ESTIMATION AND DETECTION OF MULTIVARIATE GENE REGULATORY RELATIONSHIPS

A Dissertation

by

TING CHEN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee, Ulisses de Mendonça Braga-Neto
Committee Members, Edward Russell Dougherty
Jean-Francois Chamberland-Tremblay
Thomas Wehrly
Department Head, Chanan Singh

December 2013

Major Subject: Electrical Engineering

ABSTRACT

The Coefficient of Determination (CoD) plays an important role in Genomics problems, for instance, in the inference of gene regulatory networks from gene-expression data. However, the inference theory about CoD has not been investigated systematically. In this dissertation, we study the inference of discrete CoD from both frequentist and Bayesian perspectives, with its applications to system identification problems in Genomics. From a frequentist viewpoint, we provide a theoretical framework for CoD estimation by introducing nonparametric CoD estimators and parametric maximum-likelihood (ML) CoD estimators based on static and dynamical Boolean models. Inference algorithms are developed to discover gene regulatory relationships, and numerical examples are provided to validate preferable performance of the ML approach with access to sufficient prior knowledge. To make the applications of the CoD independent of user-selectable thresholds, we describe rigorous multiple testing procedures to investigate significant regulatory relationships among genes using the discrete CoD, and to discover canalyzing genes using the intrinsically multivariate prediction (IMP) criterion. We develop practical statistic tools that are open to the scientific community. On the other hand, we propose a Bayesian framework for the inference of the CoD across a parametrized family of joint distributions between target and predictors. Examples of applications of the Bayesian approach are provided against those of nonparametric and parametric approaches by using synthetic data.

We have found that, with applications to system identification problems in Genomics, both parametric and Bayesian CoD estimation approaches outperform the nonparametric approaches. Hence, we conclude that parametric and Bayesian estimation approaches are preferred when we have partial knowledge about gene regulation. On the other hand, we have shown that the two proposed statistical testing frameworks can detect well-known gene regulation and canalyzing genes like *p53* and *DUSP1* from real data sets, respectively. This indicates that our methodology could serve as a promising tool for the detection of potential gene regulatory relationships and canalyzing genes. In one word, this dissertation is intended to serve as foundation for a detailed study of applications of CoD estimation in Genomics and related fields.

DEDICATION

To my family

ACKNOWLEDGEMENTS

First, I would like to thank my parents, Quan and Xiuhong, for their unconditional love. Without them, I would never have been able to achieve my PhD degree. I thank my financé, Qingnan, for his love and support. I thank my grandfather, Wende, for his encouragement. I thank my uncle, Jinfei, and my aunt, Xiuxiang, for their caring. I thank my financé's mother, Tianying, for treating me as her daughter.

I would like to express my deepest gratitude to my advisor, Dr. Ulisses Braga-Neto, who made my dissertation possible. I thank him for his steady guidance, inspirational discussions, and for generously supporting me financially through my doctorate program at Texas A&M University. The great experiences of working with him will definitely benefit the rest of my career. I thank Dr. Dougherty for his invaluable lessons for life and research and providing such a wonderful research environment. I thank Dr. Chamberland-Tremblay and Dr. Wehrly, for serving on my committee and for their support and useful suggestions for the improvement of this work. I thank all students and researchers, past and current, at the Genomic Signal Processing Laboratory, especially Mohammad Shahrokh Esfahani, Dr. Fang-Han Hsu, Dr. Youting Sun, Dr. Xiangfang Li, Sriram Sridharan, Jason Knight, Esmaeil Atashpaz Gargari, Dr. Lori Dalton, and Dr. Mohammadmahdi R. Yousefi, for friendship and memories. I would like to thank all the faculty and staff at Texas A&M University, especially for the administrative support provided by Ms. Tammy Carda, Ms. Jeanie Marshall, Ms. Claudia Samford and Ms. Anni Brunker.

Last but not the least, I would like to thank my former advisor at Fudan University, Professor Feng Li, for the incentive and encouragement.

NOMENCLATURE

CoD	Coefficient of Determination
MMSE	Minimum Mean-Square Error
MLE	Maximum-Likelihood Estimator
NPMLE	Nonparametric Maximum-Likelihood Estimator
IUT	Intersection-Union Test
LRT	Likelihood-Ratio Test
FWER	Family-Wise Error Rate
FDR	False Discovery Rate
MTP	Multiple Testing Procedures
CI	Confidence Interval
IMP	Intrinsically Multivariate Prediction
OBC	Optimal Bayesian Classifier

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

NOMENCLATURE vi

TABLE OF CONTENTS vii

LIST OF FIGURES x

LIST OF TABLES xvi

1. INTRODUCTION 1
   - 1.1 Background 1
   - 1.2 Contributions 3
     - 1.2.1 Frequentist Inference of the CoD 3
     - 1.2.2 Bayesian Inference of the CoD 10
   - 1.3 Organization 11

2. FREQUENTIST INFERENCE: NONPARAMETRIC COD ESTIMATION 14
   - 2.1 Discrete Prediction 14
   - 2.2 Nonparametric CoD Estimation 16
     - 2.2.1 Resubstitution CoD Estimator 18
     - 2.2.2 Leave-One-Out CoD Estimator 18
     - 2.2.3 Cross-Validation CoD Estimator 19
     - 2.2.4 Bootstrap CoD Estimator 20
   - 2.3 Performance Metrics of CoD Estimators 21
   - 2.4 Exact Moments of Non-Randomized CoD Estimators 22
     - 2.4.1 Resubstitution 23
     - 2.4.2 Leave-one-out 26
   - 2.5 Approximate Variances of Non-Randomized CoD Estimation 27
   - 2.6 Results and Discussion 29
   - 2.7 Summary 39

3. FREQUENTIST INFERENCE: PARAMETRIC MAXIMUM-LIKELIHOOD COD ESTMATION . . . 40
3.1 Stochastic Boolean Systems . . . 41
3.1.1 Predictive Power . . . 42
3.1.2 The Coefficient of Determination . . . 43
3.1.3 Estimation of the CoD . . . 44
3.2 Static Model . . . 46
3.2.1 Maximum-Likelihood Inference of the CoD . . . 48
3.2.2 Performance Analysis . . . 51
3.3 Dynamical Model . . . 57
3.3.1 Maximum-Likelihood Inference of the CoD . . . 59
3.4 Application to System Identification . . . 63
3.4.1 Static Case: Predictor Inference . . . 64
3.4.2 Dynamical Case: Network Inference . . . 70
3.5 Summary . . . 74
4. STATISTICAL DETECTION OF BOOLEAN REGULATORY RELATIONSHIPS . . . 75
4.1 Mathematical Preliminaries . . . 76
4.2 CoD Hypothesis Test . . . 78
4.3 Multiple Testing Procedure . . . 89
4.3.1 Type-I Error Rates and Power . . . 89
4.3.2 Control of the Type-I Error Rate . . . 90
4.3.3 Performance of Multiple Testing Procedures . . . 92
4.4 Case Study: Genotoxic Stress Responsive Genes . . . 97
4.4.1 Detection of Significant Regulatory Relationships . . . 98
4.4.2 Detection of Synthetic Target Genes . . . 100
4.5 Summary . . . 101
5. STATISTICAL DETECTION OF INTRINSICALLY MULTIVARIATE PREDICTIVE GENES . . . 103
5.1 Intrinsically Multivariate Prediction . . . 105
5.2 IMP Hypothesis Test . . . 108
5.3 Multiple Testing Procedures . . . 112
5.4 Results and Discussion . . . 114
5.4.1 Synthetic Data . . . 115
5.4.2 Real Data . . . 120
5.5 Summary . . . 124
6. BAYESIAN COD ESTIMATION . . . 127
6.1 Discrete Model . . . 128

6.2 Bayesian CoD Estimators 129
6.2.1 The Bayesian MMSE CoD Estimator 130
6.2.2 The Bayesian CoD Estimator Based on the Optimal Bayesian Classifier 134
6.3 Exact Moments of Bayesian CoD Estimator Based on Optimal Bayesian Classifier 136
6.4 Performance of Bayesian CoD Estimators 139
6.4.1 Performance Over One Specific Distribution 139
6.4.2 Performance Over All Distributions 142
6.5 Applications to System Identification Problems 143
6.5.1 Numerical Experiments 145
6.6 Summary 152

7. CONCLUSION 156

REFERENCES 160

APPENDIX A. RESULTS ON CONDITIONAL EXPECTATION GIVEN DISJOINT EVENTS 169

APPENDIX B. EXPRESSIONS OF BIAS AND VARIANCE OF THE ML COD ESTIMATOR IN 3-INPUT AND LOGIC MODEL 170

APPENDIX C. ASYMPTOTIC EXPRESSIONS OF BIAS AND VARIANCE OF THE ML COD ESTIMATOR FOR 10 2-PREDICTOR LOGICS . . . 174

APPENDIX D. PROOF OF PROPOSITION 2 176

APPENDIX E. PROOF OF PROPOSITION 4 178

APPENDIX F. PROOF OF PROPOSITION 5 180

LIST OF FIGURES

FIGURE Page

1.1 Example of regulatory network, equivalent logic circuit, and hypothetical sample data for the DNA synthesis pathway of the cell cycle. Adapted from Shmulevich et al. [62]. 6

2.1 Bias, variance, and RMS for several CoD estimators vs. actual CoD under a Zipf model with $c_0 = 1/2$, for $n = 40$ and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling. 31

2.2 Bias, variance, and RMS for several CoD estimators vs. number of bins ($b = 4, 8, 12$, and 16) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.6 and varying sample size. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling. 32

2.3 Bias, variance, and RMS for several CoD estimators vs. number of bins ($b = 4, 8, 12$, and 16) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.8 and varying sample size. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling. 33

2.4 Bias, variance, and RMS for several CoD estimators vs. sample size ($n = 20, 30, 40, 50$, and 60) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.6 and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling. 34

2.5 Bias, variance, and RMS for several CoD estimators vs. sample size ($n = 20, 30, 40, 50$, and 60) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.8 and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling. 35

2.6 Exact (solid line) and approximate (dashed line) variances of resubstitution CoD and leave-one-out CoD versus bin size for varying bin sizes. 37

2.7 Exact (solid line) and approximate (dashed line) variances of resubstitution CoD and leave-one-out CoD versus bin size for varying sample sizes. 38

3.1 Bias and variance versus predictive power over sample size $n = 10, 20, 30$ and 40, in a two-input AND model with $P_1 = P_2 = 0.5$ and $\gamma = 0.20$. Blue: exact results (via complete enumeration); Green: approximate results (via asymptotic approximation). 54

3.2 Bias, deviation variance, and RMS for several CoD estimators vs. predictive power with sample size $n = 60$. Top row, 2-input AND model with $P_1 = 0.8, P_2 = 0.6$ and $\gamma = 0.02$. Bottom row, 3-input AND model with $P_1 = 0.8, P_2 = 0.6$, $P_3 = 0.7, \gamma_{12} = 0.02, \gamma_{13} = 0.015, \gamma_{23} = 0.025$, and $\gamma = 0.02$. Plot key: resubstitution (red), leave-one-out (blue), cross-validation (black), 0.632 bootstrap (purple), MLE (green). The curves for resubstitution and leave-one-out are exact; the curves for cross-validation and 0.632 bootstrap are approximated via Monte Carlo sampling; the curve for the MLE is approximated via the asymptotic method described in the text. 56

3.3 Comparison between $\hat{p}$ and $\hat{p}^{\mathrm{MLE}}$ as a function of increasing sample size, for 6-variable network functions with $l = 2, 3, 4$ predictors per target and $p = 0.85$. 60

3.4 Bias, deviation variance, and RMS for several CoD estimators vs. predictive power with sample size $n = 60$. Top row: 2-input XOR; Midde row: 3-input XOR; Bottom row: 4-input XOR. All curves are approximated via Monte Carlo sampling. 62

3.5 Average percentage of logic gates correctly recovered vs. sample size: static model. 68

3.6 Average percentage of predictors correctly recovered vs. sample size: static model. 69

3.7 Average percentage of logics correctly recovered vs. time series length: dynamical model. 70

3.8 Average percentage of predictors correctly recovered vs. time series length: dynamical model 71

4.1 Statistical power vs. CoD for proposed test, in the uniform predictor case, with $d = 4$ and $\alpha = 0.05$, and varying sample size n and number of logic function minterms m. 83

4.2 Minimum sample size to achieve power = 0.8 vs. CoD for proposed test, with $\alpha = 0.05$. (a) Uniform predictor case, with $d = 4$ and varying number of logic function minterms m. (b-c) Correlated predictor case, with $d = 2$ and varying predictor covariance γ, for logic functions AND (b) and XOR (c). 83

4.3 Statistical power vs. CoD for proposed test, in the correlated predictor case with $d = 2$, and varying sample size n and predictor covariance γ. Top row: AND logic. Bottom row: XOR logic. 85

4.4 Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\mathrm{XOR}\}$; $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \bar{\mathrm{X}}_1\bar{\mathrm{X}}_2, \bar{\mathrm{X}}_1 + \mathrm{X}_2\}$, and predictive power $p = 0.85$. There is a single target to be predicted by two among $G = 24$ genes. 94

4.5 Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of initial number of genes under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\mathrm{XOR}\}$; $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \bar{\mathrm{X}}_1\bar{\mathrm{X}}_2, \bar{\mathrm{X}}_1 + \mathrm{X}_2\}$, and predictive power $p = 0.85$. There is a single target to be predicted by two among a varying number of initial genes. Sample size is fixed at $n = 40$. 95

4.6 Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of number of targets under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\mathrm{XOR}\}$; $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \bar{\mathrm{X}}_1\bar{\mathrm{X}}_2, \bar{\mathrm{X}}_1 + \mathrm{X}_2\}$, and predictive power $p = 0.85$. There is a varying number of targets D to be predicted by two among $G = 24$ genes. Sample size is fixed at $n = 40$. 96

4.7 Significant predictive relationships detected in the IR-response stress gene-expression data of [46], under the FWER- and FDR-controlling approaches. 98

4.8 Power estimates as a function of predictive power for 8 synthetic targets using both FWER- and FDR-controlling procedures, for three logic candidate sets, $K_1 = \{\text{XOR}\}$, $K_2 = \{\text{XOR, AND, NAND}\}$ and $K_3 = \{10 \text{ meaningful logics}\}$. 102

5.1 Relationship among power function, sample size and the IMP value given a 2-input stochastic XOR model with $P_1 = P_2 = 0.05, \gamma = 0.05$, for the proposed IMP test with $\alpha = 0.05$. (a) Statistical power vs. IMP over varying sample size n. (b) Minimum sample size to achieve varying power vs. IMP. 111

5.2 Examples of IMP pairs for a target and graphs of IMP pairs. (a) An example of 5 IMP pairs out of 6 predictor genes for one target; (b) Graph for IMP pairs in (a) withouth cycle (A line means the two connected genes function as an IMP pair for the target); (c) An example of 5 IMP pairs out of 5 predictor genes for one target ; (d) Graph for IMP pairs in (c) with cycle. 115

5.3 Average FWER and FDR estimates (left column) and power estimates (right column) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets and preditive power $p = 0.85$. (a) FWER and FDR estimates for model set $K_1 = \{\text{AND}\}$. (b) Power estimates for model set $K_1 = \{\text{AND}\}$. (c) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (d) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (e) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1X_2}\}$. (f) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1X_2}\}$. There is a single target to be predicted by 5 IMP pairs (as shown in Figs. 5.2) among $G = 24$ genes. 116

5.4 (a) Average FWER and FDR estimates (left column) and power estimates (right column) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets and preditive power $p = 0.85$. (a) FWER and FDR estimates for model set $K_1 = \{\text{AND}\}$. (b) Power estimates for model set $K_1 = \{\text{AND}\}$. (c) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (d) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (e) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. (f) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. There is a single test to be predicted by 5 IMP pairs (as shown in Figs. 5.2) among a vaying number of initial genes $K = 6, 8, 10, 12, 14$. Sample size is fixed at $n = 40$. 117

5.5 Number of significant IMP pairs versus target gene discovered from melanoma data set (a) using the FWER-controlling approach; (b) using the FDR-controlling approach. 121

5.6 Number of significant IMP pairs versus target gene discovered from genotoxic stress-responsive data set (a) using the FWER-controlling approach; (b) using the FDR-controlling approach. 123

6.1 Bias, variance, and RMS for several CoD estimators vs. sample size over one distribution in the 2-predictor case. Fix $p = 0.7$, $q = 1 - p = 0.3$ and $c = 0.5$. Plot key: bayesian (brown), obc (purple), resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). As comparison, we assume noninformative uniform priors for p, q and c as shown in solid brown and purple lines. In dashed lines, a beta prior for c with $\alpha = \beta = 6.0$ is specified. All results are approximated by Monte Carlo sampling method. Note that computations of the Bayesian MMSE CoD estimates associated with beta priors (in dashed lines) are exact, whereas the Bayesian MMSE CoD estimate with uniform priors for true distributions is approximated with Monte Carlo sampling method. . . . 140

6.2 Bias, variance, and RMS for several CoD estimators vs. sample size over all distributions. Top row: $b = 2$; Middle row: $b = 4$; Bottom row: $b = 8$ Plot key: bayesian (brown), obc (purple), resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). We assume uniform priors for all bin probabilities and a beta distribution $B(\alpha, \beta)$ for c, with $\alpha = b + 1$ and $\beta = b + 1$. All results are approximated by the Monte Carlo sampling method, and computations of the Bayesian MMSE CoD estimates are exact. 141

6.3 An example of probability distribution functions of beta priors for the class 0 probability c in a 2-input AND logic model for varying Δ. Set $p = 0.8$ and thus $c = \frac{2p+1}{4} = 0.65$. 147

6.4 Percentage of predictor recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.85. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method. 151

6.5 Percentage of predictor recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.75. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method. 152

6.6 Percentage of logic recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.85. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method. 153

6.7 Percentage of logic recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.75. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method. 154

LIST OF TABLES

TABLE Page

3.1 Logic gates for candidate model sets, static case, $l = 2$. 66

3.2 Logic gates for candidate model sets, static case, $l = 3$. 66

3.3 Logic gates for candidate model sets, static case, $l = 4$. 67

4.1 Distributional constraints for CoD $= 0:$ 2-input logic case 78

4.2 95% Confidence interval (CI) for the CoD based on one random sample generated from a 2-input AND logic model ($n = 100$): (a) in the general case ($P_1 = 0.8$, $P_2 = 0.6$, $\gamma = 0.05$, $d = 2$); (b) in the uniform predictor case ($m = 1, d = 2$); (c) correlated predictor case ($\gamma = 0.05$, $d = 2$) 88

4.3 Examples of detected relationships that are consistent with known biological groundtruth 99

4.4 A list of top 20 significant regulatory relationships detected in the IR-response stress gene-expression data of [46] under the FWER- and FDR-controlling approaches 100

4.5 Synthetic Relationships based on the IR-response stress gene-expression data of [46] 101

5.1 $\mathcal{D}_i$'s in $\text{IMP}_{\text{Y}}(\mathbf{X}) = 0$ for 2-Input Logics 108

5.2 A list of top 20 significant IMP pairs detected in the melanoma data 125

5.3 A list of top 20 significant IMP pairs detected in IR-response stress data under FWER- and FDR-controlling approaches 126

6.1 Logic gates for candidate model sets, static case, $l = 2$. 146

6.2 Logic gates for candidate model sets, static case, $l = 3$. 147

6.3 Specification of hyperparameters of priors for $\mathbf{p}$ using sample data drawn from the static model in the 2-predictor case 148

6.4 Specification of hyperparameters of priors for $\mathbf{q}$ using sample data drawn from the static model in the 2-predictor case 149

6.5 Specification of hyperparameters of priors for c using sample data drawn from the static model in the 2-predictor case 150

C.1 Formulas for ML CoD estimator and its bias asymptotic approximations for the five representative two-predictor logic models. 174

C.2 Formulas for ML CoD estimator and its variance asymptotic approximations for the five representative two-predictor logic models. 175

1. INTRODUCTION [*]

The coefficient of determination (CoD) plays an important role in Genomics problems, for example, in the inference of gene regulatory networks from biological data. In this dissertation, we introduce a rigorous statistical inference framework in the context of coefficient of determination from both frequentist and Bayesian perspectives. We also study its applications to the detection of gene regulatory relationships by using quantized gene-expression data. We outline in the following the content of this dissertation.

