gene transcription experiments aim to understand the behavior of genes in a certain process, such as the cell cycle, or the organism’s reaction to injury or disease. The transcription levels of genes are influenced by many factors: genes may be regulated by other genes, as well as by enzyme or protein levels in the cell, or by processes such as DNA methylation. Understanding and describing an organism’s entire gene regulatory network is an ambitious goal that is considered here in the context of time dependent microarray data.

A method is proposed that uses a state space model to represent a gene regulatory network. An algorithm is developed that estimates the optimal model parameters, as well as the behavior of hidden regulators. Based on the model parameter estimates, a criterion is proposed that describes the degree of regulation of every observed gene. Biological assumptions are incorporated to place restrictions on the model parameters, while mathematical restrictions assure statistical validity of the model. The power of the proposed method to identify regulated genes in time dependent microarray data is investigated via simulations and the algorithm is applied to several real microarray time series data sets. Recommendations are made for a minimum number of time point observations that a microarray experiment should include in
order to achieve a desired degree of statistical separation between regulated and unregulated genes.
1. INTRODUCTION

Research in the areas of genetics and molecular biology has gained increased importance. Every organism’s hereditary information is contained in its genome. A detailed analysis of an individual’s genetic material leads to a better understanding of certain phenotypic features such as disease. Since the discovery of the molecular structure of deoxyribonucleic acid (DNA) by Watson and Crick in 1953 [1], great progress has been made in discovering how information is stored in the molecule, how this information is extracted and translated into proteins, and how certain changes in the stored information relate to abnormalities.

The introduction of microarray technology in the mid 1990’s [2–5], allowed scientists to study an organism’s genome from a new perspective. It is now possible to assess the expression levels of all the genes of an organism, often many thousands, simultaneously, over different time points and under different treatment conditions. Some experiments aim to understand the behavior of genes in a certain biological process, such as the cell cycle or the organism’s reaction to injury or disease. Genes control many important cellular functions such as protein production and cell metabolism. In turn, the gene expression itself is also controlled by various factors. External inputs (such as mutation) may lead to changes in gene expression, and cell internal factors such as high protein levels or DNA methylation are able to regulate gene expression, as well. Overall, the expression levels of an organism’s genes can be modeled as a network with numerous components that influence each other to different degrees [6]. A better understanding of the complicated gene regulatory network may enable researchers to circumvent limitations posed by genes linked to specific diseases and may lead to improved treatment options.

The vast quantities of data produced by microarray technology demand the application of appropriate statistical techniques for their interpretation. In this disser-
tation a method is proposed which develops a statistical model for a gene regulatory network. While it is still important to consider individual genes and their expression levels in different treatment groups such as healthy and diseased, it is now also possible to study correlated information provided by the joint behavior of genes over different treatment conditions or over time. Traditionally, time course gene expression data have been analyzed separately for each gene. This approach, however, ignores the additional information that is provided by the joint behavior of the observed genes. To incorporate this correlation information into the analysis provides new challenges that will be addressed in this work. The statistical correlation of gene transcription at different time points is used to deduce important information about the organism’s genetic network. Specifically, a criterion will be formulated that allows the identification of regulated genes. Regulation in this context refers to gene expression that is influenced by a set of regulators that may or may not be directly observed. In order to better understand both the biology and the technology, a short overview is presented.

1.1 Genetics

DNA which is contained in the nucleus of every cell, stores genetic information. The DNA molecules are long polymers of nucleotides stored in bundles of different length called chromosomes. Each DNA molecule has a double-stranded structure that organizes itself in the form of the well known double-helix [1]. A DNA strand consists of a chain of nucleotides made up from a sugar phosphate backbone and a base. The four bases occurring in DNA are the purines adenine (A) and guanine (G) and the pyrimidines thymine (T) and cytosine (C) (Figure 1.1). Adenine and thymine are complementary and form a double hydrogen bond. Guanine and cytosine are complementary as well but form a triple hydrogen bond. The strands are anti-parallel, in that the “head” (3’) end of one strand attaches itself to the “tail” (5’) end of the other. The key to microarray technology is the affinity of DNA strands
Figure 1.1. The structure of a DNA molecule. Image courtesy: National Human Genome Research Institute.
to attach themselves to their base pair counterparts. When DNA replicates prior to mitosis and meiosis, the double strand is partially unwound by the enzyme DNA helicase. This allows another enzyme, DNA polymerase, to bind to one of the separated strands. Moving from the 5’ to the 3’ end, the polymerase enzyme attaches nucleotides that are metabolized in the cytoplasm of the cell in a sequence that is complementary to the nucleotide sequence in the original parent strand. This process is known as DNA synthesis. The resulting molecules whose base-pair sequence is complementary to that of the parent strand are referred to as complementary DNA (cDNA). The bonds between the base pairs can be broken (and the DNA strands de-naturated), when the molecule is gently heated. Combining natural DNA synthesis with artificial breaking of the strands allows the amplification of genetic material in the laboratory. With respect to microarray technology, many copies of the strands of DNA found in an organisms cell are needed.

DNA regulates many important cellular functions. One of the most important functions of DNA is to produce proteins. A protein consists of one or more polypeptides which are themselves chains of amino acids. The twenty naturally occurring amino acids are encoded in codons, or triplets of nucleotides in the DNA strand [7]. Protein production is a two step process described in the “Central Dogma” of molecular genetics (Figure 1.2) [8]. DNA is transcribed into ribonucleic acid (RNA) which is then translated by ribosomes into protein. The structure of RNA molecules is similar to that of DNA, but the sugar molecule is ribose rather than deoxyribose and the base uracil (U) replaces the base thymine. Compared to the DNA molecules the RNA molecules are relatively short chains that carry information about particular cellular functions. During the process of converting the information contained in the DNA

![Diagram](image)

**Figure 1.2. Central dogma of molecular genetics**
into protein several types of RNA molecules play different roles. In the transcription step the DNA molecule is split ("unzipped") by an enzyme called RNA polymerase. A complementary single strand of RNA called messenger RNA (mRNA) is formed. Messenger RNA transports the information contained in the DNA molecule out of the cell’s nucleus. There, the chain is read by ribosomes consisting of ribosomal RNA (rRNA). The mRNA molecules consist of a leader sequence, a start codon that initiates translation, a reading frame whose codons will be translated into amino acids, a stop codon that interrupts translation, and a trailer sequence. Transfer RNA (tRNA) molecules incorporate the encoded amino acid into a growing protein after a particular triplet of nucleotides in the reading frame of the mRNA has been read. Once a protein is fully assembled as indicated by the stop codon, the mRNA detaches from the ribosome and remains in the cell until it degrades over time. If acquired, the mRNA content of a cell represents the degree of transcription of the organism’s genes at the time of extraction. Thus, mRNA is the main source of genetic material used to measure gene expression with microarray technology.

1.2 Measuring Gene Expression

Measuring a gene’s expression levels through the amount of mRNA present in a cell determines to what degree a particular gene or chromosomal region is translated into protein. The abundance of a particular type of mRNA is quantified by allowing the mRNA to bind to its base pair sequence counterpart on a prepared microarray slide. These counterparts can be created artificially if the gene’s base pair sequence is known [4], or they can be derived from the tissue sample itself [2]. The human genome, as well as many other organisms are fully sequenced and the sequence information is available in publicly accessible databases [9]. If the organism’s genome is not fully sequenced it is still possible to measure transcription with technology that allows the experimenter to use selected pieces of the organisms DNA directly to quantify the amount of mRNA. The different technologies that can be used to
measure the abundance of a particular mRNA sequence in a cell are described in Chapter 2.

1.3 Gene Regulation

As described previously, the expression level of a gene is determined through the abundance of corresponding mRNA. Since the mRNA molecules are translated into proteins, the gene expression levels regulate the amount of protein that is being produced. In turn proteins regulate genes, as well. Proteins perform many important functions in the cell such as catalyzing metabolic reactions, maintaining structures or sensing signals. One important function is gene regulation. Certain repressor proteins are able to bind to the DNA strand and induce or transcribe genes [10]. In this way, cells are able to regulate their own protein production by intricate negative feedback mechanisms. If, for example, there is an overabundance of a certain amino acid, the amino acid starts binding to a repressor protein. The binding changes the structure of the protein which in turn will bind to the DNA strand and repress the translation of genes that encode the protein [11]. This process can be viewed as genes regulating themselves, and is often referred to as post-translational modification [12].

It is of great interest to the scientific community to better understand the gene regulation process. Understanding the connections between genes will provide insight into the cellular system. The regulation of proteins, some of which are directly linked to diseases such as Huntingtons or Parkinson [13], is of particular interest. Influencing the production of a certain protein on a cellular level might make it possible to develop treatments for many diseases. Besides being regulated by other genes and protein levels, gene transcription is also influenced by many other processes. DNA methylation [14], for instance, is now known to play a significant role in gene silencing (suppressing gene translation). DNA methylation is one example of epigenetics, a heritable change in DNA function that occurs without a change
in the DNA sequence. Even though factors such as protein abundance or methylation are not measured directly during microarray gene expression experiments, they may be responsible for changes in gene expression over time that are observed in the organism. For example, it is known that certain epigenetic alterations of gene function are connected to cancer [15]. If the promoter regions of genes that can repair DNA or repress tumors are abnormally methylated, then this may lead to a loss of function for these genes which in turn may lead to disease (cancer). It is important to understand the mechanism that leads to abnormal methylation, because it potentially allows the gene silencing process to be reversed. In gene expression microarray experiments which study the temporal development of a condition, it is therefore important to not only consider the gene expression levels themselves, but to also allow for other factors that may regulate gene expression, but that are not directly observed in the experiment.

Statistical models are needed that take potentially unobservable regulators and their influence on the gene expression levels into account. The model used in this work is based on the principles of time series analysis, and it allows the quantification of the relationship between hidden regulators and observed gene expression levels. Biological assumptions are incorporated to place restrictions on the model parameters while mathematical restrictions on the parameter space assure the statistical validity of the model.
2. MICROARRAY TECHNOLOGY

Several different microarray technologies are available to measure gene expression levels [2–5]. All of them rely on the fact that single stranded DNA will bind with its complementary sequence. Some production steps are common in the different microarray technologies: A glass or silicone slide is prepared and genetic material is affixed, in an array of spots, to the slide. A tissue sample is obtained from one or more individuals of interest and the RNA content of the sample is extracted and prepared. Complementary cDNA from the extracted RNA content of the cells is produced and amplified [16]. The cDNA in the sample is marked with fluorescent dye and allowed to bind to its partner strand on the array. The fluorescence level of a spot on the array is used as a measurement of transcript abundance for the corresponding gene.

2.1 Fabrication of Arrays

Microarray technologies differ in the way that the slides are produced and how the genetic material on a slide is obtained and applied. Some microarrays are produced commercially and are being sold through large companies such as Affymetrix® [17]. Others, that are intended for a specific experiment, are prepared, designed, and fabricated by the user. The most common technologies besides that used by Affymetrix® are spotted cDNA [2] and spotted oligonucleotide arrays [4].
2.1.1 Spotted cDNA Microarrays

Genes, or expressed sequence tags (ESTs) that represent the gene, can be isolated and amplified using polymerase chain reaction (PCR) [16]. The length of the resulting product varies depending on the gene, usually between 300 and 500 nucleotides. The PCR product is cleaned to remove PCR primers and unwanted proteins contained in the solution, and the final product is collected in a single well of a multi-welled plate (one well for each gene). Finally, a robot prints the material from the well plates onto a prepared small glass or silicone surface. Each spot on the slide will contain millions of cDNA copies from a certain gene or EST. The material on the slide is referred to as the target. One slide can accommodate thousands of target spots and can therefore potentially represent an organism’s entire genome [2].

To compare two groups of biological samples, such as healthy and diseased individuals, researchers typically obtain tissue material from the two groups and analyze the differences in gene expression using microarrays. One group is referred to as the treatment group, and the other group as the control. RNA is extracted from the cells in the samples and amplified using PCR. The PCR products are cleaned and RNA is converted into cDNA. The cDNA samples from the treatment and control groups are labeled with fluorescent dye by chemically attaching the dye molecules to the ends of the respective cDNA strands. Labeling the treatment sample with red (Cy5) and the control sample with green (Cy3) will allow later differentiation between the two samples [3]. The labeled cDNA strands are referred to as the probes in a microarray experiment. The samples are mixed and the resulting solution is hybridized onto the array. Specifically, the labeled probes in the sample are allowed to bind to their target counterparts on the slide. Unattached material is gently washed off and the slide is dried. A laser is used to excite the fluorescent dye attached to the probes. The observed intensity of red or green fluorescence in a spot on the array is used to measure how much RNA of that specific gene or EST is present in the treatment and
control group, respectively. When fluoresced, the spots that show an abundance in red represent more material from the treatment group than the control group (see Figure 2.1). In this case the corresponding gene is said to be over-expressed in the treatment group. Similarly, spots that appear green contain more material from the control group than the treatment group, and are referred to as under-expressed. If a spot is yellow, then the treatment and control group contain approximately equal amounts of this probe material (i.e., there is no difference between the groups). A spot that remains dark (i.e., not red, green, or yellow) can either be explained by a mistake in the hybridization process, or by the fact that neither the treatment nor the control group contain significant amounts of the corresponding type of RNA.

2.1.2 Spotted Oligonucleotide Microarrays

Spotted oligonucleotide arrays differ from cDNA arrays in that they do not use complete gene sequences as targets [4]. Instead, a number of short nucleotide sequences (usually 20-70 nucleotides long) are chosen to represent a certain gene or gene segment. It is possible to represent each gene by a probe set of nucleotide
sequences (Figure 2.2) on the array. The nucleotide sequences in a probe set will be spotted separately onto the array as targets. Their measurements will later be combined to yield one expression value for the gene. However, the nucleotide sequences in the probe set are not derived from cellular cDNA but are instead assembled artificially, one base pair at a time.

![Diagram](image)

**Figure 2.2.** Probe set representing a gene: Here eight short nucleotide sequences of different length represent the gene together as a set.

To manufacture the targets, slides are first coated with a silane film [18]. Linker molecules that block the attachment of new nucleotides to the chain are applied to the silane in the spots where the oligonucleotide chains will be assembled. The blocking property of these linker molecules can be altered by exposure to ultraviolet light [19]. Photolithographic masks with tiny windows corresponding to the location of the spots are used to direct the light to only those features that are to receive a

![Diagram](image)

**Figure 2.3.** Preparation of a spotted oligo microarray: blocking molecules are strategically exposed to ultraviolet light to assemble oligonucleotide sequences base by base.
certain nucleotide (Figure 2.3). Nucleotides containing one of the four bases (A, T, G, or C) and another blocker are allowed to attach themselves to the chain in the unblocked positions. A different mask is then used to remove the blocks in other positions and more base pairs are added. This process is repeated until the nucleotide chains are assembled. The microarray experiment itself can be conducted similar to a cDNA microarray experiment. RNA is extracted and prepared from either a single biological sample or from treatment and control samples. The samples are stained with fluorescent dye and hybridized onto the array. The material in the samples will attach to the oligonucleotide on the slide if the base pair sequences are complimentary. How much a particular gene is expressed is deduced from the fluorescence levels of the features that correspond to the probe set of the particular gene. Since the fluorescence levels will likely vary across a probe set it is especially important in these experiments to conduct a careful statistical analysis and apply meaningful normalization.

2.1.3 Affymetrix® Technology

The technology employed by Affymetrix® is a particular form of spotted oligonucleotide array [17]. Each gene whose expression is to be measured, is represented on the array by eleven target sequences which are each twenty-five bases long. These sequences are referred to by Affymetrix® as the “perfect match” (PM) sequences. Additionally, there are also eleven “mismatch” (MM) sequences which are identical to the perfect match sequences up to the 13th base, in which cytosine or guanine are exchanged for adenine or thymine. This system is designed to measure the extent of non-specific binding [20]. The targets on the slide are assembled one base pair at a time using the process described in Section 2.1.2. To conduct an experiment, RNA is extracted from the organism to be studied and the genetic material is amplified and labeled with the chemical biotin. The extracted RNA probe is then allowed to bind to the target material. After washing off excess material, the slide is
treated with a fluorescent stain that binds to the biotin on the RNA strands. Note, that Affymetrix® technology uses only a single dye color. Therefore, each sample treatment condition is hybridized onto its own array (Figure 2.5).

Figure 2.5. Scanned image obtained from an Affymetrix® microarray (image courtesy of Affymetrix® [17]).

2.2 Normalization of Microarray Data

The minute differences in gene expression that are to be detected in the huge amount of data collected demand a careful statistical analysis in order to yield reliable results [21, 22]. One important issue is normalization of the collected array data [23]. Both for the oligonucleotide and Affymetrix® microarray technologies the
fluorescence levels of the probes that represent a certain gene have to be statistically combined to obtain an estimate of the expression level of that gene. Furthermore, in many microarray experiments more than one slide is used to conduct the experiment.

There are two fundamentally different kinds of repetition in microarray experiments. Biological repetition is utilized if the same tissue samples from different individuals in the same treatment group (e.g., several cancer patients) are hybridized to identical microarrays. Since the gene expression levels in different individuals will vary slightly, this introduces (biological) variation in the measurement of every gene. Many experiments also produce technical replicates, where the same biological material is hybridized to repeated target spots of the same gene. The repeated spots may either be printed on the same slide, or several identical slides may be used to analyze a biological sample. Technical variation in the intensity readings is introduced, for example, through slight differences in the amount of sample material applied to a slide or non-uniformities stemming from the washing and drying procedure. To combine gene expression values obtained through intensity measurements of biological or technical repetitions, statistical normalization can be conducted. Different types of normalization may be carried out at different stages of the analysis and can be combined, if necessary. A few common normalization procedures are described below.

*Spot normalization for two-color microarrays:* As mentioned, a laser measures the red and green fluorescence intensity at several pixels for every feature on the slide. In addition, background measurements are taken between the features where no target material is supposed to be affixed. Based on the pixel-wise intensity readings, a “spot” consisting of pixels with high intensity compared to the background is defined [24]. For every spot on the array the pixel intensities are combined into two intensity readings, one for each dye color. Usually, this is done by computing the median of the pixel intensities that are interpreted as belonging to the spot and subtracting the median intensity of the background pixels [23]. This method yields one background corrected red ($R_i$) and green ($G_i$) intensity for every feature $i$ on a
two color microarray.

Affymetrix®: For the one-color arrays produced by Affymetrix®, typically eleven PM and MM measurements are available for each gene. The MM intensities are subtracted from the PM intensities to adjust for non-specific binding [20]. Normalization is conducted at the probe level and the normalized probe values are later combined to yield one expression value per gene per array [25]. If gene expression levels for two conditions such as normal and diseased are to be compared, then two slides have to be used. The expression levels of the same feature \( i \) from the two slides used for the treatment and control sample provide similar data as provided by the red and green feature measurements in a two-color array.

Global normalization: When conducting microarray experiments, most researchers are interested in the relative expression of treatment and control samples. This relative expression is most often reported as the log-fold-change of the two background corrected dye intensities for feature \( i \)

\[
M_i = \log_2 \frac{R_i}{G_i}.
\]

The base-2-logarithm assures that the values make intuitive sense: If the two dye intensities are the same \( M_i \) will be zero; if the red intensity is twice as large as the green, \( M_i \) will be 1; and if the red intensity is half as large as the green, \( M_i \) will be -1. Global normalization assures that the distribution of log-fold-change intensities is centered at zero [23]. For Affymetrix® arrays, the red and green intensities are replaced by the probe intensities obtained from the two slides used for the treatment and control samples.
Intensity dependent normalization: While the log-fold-change of the expression intensities can measure the relative expression of treatment and control samples, a measure for the absolute intensity of spot $i$ is given by

$$A_i = \frac{1}{2} \left( \log_2 R_i + \log_2 G_i \right).$$

Intensity dependent normalization fits a robust regression curve (loess [26]) to an $M$ versus $A$-plot. Ideally, an $MA$-plot should show a point cloud distributed around the $M = 0$ axis. If this is not the case, then the loess curve for the observed data is computed and the log-intensity ratios are adjusted such that the loess curve becomes the $M = 0$ axis [23, 25]. MA-plots are useful to identify non-linearity between the log-intensities and to study the relationship between differential expression and overall intensity.

Print tip group normalization: When microarrays are manufactured, the genetic material applied to the slide as targets is printed by a robot. The robot picks up material from the well plates and applies it to the slides. Usually, this is done simultaneously by several print tips arranged in a rectangular grid. The distance between the print tips corresponds to the distance between the wells in the plate. But on the much smaller microarray, the print tips will unload their genetic material in “blocks” (Figure 2.1). Specifically, the upper left print tip will print all the spots in the upper left corner of the microarray. If one of the tips is damaged such that it does not print the correct amount of material or the spots are consistently smeared, it is appropriate to conduct a print tip group normalization. For each block on the array that was printed by a different tip, either a global or intensity dependent normalization will be conducted separately [27].

