

Contents

I	Statistical Learning in Complex Systems	1
1	Introduction	3
1.1	Overview	3
1.2	Contribution of this thesis	4
2	Complex Systems	7
2.1	The challenge of learning in complex systems	7
2.2	The use of models in systems biology	8
2.3	Microarray data and their biological background	10
2.4	Microarray data sets analyzed in this thesis	12
2.4.1	Microarray time course data sets	12
2.4.2	Microarray case-control data sets	15
3	Parameter estimation in complex systems	17
3.1	Stein-type shrinkage	17
3.1.1	Stein-phenomenon	17
3.1.2	Distribution-Free Shrinkage Estimation	18
3.1.3	Construction of a Shrinkage Estimator	20
3.2	Shrinkage estimation of the covariance matrix	22
3.2.1	Variances	23
3.2.2	Correlation structure	24
3.2.3	Covariance matrix	25
3.3	Component Risk Protection by Limited Translation	25
4	Statistical Learning in High-dimensional Data Sets	27
4.1	High dimensional case-control analyses	27
4.1.1	The “Shrinkage t ” Statistic	28
4.1.2	Assessment of Quality of Gene Ranking	29
4.1.3	Performance of Gene Ranking Statistics	29
4.2	Correlation networks	31
4.3	Graphical Gaussian networks	32
4.3.1	The partial counterparts of correlation and variance	33
4.3.2	Model selection using local fdr	36

4.3.3	Inferring a graphical Gaussian network	39
5	Analysis of longitudinal data sets in complex systems	43
5.1	Dynamical correlation	43
5.1.1	The Concept of Dynamical Correlation	44
5.1.2	Regularized Inference of the Dynamical Correlation	47
5.1.3	Applications of dynamical correlation	48
5.1.4	Remarks	50
5.2	Using a vector autoregressive model for analyzing time course data	52
5.2.1	Linear regression	53
5.2.2	Shrinkage regression	54
5.2.3	Shrinkage estimation of the vector autoregressive model	56
5.2.4	VAR network model selection	57
5.2.5	Applications	58
6	Discovering causal structure in high-dimensional data	63
6.1	Causality	63
6.2	Causality in directed networks	64
6.3	Algorithm for discovering causal structure	65
6.3.1	Theoretical Basis	65
6.3.2	Discovery Algorithm	66
6.4	Results	67
6.4.1	Statistical Interpretation	67
6.4.2	Further Remarks and Properties	68
6.4.3	Application	69
6.5	Discussion	73
7	Outlook	75
i	Description of the articles	77
ii	Description of the R packages	79
iii	Definitions, Notations, and Abbreviations	81
	List of Figures	82
	Bibliography	84
II	Articles	95
A	Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach (<i>SAGMB: 6(1):9, 2007</i>)	97

A.1	Introduction	98
A.2	Distribution-Free Shrinkage Estimation	99
A.2.1	James-Stein Shrinkage Rules	99
A.2.2	Construction of Shrinkage Estimator	101
A.2.3	Positive Part Estimator and Component Risk Protection by Limited Translation	101
A.2.4	Further Remarks	102
A.3	The “Shrinkage t ” Statistic	103
A.3.1	Shrinkage Estimation of Variance Vector	103
A.3.2	Construction of “Shrinkage t ” Statistic	104
A.3.3	Other Regularized t Statistics	104
A.4	Results	104
A.4.1	Assessment of Quality of Gene Ranking	104
A.5	Discussion	109
	Bibliography	110
B	Inferring Gene Dependency Networks from Genomic Longitudinal Data:	
	A Functional Data Approach (<i>REVSTAT: 4(1):53–65, 2006</i>)	115
B.1	Introduction	115
B.2	Methods	117
B.2.1	Setup and Notation	117
B.2.2	Dynamical Correlation	117
B.2.3	Estimating Gene Association Networks Using Dynamical Correlation	119
B.3	Results	119
B.3.1	Illustrative Example	120
B.3.2	Gene Expression Time Course Data	121
B.4	Discussion	124
	Bibliography	125
C	Using Regularized Dynamic Correlation to infer Gene Dependency Networks from time-series Microarray Data (<i>Proc. of WCSB 2006, pp. 73–76</i>)	129
C.1	Introduction	129
C.2	Methods	130
C.2.1	Setup and Notation	130
C.2.2	Dynamical Correlation	131
C.2.3	Estimating Gene Association Networks Using Dynamical Correlation	134
C.3	Results	134
C.4	Conclusion	135
	Bibliography	136
D	Condition Number and Variance Inflation Factor to Detect Multicollinearity (<i>Technical report, 01/07</i>)	139
D.1	Introduction	139

D.2	Linear Regression	140
D.2.1	Linear regression based on the true model	140
D.2.2	Empirical estimation - with and without intercept	143
D.3	Detecting Multicollinearity	143
D.3.1	Multicollinearity	143
D.3.2	Variance inflation factor	144
D.3.3	Condition number	145
D.4	Simulation	146
D.4.1	Simulation Models	146
D.4.2	Results	147
D.5	Discussion	156
	Bibliography	156
E	Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process (<i>BMC Bioinformatics</i> 8 (Suppl. 2): S3)	159
E.1	Background	160
E.2	Methods	160
E.2.1	Vector Autoregressive Model	160
E.2.2	Small Sample Estimation Using James-Stein-Type Shrinkage	161
E.2.3	Shrinkage Estimation of VAR Coefficients	162
E.2.4	VAR Network Model Selection	162
E.3	Results and Discussion	163
E.3.1	Simulation Study	163
E.3.2	Analysis of a Microarray Time Course Data Set	165
E.4	Conclusions	166
	Supplementary Information	169
	Bibliography	175
F	From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data (<i>BMC Systems Biology</i>, 1:37, 2007)	179
F.1	Background	180
F.2	Methods	181
F.2.1	Theoretical basis	181
F.2.2	Heuristic algorithm for discovering approximate causal networks	183
F.3	Results and discussion	184
F.3.1	Interpretation of the resulting graph	184
F.3.2	Reconstruction efficiency and approximations underlying the algorithm	185
F.3.3	Further properties of the heuristic algorithm and of the resulting graphs	186
F.3.4	Analysis of a plant expression data set	186
F.4	Conclusions	189

Bibliography	191
G Reverse Engineering Genetic Networks using the GeneNet package (<i>R News</i> : <i>6(5):50–53, 2006</i>)	197
G.1 Prerequisites	198
G.2 Preparation of Input Data	198
G.3 Shrinkage Estimators of Covariance and (Partial) Correlation	198
G.4 Taking Time Series Aspects Into Account	199
G.5 Network Search and Model Selection	200
G.6 Network Visualization	201
G.7 Release History of GeneNet and Example Scripts	201
Bibliography	201