1.1 Background

DNA regulatory circuits can be often described by networks of Boolean logical gates updated and observed at discrete time intervals [2,9,37,38,43,44]. In a stochastic setting, the degree of association between Boolean predictors and targets can be quantified by means of the discrete CoD [31]. In classical regression analysis, the nonlinear CoD gives the relative decrease in unexplained variability when entering a variable X into the regression of the dependent variable Y, in comparison with the total unexplained variability when entering no variables. Applying this to pattern prediction, Dougherty and collaborators [31] introduced a very similar concept, that of CoD for binary random variables, which measures the predictive power of a set of predictor variables $\mathbf{X} = \{X_1, X_2, ..., X_n\} \in \{0,1\}^n$ with respect to a target variable

[*] Parts of this section are reprinted with permission from "Exact Performance of CoD Estimators in Discrete Predicition" by T. Chen, and U. Braga-Neto, 2012, *EURASIP Journal on Advances in Signal Processing (JASP), Special Issue on Genomic Signal Processing*, Article ID 487893, 13 pages, 2010. doi:10.1155/2010/487893, © 2010 EURASIP, and "Maximum-Likelihood Estimation of the Discrete Coefficient of Determination in Stochastic Boolean Systems" by T. Chen and U. Braga-Neto, 2012, *IEEE Transactions on Signal Processing*, vol. 61, no. 15, pp. 3880–3894, © 2013 IEEE.

$Y \in \{0,1\}$, as given by the simple formula:

$$\text{CoD} = \frac{\varepsilon_0 - \varepsilon}{\varepsilon_0}, \tag{1.1}$$

where ε_0 is the error of the best predictor of Y in the absence of other observations and ε is the error of the best predictor of Y based on the observation of $\mathbf{X}$. The binary CoD measures the relative decrease in prediction error when using predictor variables to estimate the target variable, as opposed to using no predictor variables. The closer it is to one, the tighter the regulation of the target variable by the predictor variables is, whereas the closer it is to zero, the looser the regulation is. The CoD will correctly produce low values in cases where the no-predictor error is already small, or when adding predictors does not contribute to a significant decrease in error.

The concept of CoD has far-reaching applications in Genomics. The CoD was perhaps the first predictive paradigm utilized in the context of microarray data, the goal being to provide a measure of nonlinear interaction among genes [31,46,47,52,62,71]. In [47,52,71], the CoD is applied to the prediction problem dealing with gene expressions quantized into discrete levels in discrete prediction. In [46,62], the CoD has its application in the reconstruction or inference of gene regulatory networks. As its classic counterpart, the binary CoD is a goodness-of-fit statistic that can be used to assess the relationship between predictor and target variables, for example, the associations between gene expression patterns in practical applications. The CoD permits biologists to focus on particular connections in the genome, and coefficient estimates are useful even if they are biased and not overly precise, because at least the estimated coefficients provide a practical means of discrimination among potential predictor sets [31].

1.2 Contributions

The contributions made in this dissertation can be summarized into two parts. First, we propose a frequentist inference framework for the estimation and testing of the discrete CoD with the applications to the system identification problems in Genomics. We enrich the existing theory of the discrete CoD by studying both nonparametric and parametric estimation of the CoD. Meanwhile, we develop novel statistic tools for the discovery of significant gene regulatory relationships by conducting multiple tests for the nonzero CoD and for the detection of significant canalyzing genes by testing the nonzero intrinsically multivariate prediction (IMP) criterion. Secondly, we discuss a Bayesian inference framework for the estimation of the CoD across a family of parametrized distributions of target and predictors from an optimization perspective, and demonstrate its applications in several groups of simulations for the recovery of gene regulatory relationships using synthetic and real gene-expression data sets.

1.2.1 Frequentist Inference of the CoD

The error of the best predictor corresponds to the optimal prediction error, also known as Bayes error, given a probability model [30, 32]. However, in practical real-world problems, the underlying probability model is unknown, and thus we arrive at the fundamental issue of how to find a good prediction error estimator in small-sample settings [10, 11]. An error estimator may be a deterministic function of the sample data, in which case it is called a non-randomized error estimator; such popular error estimators as resubstitution and leave-one-out are examples. These error estimators are random only through the random sample data. Closed-form analytical expressions for performance metrics such as bias, deviation variance, and RMS of resubstitution and leave-one-out error estimators have been given in [10, 58]. By contrast, randomized error estimators, like cross-validation and bootstrap, have "internal" random factors that affect their outcome, and thus approximate approaches, usually via Monte-Carlo sampling, are typically used to analyze their performance.

Likewise, the CoD can be estimated from sample data, so that we can speak of non-randomized CoD estimators, including the resubstitution and leave-one-out CoD estimators, and randomized CoD estimators, including bootstrap and cross-validation CoD estimators [21]. A CoD estimator is obtained by using one of the usual error estimators for the prediction error with variables, and the empirical frequency (resubstitution) estimator for the prediction error with no variables. Assuming no knowledge about the underlying probability model, we will employ the discrete histogram rule [11, 30], the most widely-used and intuitive rule for discrete prediction problems, in order to estimate prediction errors from the sample data.

We present, for the first time, an exact formulation for performance metrics of the resubstitution and leave-one-out CoD estimators, for the discrete histogram rule. Numerical experiments are carried out using a parametric Zipf model, where we compute the exact performance of resubstitution and leave-one-out CoD estimators using the previously derived formulas, for varying actual CoD, sample size, and bin size [21]. We compare these results to approximate performance metrics of randomized CoD estimators (bootstrap and cross-validation), computed via Monte-Carlo sampling. The numerical results indicate that, for moderate and large values of the actual CoD, the resubstitution CoD estimator is the least biased and least variable among all CoD estimators, especially at small number of predictors. In fact, with two predictors, the resubstitution CoD nearly dominates uniformly over all other estimators across all values of actual CoD. The leave-one-out and cross-validation CoD estimator tend to perform the worst, whereas the performance of the bootstrap

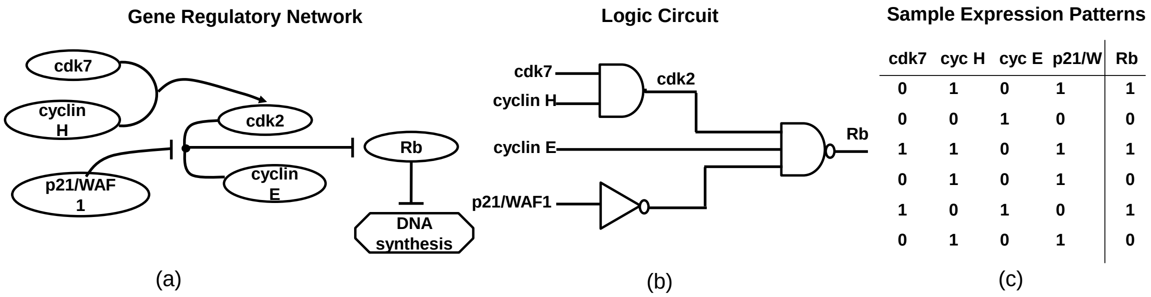
CoD estimator is intermediary, despite its high computational complexity. This indicates that, provided one has evidence of moderate to tight regulation between the genes, and the number of predictors is not too large, the CoD estimator based on resubstitution is the CoD estimator of choice [21].

Besides, we are most concerned with the feasibility of presenting a reasonable mathematical model that can incorporate prior knowledge about biological systems. This can be answered by introducing stochastic Boolean models that play a prominent role in many applications, particularly in Genomic Signal Processing [61]. Figure 1.1 displays an example of regulatory network associated with the cell cycle. Figure 1.1(a) gives gene regulatory relationships that lead to the activation or depression of DNA synthesis. Figure 1.1(b) shows a logic circuit that functions the same as the network. It is obvious that DNA synthesis occurs according to the following equation:

$$\text{DNA synthesis} \;=\; \overline{\text{Rb}} \;=\; \text{CDK7} \wedge \text{CycH} \wedge \text{CycE} \wedge \overline{\text{p21}}\,, \tag{1.2}$$

which tells that, in a healthy cell, DNA synthesis occurs only if all of the CDK7, Cyclin H and Cyclin E genes are active and the p21 gene is silenced [62].

A common task in practice is the estimation of the strength of regulation between the various components of the Boolean circuit from sample data according to partial information or even no information available about the system. Estimation and identification are complicated by the presence of *system noise.* For example, consider the expression pattern “0 1 0 1” for the predicting genes in the hypothetical sample data of Figure 1.1(c). According to eq (3.1), the state of the Rb gene should be active, and no DNA synthesis should occur. However, three instances of the “0 1 0 1” pattern are observed in the data, and only one of them behaves as the mechanistic



cdk7	cyc H	cyc E	p21/W	Rb
0	1	0	1	1
0	0	1	0	0
1	1	0	1	1
0	1	0	1	0
1	0	1	0	1
0	1	0	1	0

Figure 1.1: Example of regulatory network, equivalent logic circuit, and hypothetical sample data for the DNA synthesis pathway of the cell cycle. Adapted from Shmulevich et al. [62].

model predicts. This is the result of uncertainty in the mechanistic model, e.g., the influence of hidden or latent variables. An additional difficulty is the fact that many expression patterns may be missing due to a small number of samples. These considerations motivate the application of a stochastic approach to the problem.

As opposed to nonparametric methods, we propose a parametric maximum-likelihood estimation (MLE) approach, by introducing stochastic Boolean models for static and dynamical systems, and deriving the maximum-likelihood estimator of the CoD. In the static case, we are interested in the CoD of a Boolean target with respect to a Boolean predictor vector. In the dynamical case, we assume that there is a Markov Boolean state process, and we are interested in the CoD of each state variable with respect to the state vector at the previous time point, after the system has reached the steady state. In each case, the relationship between targets and predictors is contaminated by noise, the amplitude of which is not known and must be estimated.

The basic idea behind parametric ML estimation is to take advantage of partial knowledge about the model describing system behaviour. This information cannot be used by nonparametric approaches, which must rely purely on the sample data. In many applications, prior knowledge about the system is available, even if this knowledge is incomplete. This is common, for example, in Genomic Signal Processing applications, where it is often the case that partial knowledge about the biochemical pathways of interest is known, making the parametric approach especially suited to this domain. Naturally, as more prior knowledge becomes available, the more we expect that the parametric ML approach will outperform its nonparametric competitors.

We develop a static Boolean model associated with an arbitrary predictor vector and a dynamical Boolean model for dynamical systems in the steady state [16,22]. In the static case, analytic expressions for the asymptotic bias and variance of the ML CoD estimator are derived. Performance of the ML CoD estimator is compared to the nonparametric alternatives in terms of bias, variance, and RMS, and the results indicate that the parametric approach is to be preferred, provided that the system noise level is not too high [16].

We also consider the system identification problem [50], that is, the case where not only the system noise statistics are unknown, but also there is incomplete knowledge about the Boolean relationships in the system. This may manifest itself as partial knowledge about the logic gates regulating each target variable or about which variables are the input to each logic gate (i.e., the network "wiring"). The prior knowledge about the system is coded into a set of candidate models. In practice, the choice of models to be included in the candidate model set is a difficult one. L. Ljung states "It is here that a priori knowledge and engineering intuition and insight have to be combined with formal properties of models." [50]. Here, we consider the practical situation where partial knowledge may exist about which logic gates are present in the system, but no knowledge exists about the wiring, except for the degree of network connectivity, i.e., the number of inputs per gate. We propose inference procedures based on the parametric ML CoD estimator to recover the missing information, and investigate their performance by means of numerical experiments, which showed that identification rates converge to 100% as sample size increases, and that the convergence rate is much faster as more prior knowledge is available. For wiring identification, the parametric ML approach is compared to the nonparametric approaches, which showed that the parametric approach produces superior identification rates, though as the amount of prior knowledge is reduced, its performance approaches that of the nonparametric ML estimator, which was generally the best nonparametric approach in all our experiments.

The fact that the parametric approach in the presence of prior knowledge turns out to be superior to nonparametric alternatives is not surprising, but the amount of improvement obtained as a function of system noise level and sample size is of interest, and not obvious a-priori. One of the goals of our work is to quantify the degree of improvement achieved by the use of the parametric approach in estimation and system identification tasks.

Traditional applications of the CoD so far have been based on user-selected thresholds to decide on the presence of gene regulation between the given predictor and target genes. To address this problem, we develop a statistically rigorous tool for this inference problem, by providing a statistical test, and associated confidence interval, for a nonzero CoD between given Boolean predictors and a Boolean target. Rejection of the null hypothesis of zero CoD gives evidence for the presence of statistically-significant regulation [17].

This is done by framing the problem in the context of a stochastic logic model that naturally allows the inclusion of prior knowledge if available; e.g., knowledge about the logic gate governing the relationship sought for. For example, knowledge about a canalizing relationship [69], i.e. a logic relationship in a class of AND or OR gates (with possibly negated inputs), can be easily added. Then an Intersection-Union Test (IUT) [14] based on likelihood-ratio tests for the individual model parameters is developed by deriving its rejection region, power, p-value, and associated confidence interval.

To be useful as an inferential tool, the proposed methodology must be able to deal with the multiple testing issue created by modern gene-expression experiments that monitor thousands of genes simultaneously. We address this by describing the application of two multiple testing procedures to control the overall Type I error rate, namely the single-step Bonferroni correction and the step-up Benjamini-Hochberg procedure, for controlling the family-wise error rate (FWER) and the false discovery rate (FDR), respectively [3, 34]. The properties of the proposed statistical test and multiple testing correction procedures are assessed by both theoretical analysis and Monte-Carlo experiments, in order to analyze how FWER, FDR, average power, and the confidence interval estimates behave under FWER- and FDR-controlling procedures, for varying sample size and number of multiple tests. Furthermore, we apply the proposed methodology to real gene-expression data sets, and the proposed methodology could be verified to be a promising tool for discovery of significant gene regulatory relationships from discrete gene-expression data.

Another problem of interest is how to identify canalyzing genes from a modelling perspective. Canalizing genes are frequently found in signalling pathways, which deliver information from a variety of sources to the machinery that enacts central cellular functions such as cell-cycle, survival, apoptosis and metabolism. For example, DUSP1 antagonizes the activity of the p38 mitogen activated kinase, MAPK1 (ERK), which is a central component of the pathway by which extracellular signal-regulated kinases send mitogenic signals [15]. Therefore, DUSP1 is canalyzing when it dephosphorylates MAPK1. Martins and collaborators [52] defined the concept of *intrinsically multivariate prediction*, in which case when the controlling gene is active, it cannot be well-predicted by subsets of its predictor genes, but it can be predicted by the full set with great accuracy. Such a set of predictor genes is called Intrinsically Multivariate Predictive (IMP) set for the target gene [52]. Their work showed that DUSP1 had the largest number of IMP gene sets in related pathways, thereby providing evidence that the IMP criterion could be used as a practical tool for discovery of canalyzing genes [52]. However, applications of the IMP criterion so far have been based on user-selected thresholds to decide on the presence of gene multivariate prediction between target and predictor genes. We develop a statistically rigorous tool for this inference problem, by providing a statistical test for a nonzero IMP score between given a Boolean target and Boolean predictors. Rejection of the null hypothesis of zero IMP score gives evidence for the presence of IMP properties of statistical significance [24]. This idea is quite similar to that used for the detection of gene regulation between given predictor and target genes by testing the nonzero associated CoD [17]. Furthermore, multiple testing procedures are proposed by considering the availability of thousands of genes in gene-expression experiments. Examples of applications of IMP-based multiple testing procedure are provided using both synthetic and real data sets.

1.2.2 Bayesian Inference of the CoD

As mentioned in the frequentist perspective, nonparametric CoD estimators are defined by the discrete histogram prediction rule, while ML model-based CoD estimators are defined with respect to a parametric model. However, none of these CoD estimators are optimized based on statistical inference across a family of possible joint distributions between target and predictors, where the mass of the random parameter concentrates around true parameter values for the true target-predictor distribution.

This leads to a Bayesian approach to CoD estimation based on a parametrized family of target-predictor distributions as a function of random parameters characterized by assumed prior distributions. Such an idea was first introduced in the study of Bayesian error estimation for classification, which optimizes sample-based error estimation relative to mean-square error (MSE) between the error estimator and true error across a family of feature-label distributions [26, 27].

Following the Bayesian idea, we present the definition of one Bayesian CoD estimator in the minimum mean-square error (MMSE) sense, that is, the Bayesian MMSE CoD estimator, which minimizes the MSE with respect to the optimal CoD. Appropriate priors are specified for a exact formulation of the Bayesian MMSE CoD estimator based on discrete sample data. In addition, Dalton and Dougherty designed an optimal Bayesian classifier that minimizes the Bayesian MMSE CoD estimator over all classifiers from an arbitrary family of classifiers [28, 29]. Then we develop another Bayesian CoD estimator using the optimal Bayesian classifier, whose performance (i.e., bias, variance, RMS) can be analytically expressed. We compare the performance of the two Bayesian CoD estimators against those of the nonparametric CoD estimators, and validate the better performance of the Bayesian ones that allow the inclusion of prior knowledge. We also propose Bayesian predictor inference procedures for the recovery of gene regulatory relationships (i.e., wiring and logic gates), and compare their performance against the frequentist predictor inference algorithms based on nonparametric and parametric ML CoD estimators in Section 3.

1.3 Organization

This dissertation is organized as follows.

In Section 2, we define several nonparametric CoD estimators that are functions of nonparametric error estimators like resubstitution, leave-one-out, bootstrap and cross-validation, from a frequentist perspective. We formulate the analytical expressions of the performance metrics (i.e., bias, variance and RMS) of these CoD estimators. Furthermore, we assess their performance by using a Zipf model.

In Section 3, we propose a parametric maximum-likelihood estimation framework for the inference of the discrete CoD from sample data. We introduce stochastic Boolean models for biology systems, and deriving the maximum-likelihood estimator of the CoD given sample data drawn from the underlying distribution. We discuss the performance of ML CoD estimators based on static Boolean models and dynamical Boolean models, respectively. Furthermore, ML-based inference algorithms are developed for the identification of gene regulatory relationships in both static and dynamic cases. We validate our proposed algorithms using synthetic gene-expression data by groups of simulations.

In Section 4, we provide a statistical test for a nonzero CoD between given Boolean predictors and a Boolean target in the context of a stochastic logic model, and develop a practical statistic tool for the detection of significant gene regulatory relationships from discrete gene-expression data. We develop multiple testing procedures based on the discrete CoD, and apply our methodology to synthetic and real gene-expression data for further validation.

In Section 5, we present a rigorous statistical testing framework to investigate the property of intrinsically multivariate predictive (IMP) of canalyzing genes, by using the IMP criterion in the context of discrete CoD. Multiple testing procedures based on the IMP criterion are proposed with the applications to real gene-expression data for the detection of significant canalyzing genes.

In Section 6, we introduce a Bayesian inference framework to estimate the CoD based on a parametrized family of joint distributions of given target and predictors as a function of random parameters characterized by preassumed prior distributions. We examine the performance of well-defined Bayesian CoD estimators, and furthermore propose Bayesian predictor inference procedures with the applications to synthetic gene-expression data sets.

In Section 7, we present concluding remarks and prospects in future research.

2. FREQUENTIST INFERENCE: NONPARAMETRIC COD ESTIMATION*

The coefficient of determination (CoD) has significant applications in Genomics, for example, in the inference of gene regulatory networks [31, 46, 47, 52, 62, 71]. The CoD is closely related with the prediction error depending on the joint distribution between target and predictor variables, which, however, are usually unknown in practice. Hence, the CoD must be estimated from sample data that are drawn from the target-predictor distribution. In this chapter, we study several nonparametric CoD estimators based upon the resubstitution, leave-one-out, cross-validation, and bootstrap error estimators, from a frequentist perspective. The frequentist inference approach gives an evaluative paradigm for a repeatable randomly sampling process with unknown parameters of the true distribution remaining fixed, allowing no information prior to model specification [14]. We are mostly interested in the comparison among the performance of these nonparametric CoD estimators in such a setting, which will be addressed in this chapter.