Dye normalization: It is known that the two dyes Cy3 (green) and Cy5 (red) used to label the samples do not consistently bind to all cDNA sequences in the same manner [23]. To correct for this phenomenon, experiments are often conducted in the form of a “dye-swap”, comparing the gene expression values of a treatment and
control group using two microarray slides with alternate dye labeling (Figure 2.6). Combining the log-intensity ratios from both slides for each feature in the form \( \frac{1}{2} (M_i + M'_i) \) removes the dye bias.

![Figure 2.6. Schematic of a dye swap experiment: Treatment and control sample are labeled with red and green dyes and cross-hybridized onto two arrays.](image)

**ANOVA model:** A different approach to normalization is described by Kerr in [21]. Separately for each gene, the fluorescence log-intensity \( y \) for replication \( l \) in treatment group \( k \) on array \( i \) labeled with dye \( j \) is modeled as

\[
y_{ijkl} = \mu + A_i + D_j + T_k + \nu_{kl} + \epsilon_{ijkl}.
\]

Here, \( \mu \) represents the mean overall intensity, \( A_i \) is the array effect, \( D_j \) is the dye effect, and \( T_k \) is the treatment effect. \( \nu_{kl} \) represents the variation within treatment group \( k \) and \( \epsilon_{ijkl} \) contains the measurement error. This linear model allows convenient hypothesis testing for differential expression of a gene in different treatment groups, while simultaneously correcting for array and dye effects. The same idea can be extended, using a mixed linear model, to include all the genes spotted on the array in one global model:

\[
y_{ijkg} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{ig} + (TG)_{kg} + \epsilon_{ijkg}.
\]
In this case $G_g$ describes the global effect of gene $g$ and $AD, AG,$ and $TG$ describe interaction effects between array and dye, array and gene and treatment and gene, respectively.

### 2.3 Protein Arrays

Recently, a new technology has been developed that measures the protein content of a cell directly with a methodology similar to that of a cDNA microarray. Instead of the complementary sequence of DNA antibodies are spotted as targets onto what are referred to as protein microarrays [28]. The proteins contained in a prepared tissue sample are labeled with fluorescent dye and allowed to bind to the corresponding antibodies on the array. However, since proteins are complicated three-dimensional structures this binding process is not as uniform as it is for a DNA microarray. For example, the orientation in which the antibody is affixed to the slide plays an important role in the binding properties between protein and antibody. Similar to microarrays, protein arrays can be used to measure the protein levels in a tissue sample at different time points and under different treatment conditions [29].

As previously described, proteins play an important role in the regulation of gene expression. Measuring gene expression through the amount that a particular gene is transcribed into mRNA indirectly also measures the extent of protein production. However, there is not always a one to one correspondence between the amount of mRNA present in a cell and the amount of protein that is being produced. Post-translational modifications, such as the addition or removal of an acetate or phosphate group [30] make it difficult to measure protein abundance through cDNA microarrays alone. The method proposed in this dissertation takes hidden regulators of gene expression into account and uses a statistical model that allows for their estimation. It would be interesting to compare the estimated values of the hidden regulators, especially the shape of their temporal profiles, with measured protein levels in the cell using protein microarrays. If correspondences were observed, it
would be reasonable to assume that the observed protein is, in fact, a regulator of gene expression. Unfortunately, to date no parallel experiments have been conducted that measure gene expression and protein levels for the same biological sample over a time course of observations. However, as both gene and protein array technology improves, it is anticipated that these sorts of experiments will arise to meet the methodology that is described within this dissertation.

2.4 Summary

The large amounts of data that are now being produced in molecular biology by cDNA, oligonucleotide and protein microarray experiments demand novel statistical techniques for their analysis. Microarray experiments are usually noisy, and statistical models need to be able to partition both the biological as well as the technical variation contained in the measurements. Recently, experiments using cDNA or oligonucleotide microarrays have been conducted to study an organism’s change in gene expression over a period of time. Examples are the well studied yeast cell cycle [31,32], the bacterial cell cycle [33], bacterial light-dark transition [34], life cycle of drosophila (fruit fly) [35], and embryonal development of the rat central nervous system [36]. In these experiments, tissue samples are extracted from the organism or organisms of interest at different time points. Usually, the tissue samples are compared to a reference sample extracted at the zero time point and the log-fold-changes in gene expression are measured with microarray technology. After normalization, one log-fold-change value per gene is reported at every time point. The purpose of these experiments is to study the gene regulatory network during the observed process. Statistical methods of time series analysis [37–41] may be employed to express the complicated correlation structure of the observations.
3. TIME SERIES ANALYSIS

Time series analysis is the interpretation of repeated observations made on the same process at different time points. The correlation of the observations that arises from their sequential nature requires special statistical analysis techniques. Time series occur in many different fields such as economics, quality control, geophysics, medicine, and biology. Many methods have been developed for the analysis of time series data which have been described by a large number of authors [37-41]. In this chapter an overview of the most important concepts of time series analysis is presented. The discussion outlined here follows to a large extent that of Shumway and Stoffer in [37]. Specific methods applicable to the analysis of microarray time series data are presented in more detail.

3.1 Introduction

A time series is a set of observations \( z_t \), recorded sequentially in time \( t \) (\( t = t_1, t_2, \ldots \)). Here, only discrete time series will be considered with observations taken at specific time points. Most often the time points are equally spaced, but the theory can be extended to non-uniform sampling. The goal of time series analysis is to understand the nature of a phenomenon that is observed over time and to make predictions for future states of the system. Toward this end a model is fitted that takes the internal correlation structure of the data into account and then estimates for the model parameters are found using information provided by the observed data.

The simplest time series are those that have no internal or correlation structure. In this case observations at different time points are uncorrelated and identically distributed. The most commonly used distribution for this purpose is the normal distribution with mean zero.
Definition 3.1.1 White Noise
The time series \(\{a_t, t = t_1, t_2, \ldots\}\) with
\[ a_t \sim \text{Normal}(0, \sigma^2), \quad t = t_1, t_2, \ldots \]
independent, is called a white noise process (see Example 1.7 in [37]).

Definition 3.1.2 Stationarity
A time series \(\{z_t, t = t_1, t_2, \ldots\}\) is said to be stationary, if the distribution of the observations \(F_{z_t}(z) = P(z_t \leq z)\) does not depend on \(t\). In this case the observations have a common mean \(\mu = E(z_t)\) (see Section 1.5 in [37]).

For stationary time series that have an internal structure, a quantity of special interest is the autocorrelation function \(\gamma\) that describes the correlation of two observations \((z_t, z_{t+k})\) at different time points,
\[ \gamma_k = Cov(z_t, z_{t+k}) = E[(z_t - \mu)(z_{t+k} - \mu)]. \]

Note, that this quantity does not depend on \(t\) since the time series is assumed to be stationary.

3.2 ARMA Processes

When only a finite number of observations are available for the time series, it is often desirable to fit a finite-order parametric model to the observations. Assume for now that the observations are equally spaced in time. The time points are denoted by \(t = 1, 2, \ldots, T\). Two of the most useful representations of time series are the autoregressive (AR) and moving average (MA) processes (see Chapter 2 in [37]).

Definition 3.2.1 Autoregressive Process
A process \(\{z_t, t = 1, 2, \ldots\}\) is called autoregressive of order \(p\), denoted by AR(\(p\)), if the observations \(z_t\) are a weighted sum of the \(p\) previous observations plus an independent error term, i.e.,
\[ z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + a_t, \]
where \( \{a_t\} \) is a white noise process with constant variance \( \sigma_a^2 \). The model parameters in this case are \( \phi_1, \ldots, \phi_p \) and \( \sigma_a^2 \) (see Section 2.2 in [37]).

**Definition 3.2.2** Moving Average Process

The process \( \{z_t, t = 1, 2, \ldots\} \) is called a *moving average process of order* \( q \), denoted by MA\((q)\), if it can be modeled as a linear combination of a sequence of white noise, i.e.,

\[
z_t = a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q},
\]

where \( \{a_t\} \) is a white noise process with constant variance \( \sigma_a^2 \). In this case the model parameters are \( \theta_1, \ldots, \theta_q \), and \( \sigma_a^2 \) (see Section 2.2 in [37]).

**Example 3.2.1** Schmitt and Stephanopoulos [42] model the time series gene expression profiles of the bacterium Synechocystis using an autoregressive process. Synechocystis PCC6803 is a polyploid photosynthetic prokaryote. It can be grown on glucose and has been used to study photosynthesis related functions in bacteria [34]. In this study the light conditions under which the bacterium is grown are varied and the response in gene expression is measured at 50 subsequent time points. The gene expression levels are modeled as autoregressive processes with the current light intensity functioning as an external input. Genes are selected by their degree of correlation to the signal of light intensity and clustered into groups of similar expression profiles. The expression values of genes in a cluster are modeled as noisy linear combinations of past values plus the external input term. Model selection via Akaike’s information criterion [43] is employed to determine the optimal order of the autoregressive process.

**Definition 3.2.3** Autoregressive Moving Average Process

The process \( \{z_t, t = 1, 2, \ldots\} \) is called an *autoregressive moving average process* with parameters \( p \) and \( q \), denoted by ARMA\((p, q)\), if it can be modeled as

\[
z_t = \sum_{k=1}^{p} \phi_k z_{t-k} + \sum_{k=1}^{q} \theta_k a_{t-k} + a_t,
\]

where \( \{a_t\} \) is white noise process with constant variance \( \sigma_a^2 \). The model parameters are \( \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \) and \( \sigma_a^2 \) (see Section 2.2 in [37]).
3.3 Multivariate Time Series Analysis

In many applications, simultaneous observations on more than one variable may be made at different time points \( t (t = t_1, \ldots, t_T) \). The result is a vector valued time series \( Z_t = (z^{(1)}_t, \ldots, z^{(n)}_t)' \) where \( z^{(i)}_t \) denotes the observation on variable \( i \) at time \( t \). In this case, the goal is not only to describe the serial dependence of each component series \( z^{(i)}_t \), but also the interdependence of observations from different components \( z^{(i)}_{t_1} \) and \( z^{(j)}_{t_2} \) at different time points.

**Example 3.3.1** In a microarray experiment many simultaneous measurements are made on the expression levels of an organism’s genes. Possibly, measurements on other factors, e.g., protein levels or external stimuli such as environmental conditions, may exist as well. Recently, experimenters have not only been interested in the development of the expression of a single gene, but also in the interdependencies of gene expression measurements under changing conditions. The goal is now to describe the interdependencies of gene expression as well as regulation by external factors. For example, Beal et al. [44] obtain gene expression measurements on 88 human T-cell line genes at 10 time points. The experiment is repeated and a subset of 58 genes with consistent results in both experiments is selected for further analysis. For the selected observations a model is fitted that describes each observed gene expression value as a function of the 58 gene expressions at the previous time point plus a linear function of unobserved regulator values. From the fitted model it is concluded which genes are involved in cell proliferation as well as programmed cell death.

For a random vector the covariance matrix can give valuable insight into the interdependencies of the components. In a time series setting, this concept is extended to include the relationship between different components at different time points.
Definition 3.3.1 Autocovariance Function

If \( \{Z_t, t = t_1, t_2, \ldots \} \) is an \( n \) dimensional time series, with \( \mu_t^{(i)} = E \left( z_t^{(i)} \right) \), then the autocovariance function \( \Gamma \) of the process for a time \( t \) and a time-lag \( h \) is defined as

\[
\Gamma(t + h, t) = E \left[ (Z_{t+h} - \mu_{t+h}) (Z_t - \mu_t)' \right] = [\gamma_{ij}(t + h, t)]_{i,j=1}^n,
\]

where

\[
\gamma_{ij}(t + h, t) = E \left[ (z_{t+h}^{(i)} - \mu_{t+h}^{(i)}) (z_t^{(j)} - \mu_t^{(j)}) \right]
\]

(see Section 1.9 in [37]).

Definition 3.3.2 Stationarity

A multivariate time series is said to be stationary, if the means \( \mu_t \) and the covariance matrices \( \Gamma(t + h, t) = \Gamma(h) \) do not depend on \( t \) (see Section 1.5 in [37]).

In the context of microarray analysis, stationarity means that the interdependencies in gene expression do not change over time. This assumption is reasonable, as long as the experiment is not carried out over a very long period of time in which the organism may fundamentally or developmentally change. Most often microarray experiments are designed so that the observations are taken at relatively short intervals in time. The experimental time unit needs to be chosen short enough to capture important changes in gene expression but long enough to cover the whole process that is to be observed (e.g., cell cycle) with an affordable number of arrays. However, not all genes may respond to stimuli at the same speed. The maximum relevant biological time lag is the maximum number of observations that may lie between a cause (e.g., change in expression of regulating genes, change in external factors) and its effect (e.g., change in expression of regulated genes). This maximum relevant biological time lag depends not only on the experimental organism and the observed process, but also on the chosen experimental time unit. Often, microarray experiments are designed so that the maximum relevant biological time lag is a small multiple of the experimental time unit.
**Definition 3.3.3** Autocorrelation

For a stationary multivariate time series \(\{Z_t, t = t_1, t_2, \ldots\}\) the autocorrelation coefficient for observation components \(i\) and \(j\) which are \(h\) time units apart is defined as

\[
\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}.
\]

**3.3.1 Multivariate ARMA Processes**

Similar to the one-dimensional case, vector valued time series observations can be modeled as linear combinations of the past observations or differences of disturbance terms.

**Definition 3.3.4** Autoregressive Moving Average Process

The \(n\) dimensional time series \(\{Z_t, t = 1, 2, \ldots\}\) is said to be an \(n\)-variate autoregressive moving average process, denoted by ARMA\((p, q)\), if \(\{Z_t\}\) is a stationary solution of the difference equation

\[
Z_t = \sum_{k=1}^{p} \Phi_k Z_{t-k} + \sum_{k=1}^{q} \Theta_k A_{t-k} + A_t,
\]

where \(\Phi_1, \ldots, \Phi_p, \Theta_1 \ldots, \Theta_q\) are real \(n \times n\) matrices and \(A_t \sim MVN(0, \Sigma)\) is a vector of multivariate normal noise.

**Lemma 3.3.1** Causality Criterion

Define the polynomial

\[
\Phi(z) = I - \Phi_1 z - \cdots - \Phi_p z^p.
\]

If \(\det \Phi(z) \neq 0\) for all \(z \in \mathbb{C}\) such that \(|z| \leq 1\), then the \(n\)-variate ARMA\((p, q)\) process defined above has exactly one stationary solution. In this case the process is referred to as causal (see Section 4.12 in [37]).

**Lemma 3.3.2** Invertibility Criterion

Define the polynomial

\[
\Theta(z) = I + \Theta_1 z + \cdots + \Theta_q z^q.
\]
If det\( \Theta(z) \neq 0 \) for all \( z \in \mathcal{C} \) such that \( |z| \leq 1 \), then the \( n \)-variate ARMA\((p,q)\) process defined above is invertible, that is the perturbations \( A_t \) can be expressed as linear combinations of the observations \( Z_t \) (see Section 4.12 in [37]).

### 3.4 State Space Models

In the microarray context, the goal is to model gene expression observations as functions of some (not necessarily observable) regulators. State space models are uniquely suited for this purpose. The state of the system is a minimum set of information from the present and past that allows to predict the future behavior of the system [40]. For microarray observations, the state variables are the factors that regulate gene expression, such as protein levels in the cell or DNA methylation on the genes themselves. The observations in this case are the gene expression values as measured by a microarray.

**Definition 3.4.1 State Space Model**

The \( n \)-dimensional observations \( Z_t = \left( z_t^{(1)}, \ldots, z_t^{(n)} \right)' \) at discrete time points \( t = 1, \ldots, T \) are said to follow a state space model if they satisfy the equations:

\[
Z_t = G_t Y_t + \epsilon_t, \quad t = 1, 2, \ldots, T, \tag{3.1}
\]

\[
Y_t = F_t Y_{t-1} + \delta_t, \quad t = 1, 2, \ldots, T. \tag{3.2}
\]

Equation 3.1 is referred to as the observation equation. The gene regulation matrix \( G_t \) describes the relationship between the observations \( Z_t \) and the regulators \( Y_t \). In this context it can be assumed that the regulators \( Y_t \) cannot be observed. Equation 3.2 is often called the state equation. The system matrix \( F_t \) describes the behavior of the unobservable regulators over time (see Section 4.9 in [37]). The assumptions on model (3.1, 3.2) are:

(a) \( F_1, F_2, \ldots, F_T \) is a known set of \( m \times m \) matrices;

(b) \( G_1, G_2, \ldots, G_T \) is a known set of \( n \times m \) matrices;
(c) $Y_0$ and $(\epsilon_t, \delta_t)$ are uncorrelated for every $t$;

(d) $Y_0 \sim \text{MVN}(\mu, \Sigma)$ is the multivariate normal distribution of the initial state of
the system; and

(e) $\epsilon_t \sim \text{MVN}(0, \Sigma_{\epsilon})$ and $\delta_t \sim \text{MVN}(0, \Sigma_{\delta})$ are independent for every $t$.

**Remark 3.4.1** In many important applications, in which the regulation mechanisms
does not change over time, the matrices $F_t$ and $G_t$ do not depend on $t$. The
state equation is said to be *stable*, if all eigenvalues of the matrix $F$ lie within the
unit circle. In this case the sequence of observations $Z_t$, as well as the system states
$Y_t$ are stationary.

**Example 3.4.1** State space representations are not unique. A renumbering of the
unobservable states (i.e., of the components of $Y_t$) leads to different $F$ and $G$ matrices
in the state space representation. Two possible representations of an ARMA($p, q$)
process are as follows:

Consider the one-dimensional, causal ARMA($p, q$) process defined by

$$z_t = \sum_{k=1}^{p} \phi_k z_{t-k} + \sum_{k=1}^{q} \theta_k a_{t-k} + a_t, \quad t = 1, 2, \ldots$$

where $\{a_t\} \sim N(0, \sigma^2)$ and $\phi(z) \neq 0$ for $|z| \leq 1$. Let

$$r = \max(p, q + 1), \quad \phi_j = 0 \text{ for } j > p, \quad \theta_j = 0 \text{ for } j > q \text{ and } \theta_0 = 1.$$

Then the observation equation can be written as

$$z_t = [\theta_{r-1}, \theta_{r-2}, \ldots, \theta_0]Y_t,$$

where

$$Y_t = \begin{bmatrix} y_{t-r+1} \\ y_{t-r+2} \\ \vdots \\ y_t \end{bmatrix}.$$
and
\[ y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + a_t, \quad t = 0, 1, 2, \ldots \]

The state equation becomes
\[
Y_{t+1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_r & \phi_{r-1} & \phi_{r-2} & \cdots & \phi_1 \end{bmatrix} \begin{bmatrix} Y_t \\ a_{t+1} \end{bmatrix}, \quad t = 0, 1, 2, \ldots
\]

Another state space representation of the same ARMA\((p, q)\) process \(z_t\) defined above is the following: Let
\[
m = \max(p, q) \quad \text{and} \quad \phi_j = 0 \quad \text{for} \quad j > p.
\]

Then the state equation may be written as
\[
z_t = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} Y_t + a_t, \quad t = 0, 1, 2, \ldots
\]

Here \(Y_t\) is the unique stationary solution of
\[
Y_{t+1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_m & \phi_{m-1} & \phi_{m-2} & \cdots & \phi_1 \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-1} \\ \psi_m \end{bmatrix} + \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-1} \\ \psi_m \end{bmatrix} \cdot a_t, \quad t = 0, 1, 2, \ldots,
\]

where the \(\psi_1, \ldots, \psi_m\) are the coefficients of \(z, z^2, \ldots, z^m\) in the power series expansion of \(\theta(z)/\phi(z)\), \(|z| \leq 1\) (see Example 12.1.6 in [39]).

**Remark 3.4.2** A multivariate ARMA\((p, q)\) process can also be expressed in state space form. In this case, the system matrix \(F\) of Equation (3.2) becomes a block-diagonal matrix with components as described previously.
Remark 3.4.3 In the microarray context the state variables $Y_t$ are the unobservable regulators of gene expression. Any renumbering of these regulators will lead to a corresponding change in the system matrix $F$ (Equation 3.1), as well as a change in the gene regulation matrix $G$ (Equation 3.2).

3.5 The Kalman Recursions

Consider the time series state space model

$$Z_t = G Y_t + \epsilon_t,$$

$$Y_t = F Y_{t-1} + \delta_t,$$

where $Z_t$ are the $n$-dimensional observations and $Y_t$ are the $m$-dimensional state variables for $t = 1, \ldots, T$.

