Chapter 1

Introduction

1.1 Overview

A great challenge in science is the analysis of complex systems. Traditionally, it was only possible to successively examine small parts of these systems and assort the results. Nevertheless, due to possible interactions across all parts of the systems, knowledge of the components alone cannot lead to a full understanding of complex systems. The recent arising of new technologies however made it possible to simultaneously observe a large amount of variables as well as analyze the attained data.

The high dimensional data structure causes traditional methods no longer to be directly applicable for statistical learning. Most notably, research in systems biology stimulated the development of new methods for learning in complex systems. However, similar problems arise in various areas of scientific research like economics, finance, astronomy, meteorology or medicine.

This thesis is concerned with developing interpretable models for prediction and inference in complex systems that primarily build on a Stein-type shrinkage approach. It is based on the following seven articles:

Article A: Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *SAGMB* (Volume 6, Issue 1, Article 9, 2007); (Opgen-Rhein and Strimmer, 2006a)

Article B: Inferring Gene Dependency Networks from Genomic Longitudinal Data: A Functional Data Approach, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *REVSTAT* (Volume 4, Number 1, pp. 53–65, March 2006); (Opgen-Rhein and Strimmer, 2006b)

Article C: Using Regularized Dynamic Correlation to infer Gene Dependency Networks from time-series Microarray Data, by Rainer Opgen-Rhein and Korbinian Strimmer, refereed conference proceedings of *WCSB 2006*, pp. 73–76; (Opgen-Rhein and Strimmer, 2006e)

Article D: Condition Number and Variance Inflation Factor to Detect Multicollinearity, by Rainer Opgen-Rhein, technical report 01/07; (Opgen-Rhein, 2007)

Article E: Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *BMC Systems Biology* (Volume 8, Supplement 2, S3): Proceedings of PMSB 2006 ("Probabilistic Modeling and Machine Learning in Structural and Systems Biology"), Tuusula, Finland, 17-18 June 2006 ; (Opgen-Rhein and Strimmer, 2007b)

Article F: From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *BMC Systems Biology* (Volume 1, Article 37, 2007); (Opgen-Rhein and Strimmer, 2007a)

Article G: Reverse Engineering Genetic Networks using the **GeneNet** package, by Juliane Schäfer, Rainer Opgen-Rhein and Korbinian Strimmer, published in *R News* (Volume 6, Number 5, December 2006, pp. 50-53); (Schäfer et al., 2006b)

The articles are henceforth referred to by their character. A short description of the individual articles can be found in appendix (i).

This part I of the thesis, "Statistical Learning in Complex Systems", offers a summary of the theory and the methods of the articles. It elucidates the structure of the arguments that render learning in complex systems possible. Additionally, it provides some theory and further background for the methods developed in the articles. Part II, "Articles" provides the essentially unmodified publications, merely their layout was adjusted.

1.2 Contribution of this thesis

We will now summarize the contributions of this thesis. The new theory, methods and algorithms developed in this work are the following:

- a new Stein-type shrinkage estimator for the variance and the covariance matrix for large dimensional data sets (*Article A*)
- an extension of the shrinkage estimator, called *limited translation* for the covariance matrix, which takes into account the risk of individual components (*Article A*)
- development of a *shrinkage t* statistic for high-dimensional case-control analysis (*Article A*)
- introduction of *dynamical correlation*, a consistent extension of the concept of correlation for longitudinal data (*Article B*)
- development of a Stein-type shrinkage estimator for the dynamical correlation applicable in the "small n, large p" setting (*Article C*)

- a new interpretation and decomposition of the regression coefficient in linear regression (*Article D*)
- introduction of small sample *shrinkage regression* (*Article E*)
- a new small sample model selection method for the VAR-process (*Article E*)
- a new method of estimating partially causal networks (*Article F*)

The new methods were tested in simulations and applied to analyze the following high-dimensional data sets:

- *Arabidopsis thaliana* data set by Smith et al. (2004) in *Article E* and *Article F*
- *Escherichia coli* data set by Schmidt-Heck et al. (2004) in *Article G*
- *human T-cell* data set by Rangel et al. (2004) in *Article B* and *Article C*
- *Affymetrix spike-in study* by Cope et al. (2004) in *Article A*
- *"golden spike" Affymetrix experiment* by Choe et al. (2005) in *Article A*
- *HIV-1 infection study* by van 't Wout et al. (2003) in *Article A*

All statistical procedures described are implemented in packages of the computer program R (R Development Core Team, 2006), which is available under the terms of the GNU General Public License and can be found in the CRAN archive (`http://cran.r-project.org`). Specifically, the methods can be found in the following packages:

- *GeneNet*: Modeling and inferring gene networks (Opgen-Rhein et al., 2007)
- *corpcor*: Efficient estimation of covariance and (partial) correlation (Schäfer et al., 2006a)
- *st*: "shrinkage t" statistic (Opgen-Rhein and Strimmer, 2006d)
- *longitudinal*: Analysis of multiple time course data (Opgen-Rhein and Strimmer, 2006c)

A more detailed description of the packages can be found in appendix (ii). The packages and data sets can also be found on the following website: `http://strimmerlab.org/`.

Chapter 2

Complex Systems

2.1 The challenge of learning in complex systems

The focus of this thesis lies in gaining knowledge about complex systems, in which a large number of components interact and conjointly affect the state of the system in future points of time. The term "complex systems" itself is not explicitly defined in science and leaves room for interpretation. Generally, a complex system can be understood to be one which properties cannot be fully explained by a knowledge merely of the component parts (Gallagher and Appenzeller, 1999).

This character of complex systems renders the inference of their structure challenging, as all parts of the system have to be taken into consideration simultaneously to achieve reliable information. With the appearance of new high throughput technologies, e.g., in biology or astronomy, much progress has been made in the measurement of the components, so that – for large systems – hundreds of variables can be observed at the same time. At the same time, the number of repetitions for the observation of each item is often restricted, so that the number of variables possibly by far exceeds the number of observations. The key problem for learning in complex systems is that in this "small n, large p" paradigm standard estimation procedures tend to be unreliable or to be even inapplicable (West et al., 2000).

The methods developed in this thesis are applied to complex systems in biology which mainly appear in a discipline called systems biology. Nevertheless, they are not restricted to this scientific area. In contrast, the same data structure is encountered in various fields of contemporary research. Examples are finance, where, e.g., portfolio optimization requires the estimation of covariances between hundreds of stocks (Mantegna and Stanley, 2000; Ruppert, 2004; Ledoit and Wolf, 2003a), management, e.g., for business optimization (Huang et al., 2006; Bickel and Levina, 2006), astronomy (e.g. Kabán et al., 2006; Patat, 2003), meteorology (e.g., Murphy and Wilks, 1998; Levine and Berliner, 1999) or medicine (e.g. Efron, 2005c).

However, the development of new or improvement of established statistical methods of learning in complex systems was heavily stimulated by systems biology, where microarray

technology initialized the production of large data sets, and where new technologies like time of flight spectroscopy, proteomic devices or flow cytometry are likely to generate even larger quantities of data (Efron, 2005c).

These developments are challenges for statistical inference. Nevertheless, the consequences of multidimensional data is often avoided, either by ignoring the statistical problems arising by large scale estimation (e.g. Mantegna and Stanley (2000)) or by a deterministic reduction of the dimensionality like concentrating on prominent variables and forgoing spatial details as done for weather forecast in Pappenberger et al. (2005).

An extensive overview of the methods used in systems biology can be found in Klipp et al. (2005). In this work the focus lies on regularization techniques based on James-Stein estimation (James and Stein, 1961). Although the inference techniques introduced in the following chapters will be applied to microarray data, it is – again – to stress that they are suited for all kind of multidimensional data in the "small n, large p" paradigm.