2.1 Discrete Prediction

Let $X_1, X_2, \ldots, X_p$ be p predictor random variables, such that each X_i take on a finite number b_i of values, and $Y \in \{0, 1\}$ be the target random variable, for the discrete prediction problem. The predictors as a group can take on values in a finite space with $b = \prod_{i=1}^{p} b_i$ possible states. For analysis purposes, we establish a bijection

*Parts of this section are reprinted with permission from "Exact Performance of CoD Estimators in Discrete Predicition" by T. Chen, and U. Braga-Neto, 2012, *EURASIP Journal on Advances in Signal Processing (JASP), Special Issue on Genomic Signal Processing*, Article ID 487893, 13 pages, 2010. doi:10.1155/2010/487893, © 2010 EURASIP, and "Approximate expressions for the variances of non-randomized error estimators and CoD estimators for the discrete histogram rule" by T. Chen, and U. Braga-Neto, 2012, *Proceedings of VIII IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2009)*, Cold Spring Harbor, NY, November 2010, pp 1–4, © 2010 IEEE.

between this finite state space and a single predictor variable X taking values in the set $X \in \{1, 2, \ldots, b\}$. The variable X has a one-to-one relationship with the finite space state coded by $X_1, X_2, \ldots, X_p$: one specific value of X represents a specific combination of the values of the original predictors, i.e., a "bin" into which the data is categorized. The value b is the number of bins, which provides a direct measure of predictor complexity.

The probability model for the pair (X, Y) is specified by class prior probabilities: $c_0 = P(Y = 0), c_1 = P(Y = 1)$, and class-conditional probabilities: $p_i = P(X = i \mid Y = 0)$ and $q_i = P(X = i \mid Y = 1)$, for $i = 1, \ldots, b$, where we have the identities

$$\begin{aligned} c_0 + c_1 &= 1\,, \\ \sum_{i=1}^{b} p_i \;&=\; 1\,, \\ \sum_{i=1}^{b} q_i \;&=\; 1\,. \end{aligned} \tag{2.1}$$

Given a specific probability model, the optimal predictor for the problem is given by

$$\psi(X = i) \;=\; \begin{cases} 1, & c_1 q_i > c_0 p_i \\ 0, & \text{o.w.} \end{cases} \,. \tag{2.2}$$

with optimal error rate, also called the Bayes error [30], determined by

$$\varepsilon \;=\; \sum_{i=1}^{b} \min\{c_0 p_i, c_1 q_i\}\,. \tag{2.3}$$

If no features are provided, the optimal error rate becomes

$$\varepsilon_0 \;=\; \min\{c_0, c_1\}\,. \tag{2.4}$$

By using the simple inequality $\sum \min\{a_i, b_i\} \leq \min\{\sum a_i, \sum b_i\}$, one concludes that $\varepsilon \leq \varepsilon_0$ in all cases.

The coefficient of determination [31] is defined as (assuming that $\varepsilon_0 \neq 0$):

$$\text{CoD} \;=\; \frac{\varepsilon_0 - \varepsilon}{\varepsilon_0} \;=\; 1 - \frac{\varepsilon}{\varepsilon_0} \;=\; 1 - \frac{\sum_{i=1}^{b} \min\{c_0 p_i, c_1 q_i\}}{\min\{c_0, c_1\}} \tag{2.5}$$

Since $0 \leq \varepsilon \leq \varepsilon_0$, we have that $0 \leq \text{CoD} \leq 1$. We have $\text{CoD} = 1$ if and only if $\varepsilon = 0$, that is, there is perfect regulation between predictors and target. On the other hand, $\text{CoD} = 0$ if and only if $\varepsilon = \varepsilon_0$, that is, the predictors exert no regulation on the target.

2.2 Nonparametric CoD Estimation

In practice, the underlying probability model is unknown, and thus the CoD is not known. The need arises thus to find estimators of the CoD from i.i.d. sample data $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn from the unknown probability model distribution. All CoD estimators considered here will be of the form:

$$\widehat{\text{CoD}} \;=\; \frac{\hat{\varepsilon}_0 - \hat{\varepsilon}}{\hat{\varepsilon}_0} \;=\; 1 - \frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\,, \tag{2.6}$$

where $\hat{\varepsilon}$ is one of the usual error estimators for a selected discrete prediction rule, and $\hat{\varepsilon}_0$ is the empirical frequency estimator for the prediction error with no variables:

$$\hat{\varepsilon}_0 \;=\; \min\left\{\frac{N_0}{n}, \frac{N_1}{n}\right\}\,. \tag{2.7}$$

where N_0 and N_1 are random variables corresponding to the number of sample points belonging to classes $Y = 0$ and $Y = 1$, respectively. We assume throughout that $\hat{\varepsilon}_0 \neq 0$, that is, each class is represented by at least one sample. Note that $\hat{\varepsilon}_0$ has the desirable property of being a universally consistent estimator of ε_0 in (2.4), that is, $\hat{\varepsilon}_0 \to \varepsilon_0$ in probability (in fact, almost surely) as $n \to \infty$, regardless of the probability model.

The discrete prediction rule to be used with the error estimator $\hat{\varepsilon}$ is the discrete histogram rule, which is the "plug-in" rule for approximating the minimum-error Bayes predictor [10]. Even though we make this choice, we remark that the methods described here can be applied to any discrete prediction rule. Given the sample data S_n, the discrete histogram classifier is given by:

$$\psi_n(X = i) \;=\; I_{V_i > U_i} \;=\; \begin{cases} 1, & V_i > U_i \\ 0, & U_i \geq V_i \end{cases} \;, \quad i = 1, 2, \ldots, b\,, \tag{2.8}$$

where U_i is the number of samples with $Y = 0$ in bin $X = i$, and V_i is the number of samples with $Y = 1$ in bin $X = i$, for $i = 1, \ldots, b$.

We review next some facts about the distribution of the random vectors $\mathbf{U} = \{U_1, \ldots, U_b\}$ and $\mathbf{V} = \{V_1, \ldots, V_b\}$, which will be needed in the sequel. The variables $N_0 = \sum_{i=1}^{b} U_i$, $N_1 = \sum_{i=1}^{b} V_i$, U_i, and V_i, for $i = 1, \ldots, b$, are random variables due to the randomness of the sample data S_n (this is the case referred to as "full sampling" in [10]). More specifically, N_i is a random variable binomially distributed with parameters (n, c_i), i.e., $N_i \sim B(n, c_i)$, for $i = 0, 1$, while the vector-valued random variable (U_i, V_i) is trinomially distributed with the parameter set $(n, c_0 p_i, c_1 q_i)$, that is,

$$P(U_i = k, V_i = l) \;=\; \binom{n}{k, l, n-k-l} (c_0 p_i)^k (c_1 q_i)^l (1 - c_0 p_i - c_1 q_i)^{n-k-l}, \tag{2.9}$$

for $i = 1, \ldots, b$. In addition, the vector $\{U_1, \ldots, U_b, V_1, \ldots, V_b\}$ follows a multinomial distribution with parameters $(n, c_0p_1, \ldots, c_0p_b, c_1q_1, \ldots, c_1q_b)$, so that

$$\begin{aligned} P(U_1 = u_1, \ldots, U_b = u_b, V_1 = v_1, \ldots, V_b = v_b) \; = \\ \binom{n}{u_1, \ldots, u_b, v_1, \ldots, v_b} \; \times \; (c_0p_1)^{u_1} \ldots (c_0p_b)^{u_b} (c_1q_1)^{v_1} \ldots (c_1q_b)^{v_b} \,. \end{aligned} \tag{2.10}$$

We introduce next each of the CoD estimators considered in this chapter.

2.2.1 Resubstitution CoD Estimator

This corresponds to the choice of resubstitution [65] as the prediction error estimator:

$$\widehat{\mathrm{CoD}}_r \;=\; 1 - \frac{\hat{\varepsilon}_r}{\hat{\varepsilon}_0}, \tag{2.11}$$

where, for the discrete histogram predictor,

$$\hat{\varepsilon}_r \;=\; \frac{1}{n} \sum_{i=1}^{b} \left[U_i I_{V_i > U_i} + V_i I_{U_i \geq V_i} \right] \,. \tag{2.12}$$

The resubstitution CoD can be written equivalently as

$$\widehat{\mathrm{CoD}}_r \;=\; 1 - \frac{\sum_{i=1}^{b} \min\{\frac{N_0}{n} \times \frac{U_i}{N_0}, \frac{N_1}{n} \times \frac{V_i}{N_1}\}}{\min\left\{\frac{N_0}{n}, \frac{N_1}{n}\right\}} \,, \tag{2.13}$$

which reveals that $\widehat{\mathrm{CoD}}_r$ has the desirable property of being a universally consistent estimator of CoD in (2.5), that is, $\widehat{\mathrm{CoD}}_r \rightarrow$ CoD in probability (in fact, almost surely) as $n \rightarrow \infty$, regardless of the probability model.

2.2.2 Leave-One-Out CoD Estimator

This corresponds to the choice of the leave-one-out error estimator [48] as the prediction error estimator:

$$\widehat{\mathrm{CoD}}_l \;=\; 1 - \frac{\hat{\varepsilon}_l}{\hat{\varepsilon}_0}, \tag{2.14}$$

where, for the discrete histogram predictor (as can be readily checked),

$$\hat{\varepsilon}_l \;=\; \frac{1}{n}\sum_{i=1}^{b}\left[U_i I_{V_i \geq U_i} + V_i I_{U_i \geq V_i - 1}\right] \,. \tag{2.15}$$

The leave-one-out CoD estimator provides an opportunity to reflect on the uniform choice of the empirical frequency estimator $\hat{\varepsilon}_0$ in (3.9) as an estimator of ε_0, including here. Clearly, the empirical frequency corresponds to the resubstitution estimator of ε_0. The question arises as to whether, for the leave-one-out CoD estimator, the leave-one error estimator of ε_0 should be used instead. For $N_0 = N_1 = n/2$, we get $\hat{\varepsilon}_0 = 1/2$ with the choice of the resubstitution estimator (empirical frequency), but $\hat{\varepsilon}_0 = 1$ with the choice of leave-one-out estimator, which is a useless result. Similar problems beset other estimators of ε_0. Hence, the empirical frequency estimator is employed here as the estimator of ε_0 for all CoD estimators.

2.2.3 Cross-Validation CoD Estimator

This corresponds to the choice of the cross-validation error estimator [48,66] as the prediction error estimator. In k-fold cross-validation, sample data S_n is partitioned into k folds S_i, for $i = 1, \ldots, k$. For simplicity, we assume that k can divide n. A classifier ψ_i is designed on the training set $S_n \backslash S_i$, and tested on S_i, for $i = 1, \ldots, k$. Since there are different partitions of the data into k folds, one can repeat the k-fold cross-validation r times and then average the results. Such a process leads to the r-repeated k-fold cross-validation error estimator $\hat{\varepsilon}_{cv}$, given by

$$\hat{\varepsilon}_{cv} \;=\; \frac{1}{nr}\sum_{m=1}^{r}\sum_{i=1}^{k}\sum_{j=1}^{n/k}|Y_j^{i,m} - \psi_{i,m}(X_j^{i,m})|, \tag{2.16}$$

where $(X_j^{i,m}, Y_j^{i,m})$ represents the j-th sample point in the i-th fold for the m-th repetition of the cross-validation, for $i = 1, \ldots, k$, $m = 1, \ldots, r$ and $j = 1, \ldots, n/k$.

Based upon (2.16), the r-repeated k-fold cross-validation CoD estimator is defined by

$$\widehat{\text{CoD}}_{cv} \;=\; 1 - \frac{\hat{\varepsilon}_{cv}}{\hat{\varepsilon}_0}, \tag{2.17}$$

In order to get reasonable variance properties, a large number of repetitions may be required, which can make the cross-validation CoD estimator slow to compute.

2.2.4 Bootstrap CoD Estimator

This corresponds to the use of the bootstrap [35, 36] for the prediction error estimator. A bootstrap sample $S_n^* = \{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ consists of n equally-likely draws with replacement from the original data S_n. Some sample points from the original data may appear multiple times in the bootstrap sample, whereas other sample points may not appear at all. The actual proportion of times a sample point (X_i, Y_i) appears in S_n^* can be written as $P_i^* = \frac{1}{n}\sum_{j=1}^n I_{(X_i^*, Y_i^*) = (X_i, Y_i)}$, for $i = 1, \ldots, n$. A predictor ψ_t may be designed on a bootstrap sample S_n^{*t}, and tested on $S_n \backslash S_n^{*t}$, for $t = 1, \ldots, T$, where T is a sufficiently large number of repetitions (in this paper, $T = 100$). Then, the basic bootstrap zero estimator is given by

$$\hat{\varepsilon}_{\text{ZERO}} \;=\; \frac{\sum_{t=1}^T \sum_{i=1}^n |Y_i - \psi_t(X_i)| I_{P_i^{*t}=0}}{\sum_{t=1}^T \sum_{i=1}^n I_{P_i^{*t}=0}}, \tag{2.18}$$

The .632 bootstrap estimator then performs a weighted average of the bootstrap zero and resubstitution estimators:

$$\hat{\varepsilon}_{b632} \;=\; (1 - 0.632)\hat{\varepsilon}_r + 0.632\,\hat{\varepsilon}_{\text{ZERO}}. \tag{2.19}$$

Based on (2.18) and (2.19), the .632 bootstrap CoD estimator is then defined as

$$\widehat{\mathrm{CoD}}_{b632} \; = \; 1 - \frac{\hat{\varepsilon}_{b632}}{\hat{\varepsilon}_0}, \tag{2.20}$$

The bootstrap CoD estimator can be very slow to compute due to the complexity of $\hat{\varepsilon}_{\mathrm{ZERO}}$.

2.3 Performance Metrics of CoD Estimators

In analogous fashion to the performance metrics of prediction error estimators [11], the key performance metrics for an CoD estimator $\widehat{\mathrm{CoD}}$ are its bias,

$$\mathrm{Bias}\left[\widehat{\mathrm{CoD}}\right] \; = \; E\left[\widehat{\mathrm{CoD}} - \mathrm{CoD}\right] \; = \; E\left[\widehat{\mathrm{CoD}}\right] - \mathrm{CoD}\,, \tag{2.21}$$

the deviation variance (which in the present case is equal simply to its variance),

$$\mathrm{Var_d}\left[\widehat{\mathrm{CoD}}\right] \; = \; \mathrm{Var}\left(\widehat{\mathrm{CoD}} - \mathrm{CoD}\right) \; = \; \mathrm{Var}\left(\widehat{\mathrm{CoD}}\right)\,, \tag{2.22}$$

and the root mean-square (RMS) error,

$$\mathrm{RMS}\left[\widehat{\mathrm{CoD}}\right] \; = \; \sqrt{E\left[\left(\widehat{\mathrm{CoD}} - \mathrm{CoD}\right)^2\right]} \; = \; \sqrt{\mathrm{Var}\left[\widehat{\mathrm{CoD}}\right] \; + \; \mathrm{Bias}\left[\widehat{\mathrm{CoD}}\right]^2} \tag{2.23}$$

For a given probability model, all the performance metrics are thus obtained as a function of the expectation $E[\widehat{\mathrm{CoD}}]$ and variance $\mathrm{Var}(\widehat{\mathrm{CoD}})$.

Working further, we obtain

$$E[\widehat{\mathrm{CoD}}] \; = \; 1 \; - \; E\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right], \tag{2.24}$$

and

$$\mathrm{Var}[\widehat{\mathrm{CoD}}] \;=\; E\left[(\widehat{\mathrm{CoD}})^2\right] \;-\; \left(E[\widehat{\mathrm{CoD}}]\right)^2 \;=\; E\left[\frac{\hat{\varepsilon}^2}{\hat{\varepsilon}_0^2}\right] \;-\; \left(E\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right]\right)^2 , \qquad (2.25)$$

as can be easily checked. We conclude that all the key performance metrics for CoD estimators can be obtained from the first and second moments of $\hat{\varepsilon}/\hat{\varepsilon}_0$.

2.4 Exact Moments of Non-Randomized CoD Estimators

As mentioned in the Introduction, we can categorize CoD estimators into non-randomized and randomized, depending on whether the prediction error estimator $\hat{\varepsilon}$ is non-randomized or randomized. Non-randomized CoD estimators, such as the resubstitution and leave-one-out CoD estimators, are deterministic functions of the sample data, which makes it possible an analytical formulation of their performance metrics. On the other hand, the performance of randomized CoD estimators, such as the cross-validation and bootstrap CoD estimators, is very difficult to study analytically and is typically investigated via Monte-Carlo sampling (which is done in Section 2.6).

In this section, we will present exact expressions for the computation of the first moment $E\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right]$ and the second moment $E\left[\frac{\hat{\varepsilon}^2}{\hat{\varepsilon}_0^2}\right]$ for the case of resubstitution and leave-one-out error estimators, which suffices to compute the bias, variance, and RMS of the corresponding CoD estimator, as discussed in the previous section. These expressions are functions only of sample size, number of bins (complexity) and the probability model. We will assume throughout, for definiteness, that the sample size n is even. The case where n is odd is in fact slightly simpler and can be readily obtained in analogous fashion to the derivations presented below.

2.4.1 Resubstitution

The first moment of $\hat{\varepsilon}_r/\hat{\varepsilon}_0$ is given by

$$E\left[\frac{\hat{\varepsilon}_r}{\hat{\varepsilon}_0}\right] \;=\; E\left[E\left[\frac{\hat{\varepsilon}_r}{\hat{\varepsilon}_0} \mid \hat{\varepsilon}_0\right]\right] \;=\; \sum_{m=1}^{n/2} E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid M = m\right] P(M = m), \qquad (2.26)$$

where $M = n\hat{\varepsilon}_0$. Since $\hat{\varepsilon}_0 = \frac{1}{n}\min(N_0, N_1)$, we have $M = \min(N_0, n - N_0)$. It follows that the event $[M = m]$ is equal to the union of the disjoint events $[N_0 = m]$ and $[N_0 = n{-}m]$, for $m = 1, \ldots, n/2 - 1$, whereas $[M = n/2] = [N_0 = n/2]$. By using Proposition A in the Appendix, we can write both cases in a single expression as follows:

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid M = m\right] \;=\; & \frac{P(N_0 = m)}{P(N_0 = m) + P(N_0 = n{-}m)} E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid N_0 = m\right] I_{1 \le m < \frac{n}{2}} + \\ & \frac{P(N_0 = n{-}m)}{P(N_0 = m) + P(N_0 = n{-}m)} E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid N_0 = n{-}m\right] I_{1 \le m \le \frac{n}{2}}\,, \\ & m = 1, \ldots, n/2\,. \end{aligned} \qquad (2.27)$$

By using (2.27) in (2.26) and considering that $P(M = m) = P(N_0 = m) + P(N_0 = n{-}m)$, we obtain

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_r}{\hat{\varepsilon}_0}\right] \;=\; \sum_{m=1}^{n/2} \Bigg\{ & E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid N_0 = m\right] P(N_0 = m) I_{1 \le m < \frac{n}{2}} + \\ & E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid N_0 = n{-}m\right] P(N_0 = n{-}m) I_{1 \le m \le \frac{n}{2}} \Bigg\}, \end{aligned} \qquad (2.28)$$

where

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_r}{m/n} \mid N_0 = t\right] \;=\; & \frac{1}{m}\sum_{i=1}^{b}\left\{\sum_{l>k} kP(U_i = k, V_i = l \mid N_0 = t) + \right. \\ & \left. \sum_{k\geq l} lP(U_i = k, V_i = l \mid N_0 = t)\right\}, \end{aligned} \tag{2.29}$$

with

$$\begin{aligned} P(U_i = k, V_i = l \mid N_0 = t) \;&=\; P(U_i = k \mid N_0 = t)\, P(V_i = l \mid N_1 = n{-}t) \\ &=\; \binom{t}{k} p_i^k (1{-}p_i)^{t-k} \binom{n{-}t}{l} q_i^l (1{-}q_i)^{n-t-l}, \end{aligned} \tag{2.30}$$

for $t = m,\ n{-}m$.