The Kalman recursions [45] allow the estimation of the unobservable state variables $Y_t$ under different conditions:

- Estimating a state of the system $Y_t$ given the previous observations $Z_1, \ldots, Z_{t-1}$ is called the prediction problem. The predictions are denoted by

$$Y^{t-1}_t = E[Y_t|Z_1, \ldots, Z_{t-1}], \quad t = 1, \ldots, T. \quad (3.3)$$

- Estimating a state of the system $Y_t$ given the observations up to that state $Z_1, \ldots, Z_t$ is called the filtering problem, and these filtering estimates are denoted by

$$Y^t_t = E[Y_t|Z_1, \ldots, Z_t], \quad t = 0, \ldots, T. \quad (3.4)$$

- Estimating a state of the system $Y_t$ given a set of observations $Z_1, \ldots, Z_T$ for $T > t$ is called the smoothing problem. Smoothing estimates are denoted by

$$\hat{Y}_t = E[Y_t|Z_1, \ldots Z_T], \quad t = 0, \ldots, T. \quad (3.5)$$
In the context of microarray time series data, the goal is to obtain the smoothing estimates $\hat{Y}_0, \hat{Y}_1, \ldots, \hat{Y}_T$ which represent the states of the hidden regulators (e.g., protein levels, DNA methylation, etc.) from the observations on gene expression $Z_1, Z_2, \ldots, Z_T$. $T$ is the total number of observations available. Here, the smoothing estimates are the expected values of the state of the system over time ($t = 0, \ldots, T$) given the complete set of observations (i.e., microarray data).

The discussion of the Kalman filtering algorithm presented here follows the version described by Shumway and Stoffer in 1982, which takes the covariances of the smoothing estimates into account [46]. However, in the microarray application the dimension of the observation vector is usually very high, as thousands of genes may be observed simultaneously. The dimensionality of microarray applications presents new computational challenges. Kalman smoothing estimates are obtained in a two-step procedure by performing first a forward and then a backward pass through the data. The forward pass computes the Kalman filtering estimates from the one-step predictions of the state space variables, as well as the covariance matrices of the filter estimates. In the backward pass the filter estimates are used to determine the Kalman smoothing estimates of the unobservable state variables as well as their covariance matrices, given the complete set of observations.

For the forward recursion of the Kalman smoothing procedure define the covariance matrix of the regulator $Y_t$ as

$$P_t^T = E \left[ (Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)' \right] = Cov \left[ Y_t | Z_1, \ldots, Z_T \right].$$

Similarly, let $P_{t,t-1}^T$ denote the covariance matrix of the state space vector $Y$ at two consecutive time points

$$P_{t,t-1}^T = E \left[ (Y_t - \hat{Y}_t)(Y_{t-1} - \hat{Y}_{t-1})' \right] = Cov \left[ Y_t, Y_{t-1} | Z_1, \ldots, Z_T \right].$$

For the forward pass through the data initialize

$$Y_0^0 = \mu, \quad P_0^0 = \Sigma.$$
Then, with estimates or initial values for $F$, $G$, $\Sigma_\delta$, and $\Sigma_\epsilon$ compute the Kalman filter estimates $Y_1^{t-1}, \ldots, Y_T^{T}$ for the state variables and their covariance matrices $P_1^{t-1}, \ldots, P_T^{T}$ by the following forward recursion for $t = 1, \ldots, T$

\begin{align*}
Y_t^{t-1} &= FY_{t-1}^{t-1}, \quad (3.8) \\
P_t^{t-1} &= FP_{t-1}^{t-1}F' + \Sigma_\delta, \quad (3.9) \\
K_t &= P_t^{t-1}G' \left[ GP_t^{t-1}G' + \Sigma_\epsilon \right]^{-1}, \quad (3.10) \\
Y_t^t &= Y_t^{t-1} + K_t \left( Z_t - GY_t^{t-1} \right), \quad (3.11) \\
P_t^t &= (I - K_tG)P_t^{t-1}. \quad (3.12)
\end{align*}

The filter estimates $Y_t^t$, their covariance matrices $P_t^t$, and the prediction estimates $Y_t^{t-1}$, as well as their covariance matrices $P_t^{t-1}$, need to be stored for use in the Kalman smoothing step. It is not necessary to store the Kalman gain matrices $K_t$.

In the second step, a backward pass through the data, the Kalman smoothing estimates $\hat{Y}_0, \hat{Y}_1, \ldots, \hat{Y}_T$ and their covariance matrices $P_1^T, \ldots, P_T^T$ are computed recursively for $t = T, T-1, \ldots, 1$:

\begin{align*}
J_t &= P_{t-1}^{t-1}F' \left[ P_t^{t-1} \right]^{-1}, \\
\hat{Y}_{t-1} &= Y_{t-1}^{t-1} + J_t \left( \hat{Y}_t - Y_t^{t-1} \right), \\
P_{t-1}^T &= P_{t-1}^{t-1} + J_t \left( P_t^T - P_{t-1}^{t-1} \right) J_t'.
\end{align*}

The covariance matrices $P_{t-1,t-1}^T$ of the smoothing estimates for two successive state variables $(Y_{t-1}^T, Y_t^T)$ also need to be computed by initializing

$$P_{T,T-1}^T = (I - K_TG)FP_{T-1}^{T-1},$$

then computing (backwards) for $t = T, \ldots, 2$

$$P_{t-1,t-2}^T = P_{t-1}^{t-1}J_{t-1} + J_t \left( P_{t,t-1}^T - FP_{t-1}^{t-1} \right) J_t'.$$

This produces the smoothing estimates $\hat{Y}_0, \ldots, \hat{Y}_T$, their covariance matrices $P_0^T, \ldots, P_T^T$, and the covariance matrices of the estimates of two successive state variables $P_{1,0}^T, \ldots, P_{T,T-1}^T$. 
where $P_{t,t-1}^T$ denotes the covariance matrix of two consecutive state space vectors (Equation 3.7). In the implementation, the computation of the covariance matrices can be combined with the Kalman smoothing step to avoid having to store more than two instances of the matrices $J_t$.

### 3.6 Application to Biology

Applying the methods of time series analysis to microarray data has potential to provide valuable insight into gene regulatory networks. State space models are uniquely suited to model the hidden factors that regulate gene expression. They allow for both biological variation in the regulators (since the gene regulation process varies from individual to individual), as well as technical variation due to measurement errors. Furthermore, biological assumptions can be incorporated in a meaningful way by restricting the parameter space of the state space model.

When applied to the analysis of microarray data, state space models also provide new challenges. The number of genes that are to be analyzed simultaneously (i.e., observations in the state space model) may be extremely large. For example, Affymetrix® microarray technology, is available for many organisms [17]: The Affymetrix® human array contains over 1,000,000 oligonucleotide features that measure the expression levels of over 33,000 human genes. The size of other organism’s genomes that can be studied with Affymetrix® microarrays ranges from over 5000 genes for yeast to over 34,000 genes for mouse. Presently, in most experiments the number of available time points is small (e.g., 10 - 15 time points), due to the cost and effort that is still required to conduct a microarray experiment. For example, the well known CDC15 experiment carried out by Spellman et al. [31] to study the yeast cell cycle contains 23 observations, of which 19 are equally spaced in time. It is anticipated that experiments in the near future will contain more time points, as new versions of microarray technology become more affordable. Using this as motivation, novel statistical methods must be developed that are able to extract as much
regulatory information as possible from the available data while remaining computationally feasible. In Chapter 4, the past and current approaches to the statistical analysis of microarray time series data are discussed.
4. THE CURRENT STATE OF MICROARRAY TIME SERIES DATA ANALYSIS

Extensive work has been done in the area of time series gene expression analysis. While some of the related work is focused on identifying differentially expressed genes [47, 48], or recognizing patterns of gene expression over time [31, 49], a large body of research is aimed at understanding gene regulatory networks [50–52]. Several different mathematical models have been proposed to describe and explain changes in gene expression over time. Many of these models use the temporal profiles (i.e., gene expression as a function of time; Figure 4.1) of the gene expression patterns to extract information about the observed genes. However, studying the temporal

![Temporal Profile](image)

Figure 4.1. Temporal profiles of three genes in a microarray time series experiment.

profiles of genes separately [50, 53], does not make use of the available correlation information for different genes over time. The work presented here proposes a state space model as the basis for an algorithm that uses the technique of Kalman smoothing [45] to identify regulated genes. Specifically, hidden regulators are incorporated in the model and the temporal correlations of genes are taken into account. The proposed KM-algorithm alternately computes Kalman smoothing estimates (K-step)
and maximizes the restricted model likelihood (M-step). The algorithm is compared
to current approaches in order to demonstrate its value when analyzing microarray
data.

4.1 Mathematical Background

The Kalman filter, as described in Section 3.5, was developed by Rudolf E. Kalman
[45] around 1960. Originally, it was intended to solve the discrete-data linear filtering
problem in engineering control theory. Since then, extensive research has been
conducted in this area, and applications extend to many fields [54–56], especially the
area of autonomous navigation systems [57]. The Kalman filter allows the recursive
estimation of a series of hidden states in a noisy time series model and provides
estimates for the past, present, and future values of the hidden states. Alternately, a
priori estimates (forward in time) can be obtained for the hidden regulators, which
are then smoothed, by incorporating new observations, to obtain a posteriori esti-
mates (backward in time) for the hidden regulators given the complete set of ob-
servations. From a statistical perspective, the Kalman filter can be viewed as the
estimator that minimizes the mean squared error of a signal under the influence of
additive noise [58].

The Expectation-Maximization (EM) algorithm [59] maximizes a likelihood func-
tion to provide estimates in parametric models with missing data by alternately
computing the expected value of the unknown variables in the model and the maxi-
mum likelihood values of the model parameters. The generalized EM-algorithm [60],
replaces the maximum likelihood step, for example, by a Gauss-Newton iteration
that attempts to increase, but not necessarily maximize, the likelihood. The conver-
gegence properties of both the EM-algorithm and the generalized EM-algorithm were
investigated by Wu [61].

To fit a model to observations which are regulated by a hidden process Shumway
and Stoffer [46] employ the EM-Algorithm, and the method of Kalman smoothing
to estimate the parameters of a time series state space model. Simultaneously, they estimate the hidden process, as well as the manner in which it regulates the observations. They also propose a modification to the model that partitions the observation vector in an observed and unobserved portion in the case of missing data. Their method is applied to a relatively short (27 observations) bivariate series of economic data with the purpose of making predictions for future states of the system. In this application the unobservable state vector is known to be one-dimensional.

Wu et al. [62] generalize the algorithm described by Shumway and Stoffer to include the case where some of the model parameters are constrained. These constraints allow the incorporation of prior knowledge into the model. However, in their work the regulation matrix $G$ (in Equation 3.1) is assumed to be known. The authors successfully apply their algorithm to a short economic time series for which the mechanism of regulation is known. In their case, the straightforward maximum likelihood estimate of the regulation error covariance is not positive definite. Wu et al. [62] handle this problem by assuming some of the parameters to be fixed and by updating only the unknown model parameters.

In many biological applications, it may be possible to make general assumptions on the model. However, these assumptions should not be so stringent that they lead to fixing certain model parameters at arbitrary values. It is reasonable to assume a general structure of the data, e.g., uncorrelated measurement errors for different observations. Nevertheless, fixing a model parameter, such as the regulation matrix $G$, at some arbitrary value is not a useful endeavor since it assumes the very same biological knowledge that the experiment is intended to uncover. Another example of problem reduction that is not biologically motivated is data pre-processing. It relies on fixing arbitrary threshold values (e.g., a 2-fold gene expression change) to exclude a large portion of the observations from the final analysis. Most often there is no underlying biological reason for the choice of arbitrary thresholds, instead they are chosen solely for the purpose of reducing observations to make computations numerically feasible [50,52,53,63]. Because current methods are arbitrary and not
biologically founded or relevant, novel statistical methods are needed that rank the observations according to biologically more meaningful criteria. The proposed regulation criterion obtained through the KM-algorithm provides such an alternative by making it possible to rank all of the (thousands) observed genes in a microarray experiment according to their degree of regulation. If necessary, genes which are unlikely to be regulated in the observed process can be removed from further analysis. Highly regulated genes can be studied in more detail in follow-up experiments.

### 4.2 Time Series Microarray Applications

Since the advent of microarray technology, the analysis of time series gene expression microarray data has been explored in a variety of different applications using a variety of different statistical analyses. Typically, one goal is to identify genes that are differentially expressed between two conditions by comparing their temporal profiles. Different statistical tests have been proposed for this purpose [48, 64]. The intent is to determine whether a gene is a noisy “copy” of another gene over time, or whether two temporal patterns are sufficiently different to declare the two genes differentially expressed (Figure 4.2). Genes that share the same profile are usually

![Figure 4.2](image-url). Determination of differential expression for two temporal log-ratio gene expression profiles. (a) The observed genes are assumed to have the same function as their temporal profiles are similar. (b) The temporal profiles are not similar and the genes are declared differentially expressed over time.
assumed to share biological function as well. The log-ratios of gene expression to some baseline value (often the zero-time point) are used as the observations.

Early work in 1997 by DeRisi et al. [53] compares the temporal profiles of 6400 distinct expression sequences of *Saccharomyces cerevisiae*. A time series with 7 time point observations on 6400 features at 2-hour intervals is studied in detail. The first time point in the series is used as the reference sample to which the observations at subsequent time points are compared. Genes with similar expression profiles are manually clustered. In their analysis, only genes with large overall fold-changes (> 2) compared to the control are considered further. Genes with similar expression profiles are assumed to share the same regulatory properties and the authors deduce the function of several unknown genes through those of well studied genes in the same cluster. No statistical methods are employed to deduce the differential expression, nor to deduce the expression of a group of genes.

Swift et al. [65] extend the idea of DeRisi et al. by computing pairwise temporal correlation coefficients for pairs of observations. Pearson’s correlation coefficient and Spearman’s rank correlation are used to determine the similarity of the temporal profiles of two variables. Their method is applied to two relatively long, economical and medical time series data sets [66, 67] with 300 and 76 variables, respectively. However, for the microarray application consisting of several thousand observations this method becomes computationally infeasible due to the extremely large number of pairwise comparisons. Furthermore, correlation measures such as Pearson’s correlation coefficient disregard the temporal nature of the available data (e.g., a random reordering of the observed time points will not have any effect on the correlation results).

Cluster analysis [68] is a general statistical technique that makes use of computing an association measure for pairs of observations. It categorizes observations into groups so that the degree of association between members of the same group is strong and association between the different groups is weak. The observed data are thus organized into structures. For microarray time series analysis the association
measure determines how similar the temporal profiles of two genes are to each other (Figure 4.3). The intention is to infer the function (and possible regulation) of unknown genes in a cluster by that of known genes in the same cluster [50, 69]. However, biologically, genes that share the same function do not necessarily have similar expression profiles in an observed process. Vice versa, genes that have similar expression profiles should not automatically be assumed to share the same biological function [5].

![Graph showing Log-ratios vs Time with a cluster highlighted.](image)

**Figure 4.3.** Clustering of similar temporal gene expression profiles.

In [50], Chen et al. use clustering to infer gene regulation in the analysis of yeast microarray data. In particular, they are interested in identifying genes which act as activators or inhibitors in the regulation process. In their analysis the genes are pre-screened and only those genes whose absolute and relative expression levels exceed certain arbitrary thresholds are analyzed. Note, that this method excludes a large proportion of observed genes from the analysis. The remaining genes are clustered by computing a consensus function for genes already associated within a cluster and including new genes with expression profiles similar to the consensus function into the same cluster. The expression profiles of the genes in each cluster are averaged and smoothed by linearly interpolating peaks (Figure 4.4 (a)). However, it remains unclear whether this smoothing procedure is biologically meaningful. From the resulting averaged, smoothed curves a “gene activation” grade is computed. Referring to Figure 4.4 (b), the red cluster receives a high gene activation grade with respect to the blue cluster, since the leading edge of the red cluster’s peak slightly
Figure 4.4. (a) Smoothing of averaged gene expression profiles. (b) The red cluster is assumed to activate the blue.

precedes the leading edge of the blue cluster’s peak. The method is applied to Spellman’s *Saccharomyces cerevisiae* (CDC15) data set with 6601 features measured at 17 time points [31]. After pre-screening, 3131 features remain in the analysis and were clustered into 308 groups. A model for the gene regulatory network between these clusters was proposed based on the computed gene activation grade for each pair of clusters. This method loses a large amount of correlation information and essential biological information by pre-screening more than 50% of the available data and by employing smoothing procedures. The regulation is then modeled exclusively between clusters based solely on the fabricated peaks, even though the true biological model is likely more complex. The outcome of such an analysis is, at best, an approximation of the observed extremely complex biological process.

In [69], Jenner et al. apply the method of temporal clustering to a human herpes-virus gene expression data set. They extend Pearson’s correlation measure to include different time lags. The expression levels of 106 genes of the human herpes-virus 8 are observed at eight time points [70]. The gene expression values are discretized into three possible states: under-expressed, not expressed, and over-expressed compared to a baseline measurement. Genes with high correlation in their temporal profiles are clustered, and a model for regulation between clusters is inferred. Similar to pre-processing gene expression data according to maximum absolute or relative expression values, discretizing gene expression data also relies on introducing arbitrary
threshold values to categorize the expression values into the three possible states. Many genes, whose maximum expression change falls below these arbitrary thresholds will effectively be removed from the analysis, as they will be categorized as not expressed across the whole experiment. Especially genes with naturally low variation are excluded from further analysis, even though they may be relevant to the discovery of regulatory networks. Important information, necessary to explain biological systems as intricate as the gene regulatory network is lost in this process.

Recently, Cui et al. [71] use Kalman filtering to measure the functional connectivity between two genes from the temporal expression profiles. They propose replacing the popular and previously used similarity measures like Pearson’s correlation coefficient by a new, model based, measure. Specifically, for each pair of genes, the expression values of one gene are modeled as linear functions of the expression values of the other. The parameters of the linear model are obtained via Kalman filtering. Since the dependence parameter that models the relationship between genes may change over time, their method allows the characterization of the functional connectivity between genes over time. Cui et al. apply their method to only four genes selected from the yeast time series data set published by Cho et al. [32] that contains observations on 6220 genes at 17 time points. High computational cost due to the large number of pairwise comparisons between gene temporal profiles make this method impractical for the analysis of complete microarray data sets with thousands of genes.

4.2.1 Dynamic Bayesian Networks

To understand an organism’s entire gene regulatory network is an ambitious goal. Specifically, if the intention is to identify possible causal relationships between genes or between a gene and another component in the cell, such as a protein. Bayesian networks have recently been used to model this complex interaction structure. A Bayesian network is a directed graphical model in which the nodes represent random
variables and the arcs represent causal relationships (or dependencies) between the random variables [72]. In the microarray context (compare Figure 4.5), the nodes are the observed genes and unobservable regulators and the arcs represent regulation between genes or between genes and regulators [73]. Dynamic Bayesian networks (DBN) are directed graphical models for stochastic processes [74]. They model the dependencies between observable and unobservable components over time. Special cases of DBNs are boolean networks [75], hidden Markov models [76] and linear dynamical systems [77] such as state space models. A fundamental assumption in DBNs is that the model structure does not change over time.

In the microarray context, boolean networks are used to model genes as either on or off, and to use on/off combinations of certain genes to determine whether a third gene is on or off [69, 78]. Stochasticity can be included in these models by adding noise to the boolean regulators [79]. Hidden Markov models (HMM) use a finite number of discrete unobservable states to model the observations as random variables whose distributions depend on the current state of the system [80]. For microarray experiments, HMMs have been used in combination with clustering techniques [63] to account for the temporal nature of the observed gene expression data. Differential equation models [51] and state space models [44, 81, 82] model the observations as functions of previous observations and hidden variables.

![Bayesian network](image-url)

**Figure 4.5. Bayesian network:** The circles represent observable genes and unobservable regulators and the directed arcs describe regulation between them.
A commonality between all of these models is that they use system matrices to describe the temporal development of the unobservable states and regulation matrices to model the dependency of the observations on these states. In microarray experiments, the matrices are typically model parameters that require estimation. Both Bayesian [44] as well as frequentist methods [82, 83] can be used to estimate the parameters of a DBN model. However, in most microarray experiments only few replicated observations on the expression of each gene will be available.

Replication in microarray experiments can be biological or technical. Biological replicates are genetic material from different individuals of the same experiment. Technical replicates are repeated measurements taken on the same biological sample. In time series gene expression experiments it is not uncommon to have only one biological (possibly pooled) sample and few technical replicates per observed gene and time point. If the number of replicates (both biological and technical) is too small, it becomes impractical to determine confidence intervals for the model parameters using a Bayesian approach. However, maximum likelihood parameters for the model regulation matrices can still be found. Some results that represent the different approaches to the discovery of gene regulatory networks via the analysis of microarray data are discussed in greater detail below.