2.2 The use of models in systems biology

Generally, the interest of research lies in understanding the true nature of a system. Nevertheless, it is epistemological knowledge that this cannot be accomplished; it is only possible to observe phenomena produced by it. These observations can be used to create models of the system. Ideally, this model should be of the same complexity as the system itself to include all structures of the system and to be able to make reliable predictions in all circumstances. However, this is neither possible due to incomplete knowledge and limited computational capacities nor is it desirable: to gain an understanding of complex systems, this complexity has to be reduced, which is an important function of models. Different models concentrate on different aspects of complex systems. In this sense, a model cannot be right or wrong, it can be only more or less useful for different purposes.

We will elaborate this for genetic networks in systems biology. Important aspects are the structures (e.g. the network of gene interactions and biochemical pathways), the dynamics (behavior over time), the control methods (mechanisms controlling the state of the cell) or the design principles (the strategies of modification and construction of biological systems) (Kitano, 2002).

One possibility in modeling complex systems is to focus on different levels of scaling. An example is given in Fig. 2.1. In the left column the modeling is concentrated on describing the molecular details of the characteristics of a single gene (e.g. with differential equations). The remainder of the system (all the other genes) is ignored, as the extrapolation of this detailed model would be prohibitively complicated. The center left column focuses on the circuit of a few genes, and the center right and right column deal with increasingly larger networks of genes. Nevertheless, the details of kinetic properties of genes have to be ignored when modeling the relations among them.

The different models introduced in this thesis all try to capture the relationships among a large set of variables (here: genes), but take different points of view. Section 4.2, e.g., concentrates on correlation networks, which are extensively used in biology. They allow