The second moment of $\hat{\varepsilon}_r/\hat{\varepsilon}_0$ is given by

$$E\left[\frac{\hat{\varepsilon}_r^2}{\hat{\varepsilon}_0^2}\right] \;=\; \sum_{m=1}^{n/2} E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid M = m\right] P(M = m), \tag{2.31}$$

where $M = n\hat{\varepsilon}_0$, as before. By using Proposition 1 in the Appendix, and the same reasoning applied previously in the case of the first moment, we can write

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid M = m\right] \;=\; & \frac{P(N_0 = m)}{P(N_0 = m) + P(N_0 = n{-}m)} E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid N_0 = m\right] I_{1\leq m<\frac{n}{2}} \\ & + \frac{P(N_0 = n{-}m)}{P(N_0 = m) + P(N_0 = n{-}m)} E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid N_0 = n{-}m\right] I_{1\leq m\leq\frac{n}{2}}\,,\ m = 1, \ldots, n/2\,. \end{aligned} \tag{2.32}$$

Combining (2.32) and (2.31) leads to

$$
\begin{aligned}
E\left[\frac{\hat{\varepsilon}_r^2}{\hat{\varepsilon}_0^2}\right] \;=\; \sum_{m=1}^{n/2} \Bigg\{ & E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid N_0 = m\right] P(N_0 = m) I_{1\le m<\frac{n}{2}} + \\
& E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid N_0 = n-m\right] P(N_0 = n-m) I_{1\le m\le\frac{n}{2}} \Bigg\},
\end{aligned}
\tag{2.33}
$$

where

$$
\begin{aligned}
& E\left[\frac{\hat{\varepsilon}_r^2}{m^2/n^2} \mid N_0 = t\right] \\
& \quad = \frac{1}{m^2}\sum_{i=1}^{b}\left\{\sum_{l>k} k^2 P(U_i = k, V_i = l \mid N_0 = t) + \sum_{k\ge l} l^2 P(U_i = k, V_i = l \mid N_0 = t)\right\} \\
& \qquad + \frac{1}{m^2}\sum_{\substack{i,j=1\\ i\ne j}}^{b}\Bigg\{\sum_{l>k}\sum_{s>r} kr P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \\
& \qquad\qquad \sum_{l>k}\sum_{r\ge s} ks P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \\
& \qquad\qquad \sum_{k\ge l}\sum_{s>r} lr P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \\
& \qquad\qquad \sum_{k\ge l}\sum_{r\ge s} ls P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t)\Bigg\},
\end{aligned}
\tag{2.34}
$$

with $P(U_i = k, V_i = l \mid N_0 = t)$ as in (2.30) and

$$
\begin{aligned}
& P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) \\
& \quad = P(U_i = k, U_j = r \mid N_0 = t)\, P(V_i = l, V_j = s \mid N_1 = n-t) \\
& \quad = \binom{t}{k, r, t-k-r} p_i^k p_j^r (1-p_i-p_j)^{t-k-r} \binom{n-t}{l, s, n-t-l-s} q_i^l q_j^s (1-q_i-q_j)^{n-t-l-s}\,.
\end{aligned}
\tag{2.35}
$$

for $t = m,\, n-m$.

2.4.2 Leave-one-out

To obtain the first moment of $\hat{\varepsilon}_r/\hat{\varepsilon}_0$, one can proceed exactly as in the resubstitution case to get

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_l}{\hat{\varepsilon}_0}\right] \; = \; \sum_{m=1}^{n/2} \Bigg\{ & E\left[\frac{\hat{\varepsilon}_l}{m/n} \mid N_0 = m\right] P(N_0 = m) I_{1\leq m<\frac{n}{2}} + \\ & E\left[\frac{\hat{\varepsilon}_l}{m/n} \mid N_0 = n{-}m\right] P(N_0 = n{-}m) I_{1\leq m\leq\frac{n}{2}} \Bigg\}, \end{aligned} \tag{2.36}$$

where now

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_l}{m/n} \mid N_0 = t\right] \; = \; \frac{1}{m}\sum_{i=1}^{b} \Bigg\{ & \sum_{l\geq k} kP(U_i = k, V_i = l \mid N_0 = t) + \\ & \sum_{k\geq l-1} lP(U_i = k, V_i = l \mid N_0 = t) \Bigg\}, \end{aligned} \tag{2.37}$$

with $P(U_i = k, V_i = l \mid N_0 = t)$ as in (2.30), for $t = m,\, n{-}m$.

To obtain the second moment of $\hat{\varepsilon}_r/\hat{\varepsilon}_0$, one can again proceed as in the resubstitution case to get

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}_l^2}{\hat{\varepsilon}_0^2}\right] \; = \; \sum_{m=1}^{n/2} \Bigg\{ & E\left[\frac{\hat{\varepsilon}_l^2}{m^2/n^2} \mid N_0 = m\right] P(N_0 = m) I_{1\leq m<\frac{n}{2}} + \\ & E\left[\frac{\hat{\varepsilon}_l^2}{m^2/n^2} \mid N_0 = n{-}m\right] P(N_0 = n{-}m) I_{1\leq m\leq\frac{n}{2}} \Bigg\}, \end{aligned} \tag{2.38}$$

where now

$$
\begin{aligned}
& E\left[\frac{\hat{\varepsilon}_l^2}{m^2/n^2} \mid M = t\right] \\
& = \frac{1}{m^2}\sum_{i=1}^{b}\left\{\sum_{l \geq k} k^2 P(U_i = k, V_i = l \mid N_0 = t)\right. \\
& \qquad \sum_{k \geq l-1} l^2 P(U_i = k, V_i = l \mid N_0 = t) + \\
& \qquad \left.\sum_{l-1 \leq k \leq l} 2kl P(U_i = k, V_i = l \mid N_0 = t)\right\} + \\
& \quad \frac{1}{m^2}\sum_{\substack{i,j=1 \\ i \neq j}}^{b}\left\{\sum_{l \geq k}\sum_{s \geq r} kr P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \right. \\
& \qquad \sum_{l \geq k}\sum_{r \geq s-1} ks P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \\
& \qquad \sum_{k \geq l-1}\sum_{s \geq r} lr P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t) + \\
& \qquad \left.\sum_{k \geq l-1}\sum_{r \geq s-1} ls P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t)\right\}
\end{aligned}
\tag{2.39}
$$

with $P(U_i = k, V_i = l \mid N_0 = t)$ as in (2.30) and $P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t)$ as in (2.35), for $t = m$, $n-m$.

2.5 Approximate Variances of Non-Randomized CoD Estimation

Though the variances of the resubstitution and leave-one-out error estimators and CoD estimators could be computed exactly with the expressions derived in [10,21], it is impractical to realize these computations for large sample size or high classification complexity, given that second-order probabilities of the form $P(U_i = k, V_i = l, U_j = r, V_j = s)$ need to be calculated. In this Section, we propose an approximation method for the fast compuation of variances of both resubstitution and leave-one-out CoD estimators [20].

The variance of the resubstitution CoD estimator is given by

$$
\begin{aligned}
\mathrm{Var}[\widehat{\mathrm{CoD}}_r] \; &= \; \mathrm{Var}\left[1 - \frac{\hat{\varepsilon}_r}{\hat{\varepsilon}_0}\right] \; = \; \mathrm{Var}\left[\sum_{i=1}^{b}\left(\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0} + \frac{V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}\right)\right] \\
&= \sum_{i=1}^{b} \mathrm{Var}\left[\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0} + \frac{V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}\right] + \\
&\quad 2\sum_{i<j} \mathrm{Cov}\left[\frac{U_i I_{V_i > U_i} + V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}, \frac{U_j I_{V_j > U_j} + V_j I_{U_j \geq V_j}}{n\hat{\varepsilon}_0}\right],
\end{aligned} \tag{2.40}
$$

whereas the variance of the leave-one-out CoD estimator is formulated by substituting $V_i > U_i$ and $U_j \geq V_j$ in (2.40) with $V_i \geq U_i$ and $U_j \geq V_j - 1$, respectively. The exact expressions of the variances of resubstitution and leave-one-out CoD estimators have been formulated in [21]. Note that the covariance terms in both expressions are related with the second-order joint probabilities, and thus the application of these exact expressions become problematic regarding huge computation efforts for large sample size or bin size.

Assuming that $U_i/(n\hat{\varepsilon}_0), V_i/(n\hat{\varepsilon}_0), U_j/(n\hat{\varepsilon}_0), V_j/(n\hat{\varepsilon}_0)$ are less correlated as b increases, for $i, j = 1, \ldots, b$ and $i \neq j$, the covariance terms in (2.40) tends to zero as b increases. We drop the summation on these covariances including second-order probabilities, and the approximate expression for the variance of the resubstitution CoD estimator is given by:

$$
\begin{aligned}
\mathrm{Var}[\widehat{\mathrm{CoD}}_r] \; &= \; \sum_{i=1}^{b} \mathrm{Var}\left[\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0} + \frac{V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}\right] \\
&= \sum_{i=1}^{b} \mathrm{Var}\left[\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0}\right] + \sum_{i=1}^{b} \mathrm{Var}\left[\frac{V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}\right] - 2\sum_{i=1}^{b} E\left[\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0}\right] E\left[\frac{V_i I_{U_i \geq V_i}}{n\hat{\varepsilon}_0}\right],
\end{aligned} \tag{2.41}
$$

and the approximate expression for the variance of the leave-one-out is given by:

$$
\begin{aligned}
\mathrm{Var}[\widehat{\mathrm{CoD}}_l] \;=\; & \sum_{i=1}^{b} \mathrm{Var}\left[\frac{U_i I_{V_i \geq U_i}}{n\hat{\varepsilon}_0}\right] + \sum_{i=1}^{b} \mathrm{Var}\left[\frac{V_i I_{U_i \geq V_i - 1}}{n\hat{\varepsilon}_0}\right] + \\
& 2\sum_{i=1}^{b}\left(E\left[\frac{U_i V_i I_{V_i \geq U_i, U_i \geq V_i - 1}}{(n\hat{\varepsilon}_0)^2}\right] \;-\; E\left[\frac{U_i I_{V_i \geq U_i}}{n\hat{\varepsilon}_0}\right] E\left[\frac{V_i I_{U_i \geq V_i - 1}}{n\hat{\varepsilon}_0}\right]\right).
\end{aligned}
\tag{2.42}
$$

In order to complete the formulations in (2.41) and (2.42), we need to express the first and second moments involved, for example,

$$
\begin{aligned}
E\left[\frac{U_i I_{V_i > U_i}}{n\hat{\varepsilon}_0}\right] \;=\; & \sum_{1 \leq m < \frac{n}{2}} \sum_{i=1}^{b} \sum_{l > k} \frac{k}{m} P(U_i = k, V_i = l | N_0 = m) + \\
& \sum_{1 \leq m \leq \frac{n}{2}} \sum_{i=1}^{b} \sum_{l > k} \frac{k}{m} P(U_i = k, V_i = l | N_0 = n - m),
\end{aligned}
\tag{2.43}
$$

where $P(U_i = k, V_i = l \mid N_0 = t)$ is formed in eq. (2.30). Likewise, the other first and second moments could be formulated.

2.6 Results and Discussion

Assuming a parametric probability model in this section, we plot the exact performance metrics of the resubstitution and leave-one-out CoD estimators, by using the analytical expressions obtained in Sections 2.3 and 2.4, under varying actual CoD, sample size, and predictor complexity (number of bins). We also compare these exact performance metrics with the approximate performance metrics for cross-validation and bootstrap CoD estimators computed via Monte-Carlo sampling. The Monte-Carlo computation was carried out by drawing $M = 5000$ simulated training data sets of the required sample size from the probability model in each case, and employing sample means and sample variances to approximate the performance metrics in

Section 2.3.

The probability model used here is a parametric Zipf model [72]. The class-conditional probabilities under the parametric Zipf model are given by:

$$\begin{aligned} p_i &= \frac{K}{i^\alpha} \\ q_i &= p_{b-i+1}, \end{aligned} \tag{2.44}$$

for $i = 1, \ldots, b$, and $\alpha > 0$. The normalizing constant K is given by:

$$K = \left[\sum_{i=1}^{b} \frac{1}{i^\alpha} \right]^{-1}, \tag{2.45}$$

For simplicity, we assume that $c_0 = c_1 = \frac{1}{2}$. It can be seen easily from (2.5) that the CoD increases monotonically with α, so that large α leads to tight regulation, i.e. easy prediction, and vice-versa. There are two extreme cases. When $\alpha = 0$, there is maximal confusion between the classes, and CoD $= 0$. When $\alpha \to \infty$, there is maximal discrimination between the classes, and CoD $= 1$. Thus, varying the parameter α can traverse the probability model space continuously from easy to difficult models.

We consider here the prediction setting where each predictor variable is binary. If we employ 2, 3, and 4 predictor variables then this would correspond to bin sizes $b = 4, 8, 16$, respectively. In functional genomics applications, these cases correspond to the gene prediction problem by using $2, 3$, and 4 genes, where the activity of each gene is represented by binary gene expressions, e.g., the on-and-off switch effect of a promoter.

Figure 2.1 displays bias, variance, and RMS of the CoD estimators considered here, as a function of varying actual CoD (computed by suitable tuning the parameter

α). We recall that, in the figure, tight regulation, i.e. easy prediction, is located on the right of these plots, whereas loose regulation, i.e. difficult prediction, is located on the left.

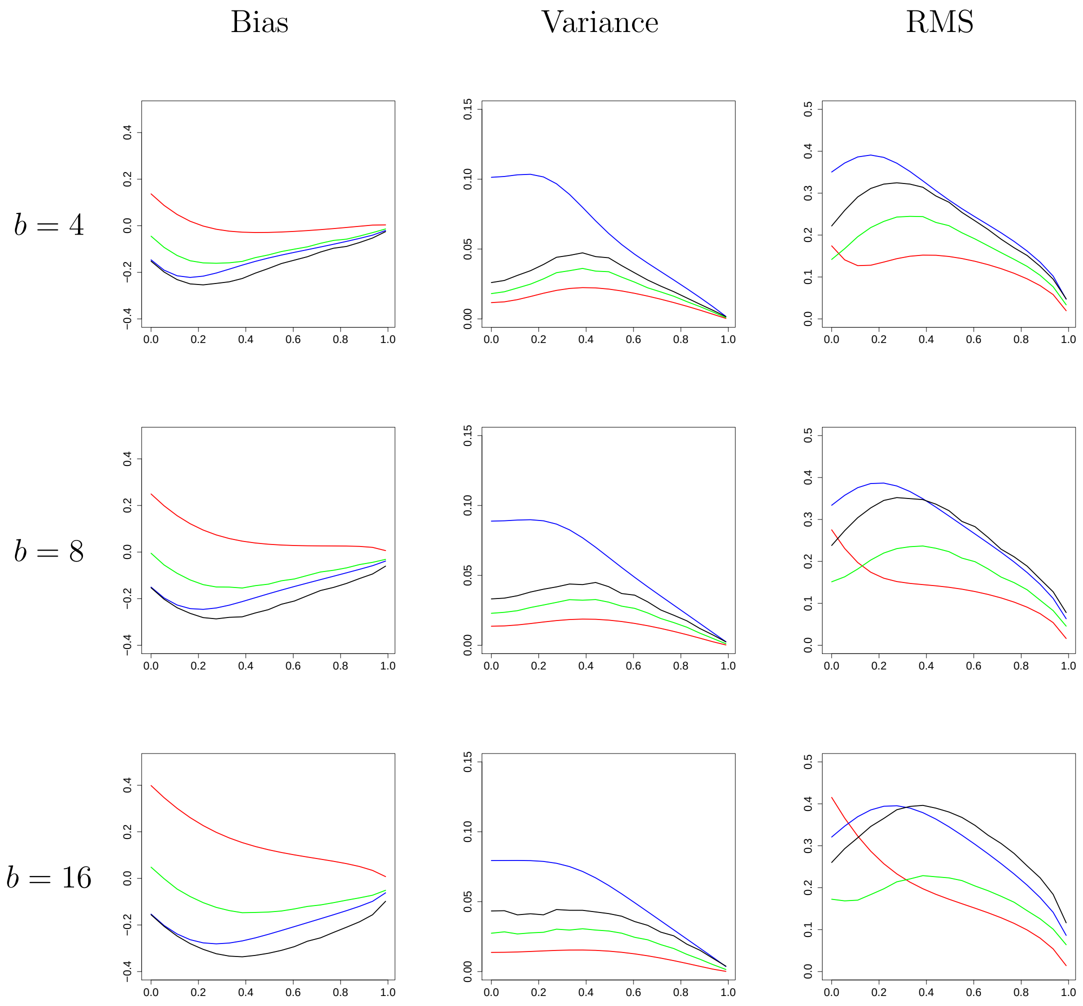


Figure 2.1: Bias, variance, and RMS for several CoD estimators vs. actual CoD under a Zipf model with $c_0 = 1/2$, for $n = 40$ and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling.

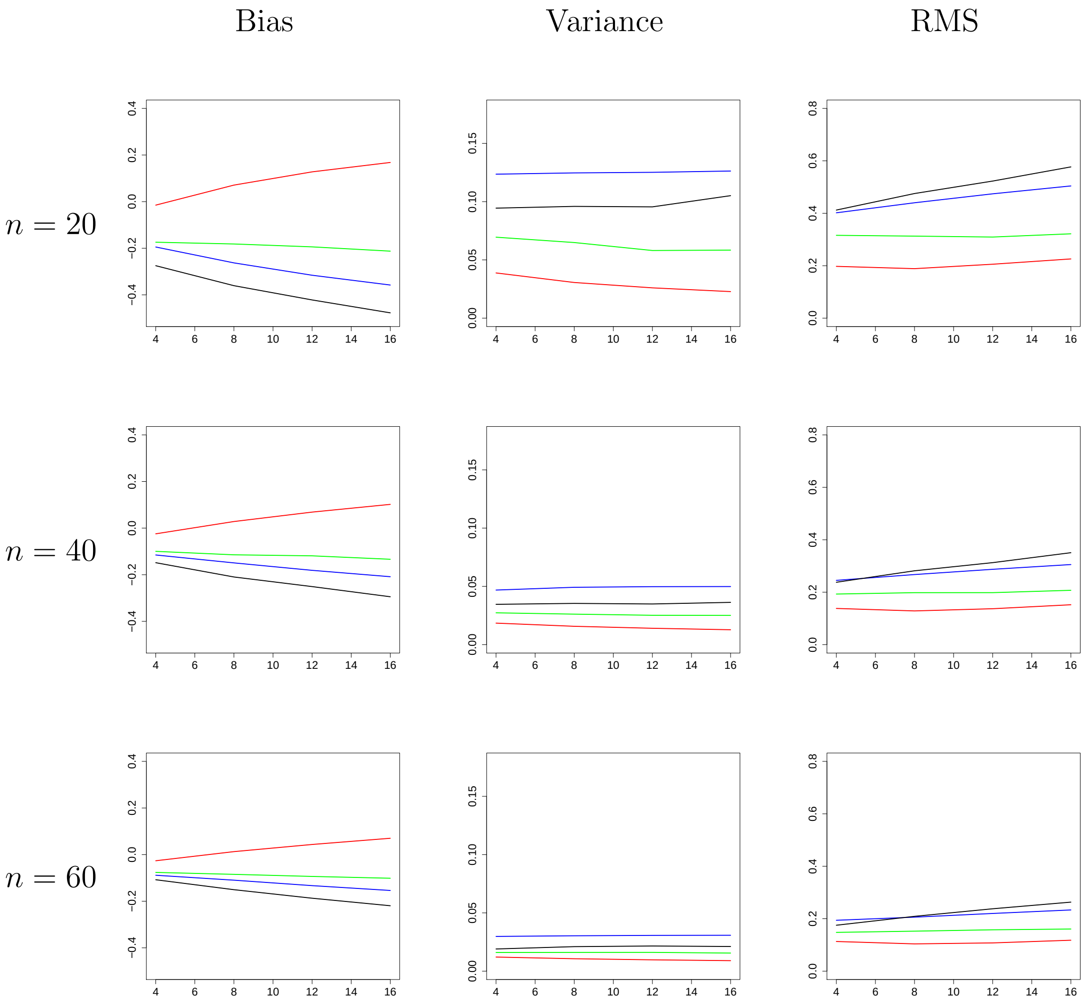


Figure 2.2: Bias, variance, and RMS for several CoD estimators vs. number of bins ($b = 4, 8, 12,$ and 16) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.6 and varying sample size. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling.