Kim et al. [84] model a gene regulatory network as a boolean system. Genes are modeled at each time point as either invariant, up-, or down-regulated compared to the reference sample. An arbitrary threshold is used to make the distinction (e.g., 2-fold expression change) which classifies a majority of genes as invariant at all time points. The significantly reduced number of genes are included in a linear model to reconstruct gene regulation. The discretized linear profile of a gene is modeled as a weighted sum of potential predictor genes. This method does not differ significantly from the “gene activation” grade computed by Chen et al. [50], except that genes are not arranged in clusters prior to inferring the regulation. Discretizing gene expression ignores important biological information (e.g., correlation, variation,
etc.). Depending on the choice of the threshold, genes with low to medium variation are excluded from the analysis.

Van Someren et al. [85] and Wessels et al. [86] compare existing approaches for modeling genetic networks from gene expression time series data. They propose criteria for model evaluation such as inferential power, predictive power, robustness, consistency, stability, and computational cost. Under these criteria they compare the pairwise correlation approaches with genetic network models, of which only some allow the inclusion of regulators other than the genes themselves. Their conclusion is that the current gene network models perform poorly if the number of available time-point observations is low, or when any noticeable amount of measurement noise is present. Furthermore, they suggest using biological knowledge to improve the performance of time-series gene expression models, and using more complex models which are able to incorporate intermediate cell products such as proteins.

Bar-Joseph [87] compares different methods for the analysis of time series microarray data such as pattern recognition and clustering through hidden Markov models and singular value decomposition with dynamic Bayesian network approaches, and lists the most important biological problems that need to be addressed by microarray time series experiments:

(i) Understanding temporal cell biology, such as the yeast cell cycle or the circadian clock in mammals;

(ii) Studying genetic interactions by using time series knockout experiments;

(iii) Observing the developmental process to better understand genetic disease; and

(iv) Understanding cell defense mechanisms in reaction to infectious disease.

Also discussed are the challenges of choosing the right sample size for experiments, and the choice between higher number of technical replicates versus longer time series. Another problematic issue with genomic time series data is the synchronization of biological material for the purpose of controlling the biological variation. For
example, when yeast or bacteria cells are studied, a mechanisms that assures that all extracted cells are at the same stage of development needs to be employed. In particular, the importance of incorporating biological information into the gene expression data in order to create meaningful network models is stressed. Bar-Joseph states that it is unlikely that a large number of time series gene knockout data sets will become available in the near future and suggests employing models into which static knockout data can be included. Also proposed is the inclusion of data from other sources, such as protein arrays to study gene-protein interactions.

Zou et al. [88] propose a DBN approach to discover signaling pathways. The authors address the problem of poor performance and excessive computational time by radically reducing the number of genes considered as potential regulators. They exclude genes whose fold-changes over time do not exceed arbitrary cutoff values (e.g., >1.2 fold-change for up-regulation and <0.7 fold-change for down-regulation). A gene is only considered as a regulator for another target gene, if the regulator undergoes a large change in expression prior to a large change in expression in the target gene. External regulators, such as proteins, are not considered. This method extends the work of Murphy et al. [52] by including the possibility of transcriptional time lags beyond the experimental time unit. Furthermore, this method does not require the gene expression values to be discretized as in the work described by Jenner et al. [69] or Kim et al. [84].

In 2005, Beal et al. [44] use a state space model to reverse engineer transcriptional networks for gene expression data, in which a large number of technical replicates is available for every gene in the experiment. Their variational Bayesian EM-algorithm requires prior distributions for the model parameters and many technical replicates in the observations. The method is applied to data that describes the response of a human T-cell line to PMA and ionomicin treatment [81]. In this data set, 44 technical replicates are available for each of 58 genes that are observed at 10 time points. In most microarray experiments, such a high level of repetition is unusual. Because microarray technology is so expensive most microarray experiments are conducted
with few technical and biological replicates. In many time series gene expression experiments only one technical replicate per gene per time point and a small number of biological replicates are available.

Wu et al. [89] also use state space models to model gene expression data. However, their approach disregards both system and measurement error. The values of the hidden regulators are found by maximum likelihood factor analysis, and the gene regulation matrix is calculated deterministically as the product of the observations and the Moore-Penrose generalized inverse of the regulator values. The authors apply their method to two microarray data sets. The Caulobacter data set from Laub et al. [33] which consists of the expression data of 1590 genes at 11 equally spaced time points and the yeast data set (CDC15) from Spellman et al. [31] consisting of 6601 genes, of which Wu et al. use the 799 genes that have been identified to be related to the cell cycle. They provide values for the hidden regulators in their work, but do not describe the regulation process in detail or make distinctions between genes according to their degree of regulation. The failure to include noise in the system severely limits its usefulness in real biological applications. Specifically when applied to larger data sets the regulatory network as derived by this method remains questionable.

Berlo et al. [90] investigate the suitability of DBNs for discovering gene regulatory interactions. They simulate gene expression time series data while varying the number of simulated genes, the number of observed time points, and the measurement error. Also varied is the number of connections between the genes, i.e., pairs of genes who regulate each other in the simulated process. The conclusion is that for a low number of observed time points (less than 50), even for only a moderate number of genes (between 50 and 500) the discovery of the regulatory process is poor. This is especially true, if the grade of connectivity between the genes is high. In their model, the simulated regulation graphs are based on discretized gene expression data. When they apply their method to a real biological data set, they remove more than 90% of the data before the analysis, “to avoid the discovery of erroneous
interactions”, because the expression values do not exceed an arbitrary threshold value. The removal of so many genes severely lowers any chances of discovering biologically relevant interactions.

4.3 Motivation of the KM-Algorithm

Overall, the existing methods for the discovery of gene regulatory networks fail to perform satisfactorily if the number of observed genes is large. In microarray experiments it is not uncommon to obtain data on several thousand genes simultaneously. For example, the plant Arabidopsis thaliana which is used as a model organism for many plant species has over 25,000 genes [91]. To apply the current techniques to such large data sets, many methods rely on heavy pre-screening of the observed genes to reduce the number of observations and thus make the problem computationally feasible. Arbitrary threshold values for absolute or relative expression are typically used to select a subset of genes for analysis. The motivation for the KM-algorithm as presented here, and as applied to microarray data is to identify regulated genes in a specific cellular process via data driven analysis. The underlying biological assumption is that not all of an organism’s thousands of genes are involved in a temporal process such as disease progression or cell defense mechanisms. Furthermore, the regulated genes should not be selected solely based on their maximum log-fold expression change. Toward this end, a method is proposed that allows the ranking of observed genes according to their degree of regulation. This ranking can be used to select single genes for knockout experiments, or to select a subset of genes for more detailed analysis.

The concept of ranking genes according to their regulation is based on a state space model in which the gene expression values function as the observations and the hidden variables can either be gene expression values themselves or measurements of other components of the cell such as proteins or enzymes. The proposed
KM-algorithm alternately computes the Kalman smoothing estimates of the hidden variables and performs a likelihood gradient ascent for the model parameters. Biological assumptions are used to place restrictions on the model parameters and mathematical restrictions are imposed on the parameter space to assure statistical validity of the model. A novel regulation criterion is developed that is based on the resulting model parameter estimates. As discussed in Section 3.4, state space models are not unique. Therefore, it is impossible to combine multiple parameter estimates in the current approaches. However, the proposed regulation criterion is flexible as it allows results for different parameter estimates of the same model to be averaged to increase its power.

The performance and power of the KM-approach are investigated via simulation. A “goodness of ranking” criterion is suggested that measures the ability of the proposed method to obtain an accurate ranking of regulated and unregulated genes. The power of the ranking criterion is studied using different simulated data sets of different length (time points) and size (gene number). The simulation results provide information from which recommendations can be made for a minimum number of time points that a microarray experiment should include in order to achieve a desired power of separation between regulated and unregulated genes. These results can be compared to previous recommendations [90].

Furthermore, simulations demonstrate that genes may be highly regulated even though their variance (over time) is low. These genes previously may have been excluded from the analysis in a data pre-processing step, because their fold-change failed to exceed the arbitrary threshold. To address this issue, the proposed regulation criterion allows for genes to be selected according to their degree of regulation in the observed process. If necessary, unregulated genes can then be excluded from further analysis and highly regulated genes can be studied in more detail.

The state space approach that is used to model the observations on gene expressions also allows the inclusion of regulators that are unobservable. Through the KM-algorithm, estimates for these unobservable regulators are obtained. An assumption
that a certain biological component, such as a particular protein, is a regulator in the observed process, can be validated by comparing the temporal profile of the protein directly with the estimated temporal profiles of the regulators. Protein array experiments are becoming increasingly popular and may soon be conducted in parallel with gene microarray time series experiments to study gene-protein interactions.

Currently, none of the existing statistical methods to model microarray time series data can be applied to extremely large data sets. Instead of removing a large portion of the observations from the analysis entirely, the KM-algorithm is complemented with a partitioning method. Since the run-time of the KM-algorithm is in the order of $O(n^2)$, where $n$ is the number of observed genes, it is not practical to apply it to an extremely large data set directly. Instead, a partitioning method is proposed that significantly reduces computational expense for the algorithm. If all regulators are, in fact, unobservable components other than the gene expressions themselves, partitioning the data has no ill effect on the performance of the KM-algorithm. If, on the other hand, some of the observed genes are among the regulators, then partitioning a large data set potentially disregards correlations between observations in different partitions. To address this issue the flexibility of the regulation criterion is exercised. Large data sets are randomly partitioned repeatedly in different ways. For each partition, the KM-algorithm is applied and the degree of regulation for each gene is determined. The regulation results of different partitions can then be combined to account for correlations between the observed genes.
5. THE KM-ALGORITHM

5.1 Overview

The KM-algorithm is proposed as a generalized EM-algorithm that alternately computes the Kalman smoothing estimates of the state variables $Y_t$ (K-step) and finds parameters that maximize the model likelihood (M-step) in a time series state space model:

\[
Z_t = G Y_t + \epsilon_t, \quad \text{observation equation,}
\]

\[
Y_t = F Y_{t-1} + \delta_t, \quad \text{state equation.}
\]

In the microarray context, $Z_t$ are the $n$ gene expression values for the observed genes at time $t$ ($t = 1, \ldots, T$) and $Y_t$ are the $m$ underlying regulators that determine the expression of some of the genes. Note that they do not necessarily have to be gene expressions themselves. They could, for example, be levels of regulatory proteins or the extent of protein degradation in the cell. The gene regulation matrix $G$ is identifiable only, if the dimension of the state space $m$ is smaller than the number of observed time points $T$. However, the number of biologically relevant regulators is typically small: for example, Wu et al. [89] have identified five regulators for the *Caulobacter* data set of Laub et al. [33] as well as for the CDC15 yeast data set of Spellman et al. [31]. Typically, not all of the genes whose expression levels are measured via microarray technology are regulated in the observed process. For example, of the more than 6000 yeast genes, only approximately 800 are known to be involved in the cell cycle [31]. For the state space model this implies that the gene regulation matrix $G$ will contain entries of small magnitude in most rows.
It is assumed that $\epsilon_t \sim MVN(0, \Sigma_\epsilon)$ and $\delta_t \sim MVN(0, \Sigma_\delta)$ are independent multivariate normal noise. $\delta_t$ represents the biological regulation variation and thus $\Sigma_\delta$ consists of the correlation information for the regulators. The measurement error in the experiment is denoted by $\epsilon_t$. Measurement errors for individual genes are assumed to be uncorrelated (i.e., $\Sigma_\epsilon$ is assumed to be a diagonal matrix). Note, that assuming the measurement errors to be uncorrelated does not imply that the gene expression values are uncorrelated. The gene expression values are correlated, as they all are modeled as linear functions of the regulators (which are themselves correlated), plus independent error terms. Assume that the initial state variable $Y_0$ is multivariate normally distributed $Y_0 \sim MVN(\mu, \Sigma)$. In practice, the matrices $F, G, \Sigma_\delta, \Sigma_\epsilon, \mu,$ and $\Sigma$ are unknown.

### 5.2 Initial Values

To implement the KM-algorithm, initial estimates for the matrices $F, G, \Sigma_\delta, \Sigma_\epsilon, \mu,$ and $\Sigma$ are required. For the system matrix $F$ (State Equation 3.2), the $m$-dimensional identity matrix is used as an initial value. A system of state space equations is said to be stable (or causal, compare Definition 3.3.1) if the (possibly complex) eigenvalues of $F$ lie within the unit circle. In this case the state space model formulated in Equations 3.1 and 3.2 has exactly one stationary solution.

The regulation criterion, which will be defined in Section 5.7, is largely based on the estimated values of the gene expression matrix $G$. Ideally, one should sample the complete range of initial values for $G$ to obtain a stable result. However, $G$ is a large $n \times m$ matrix which makes this approach infeasible. Instead, the algorithm is run several times with different random initial values of $G$ and the results are combined as described in Section 5.7.

Since the regulators are unknown, the initial value for the regulation error covariance matrix $\Sigma_\delta$ is chosen as the $m$-dimensional identity matrix. The variances of the $n$ observed gene time series components of $Z_t$ can be computed, and used as the
initial estimates in the diagonal entries of $\Sigma_{\epsilon}$. The Kalman smoothing estimates $\hat{Y}_0$ (Equation 3.5) and $P_0^T$ (Equation 3.6) obtained from the observed values $Z_t$ are used as the estimates for $\mu$ and $\Sigma$, the parameters of the initial state space distribution.

5.3 The KM-Algorithm Iterations

The KM-algorithm alternately conducts a Kalman smoothing step to extract information about the hidden regulators and a model fitting step that finds new parameter estimates that improve the model likelihood. The specifics of the Kalman smoothing step and the model fitting procedure are detailed here. The algorithm iterates these steps until a termination condition for convergence in the model likelihood (described in Section 5.5) is satisfied.

5.3.1 Kalman Smoothing Step

The observations $Z_1, \ldots, Z_T$, as well as the current estimates of $F, G, \Sigma_{\epsilon}$ and $\Sigma_{\delta}$, are used to obtain estimates of the hidden state space variables $Y_0, Y_1, \ldots, Y_T$. The Kalman recursions are applied as described in Section 3.5. Obtaining the Kalman smoothing estimates is a two-step procedure: First, a forward pass through the time series observations obtains the Kalman filtering estimates, and then a backward pass computes the Kalman smoothing estimates. In both the forward and backward pass the inverses of matrices involving $\Sigma_{\delta}, \Sigma_{\epsilon}$ and the prediction covariance matrices $P_t^{l-1}$ (compare Equation 3.8) are computed.

5.3.2 Model Fitting Step

Given the observations $Z_1, \ldots, Z_T$ and the Kalman smoothing estimates $\hat{Y}_0, \ldots, \hat{Y}_T$, it is possible to compute the maximum likelihood estimates of the parameter matrices $F, G, \Sigma_{\delta}, \Sigma_{\epsilon}, \mu, \Sigma$ by first computing the derivative of the log-likelihood function. The following matrix derivative notation will be used.
Definition 5.3.1 Let $A = (A_{ij})$ be an $n \times m$ matrix and let $f(A)$ be a real function of $A$. The first partial derivative of $f$ with respect to $A$ is defined as the $n \times m$ matrix of partial derivatives $\frac{\partial f}{\partial a_{ij}}$

$$\frac{\partial f(A)}{\partial A} = \begin{pmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1m}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f}{\partial a_{n1}} & \frac{\partial f}{\partial a_{n2}} & \cdots & \frac{\partial f}{\partial a_{nm}} \end{pmatrix}.$$ 

Let $L$ be the likelihood function of the observations and underlying state variables given $F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \Sigma$:

$$L = L(Z_1, \ldots, Z_T, Y_0, Y_1, \ldots, Y_T \mid F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \Sigma).$$

Write $L$ as the matrix expression:

$$L = f(Y_0) \prod_{t=1}^T f(Y_t \mid Y_{t-1}) \prod_{t=1}^T f(Z_t \mid Y_t)$$

$$= \frac{1}{\sqrt{2\pi|\Sigma|^{1/2}}} \exp\left(-\frac{1}{2}(Y_0 - \mu)'\Sigma^{-1}(Y_0 - \mu)\right)$$

$$\cdot \frac{1}{\sqrt{2\pi} |\Sigma_\delta|^{T/2}} \prod_{t=1}^T \exp\left(-\frac{1}{2}(Y_t - FY_{t-1})'\Sigma_\delta^{-1}(Y_t - FY_{t-1})\right)$$

$$\cdot \frac{1}{\sqrt{2\pi} |\Sigma_\epsilon|^{T/2}} \prod_{t=1}^T \exp\left(-\frac{1}{2}(Z_t - GY_t)'\Sigma_\epsilon^{-1}(Z_t - GY_t)\right).$$

The log-likelihood function (up to a constant) becomes

$$\log L(Z_1, \ldots, Z_T, Y_0, Y_1, \ldots, Y_T \mid F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \Sigma) =$$

$$-\frac{1}{2} \log |\Sigma| - \frac{1}{2} (Y_0 - \mu)'\Sigma^{-1}(Y_0 - \mu) \quad (5.1)$$

$$-\frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \sum_{t=1}^T (Y_t - FY_{t-1})'\Sigma_\delta^{-1}(Y_t - FY_{t-1}) \quad (5.2)$$

$$-\frac{T}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \sum_{t=1}^T (Z_t - GY_t)'\Sigma_\epsilon^{-1}(Z_t - GY_t). \quad (5.3)$$
This function will be maximized with respect to \( F, G, \Sigma_\delta, \Sigma_\epsilon, \Sigma, \) and \( \mu \). Note that Equations 5.2 and 5.3 are composed of summands which are, in fact, scalars. Therefore, the trace of these matrix products may be considered, and since \( \text{tr}(AB) = \text{tr}(BA) \), the order of multiplication may be exchanged, and the product expanded,

\[
\log L(Z_1, \ldots, Z_T, Y_0, Y_1, \ldots, Y_T|F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \Sigma) =
\]

\[
-\frac{1}{2} \log |\Sigma| - \frac{1}{2}(Y_0 - \mu)'\Sigma^{-1}(Y_0 - \mu)
\]

\[
-\frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \text{tr} \left[ \Sigma_\delta^{-1} \left( \sum_{t=1}^{T} (Y_t Y_t' - Y_t Y_{t-1}' F' - F Y_{t-1} Y_t' + F Y_{t-1} Y_{t-1}' F') \right) \right]
\]

\[
-\frac{T}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \text{tr} \left[ \Sigma_\epsilon^{-1} \left( \sum_{t=1}^{T} ((Z_t - G Y_t)(Z_t - G Y_t)') \right) \right].
\]

The conditional expected value of the log-likelihood (5.4) is computed given the observations \( Z_1, \ldots, Z_T \). Note, that the conditional expectation of \( Y_t Y_{t-1}' \) given the set of observations \( Z = \{Z_t, t = 1, \ldots, T\} \) is

\[
E (Y_t Y_{t-1}'|Z) = E (Y_t|Z) E (Y_{t-1}'|Z)' + \text{Cov} (Y_t, Y_{t-1}'|Z).
\]

Using the notation \( \hat{Y}_t = E (Y_t|Z) \) for the smoothing estimates and \( P_t^T = \text{Cov} (Y_t|Z) \) as well as \( P_{t,t-1}^T = \text{Cov} (Y_t, Y_{t-1}|Z) \) for the covariance matrix estimates (compare Equations 3.5, 3.6, 3.7) the conditional log-likelihood function can be written as:

\[
\log L(Z_1, \ldots, Z_T, Y_0, Y_1, \ldots, Y_T|F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \Sigma) =
\]

\[
-\frac{1}{2} \log |\Sigma| - \frac{1}{2}((\hat{Y}_0 - \mu)'\Sigma^{-1}(\hat{Y}_0 - \mu)
\]

\[
-\frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \text{tr} \left[ \Sigma_\delta^{-1} \left( \sum_{t=1}^{T} (\hat{Y}_t \hat{Y}_t' + P_t^T - (\hat{Y}_t \hat{Y}_{t-1}' + P_{t,t-1}^T)F'
\]

\[
-F(\hat{Y}_{t-1} \hat{Y}_t' + P_{t,t-1}^T) + F(\hat{Y}_{t-1} \hat{Y}_t' + P_{t,t-1}^T)F')) \right]\]

\[
-\frac{T}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \text{tr} \left[ \Sigma_\epsilon^{-1} \left( \sum_{t=1}^{T} (Z_t - G \hat{Y}_t)(Z_t - G \hat{Y}_t)' + \sum_{t=1}^{T} GP_t^T G') \right) \right].
\]

Equation (5.5) is maximized with respect to \( F, G, \Sigma_\delta, \Sigma_\epsilon, \mu, \) and \( \Sigma \). It is easy to see from (5.6) that the maximum likelihood estimate of \( \mu \) is \( \hat{\mu} = \hat{Y}_0 \), and that
\( \hat{\Sigma} = P_0^T = Var(Y_0|Z_1, \ldots, Z_T) \) can be used as an estimate for \( \Sigma \).