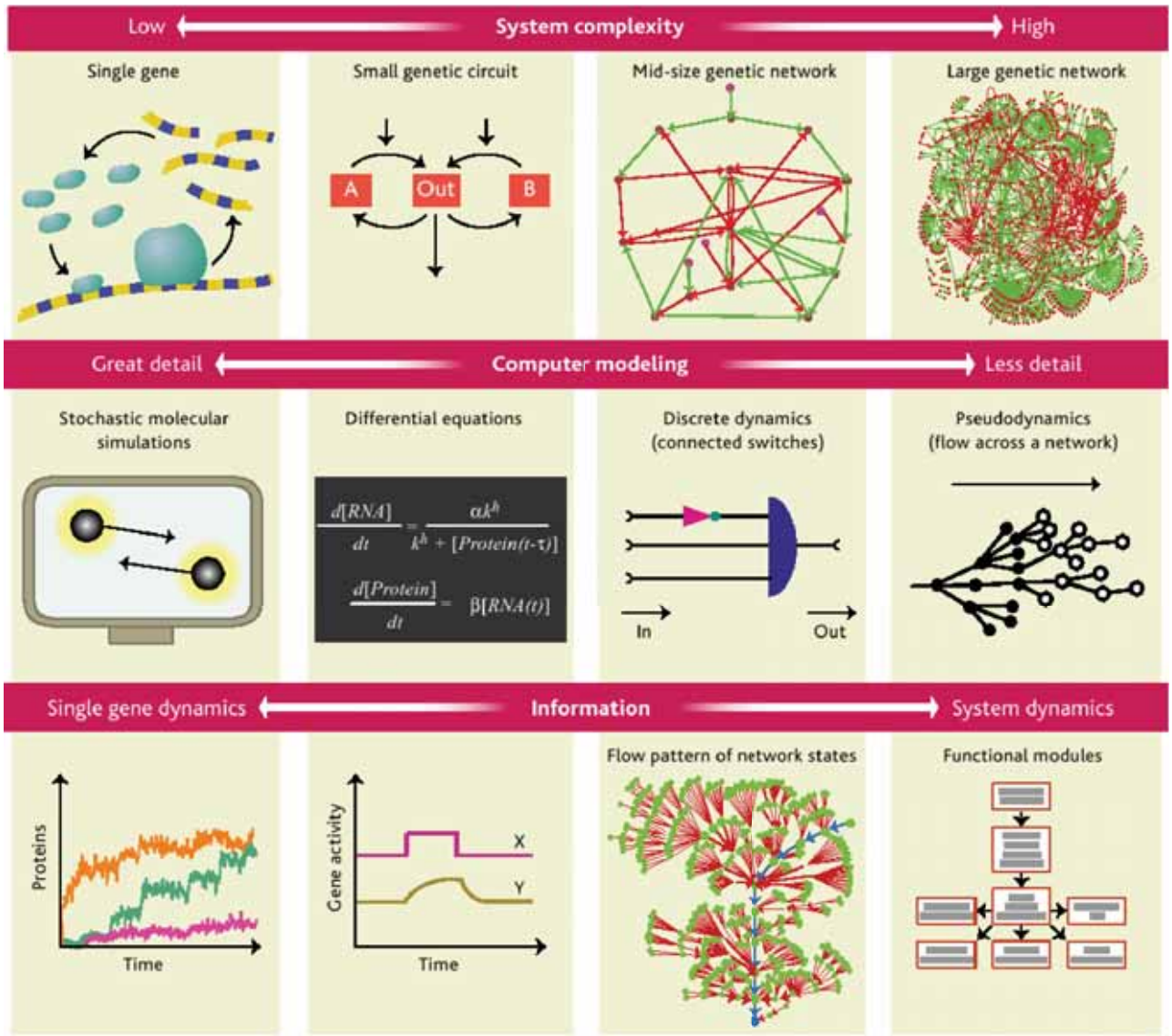

Figure 2.1: Different levels of description in models of genetic networks (figure source: Bornholdt, 2005)

identifying strongly correlated genes. This model is not wrong, but does not serve the intention of discovering functional relations between genes, which is usually the interest in systems biology. For this purpose, graphical Gaussian models will be introduced in section 4.3. Another important aspect lies in the incorporation of dynamical aspects into models as done in chapter 5 or in the identification of causal structures (chapter 6).

The next section introduces microarray data, which has a data structure typical for complex systems, as the number of variables is much larger as the number of observations. This data will be used to demonstrate the methods developed in this thesis.

2.3 Microarray data and their biological background

To understand microarray data and their connection to genetic networks, some biological background has to be provided. The particulars of the processes of molecular biology and the technical methods utilized to gain microarray data can only be indicated, for more details see, e.g., Graur and Li (2000); Gibson and Muse (2004); Klipp et al. (2005).

Cells are the basic structure and the functional units of every living system. All information needed to direct their activities and therefore to sustain life is contained in a sequence of four different bases or nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G), which compose the deoxyribonucleic acid (DNA). It consists of two long strands that have the shape of a double helix. The double-stranded structure evolves, as each type of nucleotide on one strand pairs with just one other type on the other strand: adenine forms a bond with thymine and cytosine with guanine. A part of the DNA (the coding region) consists of genes, segments that contain instructions about the synthesis and regulation of proteins. Proteins control the structure and function of a cell and are the essential parts in building complex living systems. An idea of the increasingly complex structures of life, all building up on genes, can be found in Fig. 2.2. The human genome is estimated to contain about 20,000 – 25,000 genes. However, not the number of genes is decisive: the difference between distinct species lies mainly in the regulatory program. This means that the creation, function and different amounts of gene products in the living system can only be understood by taking the activation and interaction of different genes into consideration.

The activation of genes, called gene expression, can be measured using microarray technology; the interaction between different genes has to be inferred from the gene expression of a large set of genes using statistical methods. Microarray technology is therefore the method to generate the data for statistical analysis. It builds upon the central dogma of molecular biology (Crick, 1958), which describes the process of synthesizing proteins from genes. The idea is the following:

- The first step is *transcription:* RNA polymerase, an enzyme complex, and transcription factors transfer the information of a gene to the single-stranded messenger RNA (mRNA). This often includes further processing via alternative splicing, where parts of the mRNA are removed and rearranged.

- Afterwards, *translation* takes place: ribosomes read the mRNA and build the protein by adding amino acids in the sequence given by the gene and by subsequently folding it into the correct conformation.

The term *gene expression* usually refers to the amount of mRNA. Microarray technology, which measures the gene expression, takes advantage of the fact that the single-stranded mRNA binds to oligonucleotides with complementary nucleotide sequence. A large number of gene-specific *probes* consisting of a complementary DNA (cDNA) are attached to the array surface.