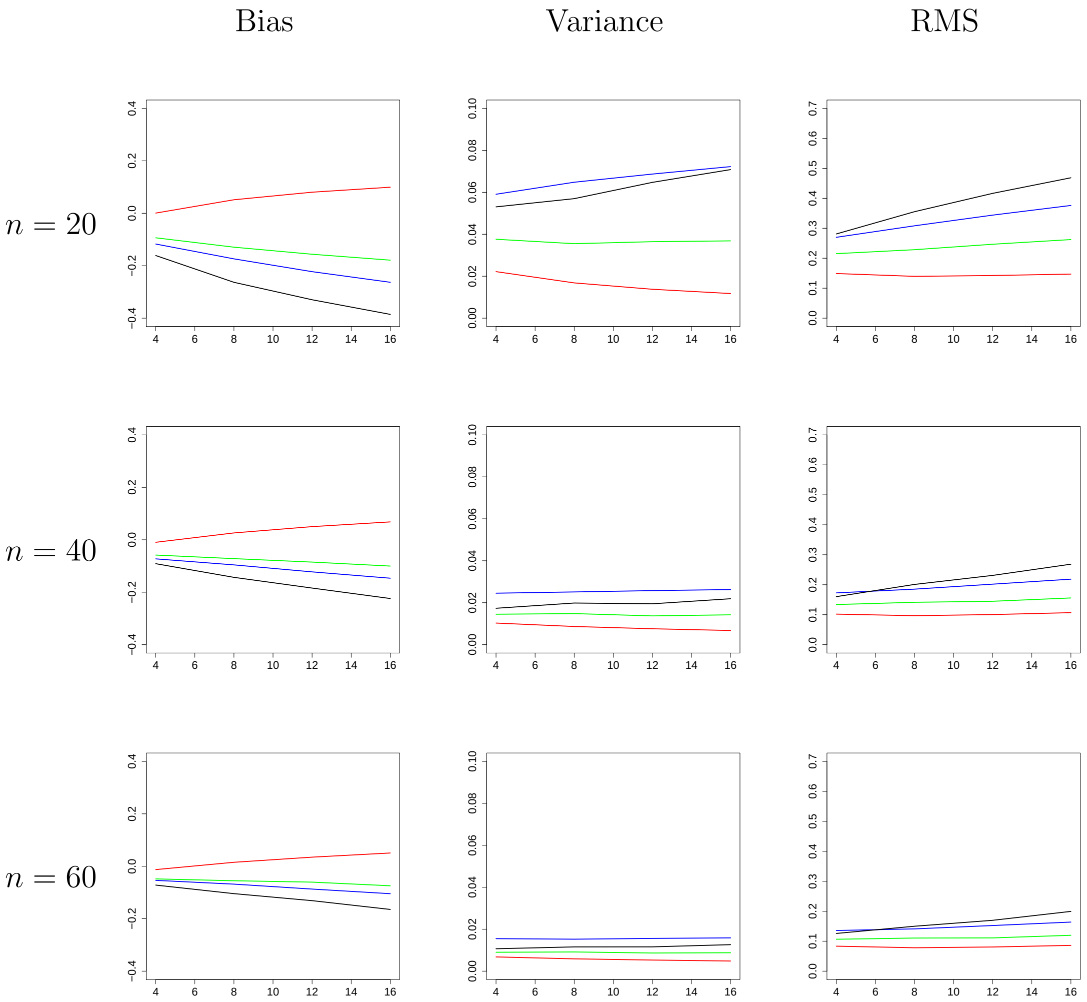


Figure 2.3: Bias, variance, and RMS for several CoD estimators vs. number of bins ($b = 4, 8, 12,$ and 16) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.8 and varying sample size. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling.

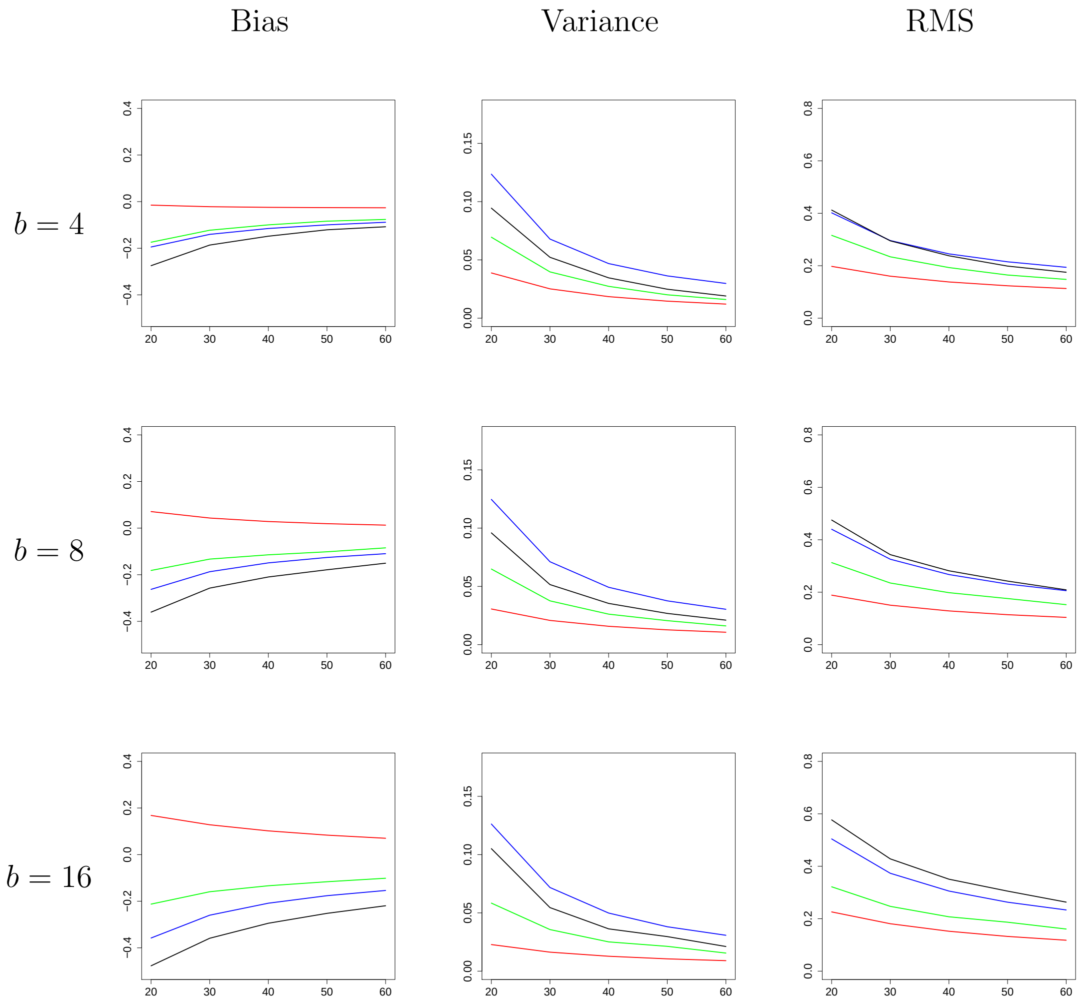


Figure 2.4: Bias, variance, and RMS for several CoD estimators vs. sample size ($n = 20, 30, 40, 50$, and 60) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.6 and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling.

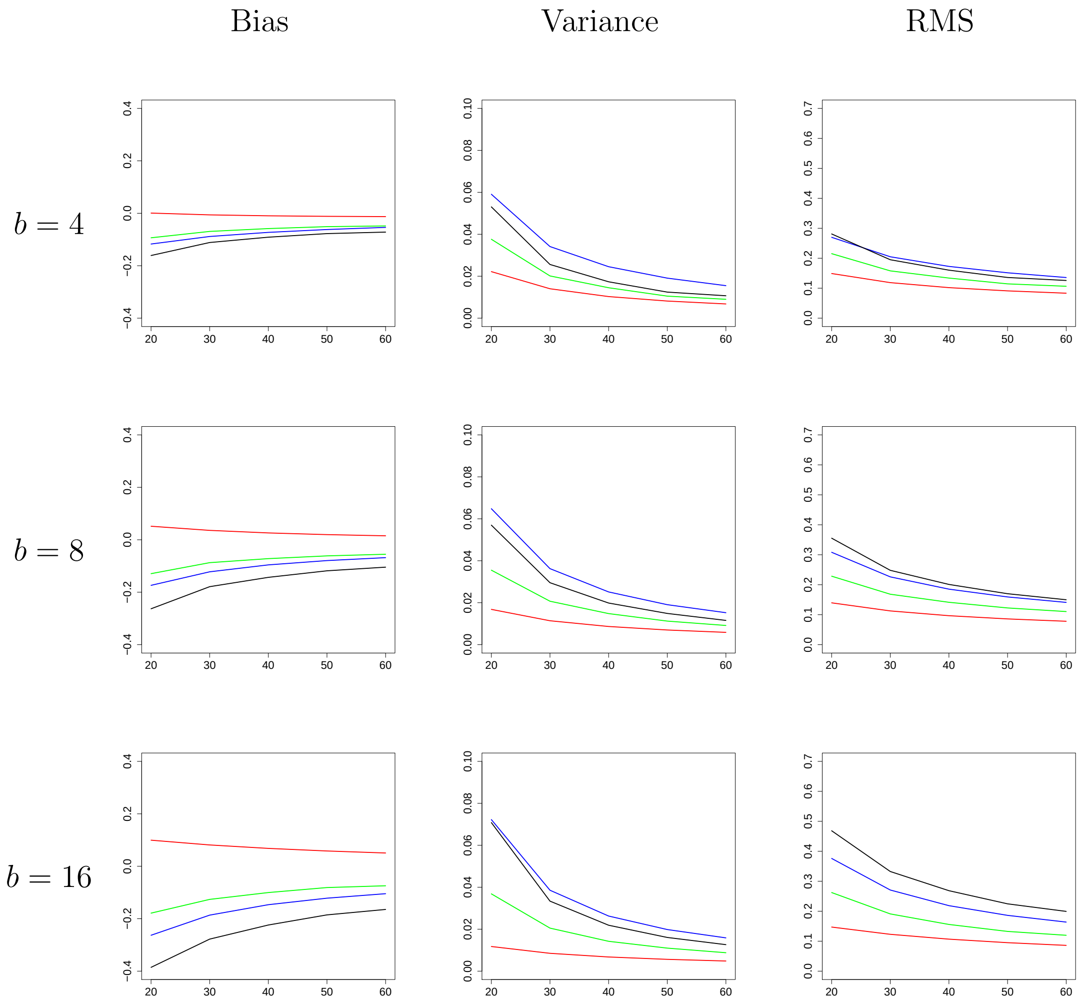
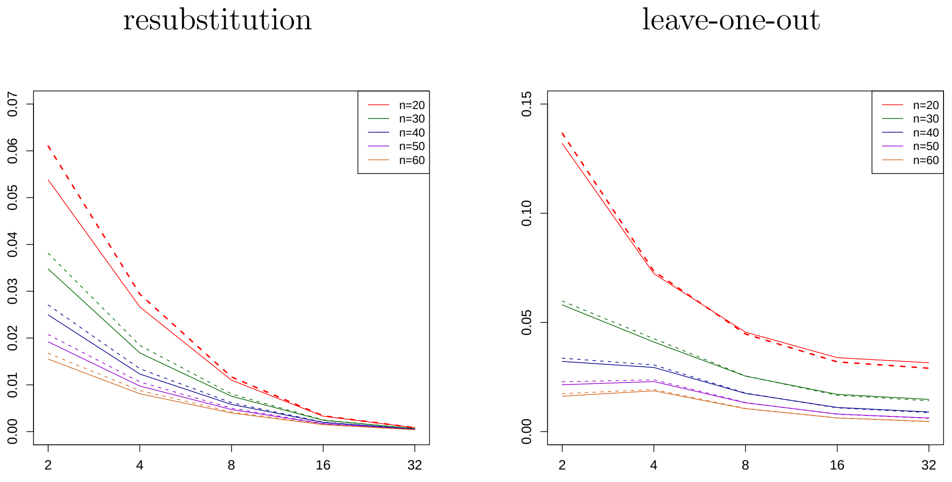


Figure 2.5: Bias, variance, and RMS for several CoD estimators vs. sample size ($n = 20, 30, 40, 50,$ and 60) under a Zipf model with $c_0 = 1/2$, for actual CoD=0.8 and varying number of bins. Plot key: resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). The curves for resubstitution and leave-one-out are exact; the curves for the other CoD estimators are approximations based on Monte-Carlo sampling.

Figure 2.1 makes apparent several facts. The resubstitution CoD is often optimistically biased, except at moderate to large CoD with $b = 4$ (two binary predictors), whereas the other estimators are generally pessimistically biased. As the number of predictors increase, the bias (in magnitude) of the resubstitution CoD increases accordingly; however its variance remains quite low in each case. The leave-one-out CoD is highly variable, in addition to being pessimistically biased. By observing the RMS, we conclude that the resubstitution CoD estimator is the best-performing estimator, except at small values of the actual CoD, beating all the other estimators, including the bootstrap. The leave-one-out CoD estimator is the worst-performing estimator for cases with small number of predictors ($b = 4$), whereas the cross-validation CoD estimator becomes the worst-performing estimator for large number of predictors and moderate actual CoD. As the number of predictors increases, the actual CoD cut-off decreases accordingly at which the leave-one-out CoD estimator starts to outperform the cross-validation CoD estimator. It is also interesting to note that, for $b = 4$, only the bootstrap beats resubstitution, and for very small actual CoD. For $b = 8$, both bootstrap and cross-validation perform better than the resubstitution, for small actual CoD. For $b = 16$, all the other CoD estimators outperform resubstitution for small actual CoD. As the number of predictors increases, the cut-off at which the resubstitution CoD estimator beats all other estimators increases.

In order to assess the performance of the resubstitution CoD estimator and the remaining CoD estimators with respect to the classifier complexity (number of predictors), we display the performance metrics as a function of varying number of bins in Figures 2.2 and 2.3, for sample size $n = 20, 40$ and 60, and moderate CoD $= 0.6$ and large CoD $= 0.80$. The bias column shows that, for CoD $= 0.60$, the resubstitution CoD is actually slightly pessimistically biased for $b = 4$ (a perhaps surprising fact, given the optimistic bias of resubstitution in discrete classification), but quickly becomes optimistically biased for larger bin sizes. In the RMS column, we can see that the resubstitution CoD always beats all other estimators, especially in the case of CoD $= 0.80$ (tight regulation), which is the more surprising when we consider that the other estimators are much more computation-intensive. It is interesting to see that the leave-one-out CoD estimator beats the more complex cross-validation CoD estimator for small number of bins and large sample size. The resubstitution CoD is the least biased and least variable among all CoD estimators, across the whole range of classifier complexity and sample size considered here, and thus it also displays the best RMS overall.

In Figures 2.4 and Figure 2.5, we examine how these performance metrics behave with varying sample sizes for $b = 4, 8, 16$, and moderate CoD $= 0.6$ and large CoD $= 0.80$. As expected, bias (in magnitude), variance and RMS all decrease as sample size increases. We can see that the resubstitution CoD is the least biased and least variable among all estimators, and thus also displays the best RMS. The cross-validation CoD estimator is the most biased, and the leave-one-out CoD estimator is the most variable, among all CoD estimators. The bootstrap CoD estimator is less variable than the cross-validation CoD estimator.


resubstitution

leave-one-out

Figure 2.6: Exact (solid line) and approximate (dashed line) variances of resubstitution CoD and leave-one-out CoD versus bin size for varying bin sizes.

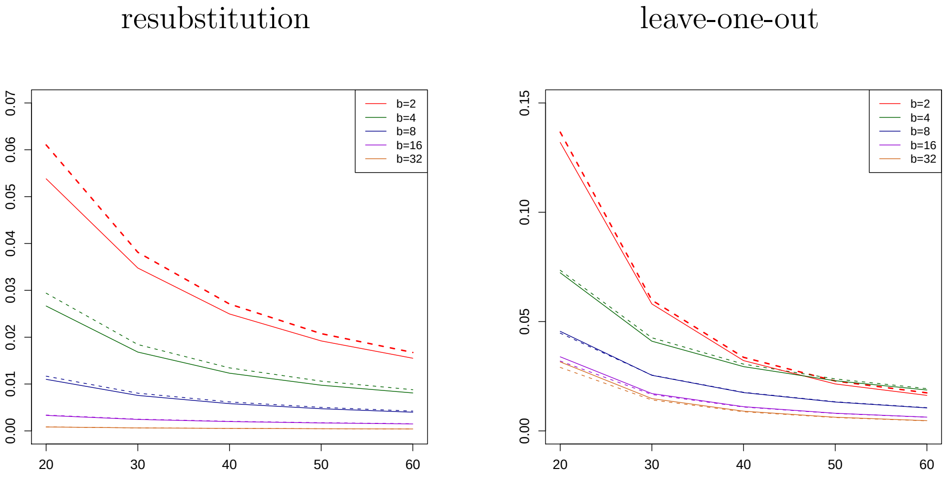


Figure 2.7: Exact (solid line) and approximate (dashed line) variances of resubstitution CoD and leave-one-out CoD versus bin size for varying sample sizes.

In addition, we run simulations for the comparison of exact variances and approximate variances (in Section 2.5) of non-randomized CoD estimators. Again, the parametric Zipf model [10] is employed here due to its simplicity and robustness. The parameter α is set to be 2.0, which corresponds to small Bayes error and large CoD. Figures 2.6–2.7 display the exact and approximate variances of the resubstitution and leave-one-out CoD estimators, respectively. We could observe that the approximations perform better for larger sample size or bin size. Also, the good accuracy of the approximations is attained while saving a lot of computation time. For instance, it takes nearly 2 hrs 20 mins to compute the exact variance for resubstitution CoD estimator but just about 5 seconds to compute the approximate one, using Eclipse (C/C++ programming tool) on Windows XP Pro Intel Duo 2.40GHz. This makes practical the analytical study of error estimation and CoD estimation for larger sample sizes and classification complexity.

2.7 Summary

This chapter has presented a comprehensive study of CoD estimators. We derived for the first time exact analytical expressions of performance metrics of the resubstitution and leave-one-out CoD estimators. Using a parametric Zipf model, we have compared the exact performance metrics of resubstitution and leave-one-out between each other and against approximate performance metrics of cross-validation and bootstrap CoD estimators. Our results lead to a perhaps surprising conclusion: under the Zipf model under consideration, the resubstitution CoD estimator is the best-performing estimator among all, for moderate to large actual CoD and not too large number of predictors. However, for small actual CoD values and high classifier complexity, the other three CoD estimators can outperform resubstitution. This indicates that, provided one has evidence of moderate to tight regulation between the genes, and the number of predictors is not too large, one should use the CoD estimator based on resubstitution.

3. FREQUENTIST INFERENCE: PARAMETRIC MAXIMUM-LIKELIHOOD COD ESTMATION[*]

The CoD is commonly estimated through nonparametric methods [23, 31, 47, 52, 61, 62, 71]. We have investigated in Section 2 the performance of four nonparametric CoD estimators, based on the resubstitution, leave-one-out, bootstrap and cross-validation error rate estimators. It was observed that, provided one has evidence of moderate to tight regulation between predictors and target, and the number of predictors is not too large, one should use the resubstitution CoD estimator, which happens to be the nonparametric maximum-likelihood estimator (NPMLE) for the unknown joint distribution between predictors and target [57].

In this chapter, we propose a parametric MLE approach, by introducing stochastic Boolean models for biology systems, and deriving the maximum-likelihood estimator of the CoD given sample data drawn from the underlying true distribution. The basic rationale behind parametric ML estimation is to take advantage of partial knowledge about the model describing system behavior. This information cannot be used by nonparametric approaches, which must rely purely on the sample data. In many applications, prior knowledge about the system is available, even if this knowledge is incomplete. This is common, for example, in Genomic Signal Processing, where partial knowledge about the biochemical pathways of interest is often known. The more prior knowledge is available, the more we expect that the parametric ML approach will outperform its nonparametric competitors. The prior knowledge about

[*] Parts of this section are reprinted with permission from “Maximum-Likelihood Estimation of the Discrete Coefficient of Determination in Stochastic Boolean Systems” by T. Chen and U. Braga-Neto, 2012, *IEEE Transactions on Signal Processing*, vol. 61, no. 15, pp. 3880–3894, © 2013 IEEE, and “Maximum Likelihood Estimation of the Binary Coefficient of Determination” by T. Chen, and U. Braga-Neto, 2012, *Proceedings of the 45th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 2011, pp 1012–1016, © 2011 IEEE.

the system is coded into a set of candidate models. We will consider in the Chapter the system identification problem [50], where not only the system noise statistics are unknown, but also there is incomplete knowledge about the Boolean relationships in the system. Specifically, we consider the practical situation where partial knowledge may exist about which logic gates are present in the system, but no knowledge exists about the wiring, except for the degree of network connectivity, i.e., the number of inputs per gate. Inference procedures will be discussed for the recovery of logic gates and wiring from sample data.