Equation (5.7) can be written as

\[
-\frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \text{tr} \left[ \Sigma_\delta^{-1} (C - BF' - FB' + FAF') \right]
\]  

(5.9)

where

\[
A = \sum_{t=1}^T \hat{Y}_{t-1} \hat{Y}'_{t-1} + P_{t-1}^T, \quad B = \sum_{t=1}^T \hat{Y}_{t-1} \hat{Y}'_{t-1} + P_{t,t-1}^T, \quad \text{and} \quad C = \sum_{t=1}^T \hat{Y}_{t} \hat{Y}'_{t} + P_{t}^T.
\]

Equation (5.9) is a quadratic matrix expression in \( F \). The matrix derivative with respect to \( F \) is

\[
\frac{1}{2} \left( \frac{\partial}{\partial F} \text{tr} \Sigma_\delta^{-1} BF' + \frac{\partial}{\partial F} \text{tr} \Sigma_\delta^{-1} FB' - \frac{\partial}{\partial F} \text{tr} \Sigma_\delta^{-1} FAF' \right) = \frac{1}{2} \left( 2 \Sigma_\delta^{-1} B - \Sigma_\delta^{-1} FA' - \Sigma_\delta^{-1} FA \right) = 0.
\]

\[
2B = F(A' + A),
\]

\[
B = FA.
\]

Therefore, the maximum likelihood estimate of \( F \) is

\[
\hat{F} = BA^{-1}.
\]

Once the maximum likelihood estimate \( \hat{F} \) of \( F \) is found, the maximum likelihood estimate of the covariance matrix \( \Sigma_\delta \) can be obtained through:

\[
\frac{\partial}{\partial \Sigma_\delta} \left( -\frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \text{tr} \left( \Sigma_\delta^{-1} (C - BF' - FB' + FAF') \right) \right)
\]

\[
= -\frac{T}{2} \left\{ \frac{1}{|\Sigma_\delta|} \Sigma_\delta^{-1} + \frac{1}{2} \Sigma_\delta^{-1} \left( C - B\hat{F}' - \hat{F}B' + \hat{F}A\hat{F}' \right) \right\} \Sigma_\delta^{-1} = 0,
\]

\[
T \Sigma_\delta^{-1} = \Sigma_\delta^{-1} \left( C - B\hat{F}' - \hat{F}B' + \hat{F}A\hat{F}' \right) \Sigma_\delta^{-1},
\]

\[
\hat{\Sigma}_\delta = \frac{1}{T} \left( C - B\hat{F}' - \hat{F}B' + \hat{F}A\hat{F}' \right) = \frac{1}{T} \left( C - BA^{-1}B' \right).
\]

Note, that the matrix \( \hat{\Sigma}_\delta \) must be symmetric, since both \( C \) and \( A \) are symmetric. However, since there is no guarantee that the maximum likelihood estimate of the
covariance matrix $\Sigma_{\delta}$ will be positive definite, there may be problems in the next iteration of Kalman smoothing.

Now consider Equation 5.8 of the conditional likelihood function

$$-\frac{T}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \text{tr} \left[ \Sigma_{\epsilon}^{-1} \left( \sum_{t=1}^{T} (Z_t'Z_t' - Z_t\dot{Y}_t'G' - G\dot{Y}_t\dot{Z}_t' + G(\dot{Y}_t\dot{Y}_t' + P_t\dot{G}')) \right) \right]$$

$$= -\frac{T}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \text{tr} \left[ \Sigma_{\epsilon}^{-1} (D - EG' - GE' + GCG') \right], \quad (5.10)$$

where $C$ is as before and

$$D = \sum_{t=1}^{T} Z_tZ_t', \quad E = \sum_{t=1}^{T} Z_t\dot{Y}_t'.$$

Maximizing Equation 5.10 with respect to $G$ yields

$$\frac{1}{2} \left( \frac{\partial}{\partial G} \text{tr} \Sigma_{\epsilon}^{-1} E G' + \frac{\partial}{\partial G} \text{tr} \Sigma_{\epsilon}^{-1} G E' - \frac{\partial}{\partial G} \text{tr} \Sigma_{\epsilon}^{-1} G C G' \right)$$

$$= \frac{1}{2} \left( 2\Sigma_{\epsilon}^{-1} E - \Sigma_{\epsilon}^{-1} G C' - \Sigma_{\epsilon}^{-1} GC \right) = 0,$$

$$2E = G(C' + C).$$

Since the matrix $C$ is symmetric, the maximum likelihood estimate of $G$ is

$$\hat{G} = EC^{-1}.$$

Using $\hat{G}$, the maximum likelihood estimate of the covariance matrix $\Sigma_{\epsilon}$ can be expressed analogous to $\Sigma_{\delta}$,

$$\hat{\Sigma}_{\epsilon} = \frac{1}{T} \left( D - E\hat{G}' - \hat{G}E' + \hat{G}C\hat{G}' \right) = \frac{1}{T} \left( D - EC^{-1}E' \right).$$

Although the resulting estimate $\hat{\Sigma}_{\epsilon}$ is symmetric, it is not necessarily a diagonal matrix. Since the model assumptions are based on measurement errors being uncorrelated, the off-diagonal elements of $\Sigma_{\epsilon}$ are replaced with zeros. The resulting diagonal matrix has only positive entries and is hence automatically positive definite. The following Section discusses the issue of when the maximum estimate of the regulation covariance matrix $\Sigma_{\delta}$ is not positive definite.
5.4 Cholesky Decomposition

As mentioned previously, the maximum likelihood estimate $\hat{\Sigma}_\delta$ may not be positive definite. When, and if this occurs, the estimate cannot be used in the preceding Kalman smoothing step. To address this issue the restricted estimate of $\Sigma_\delta$ is required such that the likelihood function (Equation 5.5) is as large as possible under the restriction of being positive definite. To achieve this solution, the parameter space is reparametrized. Instead of the parameter $\Sigma_\delta$, its Cholesky decomposition $\Gamma$ is considered, where $\Gamma$ is an upper triangular matrix with $\Sigma_\delta = \Gamma \Gamma'$ [92].

The gradient of the likelihood function (Equation 5.5) with respect to $\Gamma$ in the reparametrized model can be written as:

$$\frac{\partial}{\partial \Gamma} L = \frac{\partial}{\partial \Gamma} \left( -\frac{T}{2} \log |\Gamma|^2 - \frac{1}{2} \text{tr} [\Gamma^{-1} \Gamma'^{-1} M] \right) = -T \Gamma'^{-1} + \Gamma (\Gamma')^{-1} \frac{M + M'}{2} (\Gamma')^{-1}. \quad (5.11)$$

where $M = C - BF' - FB' + FAF'$ and $A, B,$ and $C$ as before. The cases where

$$-T \Gamma'^{-1} + \Gamma (\Gamma')^{-1} \frac{M + M'}{2} (\Gamma')^{-1} = 0$$

fails to have have a solution are precisely those where the maximum likelihood estimate of $\Sigma_\delta$ is not positive definite. For these situations, a new parameter estimate for $\Gamma$ cannot be found by maximizing the likelihood. Instead, a gradient ascent in $\Gamma$ is performed such that the model likelihood is increased as much as possible. For the next Kalman smoothing step, the current estimate of $\Sigma_\delta$ can be obtained from $\Sigma_\delta = \Gamma \Gamma'$ since it is guaranteed to be positive definite.

5.4.1 Gradient Ascent

As just described when dealing with the situation where the maximum likelihood estimate of the model parameters $F, G, \Sigma_\delta, \Sigma_e$ is not permissible, due to $\hat{\Sigma}_\delta$ not being positive definite, a gradient ascent is performed [93]. Since it is not possible to maximize the likelihood function, because the function maximum lies outside
the allowed parameter space, the likelihood is increased by following the gradient
direction. After computing the maximum likelihood estimates of the parameters
\( F, G, \Sigma_\delta, \) and \( \Sigma_\epsilon, \) the parameter \( \Sigma_\delta \) is checked for positive definiteness. If the
matrix is positive definite, then the parameter update is permissible and a gradient
ascent is not necessary. However, if the maximum likelihood of \( \Sigma_\delta \) is not positive
definite, then the parameter estimates from the previous step are used as a basis for
the gradient ascent. Since the previous estimate of \( \Sigma_\delta \) was positive definite, it is
possible to perform the Cholesky decomposition as described in Section 5.4 to find
\( \Gamma. \) The likelihood gradients with respect to the model parameters \( F, G, \Gamma, \) and \( \Sigma_\epsilon \)
are then computed as follows:

\[
\frac{\partial}{\partial F} L = \Sigma^{-1}_\delta \left( B - \frac{1}{2} F(A' + A) \right);
\]

\[
\frac{\partial}{\partial G} L = \Sigma^{-1}_\epsilon \left( E - \frac{1}{2} G(C' + C) \right);
\]

\[
\frac{\partial}{\partial \Gamma} L = -2\Gamma^{-1} + \frac{1}{2} \Gamma(\Gamma')^{-1}(C + C' + BA^{-1} + A^{-1})B'(\Gamma')^{-1}; \text{ and}
\]

\[
\frac{\partial}{\partial \Sigma_\epsilon} L = -\frac{1}{2} \Sigma^{-1}_\epsilon + \frac{1}{2} \Sigma^{-1}_\epsilon(D - EC^{-1}E')\Sigma^{-1}_\epsilon.
\]

Each previous parameter estimate \( \Gamma, \Sigma_\epsilon, F, \) and \( G \) is modified by adding a multiple
of the respective gradient to it (for example \( F_{\text{updated}} = F_{\text{original}} + g \frac{\partial}{\partial F} L \)). The step
size \( g \) (distance in gradient direction between the original and updated estimate)
is inversely proportional to the largest entry in absolute value in \( \frac{\partial}{\partial F} L. \) A steeper
likelihood function in the \( \Gamma \) direction will cause smaller steps in gradient direction
from the original parameter estimates to the updated ones [93]. However, it is
possible that \( \Gamma \) is no longer upper triangular after this gradient update. To correct
for this, all \( \Gamma \) entries below the diagonal equal are made zero. This gradient ascent
is repeated until either the likelihood no longer increases, or the potential update of
\( \Sigma_\delta \) is no longer positive definite.
5.4.2 Partial Update

Sometimes, even a small step in the gradient direction will not increase the model likelihood. In these cases, updating only the gene regulation matrix $\mathbf{G}$ and the measurement covariance matrix $\Sigma_c$, while retaining the previous estimates of $\mathbf{F}$ and $\Gamma$ may still increase the overall model likelihood. To achieve this, after the next Kalman smoothing step, the updated state variable estimates $\hat{\mathbf{Y}}_t$ change the maximum likelihood estimates of the model parameters so that an improvement of the parameters $\mathbf{F}$ and $\Sigma_\beta$ may become possible.

5.5 Termination Criterion

After each KM-algorithm iteration of Kalman smoothing and maximizing or increasing the likelihood via gradient ascent, the value of the model likelihood is computed. Iterations are continued until the relative increase in model log-likelihood is less than a pre-determined amount

$$\frac{\text{log-likelihood}_{(\text{new})} - \text{log-likelihood}_{(\text{old})}}{|\text{log-likelihood}_{(\text{old})}|} \leq c.$$  

Equation 5.12 is used for the termination criterion, because it is versatile and yields reasonable results for a range of possible values for size (gene number) and length (time points) of simulated data. The magnitude of the model log-likelihood depends strongly on the dimensions of the data set and ranges from -600 for $T = 20$ and $n = 500$ to -50.000 for $T = 100$ and $n = 2000$. A termination criterion value of $c = 0.0005$ leads to estimates of the gene regulation matrix $\mathbf{G}$ which does not improve significantly in simulations (to an accuracy of 16 digits), when more iterations are performed. The magnitude of the termination criterion was determined experimentally via simulations.

If a precise estimate of the underlying regulators is desired, then the algorithm can be run again on a smaller set of selected genes with a smaller termination criterion value (e.g., $c = 0.00001$). In rare circumstances the algorithm fails to converge
even after a large number of steps. This may be caused by an initial value for the
gene regulation matrix $G$, that causes the parameter estimates to lie in a plateau
of the likelihood surface. For these simulations, the algorithm is terminated after
100 iteration steps and restarted with a different initial value for $G$. Termination of
the KM-algorithm leads to either no estimates or final estimates of the parameters
$\hat{F}, \hat{G}, \hat{\Sigma}_\delta$, and $\hat{\Sigma}_\epsilon$, as well as the maximized log-likelihood value. A flowchart of the
KM-algorithm is presented in Figure 5.1.

5.6 Symmetries in the Likelihood Function

Recall example 3.4.1 which illustrated that the state space representation of an
ARMA process is not unique. The same is true for general state space processes
[37]. If the dimension of the state space is $m$, then there are $m!$ different ways of
renumbering the components of the state vector $Y_t$. These renumberings also lead to
a different system matrix $F$ (in which columns and rows are permuted accordingly)
and a different formulation of the gene expression matrix $G$. Since the conditional
model likelihood function

$$L = L(Y_0, Y_1, \ldots, Y_T|Z_0, \ldots, Z_T)$$

exhibits $m!$-fold symmetry, the function value of $L$ does not change under any permuta-
tion of the vector components of $Y$. Therefore, the conditional likelihood function
has many local maxima that correspond to different model parameters $F, G, \Sigma_\delta$ and
$\Sigma_\epsilon$. Two sets of maximum likelihood parameter estimates may be different, even
though they describe the same state space model.

Recall that the goal of this research is to identify the regulated genes in a time
series observation of gene expression. While every application of the algorithm yields
an estimate of the gene regulation matrix $G$, these different estimates may not be
directly comparable, since they may correspond to differently organized state
vectors. The method that is used to extract the degree of regulation for every gene
Figure 5.1. Flowchart of the proposed KM-algorithm. Here, $Z_t$ are the $n$ dimensional observations and $Y_t$ are the $m$ dimensional unobservable regulators. $F, G, \Sigma_\delta, \Sigma_\epsilon$ are the parameters of the state space model in Equations 3.1 and 3.2.
is independent of the order of the hidden regulators and is described in the next section.

5.7 Regulation Criterion

In order to extract the regulation information for every gene from the estimates of the model parameters of the state space model a criterion needs to be defined. In actuality, gene $k$ is regulated if the gene regulation matrix $G$ of the true underlying model has non-zero entries in the $k^{th}$ row. The estimated gene regulation matrix $\hat{G}$, however, will contain mostly non-zero entries. Therefore, the magnitude of these entries needs to be considered in a criterion for regulation. However, since the magnitude of the row entries of the estimated gene regulation matrix $\hat{G}$ is also influenced by other factors, a modification of the criterion becomes necessary. Both large positive and large negative entries are an indicator of regulation activity. They can be interpreted as up- or down-regulation of the particular gene by a regulator, respectively.

Genes that have a large variance over the time course of observations, also tend to have larger gene regulation matrix entries. Hence, the magnitude of the entries in the gene regulation matrix is not limited to the degree of regulation. To correct for this effect, a regulation criterion as applied here needs to be standardized with respect to the observed gene variation. Let $V_i = Var \left(Z_t^{(i)}, t = 1, \ldots, T \right)$, $i = 1, \ldots, n$ be the observed variance of gene $i$ over time points. Let $\hat{G} = (\hat{g}_{ij})_{i=1,\ldots,n}^{j=1,\ldots,m}$ be the estimated gene regulation matrix resulting from one successful implementation of the KM-algorithm. Define the degree of regulation for every gene $i$ as

$$R_i = \frac{\sum_{j=1}^{m} \hat{g}_{ij}^2}{V_i}, \quad i = 1, \ldots, n.$$

The magnitude of regulation is represented by the sum of squared row entries of the gene regulation matrix, and this value is standardized by the variance over time. This definition of regulation is independent of the ordering of the hidden variables in the
state space. The regulation criterion values \( R_i \) extracted from several estimated gene regulation matrices, obtained by running the algorithm with different initial values for \( G \), may be averaged using this criterion. Note, that since this method extracts the regulation information for every observed gene, it is not necessary to exclude genes with low variance over time \emph{a priori} as is done in many other approaches (i.e., arbitrary thresholds). In fact, simulations will show in Chapter 6 that even genes with low variation over time may be regulated in a particular process.

5.8 Numerical Computation of the KM-Algorithm

The KM-algorithm is implemented in the Matlab computing environment [94]. The most computationally challenging elements of the KM-algorithm are the inverse and the Cholesky decomposition of the \( m \times m \) regulation covariance matrices \( \Sigma_\delta \). Since the dimension \( m \) of the state space is usually moderate, the computations remain numerically feasible. For the large dimensional \( n \times n \) error covariance matrix \( \Sigma_\epsilon \), the determinant and inverse must be computed as well. Since the model assumes this matrix to be diagonal, its inverse is obvious. The determinant, which is needed in the computation of the model log-likelihood, provides a greater challenge. The entries in the diagonal of the error covariance matrix \( \Sigma_\epsilon \) represent the variance of the measurement errors for the respective genes. If a gene is unregulated, these variances will also reflect the temporal variance of the gene. Thus, the observable temporal variance of a gene provides a good estimate for the error variance. However, for any one microarray experiment where thousands of genes may be observed simultaneously the variance over time of many of the genes is small (on the magnitude of 0.1). The product of diagonal elements often lies outside the range of double floating point precision \( 10^{-324} \) [95]. Numerically, the result is zero which makes it impossible to compute the inverse. However, since the inverse of the determinant is needed in the computation of the log-likelihood, a modification of this procedure is necessary.
To allow for a meaningful computation of the determinant of the error covariance matrix, small ($< 1$) entries in the diagonal of $\Sigma_e$ are replaced by 1. Effectively, the error component is consistently overestimated for genes with low temporal variation. The regulation criterion proposed in Section 5.7 relies heavily on the gene expression matrix $G$. While changing the estimate of $\Sigma_e$ will lead to slightly different Kalman smoothing estimates in the K-step of the algorithm, the effect on the gene regulation matrix is small. In Chapter 6 it will be shown that the identification of regulated genes performs well on simulated data sets, even if the measurement errors are overestimated to enable computation.

5.9 Unequally Spaced Observations

The methodology described so far applies only to measurements taken at equally spaced time points. However, this methodology can be extended to the case where the observations are spaced unequally in time. Specifically, the proposed KM-algorithm can be modified for situations where the microarray data are based on a non-uniform sampling.

Wu et al. [89] consider a state space model for gene expression data and suggest a method to extend the model to unequally spaced time points. Let $\Delta t$ be the largest common time unit among all the measured time intervals (Figure 5.2). That means that every time step can be represented as an integer multiple of $\Delta t$. To extend the theory developed in this work, consider the state space model in which

![Figure 5.2](image)

Figure 5.2. $\Delta t$ is the highest common factor between measured time points.
the observation equation (3.1) is unchanged, and the state equation (3.2) is modified to take the length of the time step into account.

\[ Z_{tk} = GY_{tk} + \epsilon_{tk} \quad k = 1, \ldots, T \]
\[ Y_{tk} = F^{jk}Y_{tk-1} + \delta_{tk} \quad \text{where } j_k = \frac{t_k-t_{k-1}}{\Delta t} \text{ is an integer for } k = 1, \ldots, T. \]

In this case the Kalman smoothing estimates of the hidden regulators can be obtained in a similar manner as the process described in Section 3.5. The system matrix \( F \) must be replaced with the appropriate \( F^{jk} \) in both the forward and backward recursion of the Kalman smoothing step. The terms of the conditional likelihood function that depend on the gene regulation matrix \( G \) and the distribution of the measurement error \( \epsilon_{tk} \) remain the same. However, the terms of the conditional likelihood function that depend on \( F \) will no longer be a quadratic expression in \( F \) (as in Equation (5.9)). Instead, they will be a polynomial expression in \( F \). Therefore, the solution of the matrix equation

\[ \frac{\partial}{\partial F} L (Y_0, Y_1, \ldots, Y_T \mid Z_1, \ldots, Z_T) = 0, \]

may not exist in closed form. A gradient ascent in the model parameter \( F \) that increases the model likelihood, however, is still possible, and in turn allows the application of the modified KM-algorithm.

A different approach to dealing with unequally spaced time points is presented by Jones and Tryon [96]. They define a continuous time state space model

\[ dY(t) = FY(t)dt + d\delta(t), \]

where \( d\delta(t) \) is assumed to be continuous time Gaussian white noise with instantaneous covariance matrix \( \Sigma_\delta \),

\[ E [d\delta(t)d\delta(t)'] = \Sigma_\delta. \]

In this case, the observation equation (3.1) becomes

\[ Z(t) = GY(t) + \epsilon(t), \]
where \( \epsilon(t) \) is a random vector of measurement errors which is normally distributed with mean zero and covariance matrix \( \Sigma_\epsilon \), independent of the noise \( d\delta(t) \). The homogeneous portion of the state equation (3.2) is the differential matrix equation
\[
\frac{d}{dt} Y(t) = FY(t),
\]
which is solved by
\[
Y(t + \eta) = \exp(F\eta)Y(t).
\]
The matrix exponential in this expression is defined in terms of its power series
\[
\exp(F\eta) = I + \sum_{k=1}^{\infty} \frac{(F\eta)^k}{k!}.
\]
The authors describe how the Kalman filter can be applied to the continuous time state space model to derive the smoothing estimates of the hidden regulators \( Y_{t_0}, Y_{t_1}, \ldots, Y_{t_T} \). As in the approach of Wu et al. [89], the maximum likelihood estimates of the model parameters can then be obtained from a gradient ascent performed on the conditional likelihood function. In this case, the likelihood will be a function of the model parameter \( F \) in the form \( \exp(F\eta) \), where the values of \( \eta \) depend on the distance between observations.