3.1 Stochastic Boolean Systems

Stochastic Boolean models play a prominent role in many applications, particularly in Genomic Signal Processing [61]. Figure 1.1 displays an example of regulatory network associated with the cell cycle. Figure 1.1(a) depicts the activation and suppression relationships between the various genetic switches, which lead to the activation or not of DNA synthesis, a necessary preparatory step for cell division and a tightly regulated mechanism in normal cells — this mechanism is often found to be out of control in cancerous cells, due to deleterious gene mutations. We can see in Figure 1.1(b) that this network, or pathway, corresponds to a logic circuit:

$$\text{DNA synthesis} \;=\; \overline{\text{Rb}} \;=\; \text{CDK7} \wedge \text{CycH} \wedge \text{CycE} \wedge \overline{\text{p21}}\,. \tag{3.1}$$

In other words, in a healthy cell, DNA synthesis occurs only if all of the CDK7, Cyclin H and Cyclin E genes are active and the p21 gene is silenced [62].

A common task in practice is the estimation of the strength of regulation between the various components of the Boolean circuit from sample data. In addition, it is often the case that only partial information (or even no information) is available about the system, which must also be identified from the sample data. Estimation and identification are complicated by the presence of *system noise*. For example, consider the expression pattern “0 1 0 1” for the predicting genes in the hypothetical sample data of Figure 1.1(c). According to eq (3.1), the state of the Rb gene should be active, and no DNA synthesis should occur. However, three instances of the “0 1 0 1” pattern are observed in the data, and only one of them behaves as the mechanistic model predicts. This is the result of uncertainty in the mechanistic model, e.g., the influence of hidden or latent variables. An additional difficulty is the fact that many expression patterns may be missing due to a small number of samples. These considerations motivate the application of a stochastic approach to the problem, which is described in the next subsection. Our stochastic model does not attempt to include the effects of *observation noise*, that is, inaccuracies intrinsic to the observation of the expression patterns (e.g., microarray noise). For that purpose, more complex state-space models are necessary [63].

3.1.1 Predictive Power

Let $Y \in \{0,1\}$ be the Boolean (i.e., binary) target output to be predicted (in the previous example, Y indicates the presence or not of DNA synthesis), and let $\mathbf{X} = (X_1, \ldots, X_d) \in \{0,1\}^d$ be a set of d Boolean predictors (in the previous example, these indicate the activation status of the CDK7, Cyc H, Cyc R, and p21 genes). Let f be a proposed mechanistic model for the relationship between Y and $\mathbf{X}$. In accordance with the previous discussion, we define the *predictive power* p of the model as

$$p \,=\, P(Y = f(\mathbf{X}))\,. \tag{3.2}$$

If $p = 1$, there can be no inconsistencies between the model and the sample data, i.e., the target is predicted deterministically, whereas if $p = \frac{1}{2}$, there is a maximum amount of indeterminacy, and the model is essentially useless. Intermediates values of p in this range will produce variable amounts of inconsistency between the model and the observed sample data. The mean-squared error (MSE) of the model, denoted here by ε, is given by

$$\varepsilon \,=\, E\left[(Y - f(\mathbf{X}))^2\right] \,=\, P(Y \neq f(\mathbf{X})) \,=\, 1 - p\,. \tag{3.3}$$

It is a well known result [30] that, given the joint distribution between $\mathbf{X}$ and Y, the minimum MSE (MMSE) model, or predictor, is given by

$$f^*(\mathbf{X}) \,=\, I\left(P(Y = 1 \mid \mathbf{X}) > \frac{1}{2}\right)\,, \tag{3.4}$$

with MMSE

$$\begin{aligned} \varepsilon^* \,=\, 1 - p^* \,&=\, E\left[\min\{P(Y = 0 \mid \mathbf{X}), P(Y = 1 \mid \mathbf{X})\}\right] \\ &= \sum_{\mathbf{x} \in \{0,1\}^d} \min\{P(Y = 0, \mathbf{X} = \mathbf{x}), P(Y = 1, \mathbf{X} = \mathbf{x})\}\,. \end{aligned} \tag{3.5}$$

In the previous equations, $I(\cdot)$ denotes the usual indicator function, and p^* denotes the predictive power of the optimal model.

3.1.2 The Coefficient of Determination

Following [31], we define the following measure of association between $\mathbf{X}$ and Y:

$$\text{CoD} \,=\, \frac{\varepsilon_0 - \varepsilon^*}{\varepsilon_0} \,=\, 1 - \frac{\varepsilon^*}{\varepsilon_0}\,, \tag{3.6}$$

where $\varepsilon_0 = \min\{P(Y = 1), P(Y = 0)\}$ is the MMSE of the optimal constant predictor $f_0 = I(P(Y = 1) > \frac{1}{2})$. It can be shown quite easily that $\varepsilon^* \leq \varepsilon_0$, so that $0 \leq \text{CoD} \leq 1$. Moreover, in case $\varepsilon^* = \varepsilon = 0$, we define $\text{CoD} = 1$. In analogy to the classical regression case, this measure is called the *coefficient of determination.* Note that $\text{CoD} = 1 \Leftrightarrow \varepsilon^* = 0 \Leftrightarrow p^* = 1$, in which case Y is deterministically predicted by $\mathbf{X}$, whereas $\text{CoD} = 0 \Leftrightarrow \varepsilon^* = \varepsilon_0 > 0$, i.e., the predictor set $\mathbf{X}$ offers no improvement in prediction accuracy over the constant predictor.

3.1.3 Estimation of the CoD

In practice, the probability structure of the problem is unknown or only partially known, and one attempts to infer the underlying prediction relationships from i.i.d. sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ drawn from the underlying probability model. The broad class of CoD estimators considered here are obtained by employing estimators $\hat{\varepsilon}$ of ε^* and $\hat{\varepsilon}_0$ of ε_0 in (3.6):

$$\widehat{\text{CoD}} = \frac{\hat{\varepsilon}_0 - \hat{\varepsilon}}{\hat{\varepsilon}_0} = 1 - \frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\,. \tag{3.7}$$

It is assumed that $0 \leq \hat{\varepsilon}, \hat{\varepsilon}_0 \leq 1$. By definition, if $\hat{\varepsilon} = \hat{\varepsilon}_0 = 0$, then $\widehat{\text{CoD}} = 1$, whereas if $\hat{\varepsilon} > \hat{\varepsilon}_0$ (including the case $\hat{\varepsilon}_0 = 0$), then $\widehat{\text{CoD}} = 0$.

3.1.3.1 Nonparametric Maximum-Likelihood CoD Estimation

If no information is available about the probability model that generates the data, $\hat{\varepsilon}$ and $\hat{\varepsilon}_0$ can be derived by empirical frequency estimators, i.e., the *nonparametric maximum-likelihood estimators* (NPMLE) of the discrete distribution [57]. Let $N_0 = \sum_{i=1}^n I(Y_i = 0)$, $N_1 = \sum_{i=1}^n I(Y_i = 1) = n - N_0$, $U(\mathbf{x}) = \sum_{i=1}^n I(\mathbf{X}_i = \mathbf{x}, Y_i = 0)$, and $V(\mathbf{x}) = \sum_{i=1}^n I(\mathbf{X}_i = \mathbf{x}, Y_i = 1)$, for $\mathbf{x} \in \{0,1\}^d$. Then the NPMLEs $\hat{\varepsilon}$ and $\hat{\varepsilon}_0$ are given by

$$\begin{aligned} \hat{\varepsilon} &= \sum_{\mathbf{x} \in \{0,1\}^d} \min\{\hat{P}(Y = 0, \mathbf{X} = \mathbf{x}), \hat{P}(Y = 1, \mathbf{X} = \mathbf{x})\} \\ &= \sum_{\mathbf{x} \in \{0,1\}^d} \min\left\{\frac{U(\mathbf{x})}{n}, \frac{V(\mathbf{x})}{n}\right\} \end{aligned} \tag{3.8}$$

and

$$\hat{\varepsilon}_0 \;=\; \min\{\hat{P}(Y=1), \hat{P}(Y=0)\} \;=\; \min\left\{\frac{N_1}{n}, \frac{N_0}{n}\right\}. \tag{3.9}$$

leading to the NPML CoD estimator:

$$\widehat{\text{CoD}}^{\text{NPML}} \;=\; 1 - \frac{\sum_{\mathbf{x}\in\{0,1\}^d} \min\{U(\mathbf{x}), V(\mathbf{x})\}}{\min\{N_0, N_1\}}. \tag{3.10}$$

It is easy to show that $\widehat{\text{CoD}}^{\text{NPML}}$ has the desirable property of being a universally consistent estimator of CoD in (3.6), that is, $\widehat{\text{CoD}}^{\text{NPML}} \to$ CoD in probability (in fact, almost surely) as $n \to \infty$, regardless of the probability model. We remark that the estimator $\hat{\varepsilon}$ in (3.8) is also known in the Pattern Recognition literature as the *resubstitution* estimator, and thus the NPML CoD estimator has been called the resubstitution CoD elsewhere [21].

3.1.3.2 Nonparametric Resampling-Based CoD Estimation

Nonparametric resampling-based CoD estimation is a variation of NPMLE, where the same estimator $\hat{\varepsilon}_0$ is used for ε_0, but the MMSE ε^* is estimated using a resampling method, e.g., the *leave-one-out* [48], the *cross-validation* [66], and the *0.632 bootstrap* [36] estimators. The case of leave-one-out is the most basic one and exemplifies well the other resampling methods: the MMSE is estimated by leaving one sample data point out, estimating what the optimal predictor would be based on the remaining $n-1$ sample points using a NPMLE approach, and applying that to the left-out sample. The process is repeated with each of the n sample points and the estimator $\hat{\varepsilon}$ is the number of errors made divided by n. It can be shown that this leads to the leave-one-out CoD estimator:

$$\widehat{\text{CoD}}^{\text{LOO}} \;=\; 1 - \frac{\sum_{\mathbf{x}\in\{0,1\}^d} U(\mathbf{x})\, I(A(\mathbf{x})) + V(\mathbf{x})\, I(B(\mathbf{x}))}{\min\{N_0, N_1\}}, \tag{3.11}$$

where $A(\mathbf{x})$ and $B(\mathbf{x})$ are equivalent to $U(\mathbf{x}) \leq V(\mathbf{x})$ and $U(\mathbf{x}) \geq V(\mathbf{x}) - 1$, respectively. See Chapter 2 for details about the cross-validation and .632 bootstrap CoD estimators.

3.1.3.3 Parametric Maximum-Likelihood CoD Estimation

The previous CoD estimators utilize nonparametric estimators $\hat{\varepsilon}$ and $\hat{\varepsilon}_0$, which may have a large data requirement for high accuracy. It is often the case that at least partial information is available about the phenomenon in question that might reduce the data requirement, and the nonparametric approach cannot take advantage of this fact. For example, the mechanistic model of DNA synthesis discussed previously has been uncovered by many painstaking experiments in the Cell Biology literature, even though the presence of noise and latent variables will mean that its predictive power is not perfect. This a-priori knowledge can be captured by means of a statistical model, where parts of the model that are unknown are coded by a finite, small number of parameters that can be estimated from sample data in an optimal way, e.g., by employing the principle of *maximum likelihood* (ML) [14]. By expressing ε and ε_0 in terms of these parameters, ML estimators $\hat{\varepsilon}$ and $\hat{\varepsilon}_0$, and thus $\widehat{\mathrm{CoD}}$, are obtained by plugging in the ML estimators of the model parameters. This approach will be pursued in the next sections, where we consider separately models for the static and dynamical cases.

3.2 Static Model

For a target variable $Y \in \{0,1\}$ and predictor variables $\mathbf{X} = (X_1, \ldots, X_d) \in \{0,1\}^d$, we study the following nonlinear model:

$$Y = f(\mathbf{X}) \, \oplus \, N \,, \tag{3.12}$$

where $f : \{0,1\}^d \to \{0,1\}$ is a Boolean function, the symbol "$\oplus$" indicates modulo-2 addition, and $N \in \{0,1\}$ is a noise random variable. The predictor $\mathbf{X}$ is a random vector, the distribution of which is assumed to be arbitrary, whereas the target Y is a random variable, the distribution of which is determined by (3.12). The distribution of N is determined by a parameter p, such that $P(N = 1) = 1 - p$. Notice that one can assume $p \geq \frac{1}{2}$ without loss of generality, since if $p < \frac{1}{2}$ one can employ an equivalent model with negated Boolean function $\bar{f}$ and noise parameter $1 - p \geq \frac{1}{2}$. The noise variable N is assumed to be independent of the predictor vector $\mathbf{X}$. The modulo-2 addition behaves as a XOR operation, which flips the state of the target Y when $N = 1$, and leaves it unaltered when $N = 0$; the value $1 - p$ measures therefore the amplitude of the noise. If $p = 1$, the system is noiseless and prediction is deterministic, while if $p = \frac{1}{2}$, there is maximum indeterminacy in the state of the target given the state of the predictors. We remark that the extension of this model to the case of multivariate target $\mathbf{Y}$ can be readily accomplished, by essentially considering multiple versions of (3.12), one for each component of $\mathbf{Y}$.

From the previous discussion, it is apparent that p must be related to the predictive power of the model. In fact, p is itself the optimal predictive power. To see that, note that

$$\begin{aligned} P(Y = 1 \mid \mathbf{X}) &= P(f(\mathbf{X}) = 1, N = 0 \mid \mathbf{X}) + P(f(\mathbf{X}) = 0, N = 1 \mid \mathbf{X}) \\ &= I(f(\mathbf{X}) = 1)\, p + I(f(\mathbf{X}) = 0)(1 - p)\,, \end{aligned} \tag{3.13}$$

where we used the assumption that N is independent of $\mathbf{X}$. From the fact that $p \geq \frac{1}{2}$, it follows that the optimal predictor of Y given $\mathbf{X}$ is $f^*(\mathbf{X}) = I\left(P(Y = 1 \mid \mathbf{X}) > \frac{1}{2}\right) = f(\mathbf{X})$, with optimal predictive power $p^* = P(Y = f(\mathbf{X})) = P(N = 0) = p$, and MMSE $\varepsilon^* = 1 - p^* = 1 - p$. In other words, f itself is the optimal predictor for this model, p is the optimal predictive power, and $1-p$ is the MMSE.

3.2.1 Maximum-Likelihood Inference of the CoD

The CoD according to model (3.12) is given by

$$\begin{aligned}\text{CoD} &= 1-\frac{\varepsilon^*}{\varepsilon_0} = 1-\frac{1-p}{F(P(Y=1))}\\ &= 1-\frac{1-p}{F\left(\sum\limits_{\mathbf{x}\in\{0,1\}^d} P(Y=1\mid\mathbf{X}=\mathbf{x})P(\mathbf{X}=\mathbf{x})\right)}\\ &= 1-\frac{1-p}{F\left(\sum\limits_{\mathbf{x}\in\{0,1\}^d}[p+I(f(\mathbf{x})=0)(1-2p)]P(\mathbf{X}=\mathbf{x})\right)},\end{aligned}\tag{3.14}$$

where $F:[0,1]\to[0,1]$ is a fixed functional given by $F(x)=\min\{x,1-x\}$. *Assuming that f is known*, a Maximum-Likelihood Estimator (MLE) to the CoD can be obtained by deriving MLEs of the predictive power p and of the parameters of the distribution $P(\mathbf{X}=\mathbf{x})$, and plugging those back into (3.14). The assumption of known f corresponds to a model-based approach, which introduces a degree of regularization into the inference problem by incorporating a-priori knowledge. However, the assumption of known f will be relaxed later to reflect the presence of *incomplete* a-priori knowledge; see Section 3.4.

Before we can proceed, we need to introduce a parametrization of the predictor distribution $P(\mathbf{X}=\mathbf{x})$. Ideally, this parametrization will single out the marginal probability parameters $P(X_i=1)=P_i$, for $i=1,\ldots,d$, called here the *predictor biases*, as well as the covariance structure among the predictors. This can be accomplished in different ways.

One possibility is to employ the theory of multivariate cumulants, which has a long and distinguished history in Signal Processing [53]. The cumulants of the joint distribution of $\mathbf{X} = (X_1, \ldots, X_d)$ are the coefficients in the Taylor series expansion around the origin of the multivariate cumulant generating function $K(\xi_1, \ldots, \xi_d) = \log E\left[e^{\xi_1 X_1 + \cdots + \xi_d X_d}\right]$. First-order cumulants are given simply by $g(i) = E[X_i] = P_i$, for $i = 1, \ldots, d$, giving the biases. On the other hand, higher-order cumulants can be interpreted as the "covariance" among two or more variables variables; e.g., it can be shown that $g(i,j) = \mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \mathrm{Cov}(X_i, X_j)$, for $i, j = 1, \ldots, d$. We will not pursue this parametrization further here.

We will employ instead a slightly different approach. Let $J_d = \{1, \ldots, d\}$. For an arbitrary subset of indices $\{i_1, \ldots, i_r\} \subseteq J_d$, define

$$\gamma(i_1, \ldots, i_r) \;=\; E[X_{i_1} \cdots X_{i_r}] - E[X_{i_1}] \cdots E[X_{i_r}]\,. \tag{3.15}$$

Note that $\gamma(i,j) = E[X_i X_j] - E[X_i]E[X_j]$ is the covariance between X_i and X_j. One can show that

$$\begin{aligned} P(X_1 = x_1, \ldots, X_d = x_d) \;=\; & \prod_{i=1}^{d} P_i^{x_i}(1-P_i)^{1-x_i} + \\ (-1)^{x_1+\cdots+x_d} \sum_{\{i_1,\ldots,i_r\}\subseteq J_d} (-1)^r & \prod_{k \in J_d \setminus \{i_1,\ldots,i_r\}} (1-x_k)\gamma(i_1, \ldots, i_r)\,. \end{aligned} \tag{3.16}$$

For instance, in the case of $d = 2$ predictors, the distribution $P(X_1 = x_1, X_2 = x_2)$ is parametrized by the predictor biases P_1, P_2 and the covariance $\gamma(1,2) = \mathrm{Cov}(X_1, X_2)$:

$$P(X_1 = x_1, X_2 = x_2) \;=\; \prod_{i=1}^{2} P_i^{x_i}(1-P_i)^{1-x_i} \;+\; (-1)^{x_1+x_2}\gamma(1,2)\,. \tag{3.17}$$

This parametrization allows one to easily to impose meaningful constraints such as unbiased predictors, $P_1 = P_2 = 0.5$, or independent predictors, $\gamma(1,2) = 0$, or both, in which case the predictor distribution becomes uniform over the predictor states.

In the case of $d = 3$ predictors, the distribution $P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ is parametrized by the predictor biases P_1, P_2, P_3, and the four parameters:

$$\begin{aligned} &\gamma(1,2) = \mathrm{Cov}(X_1, X_2)\,, \quad \gamma(1,3) = \mathrm{Cov}(X_1, X_3)\,, \\ &\gamma(2,3) = \mathrm{Cov}(X_2, X_3)\,, \gamma(1,2,3) = E[X_1X_2X_3] - E[X_1]E[X_2]E[X_3]\,, \end{aligned} \tag{3.18}$$

such that

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \; = \; \prod_{i=1}^{3} P_i^{x_i}(1-P_i)^{1-x_i} \;\; + (-1)^{x_1+x_2+x_3} \times \\ &[\,(1-x_1)\gamma(1,3) + (1-x_2)\gamma(1,3) + (1-x_3)\gamma(1,2) - \gamma(1,2,3)\,]\;. \end{aligned} \tag{3.19}$$

This parametrization allows one to obtain simple expressions for the CoD as a function of the model parameters in many cases of interest. For example, under an AND model, with an arbitrary number of predictors d, it follows easily from (3.14) that

$$\mathrm{CoD}_{\mathrm{AND}^d}(p, P_1, P_2, \ldots, P_d, \gamma) \; = \; \begin{cases} 1 - \dfrac{1-p}{(1-p) + (P_1P_2\cdots P_d + \gamma)(2p-1)}, \\ \qquad\qquad\qquad\qquad\qquad P_1P_2\cdots P_d + \gamma \leq \frac{1}{2} \\ 1 - \dfrac{1-p}{p - (P_1P_2\cdots P_d + \gamma)(2p-1)}, \text{ o.w.,} \end{cases} \tag{3.20}$$

where $\gamma = \gamma(1, \ldots, d)$.