5.10 Missing Data

In many microarray experiments, some gene expression values are missing [33,35]. Reasons for missing data include manufacturing or hybridization problems for single spots on the array or difficulties with scanning the image. Statistical methods exist for estimating missing observations in microarray experiments [97–99]. The most commonly used method is to replace missing log\(_2\) signal ratios with zeros [100]. Another popular method is to take row averages, thus replacing a missing gene expression value with the average of available observations on this gene under different conditions [98]. When the experimental details include a time series the conditions would be the different time points at which the gene is observed. Neither replacing
missing values by zeros nor averaging over time points takes the correlation structure of the data into account. Troyanskaya et al. [98] suggest two alternatives: one based on singular value decomposition (SVD), and the other based on k-nearest neighbor (KNN) averages. These methods are compared to those predominantly used in practice. The SVD method attempts to find a set of mutually orthogonal expression patterns, such that all gene expression patterns in the data set can be modeled as linear combinations of these “eigengenes.” The “eigengenes” are sorted by their corresponding eigenvalues and the genes corresponding to the k largest eigenvalues are selected. A gene with missing values is regressed against the selected “eigengenes” and the regression coefficients are used to reconstruct the missing value. The KNN method selects the k genes, whose expression profiles are most similar to the gene with missing values as measured by the euclidean distance of the genes. The missing value is obtained as a weighted average of the expression values in the k neighboring genes. The weights are the degrees of similarity that the neighboring genes have with the gene that is missing the observation.

For time series gene expression experiments (with moderate to low noise), Troyanskaya et al. conclude that the SVD method yields better results, especially when the expression data exhibit strong patterns over time. For non-time series or noisy gene expression data, the KNN method shows more robust results than SVD. Both the SVD and the KNN methods outperform the common practice of filling missing values with zeros or computing row averages.

Specifically for the case of time series state space models, Shumway and Stoffer [37,46] have developed a modification that allows for missing values in the observed time series. At a given time point t, they split the observation vector Z_t into the observed part Z_t^{(1)} and unobserved or missing part Z_t^{(2)}

\[
Z_t = \begin{pmatrix} Z_t^{(1)} \\ Z_t^{(2)} \end{pmatrix} = \begin{pmatrix} G^{(1)} \\ G^{(2)} \end{pmatrix} Y_t + \begin{pmatrix} \epsilon_t^{(1)} \\ \epsilon_t^{(2)} \end{pmatrix}.
\]
The error covariance matrix is split similarly,
\[
\Sigma_\epsilon = \begin{bmatrix}
\Sigma_{11}^\epsilon & \Sigma_{12}^\epsilon \\
\Sigma_{21}^\epsilon & \Sigma_{22}^\epsilon
\end{bmatrix}.
\]
Shumway and Stoffer show that if the replacements
\[
Z_t = \begin{pmatrix} Z_t^{(1)} \\ 0 \end{pmatrix}, \quad G = \begin{bmatrix} G^{(1)} \\ 0 \end{bmatrix}, \quad \Sigma_t = \begin{bmatrix} \Sigma_{11}^{\epsilon} & 0 \\
0 & \Sigma_{22}^{\epsilon}
\end{bmatrix}
\]
are made in the Kalman filter at update \( t \), then the result is the “missing data” Kalman smoothing estimate. Effectively, the missing observations can be replaced with zeros, and the corresponding rows of the gene expression matrix should be replaced by zeros as well in each Kalman filtering step. Since this method assumes the measurement errors as uncorrelated, the splitting of the covariance and replacing the off-diagonal block-matrices with zeros has no additional effect.

5.11 Model Selection

The goal of model selection is to choose the most appropriate model that describes the observations in an optimal way. As applied here, the task is to choose the appropriate dimension of the state space, i.e., the number of hidden regulators that are to be considered. The most commonly used model selection criteria are the Akaike information criterion (AIC) [101] and Schwarz’s Bayesian information criterion (BIC) [102]. Both criteria weigh the value of the model log-likelihood against the number of parameters in the model. More complex models usually are reflected in higher likelihood values. In addition to the log-likelihood term, both the AIC and the BIC criterion also contain a penalty term that is increasing in the number of model parameters. To reward parsimony, lower model dimension leads to improved AIC and BIC scores. However, both the AIC and BIC as well as other common model selection methods fail in the analysis of microarray data. Not only is the number of observations extremely large (\( T \cdot n \), where \( n \) is the number of observed genes, and \( T \) is the number of available time points), the number of model parameters (for the state
space model in Equations 3.1 and 3.2: $3m^2 + nm + n + m$, where $m$ is the selected state space dimension) is very large, as well. They both are a function of the number of observed genes $n$, which typically is on the order of thousands for a microarray experiment. In this case, the penalty term of the common model selection criteria entirely dominates the model log-likelihood term and leads to selection of erroneous models.

To solve the model selection dimensionality problem, a time series method of model selection specifically designed for state space models will be employed [103]. It is based entirely on the autocovariances between the observations. When implemented it becomes unnecessary to run the algorithm a priori for a range of different model dimensions which considerably shortens computation time. Consider the block-Hankel matrix

$$
\mathbf{H} = \begin{pmatrix}
\hat{\Gamma}_1 & \hat{\Gamma}_2 & \cdots & \hat{\Gamma}_p \\
\hat{\Gamma}_2 & \hat{\Gamma}_3 & \cdots & \hat{\Gamma}_{p+1} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\Gamma}_p & \hat{\Gamma}_{p+1} & \cdots & \hat{\Gamma}_{2p-1}
\end{pmatrix},
$$

(5.13)

where

$$\hat{\Gamma}_i = \frac{1}{T} \sum_{i=1}^{T-i} Z_{t+i} Z'_{t+i}.$$
**Theorem 5.11.1** (Kronecker)

In the absence of error, the rank of the matrix $H$ in (5.13) equals the number of states $m$ required to characterize the time series $Z_t$ [103–105].

The microarray measurements $Z_t$ considered in this work contain both biological as well as technical measurement errors. Therefore, the rank of the matrix $H$ computed from the observations will not be exactly equal to the state space dimension $m$. However, it is possible to perform a singular value decomposition (SVD) of the matrix $H$ and to consider the magnitudes of the singular values of $H$ in order to select a model dimension $m$. If the signal to noise ratio in the observations is large, then a SVD will yield $m$ singular values of large magnitude. Numerical algorithms for performing singular value decompositions on very large matrices exist [92]. They are not time extensive, but require a large amount of available memory (e.g., 1.6 GB for a $6000 \times 6000$ matrix). Simulation results for the proposed model selection method will be described in Section 6.7.
6. SIMULATED DATA AND KM-ALGORITHM RESULTS

In order to evaluate the proposed KM-algorithm under different circumstances, data sets are simulated to represent microarray gene expression experiments. The influence of variables such as the number of available observations \(T\) and the number of observed genes \(n\) on the ability to identify regulated genes is investigated. All data sets are simulated with the same underlying state space structure. The \(m\)-dimensional regulators are simulated first. Their development over time is controlled by the \(m \times m\) system matrix \(\mathbf{F}\). The choice of the system matrix \(\mathbf{F}\) together with the regulation covariance matrix \(\mathbf{\Sigma}\) allows, for example, the simulation of the regulators as AR processes of a certain order. The observations are then simulated as linear combinations of the regulators at each respective time point together with an additive measurement error term. Since genes may be regulated or not, the regulated genes are those which have at least one non-zero factor in the linear combination of regulators. The unregulated genes are modeled as multivariate normal noise with different variances over time.

**Example 6.0.1** Consider the following setup. The regulation is modeled in the form of five independent AR\((p)\) processes of varying degree \(p\) \((p = 1, 2, 3)\).

\[
x_t^{(1)} = -0.4x_{t-1}^{(1)} + a_t^{(1)}, \quad (p = 1)
\]

\[
x_t^{(2)} = 0.6x_{t-1}^{(2)} + a_t^{(2)}, \quad (p = 1)
\]

\[
x_t^{(3)} = 0.3x_{t-1}^{(3)} + 0.1x_{t-2}^{(3)} + a_t^{(3)}, \quad (p = 2)
\]

\[
x_t^{(4)} = 0.4x_{t-1}^{(4)} - 0.3x_{t-2}^{(4)} + 0.5x_{t-3}^{(4)} + a_t^{(4)}, \quad (p = 3)
\]

\[
x_t^{(5)} = 0.7x_{t-1}^{(5)} + 0.2x_{t-2}^{(5)} - 0.1x_{t-3}^{(5)} + a_t^{(5)}, \quad (p = 3)
\]
where the $a_t^{(i)} \sim N(0, 1), (i = 1, \ldots, 5)$ represent the biological regulation variation. The ten regulators $Y_t = (Y_t^{(1)}, \ldots, Y_t^{(10)})$ represent the present and past $p$ states of these regulating AR($p$) processes. For example,

$$
\begin{pmatrix}
Y_t^{(1)} \\
Y_t^{(2)}
\end{pmatrix}
= 
\begin{pmatrix}
-0.4 & 1 \\
0.1 & 0
\end{pmatrix}
\begin{pmatrix}
Y_{t-1}^{(1)} \\
Y_{t-1}^{(2)}
\end{pmatrix}
+ 
\begin{pmatrix}
\epsilon_t^{(1)} \\
\epsilon_t^{(2)}
\end{pmatrix},
$$

where

$$
\begin{pmatrix}
\epsilon_t^{(1)} \\
\epsilon_t^{(2)}
\end{pmatrix}
\sim MVN
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\right).
$$

To simulate all ten regulators simultaneously, the system matrix $F$ can be written as a block-diagonal matrix with entries as in the system matrix in the 2-dimensional example above. The regulation covariance matrix can be written in block-diagonal form as well. The regulators are chosen to represent a wide variety of different circumstances. Positive and negative autocorrelation, as well as small and large variance over time (Figure 6.1), can be considered.

To simulate regulated and unregulated genes, the gene regulation matrix $G$ is formulated to contain non-zero entries only in the first $m$ rows. The remaining $n - m$ rows are zero.

$$
G = 
\begin{pmatrix}
* & * & \cdots & * \\
: & : & \cdots & : \\
* & * & \cdots & * \\
0 & 0 & \cdots & 0 \\
: & : & \cdots & : \\
0 & 0 & \cdots & 0
\end{pmatrix}
\begin{cases}
m \text{ regulated} \\
(n - m) \text{ unregulated}
\end{cases}
$$

The measurement error for the regulated genes is independent Gaussian noise with variance 0.1. The unregulated genes are simulated to be independent Gaussian noise. Their variances are simulated in different magnitudes to assure that the variance over time of the unregulated genes covers the range of variances of the regulated genes. However, as is the case in most biological experiments, many of the observed genes (in
Figure 6.1. Regulators from Example 6.0.1 in a simulation for \( T = 40 \) time points. (a) Regulators \( Y_1 \) and \( Y_8 \) with different autocorrelation; (b) Regulators \( Y_2 \) and \( Y_9 \) with different variance.

In fact, half of the unregulated genes are simulated with a small variance (\( \sim 0.1 \)) over time. Without loss of generality, the regulated genes are the observations numbered \( 1, \ldots, m \), while the unregulated genes are numbered \( m + 1, \ldots, n \).
Example 6.0.2 To investigate the influence that the magnitude of the entries in the gene regulation matrix $G$ has when compared to the variance of the simulated regulators, the simulated genes require different structures. Since the simulated regulated genes are linear combinations of the hidden regulators, the factors in the linear combination determine the magnitude of entries in the rows of the gene regulation matrix $G$ as well as the temporal variance of the simulated gene. To simulate twenty regulated genes with different behaviors, the regulation structure can be chosen in the following way:

$$
\begin{align*}
Z_1 &= Y_6, & Z_{11} &= 0.64Y_3 - 0.64Y_5, \\
Z_2 &= 1.5Y_6, & Z_{12} &= Y_4, \\
Z_3 &= -Y_6, & Z_{13} &= Y_5, \\
Z_4 &= Y_4 + Y_6 + Y_9, & Z_{14} &= 0.5Y_5, \\
Z_5 &= 0.8Y_1 + 0.8Y_3, & Z_{15} &= Y_9, \\
Z_6 &= 2Y_7 + 4.5Y_{10}, & Z_{16} &= Y_{10}, \\
Z_7 &= 0.1(Y_1 + Y_2 + Y_3 + Y_5 + Y_8), & Z_{17} &= 0.5Y_2 - Y_6 + 3Y_{10}, \\
Z_8 &= 0.3(Y_1 + Y_2 + Y_3 + Y_5 + Y_8), & Z_{18} &= 0.8Y_2 + 0.1Y_6 - Y_{10}, \\
Z_9 &= Y_1 - 0.5Y_2, & Z_{19} &= 0.6Y_8, \\
Z_{10} &= 5Y_4 - 5Y_9, & Z_{20} &= -0.6Y_8.
\end{align*}
$$

Here, for example, regulator $Y_6$ regulates gene $Z_1$, whereas gene $Z_4$ is regulated by regulators $Y_4, Y_6$ and $Y_9$. The sum of squares of entries in the rows of the corresponding gene regulation matrix $G$ (whose estimate will be used in the regulation criterion) ranges from 0.05 (for gene $Z_7$) to 25 (for gene $Z_{10}$). Since the regulators themselves have different variances, the variances of the regulated genes vary too. In order to be able to compare the effects of regulator variance over time and magnitude of the entries in the rows of the gene regulation matrix, some of the genes have been simulated to have comparable variance but from different sources. The effects of magnitude of entries in the gene regulation matrix can be studied by comparing regulation results for genes which are regulated by a single regulator, but with different magnitudes (e.g., $Z_1, Z_2$ and $Z_3$, or $Z_{13}$ and $Z_{14}$). The effects of regulator variance
can be studied directly by comparing genes who are simulated to be equivalent to a regulator (e.g., \(Z_1, Z_{12}, Z_{13}, Z_{15},\) and \(Z_{16}\)).

### 6.1 Data Sets used in the Simulation

To compare the performance of the proposed method on data sets of different size \(n\) (gene number) and length \(T\) (time points), nine data sets for different values of \(T\) and \(n\) are simulated. For comparability, all of the data sets have the same structure as described in the Examples 6.0.1 and 6.0.2. For all possible combinations of three different time lengths

\[
T = 20, 40, 100
\]

and three different gene numbers

\[
n = 500, 1000, 2000
\]

nine data sets are created. For each data set, the ten \((m = 10)\) underlying regulators described in Example 6.0.1 were used to model twenty regulated genes (as described in Example 6.0.2). Note that all eigenvalues of the \(F\) matrix described in Example 6.0.1 lie within the unit circle. This assures that the regulators \(Y_t\) and consequently the observations \(Z_t\) are stationary time series observations. Most biological gene expression time series observations, such as the yeast cell cycle for example, fall into this category.

### 6.2 Model Likelihood

Dempster et al. [59] showed that the model likelihood increases monotonously under a generalized EM-algorithm. For the time series scenario considered here the model likelihood increases dramatically in the first few iterations and then stabilizes rather quickly. Usually, after ten or fewer iterations of the algorithm, the value of the model log-likelihood increases by less than 0.05%. The estimate of the gene regulation matrix \(G\) does not change (in any of the first significant 16 digits) if the
Figure 6.2. Increase in the model log-likelihood for different data sets: (a) Real microarray data sets. The *Caulobacter crescentus* data set from Laub et al. [33] as well as the *Synechocystis* data set from Gill et al. [34]. (b) Simulated data with $n=500$ genes.
termination criterion \((c)\) is changed from 0.05\% to 0.001\%. Small changes do occur, however, in the Kalman smoothing estimates of the hidden regulators which lead to continued small changes in the model log-likelihood. If the KM-algorithm is run with the purpose of identifying regulated genes, then faster run-time through fewer iterations is preferable to more precise regulator estimates. Once the regulated genes have been identified, the algorithm can be run again on a smaller set of genes with a smaller termination criterion \((c)\) to obtain estimates for the hidden regulators. Note that the absolute value of the model likelihood depends on the number of observations made (Figures 6.2 and 6.3). If thousands of genes are observed simultaneously at many time points, then the model likelihood is a product of thousands of terms, many of which may be quite small. In these cases, the initial values of the model likelihood are large negative numbers and the maximum likelihood for the proposed model parameters is a large negative (or small positive) number as well. Since the magnitude of the log-likelihood depends so strongly on the length (number of time
points) and size (number of genes) of the simulated data set, the termination criterion is formulated as a percentage rather than an absolute increase.

6.3 Regulation Criterion

The model estimates obtained by the KM-algorithm will be used to determine the degree of regulation for each observed gene. The most important parameter in this application of the algorithm is the gene regulation matrix $G$. However, genes that tend to have larger expression values also tend to have larger variances over time, as well as larger entries in the gene regulation matrix (Figure 6.4). Therefore, the temporal variance of a gene must be taken into account. An estimate for this temporal variance can be easily obtained from the observations. According to the proposed model a gene is regulated, if it has (large) non-zero entries in the gene regulation matrix. The proposed regulation criterion is the sum of the squared gene regulation matrix row entries standardized by the temporal variance of the gene (Figure 6.5).

$$R_i = \frac{\sum_{j=1}^{T} g_{ij}^2}{\text{Var}(Z^{(i)})},$$

The genes for which the value of the regulation criterion is large, are those most likely to be regulated in the observed biological process. For the discovery of regulatory networks, as previously discussed it is often necessary to reduce the dimension of the data set by excluding certain genes from the analysis. The more popular methods [50,84,88] often exclude between 20% and 90% of the observations by arbitrary pre-screening. Excluding genes according to their degree of regulation, as measured by the proposed regulation criterion, assures that the genes likely to be involved in the observed process are not excluded from the analysis. If resources are available to study single genes in more detail, for example by knock-out experiments, then this effort should be focused on the genes which are determined to be the ones most likely to be regulated in the observed process.
Figure 6.4. Observations on $n = 1000$ genes at $T = 100$ time points are simulated according to the model in Example 6.0.1. The KM-algorithm is implemented five times with different initial values. This plot shows the averaged sum of squares of the row entries in the gene regulation matrices $G$ against the temporal variance of each gene. The simulated regulated genes are plotted as red stars and the unregulated genes as black dots.

6.4 Goodness of Ranking

To compare the results from applying the KM-algorithm to data sets of different size (gene number) and length (observed time points), a measure is needed that compares the ranking of genes according to their degree of regulation with the true underlying regulation structure. A good ranking result is one in which the gene regulation criterion assigns high values to the regulated genes and low values to the unregulated genes. If the genes are sorted by the value of their regulation criterion (Figure 6.6), the regulated genes should be highly ranked (i.e., at the top of the list).
Figure 6.5. Observations on \( n = 1000 \) genes at \( T = 100 \) time points are simulated according to the model in Example 6.0.1. The KM-algorithm is implemented five times. The averaged proposed regulation criterion is plotted against the temporal variance of each gene. The simulated regulated genes are plotted as red stars and the unregulated genes as black dots.

For simulated data it is known which of the genes are regulated. A ranking resulting from application of the KM-algorithm can therefore be evaluated by the position of those genes. The proposed novel “goodness of ranking” (GR) measure computes the average position of the regulated genes and compares it to the average position of the unregulated genes. Suppose there are \( l \) regulated genes and let \( a_i, i = 1, \ldots, l \) be the ranking positions of the regulated genes as obtained by the KM-algorithm. Then

\[
GR = \frac{1}{l} \left( \frac{n(n+1)}{2} - \frac{(n-l)(n-l+1)}{2} \right) - \frac{n+1}{2}
\]
Figure 6.6. Principle behind the proposed goodness of ranking measure to evaluate the performance of the regulation ranking.

\[ = 2l \left( \frac{\bar{a} - \frac{n+1}{2}}{n(n+1) - (n-l)(n-l+1) - l(n+1)} \right). \]

For an average ranking, the mean position of the regulated genes is \( \frac{n+1}{2} \). For the optimal ranking, in which the \( l \) regulated genes appear on top of the list of \( n \) observed genes the sum of their positions is \( \frac{n(n+1)}{2} - \frac{(n-l)(n-l+1)}{2} \). The measure is standardized to yield a score of one for a perfect ranking and a score of zero for an average random ranking. Note, that the measure can take on negative values when the regulated genes are ranked toward the bottom of the list.

### 6.5 Power of Regulation Criterion

To study the power of the proposed KM-algorithm via simulations, the KM-algorithm is applied each of 1000 times to three sets of simulated time series gene expression data of different length. For comparability, all data sets are simulated with the same number of genes \( (n = 500) \) and the same structure described in Examples 6.0.1 and 6.0.2. The data sets are simulated to have \( T = 20 \), \( T = 40 \), and \( T = 100 \) observations, respectively. In all three cases, 20 of the simulated genes are regulated and the remaining 480 genes are unregulated.
The regulation criterion as described in Section 6.3 derives part of its power from the fact that it is independent of the formulation of the state space model. Hence, the KM-algorithm can be run several times and the regulation criterion results may be averaged to increase the power of identifying regulated genes. In Figure 6.7 the GR results are shown for combining the regulation criterion results of \( k \) runs of the KM-algorithm, each with a different initial value for the gene regulation matrix \( G \). For example, the KM-algorithm is run 1000 times on the simulated data set with \( T = 20 \) and \( n = 500 \). The resulting 1000 estimates of the gene regulation matrix \( G \) are grouped into 1000 groups of \( k = 1 \), or 500 groups of \( k = 2 \), or 333 groups of \( k = 3 \), etc. For each group, the gene regulation criteria are averaged to yield one ranking

![Figure 6.7](image_url)

Figure 6.7. Goodness of ranking obtained from averaging \( k \) independent regulation criterion results to assess the power of the KM-algorithm. Implementing the KM-algorithm more often and averaging the results provides more consistent and improved ranking results.
of all 500 observed genes according to their (estimated) degree of regulation. This result is compared to the true underlying structure of the simulated data set and the GR measure for the goodness of ranking is computed. This approach yields 1000 GR measurements for $k = 1$, 500 GR measurements for $k = 2$, 333 GR measurements for $k = 3$, etc. For each $k$, the mean of the GR measurements is computed together with its standard error. This procedure is repeated for the other data sets with $T = 40$ and $T = 100$ observations and the results are presented in Figure 6.7. There is a trade off between combining more results to increase power and increased run-time of the KM-algorithm. The increase in run-time is approximately linear in $k$ (run-times may differ slightly with different initial values). However, the increase in power is less than linear in $k$ (compare Figure 6.7). A value that balances increased run-time against increased power is approximately $k = 5$.

Clearly, the KM-algorithm performs better if the data set contains more time point observations. Depending on the degree of separation between regulated and unregulated genes an experiment intends to achieve, it should include a certain minimal number of observations in time. The state of microarray experiments today is limited to the order of 20-40 time points by the cost of the technology. However, it is expected in the near future, that the cost of this technology will be more affordable, thus allowing experiments of a magnitude that best exploits the KM-algorithm approach. This simulation illustrates that the identification of regulated genes in a data set with $T = 40$ observations is clearly superior to the case of only $T = 20$ observations. In fact, the power of separation between regulated and unregulated genes as indicated by the GR-measure is on average more than twice as high in this case. Microarray experiments with time point observations on the order of $T = 40$ are currently being conducted and are expected to become increasingly common.

Simulations such as the one here provide valuable information that can be used when designing microarray experiments. Recommendations can be made to biologists regarding the minimum number of time points and the maximum time between
observations in a time dependent microarray experiment. While these factors certainly also depend on the biological process that is to be observed, clear guidelines (e.g., the minimum number of time points) can help to design more efficient experiments. The analysis of the data derived from those experiments will allow the discovery of genetic networks with a higher degree of accuracy.

6.6 Gene Variance and Regulation

Currently, many methods [50, 84, 88] that are employed to discover gene regulatory networks utilize some sort of pre-screening prior to the analysis. Often, genes are excluded according to criteria such as that their maximum observed expression over time does not exceed an arbitrary cut-off value. Thus, genes with naturally low variation over time have a higher chance of being excluded \textit{a priori} from the analysis. Simulations show that even genes with a low variance over time, whose expression values are consistently small, may be highly likely to be regulated in a certain biological process. These genes should not be excluded from an attempt to reconstruct regulatory networks, as they may be highly relevant to the discovery process. In Figure 6.8, the results of applying the KM-algorithm to the nine simulated data sets described in Section 6.1 are shown. For comparability of the results, the KM-algorithm is run five times for each data set at the true state space dimension \( m = 10 \). The regulation results are averaged, and the resulting regulation criterion value is plotted against the maximum absolute expression value over the time course of observations for each gene. Maximum absolute expression fold-change is currently used as a measure to pre-screen genes prior to analysis. A popular and arbitrary threshold value [53, 84] is 2-fold absolute expression change. Selecting genes according to their maximum absolute expression change alone selects the genes in Figure 6.8 that are to the right of some arbitrarily chosen threshold value (e.g., maximum absolute expression change of 2). In all cases, this would disregard the regulated genes (red stars in Figure 6.8) with low overall fold-change. Selecting genes ac-
Figure 6.8. For nine simulated data sets of different size $n(500, 1000, 2000)$ and length $T(20, 40, 100)$, the KM-algorithm is implemented five times each. The regulation results are averaged and the combined regulation criterion value for each gene is plotted on the $y$-axis. On the $x$-axis, the maximum absolute expression value of the gene over the time course of observations is plotted. Regulated genes are shown as red stars and unregulated genes are shown as black dots.
According to the proposed KM-algorithm based regulation criterion takes the internal correlation structure of the data into account. Gene selection based solely on the proposed regulation criterion would select genes with high regulation criterion values. Especially for the longer time series \(T = 100\) this selection method results in significantly more regulated genes being selected for further analysis. For shorter time series a combination of both the new regulation criterion approach and the old maximum absolute gene expression cutoff is recommendable. Genes could be selected for further analysis if either their regulation criterion value is high or their maximum absolute expression value exceeds a threshold. Current methods remove a certain percentage of observed genes from further analysis (e.g., 50\% or 80\%) based on their maximum absolute expression value. If the same percentage of genes were to be removed through selection by the proposed method then the KM-algorithm approach would assure that fewer regulated genes were among them.

6.7 Model Selection

The model selection method presented in Section 5.11 is applied to the nine data sets described in Section 6.1 and its performance investigated. The data sets have different length \(T = 20, 40\) and 100\) and size \(n = 500, 1000\) and 2000\). For comparability they all have the same structure described in Examples 6.0.1 and 6.0.2. Particularly, the true state space dimension for each of the simulated data sets is \(m = 10\). The maximum relevant biological time lag \(p\) is the amount of time that can lie between change in a regulator and the resulting change in the gene that is being regulated. It depends not only on the observed biological process, but also on the time between observations in the experiment. In the case of the nine data sets used here for simulation, the true maximum relevant biological time lag is \(p = 3\), since the hidden regulators are modeled as \(AR(p)\) processes with rank of up to \(p = 3\). Most microarray experiments are designed so that the value of \(p\) does not exceed 3. To study the influence of the choice of the maximum relevant time lag \(p\) on the
model selection process, model selection is conducted for the nine data sets with three different values of \( p \) (\( p = 1, 2, 3 \)). For each data set and a value of \( p \) the Hankel matrix

\[
H = \begin{pmatrix}
\hat{\Gamma}_1 & \hat{\Gamma}_2 & \cdots & \hat{\Gamma}_p \\
\hat{\Gamma}_2 & \hat{\Gamma}_3 & \cdots & \hat{\Gamma}_{p+1} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\Gamma}_p & \hat{\Gamma}_{p+1} & \cdots & \hat{\Gamma}_{2p-1}
\end{pmatrix},
\]

with

\[
\hat{\Gamma}_i = \frac{1}{T} \sum_{t=1}^{T-i} Z_{t+i} Z_t'
\]

is computed from the observations (compare Definition 3.3.1 in Section 3.3). A singular value decomposition (SVD) for \( H \) is performed

\[
H = USV',
\]

where \( S \) is a diagonal matrix, whose diagonal entries \( \lambda_1, \lambda_2, \ldots, \lambda_{np} \) are the singular values of \( H \). The singular values are ordered by size \( \lambda_1 > \lambda_2 > \cdots > \lambda_{np} \) and can be scaled by the value of the largest singular value:

\[
1, \frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_1}, \ldots, \frac{\lambda_i}{\lambda_1}, \ldots, \frac{\lambda_{np}}{\lambda_1}.
\]

The purpose of scaling is to make the singular value magnitudes comparable for different data sets. Note, that in general singular values depend on the units that the observations (gene expressions) are measured in and may not be directly comparable for biological experiments on different technological platforms unless scaled. Figure 6.9 shows the scaled magnitudes of the 100 largest singular values for each data set computed for three different values of the biological time lag (top \( p = 1 \), middle \( p = 2 \), bottom \( p = 3 \)). Note that each curve in these graphs has a characteristic “S” shape. The magnitudes decrease rapidly for the first singular values, then they decrease more moderately, and finally they sharply decrease to zero. In fact, if a microarray experiment contains \( T \) time point observations, then only the first \( T - 1 \) singular values will be different from zero.
Figure 6.9. Magnitude of singular values for nine data sets of different size $n$ and length $T$ for lag-values $p = 1$ (top), $p = 2$ (bottom). The solid line represents 80% of the largest singular value. The dotted lines indicate the selected model dimension for each data set.
Figure 6.10. Magnitude of singular values for nine data sets of different size \( n \) and length \( T \) for lag-values \( p = 3 \). The solid line represents 80\% of the largest singular value. The dotted lines indicate the selected model dimension for each data set.

This stems from the fact that the gene regulation matrix \( \mathbf{G} \) is unidentifiable in a state space model in which the dimension of the state space \( m \) is larger than or equal to the number of available observations \( T \). Each data set of simulated observations is represented by one curve in Figures 6.9 and 6.10. Selecting the most appropriate state space dimension for the data set means identifying the number of singular values of comparatively large magnitude for each data set. A good estimate would be provided by locating the first reflection point of the singular value magnitude curve for each data set and using the corresponding singular value number as an estimate for \( m \). However, since none of the curves are absolutely smooth, this reflection point may not always be easy to find. Instead, a model selection method is needed that can be used in practice by biologists. While ad hoc, note that the reflection points for all simulated data sets occur at a singular value magnitude of approximately 0.8. Counting the number of singular values for each data set, whose magnitude is greater
than or equal to 80% of that of the largest singular value, leads to an acceptable estimate for the model dimension $m$. In general, using an 80% cutoff value selects the number of singular values of comparably large magnitude, which can be used as an estimate for the state space dimension $m$.

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th>$T = 40$</th>
<th>$T = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 500$</td>
<td>7</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>6</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>$n = 2000$</td>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>23</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 6.1
Selected model dimensions for nine simulated data sets using different maximum relevant biological time lags $p$. Top $p = 1$, middle $p = 2$, bottom $p = 3$. The true model dimension in each case is $m = 10$.

Table 6.1 lists the selected model dimensions according to this method for the biological time lags $p = 1, 2, 3$. Recall, that the true state space dimension for each data set is $m = 10$.

6.7.1 Stability of Regulation Ranking with Respect to Model Selection

If the KM-algorithm is run for a different state space dimension $m$, then the regulation criterion values obtained by the algorithm will vary slightly. However, results are consistent, even if the algorithm is implemented for another state space dimension than the one used to generate the data. In Figures 6.11 and 6.12, ranking results as obtained from implementing the algorithm with the true state space dimension ($m = 10$) are compared to the results obtained from implementing the algorithm
Figure 6.11. Quantile plots comparing the ranking results obtained with the true state space dimension to the results obtained from implementing the KM-algorithm with the state space dimension obtained through model selection. Top: Comparing selected state space dimension $m = 4$ to true state space dimension $m = 10$ for $n = 500$ genes simulated at $T = 100$ time points. Bottom: Comparing selected state space dimension $m = 9$ to true state space dimension $m = 10$ for $n = 1000$ genes simulated at $T = 100$ time points.
Figure 6.12. Quantile plots comparing the ranking results obtained with the true state space dimension to the results obtained from implementing the KM-algorithm with the state space dimension obtained through model selection. Comparing selected state space dimension \( m = 18 \) to true state space dimension \( m = 10 \) for \( n = 2000 \) genes simulated at \( T = 100 \) time points.

with the state space dimension selected by the autocovariance method using maximum relevant biological time lag \( p = 1 \). For the three data sets of length \( T = 100 \) and size \( n = 500, 1000 \) and \( 2000 \) described in Section 6.1, the KM-algorithm is implemented five times with different initial values for the gene regulation matrix \( G \) at both the true state space dimension and the one obtained through model selection. The regulation criterion results from the five runs are averaged and the performance of the method at the true state space dimension is compared to that at the selected one. Figures 6.11 and 6.12 show three quantile-plots for the standardized regulation criterion results of the two implementations for each data set. Clearly, the regulation results are most consistent if the selected state space dimension \( (m = 9) \) is close to the true dimension of the state space \( (m = 10) \) (Figure 6.11, bottom). In this case, the point cloud is distributed close to the diagonal. If the selected state space dimension either underestimates (Figure 6.11, top) or overestimates (Figure 6.12)
the true state space dimension by a wider margin, then the point cloud is spread out further from the diagonal in the quantile plot. However, genes that receive high regulation criterion values at the true state space dimension also consistently receive high regulation criterion values at the selected model dimension.

6.7.2 Stability of GR with Respect to Model Selection

Here, the stability of gene ranking as determined by the KM-algorithm with respect to model selection as a function of maximum relevant biological time lag $p$ is investigated. The results are compared using the goodness of ranking measure described in Section 6.4. Different choice of maximum relevant biological time lag $p$ leads to different autocovariance matrices $H$ that are being considered in the model selection process. This may lead to a different optimal state space dimension $m$ that is chosen for a data set. Figure 6.13 illustrates that the identification of regulated genes does not significantly change if a different maximum relevant biological time lag $p$ is chosen. The GR measure is used to quantify the ranking results. Specifically, Figure 6.13 shows the GR-measure of a ranking outcome produced by the KM-algorithm for the nine different data sets described in Section 6.1 in nine graphs. In each graph, the value of GR obtained from applying the KM-algorithm at the true model dimension $m = 10$ is compared to applying the KM-algorithm at the state space dimension $m$ chosen by model selection with maximum biological time lag $p$ ($p = 1, 2, 3$). In fact, the magnitude of the GR measure is similar for each data set, independent of which maximum relevant time lag $p$ was used for model selection. Even though the true maximum relevant biological time lag value $p$ for a microarray experiment may be unknown, and an estimate is provided by the experimenter, its choice is not essential for the performance of the KM-algorithm method.
Figure 6.13. Goodness of ranking (GR) obtained from running the KM-algorithm five times at state space dimension \( m \) and averaging the results. Shown here are GR results for the true dimension \( (m = 10) \) and the dimensions obtained by model selection with maximum relevant biological time lag \( p = 1, 2, 3 \).

6.7.3 Stability of GR with Respect to Initial Value Choice

A different choice of initial value for the gene regulation matrix \( G \) leads to slightly different parameter estimates for the state space model in Equations (3.1) and (3.2). To study the effect that the choice of initial value has on the regulation estimation, the order of observed genes in the data set with \( n = 500 \) genes observed at \( T = 100 \) time points is permuted one hundred times. The KM-algorithm is applied five times
to each gene order permutation with the same five initial values for the gene regulation matrix $G$. The regulation criterion results are averaged and the goodness of ranking (GR) measure is computed for each gene permutation. Effectively, permuting the gene order has the same effect as permuting the rows of the matrix that is being used as an initial value for the gene regulation matrix $G$. Figure 6.14 shows the resulting GR values for the one hundred permutations of gene order. The GR magnitude is comparable for the different permutations. While the choice of a different initial value for the gene regulation matrix $G$ leads to slightly different regulation criterion results, the magnitude of quality of the ranking result remains the same. This result also demonstrates that the order in which the genes are recorded has no influence on the distinction between regulated and unregulated genes.

Figure 6.14. Goodness of ranking (GR) obtained from averaging the regulation criteria resulting from five implementations of the KM-algorithm. The same five initial gene regulation matrices $G$ are used on a data set with $T = 100$ observations on $n = 500$ genes in which the order of genes is permuted randomly one hundred times.
6.8 Run-Times

The computer code for the KM-algorithm is written in Matlab [94]. The applications were implemented on an IBM 375 MHz POWER3-II processor with 768 MB of main memory. Run-times differ slightly with the choice of the initial value for the gene regulation matrix $G$. Besides the number of observed genes $n$ and the number of time points $T$, the run-times also depend on the selected value of the termination criterion for the algorithm (see Section 5.5). In Table 6.2, the run-times in minutes for one run of the KM-algorithm for different values of $T$ and $n$ are shown.

<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th>$T = 40$</th>
<th>$T = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 500$</td>
<td>1.5 - 3 minutes</td>
<td>2.5 - 4 minutes</td>
<td>5 - 7 minutes</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>10 - 20 minutes</td>
<td>16 - 22 minutes</td>
<td>45 - 60 minutes</td>
</tr>
<tr>
<td>$n = 2000$</td>
<td>75 - 90 minutes</td>
<td>130 - 150 minutes</td>
<td>350 - 450 minutes</td>
</tr>
</tbody>
</table>

Table 6.2
Run-times for the KM-algorithm with a termination criterion value of $c = 0.0005$.

6.9 Partitioning of Large Data Sets

The data sets resulting from real biological microarray experiments are often very large in size. Some plant genomes contain on the order of 26,000 genes which are spotted on one or more arrays. The KM-algorithm involves matrices of order $n$ (number of observe genes), and its run-time is on the order of $O(n^2)$. For very large matrices, this implies substantial computation time and it requires large amounts of available memory for computation. When dealing with extremely large data sets an alternative approach is suggested. This alternative involves randomly splitting a large data set into several smaller ones. Figure 6.15 details how the observed genes in the large data set (a) are randomly permuted in order (b). The permuted data set is split into several smaller parts (c). Model selection is conducted by computing the
autocovariance matrix for each smaller data set separately and performing singular value decompositions. The resulting singular values for the smaller sets are combined and the number of singular values of large magnitude in each set determines the model dimension \((d)\). For each smaller data set, the KM-algorithm is run several times at estimated state space dimension \(m_i\) with different initial values for the gene expression matrix \(G\) \((e)\). Regulation results are combined, first the regulation criterion results from step \((e)\) for each smaller set are averaged, and finally the results for the smaller sets are combined to yield one regulation criterion value for each gene in the large data set. The genes are returned to their original order \((f)\) and the thus obtained regulation result may be combined with other similarly obtained results for different gene permutations to yield one final regulation result \(R\) \((f)\). This partitioning method can be implemented considerably faster than applying the KM-algorithm to the large data set as a whole. Especially, if a computer cluster is

![Figure 6.15](image-url)
utilized in which several computations can be carried out at the same time, as the computations in parts (b)-(f) are entirely parallel.

Unfortunately, splitting a large data set of gene expression observations into multiple parts ignores the correlations between observations in disjoint sets. If gene expression is regulated by hidden factors that are not genes (such as protein levels), then splitting has no influence on the regulation structure. However, if some of the regulators are the genes themselves, then information may be lost through the splitting procedure. To minimize the loss of observation through the splitting procedure, it is recommendable to repeat the splitting for randomly permuted observations several times. Since the KM-algorithm performs better on sets with fewer genes (compare Figure 6.13), the results obtained from the smaller data sets may actually be more powerful in identifying regulated genes than applying the method to the large data set as a whole.

Example 6.9.1 Three of the data sets described in Section 6.1 contain observations on \( n = 2000 \) genes. These data sets have been analyzed as a whole using the KM-algorithm and the same data sets were analyzed again with the partitioning method. For this purpose, the 2000 genes have been randomly permuted in order five different times. For each of the five permutations, the observations are split into four smaller data sets of \( n' = 500 \) genes each. Model selection is performed for each of the \( 5 \times 4 = 20 \) smaller data sets separately, with maximum biological relevant time lag \( p = 3 \). The singular values obtained from SVD of the autocovariance matrix \( \mathbf{H} \) are ordered by size and scaled. The number of singular values of magnitude greater than 0.8 is counted and used as an estimate for state space dimension \( m \). On each of the 20 smaller data sets, the KM-algorithm is run three times at the selected model dimension \( m \). The results are combined to yield one final regulation result for each gene in the large data set. A comparison of the goodness of ranking (GR) result for the three data sets with \( n = 2000 \) genes and \( t = 20, 40, \) and 100 observations, respectively, can be found in Table 6.3. For the shorter data sets \( (T = 20 \) and \( T = 40 \) observations), the results obtained via the partitioning method significantly
<table>
<thead>
<tr>
<th></th>
<th>$T = 20$</th>
<th>$T = 40$</th>
<th>$T = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR for complete data set ($m = 10$)</td>
<td>0.0421</td>
<td>0.1543</td>
<td>0.6434</td>
</tr>
<tr>
<td>GR for partitioning method</td>
<td>0.2013</td>
<td>0.2982</td>
<td>0.5349</td>
</tr>
</tbody>
</table>

Table 6.3
Comparison of regulation results for three data sets with $n = 2000$ genes and $T = 20, 40,$ and $100$ observations, respectively. The regulation results obtained by applying the KM-algorithm to the large data set as a whole are compared with the results from partitioning the data set into smaller parts and applying the KM-algorithm to the smaller parts separately.

exceed those from analyzing the data set as a whole. One explanation is that the partitioning method uses three runs of the algorithm for every smaller data set for five partitions which results in effectively fifteen regulation criterion values for every gene that are averaged as opposed to only five values that were used for the analysis of the large data set as a whole. Note, however, that the fifteen regulation criteria are not equivalent to fifteen implementations of the KM-algorithm on the complete data set, because partitioning the data disregards potential correlations between genes in different partitions. Additionally, Figure 6.13 shows that the average GR measure for analysis of short time series is higher, the smaller the data set. In data sets, where the ratio of available time points to observed genes is larger, the identification of regulated genes is expected to perform better.