Now, given i.i.d. sample data $S_n = \{(X_{11}, \ldots, X_{1d}, Y_1), \ldots, (X_{n1}, \ldots, X_{nd}, Y_n)\}$, the MLE of the predictive power $p = P(Y = f(X_1, \ldots, X_d)$ and the parameters in the (unconstrained) model (3.17) are obtained by substituting empirical frequencies for probabilities and, equivalently, sample means for expectations, leading to

$$\hat{p} \,=\, \frac{1}{n}\sum_{i=1}^{n} I(f(X_{i1},\ldots,X_{id}) = Y_i)\,, \hat{P}_i \,=\, \frac{1}{n}\sum_{j=1}^{n} X_{ji}\,, \quad i=1,\ldots,d\,,$$

$$\hat{\gamma}\,(i_1,\ldots,i_r) \,=\, \frac{1}{n}\sum_{j=1}^{n} X_{ji_1}\ldots X_{ji_r} - \frac{1}{n^r}\sum_{j=1}^{n} X_{ji_1}\cdots\sum_{j=1}^{n} X_{ji_r}\,, \quad (i_1,\ldots,i_r)\subseteq J_d\,. \tag{3.21}$$

Notice that there are 2^d parameters to be estimated in the model. It is easy to show that $\hat{p}$ and $\hat{P}_i$, for $(i = 1, 2, \ldots, d)$, are minimum-variance unbiased, with $\text{Var}[\hat{p}] = \frac{1}{n}p(1-p)$, $\text{Var}[\hat{P}_i] = \frac{1}{n}P_i(1-P_i)$, for $i = 1, 2, \ldots, d$. However, $\hat{\gamma}\,(i_1,\ldots,i_r)$ is biased. For example, for $d = 2$, $E[\hat{\gamma}(1,2)] = \frac{n-1}{n}\gamma$, and for $d = 3$, $E[\hat{\gamma}(1,2,3)] = \frac{n^2-1}{n^2}\gamma$. It is well-known that, under certain minimal regularity conditions, which are satisfied in our case, MLEs are asymptotically unbiased, asymptotically efficient, and consistent [14, Thm. 10.1.6], so that all the estimators defined previously have these properties. Finally, the ML CoD estimator $\widehat{\text{CoD}}^{\text{ML}}$ is obtained by plugging in the estimators in (3.21) back into equations (3.14) and (3.16). In practice, the estimators are plugged into simplified expressions for specific models, such as (3.20) .

3.2.2 Performance Analysis

Regarding the performance of a CoD estimator $\widehat{\text{CoD}}$, the quantities of interest are the bias, variance, and RMS, given by $\text{Bias}[\widehat{\text{CoD}}] = E[\widehat{\text{CoD}}] - \text{CoD}$, $\text{Var}[\widehat{\text{CoD}}]$, and $\text{RMS}[\widehat{\text{CoD}}] = \sqrt{\text{Bias}[\widehat{\text{CoD}}]^2 + \text{Var}[\widehat{\text{CoD}}]}$, respectively, which should be as small as possible for best performance. For a general CoD estimator, these quantities can be computed exactly via *complete enumeration* [1]. This requires a large amount of time and computation, being applicable only if the sample size and number of variables is small (but see [21] for exact expressions for the NPML and LOO CoD estimators, which avoid complete enumeration). For the ML CoD estimator, we obtain here asymptotic expressions for its bias, variance, and thus RMS. These expressions are asymptotically exact as the sample size increases, but they are also accurate for moderate finite sample sizes.

3.2.2.1 Bias

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{2^d})$be the vector of model parameters, e.g., $\theta_1 = p$, $\theta_2 = P_1, \ldots, \theta_{d+1} = P_d$, $\theta_{d+2} = \gamma(1,2), \ldots, \theta_{2^d} = \gamma(1, \ldots, d)$, and let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$ be the vector of corresponding ML parameter estimators, given by (3.21). We have $\text{CoD} = \text{CoD}(\boldsymbol{\theta})$ and $\widehat{\text{CoD}}^{\text{ML}} = \text{CoD}(\hat{\boldsymbol{\theta}})$. Assuming differentiability at $\boldsymbol{\theta}$, one can employ a Taylor series expansion to obtain:

$$\widehat{\text{CoD}}^{\text{ML}} - \text{CoD} = \sum_{i=1}^{2^d} \frac{\partial\, \text{CoD}(\boldsymbol{\theta})}{\partial \theta_i} \, (\hat{\theta}_i - \theta_i) \, + \, o_P(1) \, . \tag{3.22}$$

where $o_P(1)$ indicates a term that goes to zero in probability as $n \to \infty$, since $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in probability — the latter convergence necessarily occurs because $\hat{\boldsymbol{\theta}}$ is consistent, as discussed at the end of the previous section. Taking expectations on both sides then leads to

$$\text{Bias}[\widehat{\text{CoD}}^{\text{ML}}] = \sum_{i=1}^{2^d} \frac{\partial\, \text{CoD}(\boldsymbol{\theta})}{\partial \theta_i} \, \text{Bias}[\hat{\theta}_i] \, + \, o(1) \, , \tag{3.23}$$

where $o(1)$ is a negligible term as $n \to \infty$. Since $\hat{p}$ and $\hat{P}_i$, for $i = 1, 2, \ldots, d$, are unbiased, we further obtain the simplified expression

$$\text{Bias}[\widehat{\text{CoD}}^{\text{ML}}] = \sum_{\{i_1, \ldots, i_r\} \subseteq J_d} \frac{\partial\, \text{CoD}(\boldsymbol{\theta})}{\partial \gamma(i_1, \ldots, i_r)} \, \text{Bias}[\hat{\gamma}(i_1, \ldots, i_r)] \, + \, o(1) \, , \tag{3.24}$$

and the bias of the ML CoD estimator is a function of the bias of the ML covariance estimators. For the two-predictor AND model, for instance, this produces, by discarding the vanishing term:

$$\text{Bias}\left[\widehat{\text{CoD}}^{\text{ML}}_{\text{AND}^2}\right] \approx \begin{cases} \dfrac{-(1-p)(2p-1)\gamma}{n[(1-p)+(P_1P_2+\gamma)(2p-1)]^2}, & \text{if } P_1P_2+\gamma < \dfrac{1}{2} \\ \dfrac{(1-p)(2p-1)\gamma}{n[p-(P_1P_2+\gamma)(2p-1)]^2}, & \text{if } P_1P_2+\gamma > \dfrac{1}{2} \end{cases}. \tag{3.25}$$

Hence, the estimator is optimistic if $P_1P_2 + \gamma < \frac{1}{2}$, and pessimistic if $P_1P_2 + \gamma > \frac{1}{2}$. If $P_1P_2 + \gamma = \frac{1}{2}$, then the CoD is not differentiable at $\boldsymbol{\theta}$ and the approximation cannot be applied; however, in this case we obtain directly from (3.20) that CoD $=$ $2p - 1$, with $\widehat{\text{CoD}}^{\text{ML}} = 2\hat{p} - 1$, so that $\text{Bias}[\widehat{\text{CoD}}^{\text{ML}}_{\text{AND}^2}] = 0$, and the estimator is unbiased for all n (this is an exact result). Equation (3.25) also allows us to conclude that the bias becomes small for p close to the extreme values $p = \frac{1}{2}$ and $p = 1$. Moreover, the bias vanishes as $n \to \infty$, regardless of p and the other parameters. A corresponding expression for the bias of the 3-input AND logic model can be found in the Appendix B, with similar conclusions.

3.2.2.2 Variance

Using again the Taylor series expansion (3.22), one obtains

$$\begin{aligned}\text{Var}(\widehat{\text{CoD}}^{\text{ML}}) &= \text{Var}(\text{CoD}(\hat{\boldsymbol{\theta}})) = E\left[\left(\text{CoD}(\hat{\boldsymbol{\theta}}) - E[\text{CoD}(\hat{\boldsymbol{\theta}})]\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^{2^d} \frac{\partial\,\text{CoD}(\boldsymbol{\theta})}{\partial\theta_i}(\hat{\theta}_i - E[\hat{\theta}_i])\right)^2\right] + o(1) \\ &= \sum_{i=1}^{2^d}\left(\frac{\partial\,\text{CoD}(\boldsymbol{\theta})}{\partial\theta_i}\right)^2 \text{Var}(\hat{\theta}_i) + 2\sum_{\substack{i,j=1\\ i<j}}^{2^d} \frac{\partial\,\text{CoD}(\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial\,\text{CoD}(\boldsymbol{\theta})}{\partial\theta_j}\,\text{Cov}(\hat{\theta}_i, \hat{\theta}_j) + o(1)\,.\end{aligned} \tag{3.26}$$

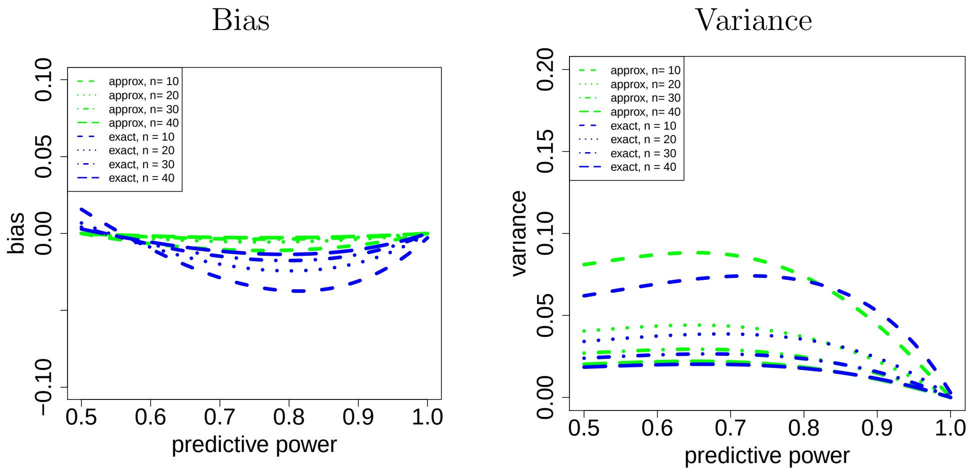


Figure 3.1: Bias and variance versus predictive power over sample size $n = 10, 20, 30$ and 40, in a two-input AND model with $P_1 = P_2 = 0.5$ and $\gamma = 0.20$. Blue: exact results (via complete enumeration); Green: approximate results (via asymptotic approximation).

Notice that this expression requires the computation of the entire covariance matrix $\Sigma(\hat{\boldsymbol{\theta}})$, i.e., the variances of the individual estimators $\hat{\theta}_i$ and the covariances between all pairs of estimators $\hat{\theta}_i, \hat{\theta}_j$. For the two-predictor case, it can be verified that these are given by:

$$
\begin{aligned}
\mathrm{Var}(\hat{p}) &= \frac{1}{n}p(1-p)\,, \quad \mathrm{Var}(\hat{P}_1) = \frac{1}{n}P_1(1-P_1)\,, \quad \mathrm{Var}(\hat{P}_2) = \frac{1}{n}P_2(1-P_2) \\
\mathrm{Var}(\hat{\gamma}) &= \frac{n-1}{n^2}P_1P_2(1-P_1)(1-P_2) + \frac{(n-1)^2}{n^3}(1-2P_1)(1-2P_2)\gamma - \\
&\quad \frac{(n-1)(n-2)}{n^3}\gamma^2\,, \\
\mathrm{Cov}(\hat{p},\hat{P}_1) &= \mathrm{Cov}(\hat{p},\hat{P}_2) = \mathrm{Cov}(\hat{p},\hat{\gamma}) = 0\,, \\
\mathrm{Cov}(\hat{P}_1,\hat{P}_2) &= \gamma/n\,, \quad \mathrm{Cov}(\hat{P}_1,\hat{\gamma}) = \frac{n-1}{n^2}(1-2P_1)\gamma\,, \\
\mathrm{Cov}(\hat{P}_2,\hat{\gamma}) &= \frac{n-1}{n^2}(1-2P_2)\gamma\,.
\end{aligned}
\tag{3.27}
$$

For the two-predictor AND model, for instance, this produces, by discarding the vanishing term:

$$\mathrm{Var}\left(\widehat{\mathrm{CoD}}^{\mathrm{ML}}_{\mathrm{AND}^2}\right) \approx$$

$$\begin{cases} \dfrac{1}{n[(1-p)+(P_1P_2+\gamma)(2p-1)]^4}\Bigg[(P_1P_2+\gamma)^2p(1-p) \;+\; (1-p)^2(2p-1)^2 \times \\ \times\Bigg(P_1P_2^2(1-P_1)+P_1^2P_2(1-P_2)+2P_1P_2\gamma+\dfrac{n-1}{n}P_1P_2(1-P_1)(1-P_2)+ \\ +\dfrac{(n-1)^2}{n^2}(1-2P_1)(1-2P_2)\gamma-\dfrac{(n-1)(n-2)}{n^2}\gamma^2+ \\ 2\dfrac{n-1}{n}(P_1+P_2-4P_1P_2)\gamma\Bigg)\Bigg], \qquad \text{if } P_1P_2+\gamma<\frac{1}{2} \\ \dfrac{1}{n[\,p-(P_1P_2+\gamma)(2p-1)]^4}\Bigg[(P_1P_2+\gamma)^2p(1-p) \;+\; (1-p)^2(2p-1)^2 \times \\ \times\Bigg(P_1P_2^2(1-P_1)+P_1^2P_2(1-P_2)+2P_1P_2\gamma+\dfrac{n-1}{n}P_1P_2(1-P_1)(1-P_2)+ \\ +\dfrac{(n-1)^2}{n^2}(1-2P_1)(1-2P_2)\gamma-\dfrac{(n-1)(n-2)}{n^2}\gamma^2+ \\ 2\dfrac{n-1}{n}(P_1+P_2-4P_1P_2)\gamma\Bigg)\Bigg], \qquad \text{if } P_1P_2+\gamma>\frac{1}{2} \end{cases} . \tag{3.28}$$

If $P_1P_2+\gamma=\frac{1}{2}$ then, as mentioned previously, the CoD is not differentiable at $\boldsymbol{\theta}$ and the approximation cannot be applied; however, in this case $\widehat{\mathrm{CoD}}^{\mathrm{ML}}=2\hat{p}-1$, so that $\mathrm{Var}[\widehat{\mathrm{CoD}}^{\mathrm{ML}}_{\mathrm{AND}^2}]=4\mathrm{Var}[\hat{p}]=\frac{4}{n}p(1-p)$. Equation (3.28) allows us to conclude that the the variance of the ML CoD estimator vanishes as $n\to\infty$. A corresponding expression for the variance of the 3-input AND logic model can be found in the Appendix B, and similar conclusions apply. In addition, Tables C.1 and C.2 in Appendix C lists the bias and variance asymptotic expressions for five 2-predictor logics: AND, XOR, OR, $X_1\bar{X}_2$, and $\bar{X}_1X_2$. The remaining five useful 2-predictor logics are negations of these, and it can be easily verified that the expression for the CoD and its bias and variance asymptotic approximations are the same for a logic

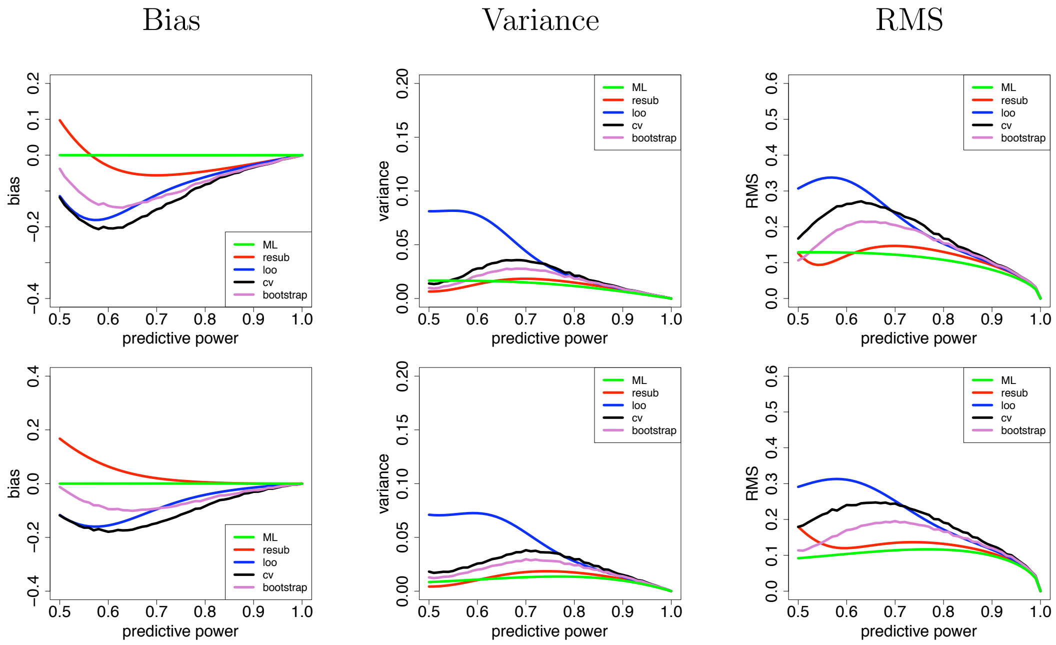


Figure 3.2: Bias, deviation variance, and RMS for several CoD estimators vs. predictive power with sample size $n = 60$. Top row, 2-input AND model with $P_1 = 0.8, P_2 = 0.6$ and $\gamma = 0.02$. Bottom row, 3-input AND model with $P_1 = 0.8, P_2 = 0.6$, $P_3 = 0.7, \gamma_{12} = 0.02, \gamma_{13} = 0.015, \gamma_{23} = 0.025$, and $\gamma = 0.02$. Plot key: resubstitution (red), leave-one-out (blue), cross-validation (black), 0.632 bootstrap (purple), MLE (green). The curves for resubstitution and leave-one-out are exact; the curves for cross-validation and 0.632 bootstrap are approximated via Monte Carlo sampling; the curve for the MLE is approximated via the asymptotic method described in the text.

and its negation (except that $\hat{p}$ is computed differently in each case, naturally).

Figure 3.1 illustrates the accuracy of the preceding approximations by comparing them to the exact values computed by complete enumeration, across the entire range of possible predictive values, for a 2-predictor AND model with $P_1 = P_2 = 0.5$ and $\gamma = 0.20$. Complete enumeration is here possible due to the small sample sizes considered, namely, $n = 10, 20, 30, 40$. The plots show that for sample sizes as small as $n = 30$, the results produced by the asymptotic approximation are essentially equal to the exact values, especially in the case of the variance, across the entire range of predictive power. We may therefore expect that the approximations will be very accurate for larger sample sizes, for which exact computation via complete enumeration is not possible. We also gather from the previous plots that the bias of the ML CoD estimator is very small, being essentially zero for $n = 40$ and larger sample sizes, also in agreement with the asymptotic approximation.

3.2.2.3 Comparison with Nonparametric CoD Estimators

Here we compare the performance of the parametric ML against that of the nonparametric ML (resubstitution) and resampling-based (leave-one-out, cross-validation, and 0.632 bootstrap) CoD estimators. Figure 3.2 displays results for the 2- and 3-predictor AND models, for varying predictive power, given a sample size $n = 60$. For the ML CoD estimator, the accurate asymptotic expressions developed in the previous section are used, whereas for the resubstitution and leave-one-out CoD estimators, exact formulas developed in [21] are used. For the cross-validation and 0.632 bootstrap CoD estimators, approximations based on Monte Carlo sampling are used (hence the plot jitter in the case of these estimators). One can see that the ML approaches have a clear advantage over the other estimators, being similar to each other in variance and RMS. However, the parametric MLE has the least bias, and the least RMS if the predictive power is not too small. The parametric MLE performs better in the 3-input than in the 2-input case, since nonparametric estimation becomes more difficult in higher-dimensional spaces, where the model information used by the parametric MLE becomes more important; this advantage can be expected to increase with 4 or more inputs.