Overall, the partitioning method is recommended for the analysis of microarray data sets with a large number of observations ($> 2000$ genes). It greatly reduces computational expense and makes even the analysis of very large microarray data sets computationally feasible. The drastic reduction in computation time for each smaller data set allows more repetitions to be performed in the same time it would take to analyze the large data set as a whole, which improves the ability to accurately identify regulated genes.
7. APPLICATION TO REAL MICROARRAY DATA

Many experiments have been conducted for the purpose of understanding the development of a biological process over time [31–36]. Typically gene expression measurements are taken at consecutive time points. For some of these experiments, the microarray data is publicly available [106–108]. Since the results of the original findings for these experiments have been published as well, it is possible to compare the genes which are identified as regulated by the KM-algorithm with genes that have previously been found to play a role in the observed process. The proposed method to identify regulated genes is applied to the following well known experiments.

7.1 Caulobacter Data Set

Laub et al. [33] studied the genetic network controlling the cell cycle of the bacterium *Caulobacter crescentus*. The goal of this experiment was to identify genes that play a relevant role in the cell cycle. The authors suggest that genes with similar temporal expression profiles encode for proteins whose functions are related.

Spotted cDNA microarrays representing 2966 genes of the *Caulobacter* genome (about 90%) were employed. The arrays were manufactured by the experimenters themselves and open reading frames, which are the elements of DNA that are translated into protein, were used as targets. RNA was harvested from synchronized bacteria cells every 15 minutes during the 150 minute cell cycle. The samples were hybridized to microarrays using a mixed pool of unsynchronized reference cells as a control. A discrete cosine transform algorithm was employed and arbitrary threshold values were used to identify 553 genes whose expression patterns vary significantly in a cyclic manner over the cell cycle. Note, that this method of pre-processing the data will potentially discard regulated genes with small temporal variation. Clustering
techniques were employed for the temporal gene expression profiles of the 553 selected genes to yield 14 groups of genes with different function e.g., DNA replication, cell envelope metabolism, or cell division.

To apply the KM-algorithm, we use the same set of genes 1590 genes that has been analyzed by Wu et al. [89]. The available information consists of one expression ratio (sample vs. unsynchronized control) at each time point for each gene. The expression ratios obtained from [106] are log-transformed. To select the most appropriate model dimension, the Hankel matrices $H$ (compare Equation 5.13) are computed for different maximum relevant biological time lags ($p = 1, 2, 3$). Singular value decomposition is performed for the Hankel matrices and the number of eigenvalues of comparatively large magnitude is used as an estimator for the model dimension. Figure 7.1 shows that the model selection results do not differ signif-

![Figure 7.1. Magnitude of singular values for the Caulobacter data set obtained by singular value decomposition of the Hankel matrix $H$ for different maximum relevant biological time-lags $p = 1, 2, 3$.](image)

antly for different choices of the maximum relevant biological time lag $p$. The selected model dimension for the *Caulobacter* data set is determined to be $m = 4$. For the same data set, Wu et al. [89] have selected model dimension $m = 5$ using factor analysis for a state space model of gene regulation. After determination of the state space dimension the KM-algorithm is run five times with the selected number
Figure 7.2. For the genes in the Caulobacter data set [33], the regulation criterion value is plotted against the maximum absolute log-fold expression value over the time course of observations. Genes that had been determined to be cell-cycle regulated previous to the analysis of Laub et al. are shown in green. Genes that were found to be regulated by Laub et al. are shown in red.
Figure 7.3. For the genes in the Caulobacter data set [33], the regulation criterion value is plotted against the temporal variance of the gene. Genes that had been determined to be cell-cycle regulated previous to the analysis of Laub et al. are shown in green. Genes that were found to be regulated by Laub et al. are shown in red.
(m = 4) of hidden regulators. For each gene, the regulation criterion is computed and the results from the five runs are averaged. This provides a ranking of all 1590 genes according to their degree of regulation. A few genes have very small temporal variance over time (in the order of $10^{-15}$), thus their regulation criterion values are very large (in the order of $10^{33}$). They have been omitted from Figures 7.2 and 7.3.

It is apparent in Figures 7.2 and 7.3 that the genes which have been determined to be regulated by Laub et al. (plotted in red) as well as the genes which have been found to be regulated previous to the study by Laub et al. (plotted in green) are predominantly those with large maximum gene expression and thus with high temporal variance. Of the genes with high regulation criterion values but more moderate temporal variance, only few have so far been identified as regulated in cell-cycle of the bacterium *Caulobacter crescentus*. Biological follow up experiments are needed to study the genes with high regulation criterion values in more detail. Of special interest in this endeavor would be genes which have large regulation criterion values as well as large maximum absolute gene expression (upper right quadrant of Figure 7.2). It would also be interesting to conduct follow up experiments on the genes with small regulation criterion values who have been determined to be cell-cycle regulated based solely on their large temporal variance to either confirm or rule out those gene’s involvement in the bacterial cell-cycle.

### 7.2 Synechocystis Data Set

Gill et al. [34] studied the gene expression of the bacterium *Synechocystis* (strain PCC 6803) while undergoing a light-to-dark transition. Cells of the photosynthetic bacterium were grown and then exposed to changing light conditions. At 7 equally spaced (30 min) time points during this transition, RNA samples were obtained. The samples were hybridized onto a full-genome spotted cDNA microarray with 3169 genes as well as a partial-genome cDNA microarray with 88 genes. Both arrays
were manufactured by the experimenters. The partial-genome arrays were used for calibration and validation of the experiment.

At each time point, twenty biological replicates of RNA samples were obtained and hybridized to identical slides. The high degree of repetition allowed for effective normalization of the many slides used in this experiment and provided a better estimate for the mean gene expression value of the bacterium *Synechocystis* during a light-dark transition. Hierarchical clustering, as well as Fisher discriminant analysis were employed by Gill et al. [34]. Based on the averaged gene expression values at every time point, 783 genes were found to be involved in the cell’s response to light in the whole-genome experiment. These genes were classified into four classes based upon the delay of their response to the changing light conditions.

![Graph of singular values](image)

Figure 7.4. Magnitude of singular values of the Bacteria data set obtained by singular value decomposition of the Hankel matrix $H$ for maximum relevant biological time-lags $p = 1, 2$.

To determine the optimal state space dimension $m$ for the KM-algorithm, the Hankel matrices $H$ (compare Equation 5.13) are computed for maximum biological time lags ($p = 1, 2$). Singular value decomposition is performed for the Hankel matrices and the magnitude of the singular values is plotted in Figure 7.4. The
selected model dimension for the Bacteria *Synechocystis* data set is \( m = 4 \). The model selection results are similar for the maximum biological relevant time lags \( p = 1 \) and \( p = 2 \). For \( p = 3 \), the singular value decomposition of \( \mathbf{H} \) (which has dimension \( 7836 \times 7836 \)) exceeded the 768 MB of available memory on the IBM machine on which the application was run.

The KM-algorithm is applied five times with state space dimension \( m = 4 \) to the log-ratio expression values of \( n = 2612 \) genes at \( T = 7 \) time points for which measurements were reported. The available data [108] is already normalized, so that the temporal variance of the expression values for each gene is equal to one. Figure 7.5 shows the values of the regulation criterion obtained from the normalized data plotted against the maximum absolute expression value for each of the 2612 genes. Important information has been lost through the data normalization step that changes the temporal gene variances. The maximum absolute expression of the normalized data can no longer be used as a reliable measure for the temporal variance of the gene.

It becomes apparent in Figure 7.5 that the regulation criterion values of the genes identified by Gill et al. [34] vary by class. The genes with late response times (class IV plotted in yellow) tend to have significantly smaller regulation criterion values than those with early response times (classes I and II plotted in green and red). The genes with medium response times correspondingly also tend to have medium regulation criterion values (class III plotted in blue). However, as judged by the regulation criterion, there is no obvious differentiation between unregulated and regulated genes as identified by Gill et al. Most of the genes that have large regulation criterion values as well as large maximum absolute gene expression values have been identified to be cell cycle regulated in class II by Laub et al. Especially for the currently unidentified genes with large regulation criterion values but with moderate to small absolute gene expression value it would be interesting to conduct validating biological follow up experiments. However, since the available time series
Figure 7.5. For each gene in the Bacteria *Synechocystis* data set the regulation criterion is plotted against the maximum absolute expression value. Genes classified into response classes by Gill et al. [34] are shown in green, red, blue, and orange, respectively. Genes that were not found to be involved in the light-dark reaction of the cell by Gill et al. are plotted in black.
was short (it consisted of only 7 observations), it is also possible that the quality of regulation ranking produced by the KM-algorithm is poor.

### 7.3 Spellman CDC15 Yeast Data Set

The yeast cell cycle has been intensively studied in gene microarray experiments [31,32]. The experiments conducted by Spellman et al. [31] are among the most well known. In these experiments, yeast cells were grown and their developmental stages were synchronized by three different methods. In one of the experiments described in [31], mRNA from a CDC15 temperature sensitive yeast mutant was harvested every ten minutes at 19 time points under growing conditions while a control sample was harvested under conditions that prohibited cell growth. The samples were hybridized onto spotted cDNA microarrays manufactured by the experimenters themselves using open reading frames (ORFs) to represent the genes. Most ORFs were spotted only once on the array, but some were spotted in duplicate or triplicate. There is only one biological replicate in this experiment. Spellman et al. used a Fourier transformation and the correlation of gene temporal profiles with those of known regulating genes to classify 799 genes as regulated in a cell cycle-dependent manner.

The KM-algorithm is applied to the data available from [107] after \( \log_2 \) transformation of the expression ratios. For the ORFs that were spotted more than once the available values are averaged to yield one log-fold change value per ORF and time point. Since the number of distinct ORFs spotted on the arrays at each time point is large \( (n = 6308) \), the partitioning method described in Section 6.9 is used to analyze this data set. Ten independent times the observations are randomly partitioned into nine subsets of 630 genes and one subset of 638 genes, respectively. Model selection via singular value decomposition of the Hankel matrix is applied to each smaller data subset and the largest singular values from each of the ten subsets are ranked. Using the 80% cutoff, the singular values of largest magnitude are selected. The number of singular values of large magnitude from a particular subset is used as an estimate
for the state space dimension $m$ of the subset. With the exception of one subset for which the selected model dimension was $m = 0$, all other subsets were determined to have state space dimension $m = 1$. Since the algorithm cannot be run for state space dimension $m = 0$, the KM-algorithm was subsequently applied to each data subset with state space dimension $m = 1$ three times with different initial values for the gene regulation matrix $G$.

The regulation criterion results from the three applications of the KM-algorithm were averaged and the results for all ten subsets combined and the order of ORFs changed back to its original state. Finally, the ten regulation criterion results for the large ($n = 6308$) data set obtained from different ways of partitioning were averaged. Since the application of the KM-algorithm is repeated for different data partitions, effectively the KM-algorithm is applied thirty times for each gene. This yields one regulation criterion value for each of the observed $n = 6308$ genes. In Figure 7.6 the regulation criterion results are plotted against the temporal variance of a gene and in Figure 7.7 the same regulation criterion results are plotted against the maximum absolute expression value of a gene.

Since the regulation criterion of Spellman et al. [31] is based on a Fourier transformation, their regulation criterion does not solely depend on the maximum absolute gene expression or temporal gene variance. In fact, their validation experiments show that cell-cycle regulated genes cannot be identified by large maximum gene expression or large temporal gene variance alone. In Figures 7.6 and 7.7 the top three genes identified by the KM-algorithm as highly regulated have only moderate maximum gene expression and small temporal variance, yet they have been identified by Spellman et al. to be cell cycle regulated. Since the biology underlying any regulation network e.g., the cell-cycle, is likely extremely complex, a promising approach to identifying regulated genes is to combine new methods such as the proposed KM-algorithm with existing ones. Genes that are identified as regulated by several independent methods are the most promising candidates for in depth follow-up experiments.
Figure 7.6. Regulation criterion value obtained by the KM-algorithm plotted against the temporal variance for each ORF representing a gene in Spellman’s CDC15 yeast data set. The genes found to be regulated in a cell cycle-dependent manner by Spellman are plotted in red, with the genes found to be directly involved in cell cycle processes plotted in green.
Figure 7.7. Regulation criterion value obtained by the KM-algorithm plotted against the maximum absolute expression value for each ORF representing a gene in Spellman’s CDC15 yeast data set. The genes found to be regulated in a cell cycle-dependent manner by Spellman are plotted in red, with the genes found to be directly involved in cell cycle processes plotted in green.
8. SUMMARY AND FUTURE WORK

Discovering gene regulatory networks from time series microarray data is an important field in which many unanswered questions remain. The continuously improving technology allows to conduct increasingly complex biological experiments in which more time point observations can be taken on numerous genes simultaneously. To construct a meaningful regulatory network between all the observed genes as well as unobserved regulators, complex statistical models need to be utilized. Dynamic Bayesian networks provide a flexible foundation for building genetic networks. They allow to model the complicated dependence structure between genes as well as observable and unobservable regulators. This work demonstrates that state space models, which are a particular class of dynamic Bayesian networks, are well suited to represent observations on gene expression as well as hidden factors which influence gene expression and their temporal dependence.

However, estimating the parameters of a Bayesian network is an $np$-hard problem [109, 110]. Therefore, regardless of the method used to estimate the model parameters, the number of observations first needs to be drastically reduced in many applications. Until now, the commonly used methods for this purpose rely solely on the temporal profiles for each gene. Genes are selected for further analysis based upon either their temporal variance over the time course of observations or based upon their maximum expression ratio relative to the control sample in the experiment. While large expression changes over a time course of observations are an indication that a gene may be regulated, the maximum expression change or temporal variance should not be the only criterion used for selection of regulated genes. Simulations have shown that genes may be highly regulated, even though their change in expression compared to the control is small. The proposed regulation criterion provides an alternative for the selection of regulated genes. Especially in combination with
existing methods it allows for the selection of a subset of genes which are suspected to be regulated in a particular observed process. Subsequently, for the reduced data set which contains the regulated genes the structure of the Bayesian network can be inferred.

The proposed KM-algorithm is based on a state space model for gene regulation. The model is fitted by alternately computing the Kalman smoothing estimates of the hidden regulators and the restricted maximum likelihood estimates of the model parameters. The novel approach of computing the Cholesky decomposition of the error covariance matrix assures that the corresponding parameter estimate is positive definite. Model selection for the dimension of the hidden state space is performed efficiently by considering the autocovariances of the observations. For large microarray data sets with very many simultaneous observations on gene expression, an effective partitioning method is proposed that considerably shortens computation time. The power of the proposed algorithm to identify regulated genes among a large number of observations was studied via simulations. It has been shown that the proposed KM-algorithm performs well on time series that are of moderate ($T = 40$) number of time points, and only improves as the number of time points increases. The benefit of this improvement will be fully appreciated as the number of time points is no longer limited by the cost of performing microarray analysis.

8.1 Future Directions

It is to be expected that a continued large effort will be directed towards the discovery of gene regulatory networks for many organisms under different circumstances. The KM-algorithm provides a computationally feasible method to select genes for constructing such regulatory networks. However, this method can also be applied to areas other than microarray technology.

Gene regulatory networks are only one example of a more general biological pathway. In addition to the genome applications presented here, other experiments
involve the study of an organism’s metabolome or proteome over time [111, 112]. Analogous to microarray applications, the KM-algorithm can be applied to these “omic” studies for the purpose of identifying regulated variables in the study of biological pathways [6].

8.1.1 Regulator Profiles

In addition to the model parameters of the state space model, the KM-algorithm also produces the Kalman smoothing estimates of the underlying hidden regulators. If experiments contain additional observations on variables which are suspected to control gene expression, then the temporal profiles of these variables can be compared to the ones estimated by the KM-algorithm. Especially in cases where gene-protein interactions are studied it would be interesting to combine the results of gene microarray and protein microarray experiments.

However, since the regulators in state space models are noisy and unobservable, statistical methods must be developed to compare the estimated temporal profiles of regulators with actually observed variables. Furthermore, for several applications of the KM-algorithm on the same data set with different initial values, the order of the hidden regulators may change. Therefore, it is not trivial to combine the estimates of temporal regulator profiles from different applications of the KM-algorithm on the same data set, particularly if the state space dimension is high.

8.1.2 Partitioning Method

The partitioning method has been shown in simulations to yield reliable results for the identification of regulated genes in large data sets. Since the partitioning method is flexible in the number and size of partitions and the number of implementations with different initial values on each subset, more extensive power studies would be useful to determine the optimal settings as a function of size of data set and number of time point observations. Especially for short time series, where only few time
point observations are available, the partitioning method may, in fact, be preferable to analyzing the complete data set as a whole.

8.1.3 Model Selection Method

The model selection method utilized to determine the dimension of the state space provides a viable alternative to existing procedures. Common model selection criteria, such as AIC and BIC [101,102], do not work well in the context of microarray applications, because of the extremely large number of both observations and model parameters. Furthermore, they are not applicable in models that contain hidden variables, such as state space models. In these cases the proposed model selection method based on the Hankel matrix of autocovariances may be applied. While the cutoff method for the magnitude of singular values currently used can be improved by procedures that are based on the shape of the singular value magnitude curve, the principle of using model selection methods based on the autocovariances of the time series gene expression observations can be applied in many other microarray applications.

8.1.4 Biological Validation

When the proposed KM-algorithm is applied to existing data sets of gene expression, some genes that have not previously been found to be regulated in the observed process nevertheless receive a high regulation criterion value. It would be interesting to investigate these particular genes in follow up studies, for example by gene knock-out experiments.

On the other hand, some studies have identified genes as regulated solely based on their maximum expression or temporal variance without considering the data correlation structure. If the KM-algorithm is applied to these data sets and some of the genes that have been found to be regulated receive very low regulation criteria,
then it would also be worthwhile to conduct more in depth analysis to assure that the previous findings are indeed biologically relevant.

8.2 Conclusions

The proposed KM-algorithm provides a novel approach to selecting genes according to their degree of regulation in large networks. Discovering gene regulatory networks from noisy microarray data is a very challenging problem. Statistical selection of regulated genes in microarray experiments in a biologically meaningful way allows existing methods for the estimation of regulatory networks to perform with a higher degree of accuracy. Since it has been shown that the KM-algorithm performs reliably for microarray time series applications with a moderate number \((T = 40)\) of time points, even if the number of observed genes is very large, it is anticipated that it can be utilized in many important applications in the future.
LIST OF REFERENCES
LIST OF REFERENCES


licka, T.G. Wolfsberg and A.E. Gabrielson, D. Landsman, D.J. Lockhart, and 

Global analysis of the genetic network controlling a bacterial cell cycle. 

and G. Stephanopoulos. Genome-wide dynamic transcriptional profiling of 
the light-to-dark transition in Synechocystis sp. strain PCC 6803. *Journal of 

sion during the life cycle of drosophila melanogaster. *Science*, 297:2270–2276, 
2002.

R. Somogyi. Large-scale temporal gene expression mapping of central nervous 
system development. In *Proceedings of the National Academy of Sciences, 


[38] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis - Forecast-


[42] W.A. Schmitt and G. Stephanopoulos. Prediction of transcriptional profiles of 
Synechocystis PCC 6803 by dynamic autoregressive modelling of DNA mi-


approach to reconstructing genetic regulatory networks with hidden factors. 


[46] R.H. Shumway and D.S. Stoffer. An approach to time series smooting and fore-
264, 182.


VITA
VITA

Martina Muehlbach Bremer was born February 28, 1975 in Hannover, Germany. She received a B.S. in Mathematics and Physics from the University of Hannover, Germany in 1996, a M.S. in Applied Mathematics from Purdue University in 1999, a Diploma in Mathematics and Physics from the University of Hannover, Germany in 2001, and the 'Staatsexamen' (education degree) in Mathematics and Physics from the University of Hannover, Germany in 2002. She has been a graduate student in the Mathematics department at Purdue University from 1997 until 1999 and joined the Statistics Department at Purdue University in 2002.