3.3 Dynamical Model

We assume a vector stochastic process $\{\mathbf{X}_k; k = 0, 1, \ldots\}$, where $\mathbf{X}_k \in \{0, 1\}^d$ is a Boolean vector of size d representing the system state $\mathbf{X}_k = (\mathbf{X}_k(1), \ldots, \mathbf{X}_k(d))$ at time point k. We study the following nonlinear model:

$$\mathbf{X}_k \;=\; \mathbf{f}\left(\mathbf{X}_{k-1}\right) \;\oplus\; \mathbf{n}_k\,, \tag{3.29}$$

for $k = 1, 2, \ldots$. Here, "$\oplus$" indicates component-wise modulo-2 addition, $\mathbf{f} : \{0,1\}^d \to \{0,1\}^d$ is an arbitrary *network function*, which expresses a logical relationship between the system variables at consecutive time points, and $\{\mathbf{n}_k; k = 1, 2, \ldots\}$ is a white noise process, with $\mathbf{n}_k \in \{0,1\}^d$. The noise process is "white" in the sense that the noise at distinct time points are independent random variables. It is also assumed that the noise process is independent of the state process. The network function can be written in terms of its components, $\mathbf{f} = (f_1, f_2, \ldots, f_d)$, where each component $f_i : \{0,1\}^d \to \{0,1\}$, $i = 1, \ldots, d$, is a Boolean function expressing a logical relationship between $\mathbf{X}_k(i)$ and the previous state vector $\mathbf{X}_{k-1}$.

Under model (3.29), it is clear that $\{\mathbf{X}_k; k = 0, 1, \ldots\}$ is a Markov chain. Furthermore, it is a time-homogeneous Markov Chain if the noise process is identically distributed, i.e., $\mathbf{n}_k$ has the same distribution for all $k = 1, 2, \ldots$ which is assumed here. We make the additional assumption that the noise components $\mathbf{n}_k(i)$ are independent, with $P(\mathbf{n}_k(i) = 1) = 1 - p$, for $i = 1, \ldots, d$, for a parameter $\frac{1}{2} \leq p < 1$. In a similar fashion to the static model previously considered, one can assume $p \geq \frac{1}{2}$ without loss of generality, with $1 - p$ giving the amplitude of the noise; i.e. how often the state vector will be perturbed by flipping its components. Notice that components are flipped independently; it is only the rate of flipping that is assumed to be the same for all components. Under this noise distribution, the model (3.29) has been known in the literature as the "Boolean Network with perturbation" model [51].

The *transition matrix* $M = [M_{ij}]$ of the corresponding Markov Chain is given by

$$\begin{aligned} M_{ij} \; &= \; P(\mathbf{X}_k = \mathbf{x}^i \mid \mathbf{X}_{k-1} = \mathbf{x}^j) \; = \; P\left(\mathbf{n}_k \; = \; \mathbf{x}^i \oplus \mathbf{f}(\mathbf{x}^j)\right) \\ &= \prod_{k=1}^{d} p^{1-\mathbf{x}^i(k) \oplus f_k(\mathbf{x}^j)} (1-p)^{\mathbf{x}^i(k) \oplus f_k(\mathbf{x}^j)} \,, \end{aligned} \tag{3.30}$$

for $i, j \; = \; 1, \ldots, 2^d$, where $(\mathbf{x}^1, \ldots, \mathbf{x}^{2^d})$ is an arbitrary enumeration of the state vectors. It is clear that M is the transition matrix of an *ergodic* Markov chain [59]. Let $\boldsymbol{\pi}$ be the stationary probability distribution vector, with $\boldsymbol{\pi}(i) = P(\mathbf{X}_k = \mathbf{x}^i)$, for $i = 1, \ldots, 2^d$. We have $M\boldsymbol{\pi} = \boldsymbol{\pi}$. It can be shown that $\boldsymbol{\pi}$ can be computed explicitly as [41]

$$\boldsymbol{\pi} = (\mathbf{1}\mathbf{1}^T + I - M)^{-1} \, \mathbf{1} \,, \tag{3.31}$$

where I is the $2^d \times 2^d$ identity matrix, and $\mathbf{1} = (1, \ldots, 1)$ has length 2^d. From eqs. (3.30) and (3.31), we gather that $\boldsymbol{\pi}$ is a function of only the network function $\mathbf{f}$ and the noise parameter p, a fact that will be important in the next section.

3.3.1 Maximum-Likelihood Inference of the CoD

Consider the vector $\mathbf{CoD}$, where $\mathbf{CoD}(i)$ is the individual CoD of variable $\mathbf{X}_k(i)$ with respect to the preceding state $\mathbf{X}_{k-1}$, for $i = 1, \ldots, d$. The MLE of $\mathbf{CoD}$ is defined here as the vector $\widehat{\mathbf{CoD}}$ consisting of the MLEs of the individual CoDs. We derive in this section an accurate approximation to $\widehat{\mathbf{CoD}}$, under the assumption of stationarity, i,e, we assume that the system is in the steady state. In other words, we assume that the system has already been allowed to evolve "for a long time," so that the process $\{\mathbf{X}_k; k = 0, 1, \ldots\}$ is identically distributed according to the stationary distribution $\boldsymbol{\pi}$ of the Markov chain. Due to this, $\mathbf{CoD}$ is itself time-invariant and does not depend on k. Notice that, while identically distributed, $\{\mathbf{X}_k; k = 0, 1, \ldots\}$ is *not* independent.

According to the model (3.29),

$$\begin{aligned}\mathbf{CoD}(i) &= 1 - \frac{\varepsilon^*}{\varepsilon_0} = 1 - \frac{1-p}{F(P(\mathbf{X}_k(i)=1))} \\ &= 1 - \frac{1-p}{F\left(\sum_{j=1}^{2^d} P(\mathbf{X}_k(i)=1 \mid \mathbf{X}_{k-1}=\mathbf{x}^j)P(\mathbf{X}_{k-1}=\mathbf{x}^j)\right)} \\ &= 1 - \frac{1-p}{F\left(\sum_{j=1}^{2^d} \left[I(f_i(\mathbf{x}^j)=1)\,p + I(f_i(\mathbf{x}^j)=0)(1-p)\right]\boldsymbol{\pi}(j)\right)}\,,\end{aligned} \tag{3.32}$$

for $i = 1, \ldots, d$. Since $\boldsymbol{\pi}$ is a function of only $\mathbf{f}$ and p, and $\mathbf{f}$ is assumed to be known, it follows that only the MLE $\hat{p}$ of p is needed to obtain the MLE of $\mathbf{CoD}(i)$. In particular, it is not necessary to estimate any of the bias and covariance parameters present in the static case, discussed in Section 3.2.

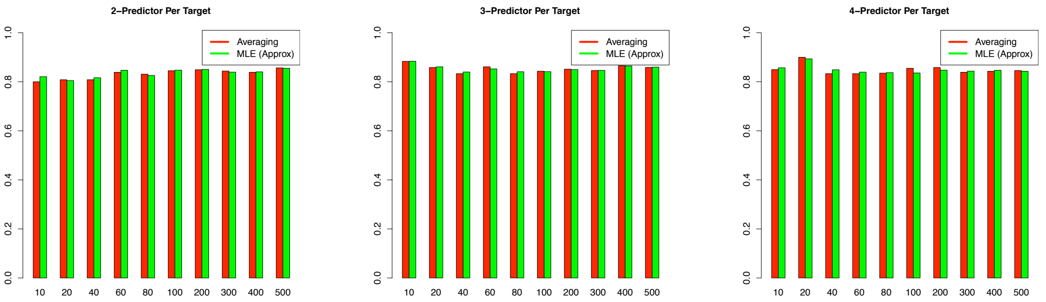


Figure 3.3: Comparison between $\hat{p}$ and $\hat{p}^{\mathrm{MLE}}$ as a function of increasing sample size, for 6-variable network functions with $l = 2, 3, 4$ predictors per target and $p = 0.85$.

Let $S_n = \{\mathbf{X}_m = \mathbf{x}^{i_m}, \ldots, \mathbf{X}_{m+n} = \mathbf{x}^{i_{m+n}}\}$ be an observation of the stationary process $\{\mathbf{X}_k; k = 0, 1, \ldots\}$, consisting of $n+1$ consecutive observations, comprising n state transitions. The likelihood function is:

$$L(p \mid S_n) \;=\; \boldsymbol{\pi}(i_m)\, M_{i_{m+1}i_m} \;\cdots\; M_{i_{m+n}i_{m+n-1}}\,. \tag{3.33}$$

There appears to be no simple analytical solution to the maximization of this likelihood function, as it involves the complex matrix inversion in (3.31). However, by noting that

$$p = P(\mathbf{X}_k(1) = f_1(\mathbf{X}_{k-1})) = P(\mathbf{X}_k(2) = f_2(\mathbf{X}_{k-1})) = \cdots = P(\mathbf{X}_k(d) = f_d(\mathbf{X}_{k-1})), \tag{3.34}$$

the following estimator immeditely presents itself

$$\hat{p} = \frac{1}{dn} \sum_{j=1}^{d} \sum_{k=1}^{n} I\left(\mathbf{x}^{i_{m+k}}(j) = f_j(\mathbf{x}^{i_{m+k-1}})\right) \tag{3.35}$$

We can actually show that $\hat{p} \to \hat{p}^{\mathrm{MLE}}$, the MLE of parameter p, in probability as $n \to \infty$. But $\hat{p}$ is also quite accurate for finite sample sizes, as shown in Figure 3.3, which plots $\hat{p}$ and $\hat{p}^{\mathrm{MLE}}$ as a function of increasing sample size, for 6-variable network functions with $l = 2, 3, 4$ predictors per target and $p = 0.85$. Here, $\hat{p}^{\mathrm{MLE}}$ is computed, to a good approximation, via numerical maximization of $L(p \mid S_n)$ in (3.33).

An accurate approximation to the MLE $\widehat{\mathbf{CoD}}(i)$ is then obtained by plugging $\hat{p}$ in (3.30), (3.31), and (3.29). A comment on (3.31): since this involves matrix inversion of a potentially very large matrix, an alternative to find the stationary distribution is to use the fact that each row of $\lim_{k\to\infty} M^k$ is equal to $\boldsymbol{\pi}$ [59]. The procedure adopted here is to increase k until $||M^k - M^{k-1}||$ is smaller than a certain pre-specified tolerance, and then read $\boldsymbol{\pi}$ off the resulting matrix.

3.3.1.1 Comparison with Nonparametric CoD Estimators

As in the static case, it is of interest to study how accurate the MLE developed in the previous section is, as compared to nonparametric alternatives. We again consider the nonparametric ML (resubstitution) and resampling-based (leave-one-out,

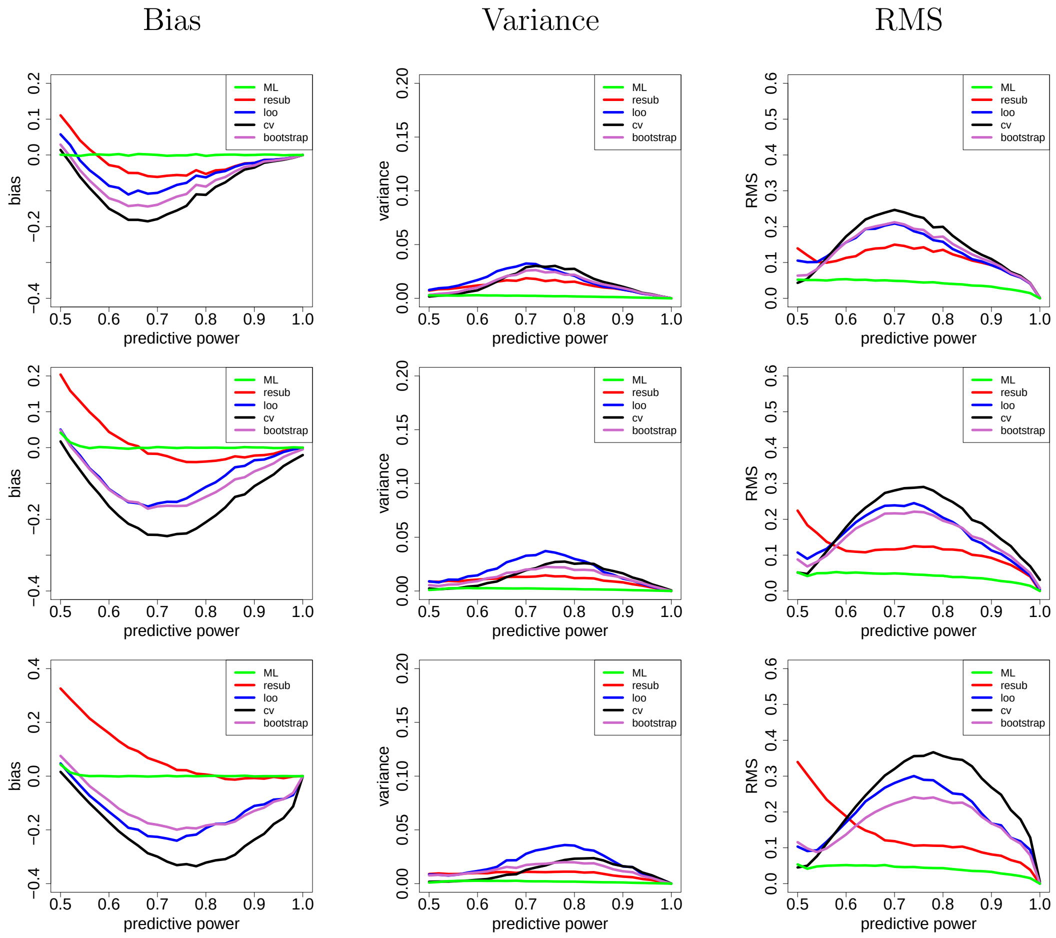


Figure 3.4: Bias, deviation variance, and RMS for several CoD estimators vs. predictive power with sample size $n = 60$. Top row: 2-input XOR; Midde row: 3-input XOR; Bottom row: 4-input XOR. All curves are approximated via Monte Carlo sampling.

cross-validation, and 0.632 bootstrap) CoD estimators. The bias, variance and RMS of estimation for the vector target are defined as the averages of the corresponding quantities for the estimators of each individual target. In the dynamical case, it is not desirable to consider systems containing only AND logics, as the underlying Boolean network converges quickly to the single attractor state $00 \cdots 0$. In the case where the noise is small, i.e., p is close to 1, the stationary distribution of the associated Markov process will assign large probability to this single state, which renders the comparison among the several CoD estimators problematic. Here we consider instead networks of XOR logics, which produce much less peaked stationary distributions for large p (see Supplementary Information). Figure 3.4 displays results for XOR models with $l = 2, 3, 4$ predictors per logic gate, for varying predictive power, given a sample containing $n = 60$ transitions. All results are approximations based on Monte Carlo sampling (hence the plot jitter). One can see that the general behavior is similar to that obtained in the static case with AND gates, c.f. Figure 3.2, except that now the MLE has an even bigger advantage over the other estimators. This can be explained by the fact that the MLE takes advantage here of the additional modeling assumption of a fixed p for all targets, whereas the nonparametric estimators, being unable to take advantage of any modeling assumptions, estimate p "anew" for each of the targets.

3.4 Application to System Identification

In this section we consider the system identification problem [50], that is, the case where incomplete knowledge about the network function is available, in the form of partial knowledge about the logic gates regulating each target variable, but no knowledge about the input variables to each logic gate (i.e., the network "wiring"). We propose inference procedures based on the parametric ML CoD estimator to recover the missing information, and investigate their performance by means of simulation. In the case of network wiring recovery, we compare the performance of the ML approach against the use of nonparametric CoD estimators, which are not capable of taking advantage of the available partial information.

We consider separately the static and dynamical cases. In both cases, the prior knowledge about the system The simulated numerical examples in this section take this into account by considering nested sets of candidate models, from more (smaller set) to less (larger set) informative. This allows us to examine the impact of the amount of prior knowledge has on inference accuracy.

3.4.1 Static Case: Predictor Inference

We consider here inference of the Boolean function f, or predictor, in model (3.12). It is assumed that the true predictor f is unknown but is a member of a candidate model set F containing several Boolean functions. For simplicity, it is assumed here that each predictor f in F depend on the same number l of essential predictive variables, or inputs, but the approach can be extended to remove this assumption. Each predictor f in F is thus specified by (1) a Boolean function $g : \{0,1\}^l \to \{0,1\}$, or logic gate, and (2) the indices for the predicting variable set $\{i_1, \ldots, i_l\} \subset \{1, \ldots, d\}$, or wiring, such that

$$f(\mathbf{X}) \;=\; f(X_1, \ldots, X_d) \;=\; g\left(X_{i_1}, \ldots, X_{i_d}\right) . \tag{3.36}$$

The total number of possible predictors is therefore $2^l \times \binom{d}{l}$.

Here we assume that the model set F consists of a number c of possible logic gates and arbitrary wiring of connectivity l. This reduces the number of all possible networks to $c \times \binom{d}{l}$. The parameter c is inversely related to the amount of prior knowledge available; the smaller c is, the more is known about the system, and vice-versa.

We propose the following predictor inference procedure to select a predictor from F.

1. For each logic gate, pick the wiring that produces the largest ML CoD estimate. Ties, if any, are broken randomly.

2. Among the c candidate predictors obtained from the previous step, select the

one that presents the largest predictive power estimate. Ties, if any, are broken randomly.

The previous procedure provides heuristics for the application of the ML CoD and ML predictive power estimators to predictor inference. Its effectiveness is assessed in the sequel by means of numerical experiments.

3.4.1.1 Numerical Experiments

We let $d = 8$ and set up three groups of experiments, corresponding to $l = 2, 3, 4$. A set of $k = 8$ models are considered in each case, each model being obtained by a random wiring assignment $\{i_1, \ldots, i_l\}$ and a choice of one of two logic gates:

- $l = 2$: $g_1(X_{i_1}, X_{i_2}) = X_{i_1}X_{i_2}$; $g_2(X_{i_1}, X_{i_2}) = X_{i_1} \oplus X_{i_2}$.
- $l = 3$: $g_1(X_{i_1}, X_{i_2}, X_{i_3}) = \overline{X_{i_1}}X_{i_3} + X_{i_2} \oplus X_{i_3}$; $g_2(X_{i_1}, X_{i_2}, X_{i_3}) = X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$.
- $l = 4$: $g_1(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}) = \overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$; $g_2(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}) = X_{i_1} \oplus X_{i_2} \oplus X_{i_3} \oplus X_{i_4}$.

Furthermore, we consider three different values of predictive power, $p = 0.65$, $p = 0.75$, and $p = 0.85$.

To set up the inference problem, we consider, for each value of l, three candidate model sets $F_l^1 \subset F_l^2 \subset F_l^3$, each containing all $\binom{8}{l}$ possible predictor variable assignments $\{i_1, \ldots, i_l\}$, for $l = 2, 3, 4$, and the logic gates depicted in Tables 3.1–3.3. As mentioned previously, the nesting of the candidate model sets allows us to assess the impact of a decreasing amount of prior knowledge about the system.

For each number of inputs l, predictive power p, and sample size n, a total of $r = 100$ datasets are drawn from each model. The proposed inference procedure is applied, and two performance measures are recorded for each of the three candidate model sets: the average rate of correct logic gates recovered and the average rate

Table 3.1: Logic gates for candidate model sets, static case, $l = 2$.

F_2^1	F_2^2	F_2^3
$X_{i_1}X_{i_2}$	$X_{i_1}X_{i_2}$	$X_{i_1}X_{i_2}$
$X_{i_1} \oplus X_{i_2}$	$X_{i_1} \oplus X_{i_2}$	$X_{i_1} \oplus X_{i_2}$
	$X_{i_1}\overline{X_{i_2}}$	$X_{i_1}\overline{X_{i_2}}$
	$\overline{X_{i_1}} + X_{i_2}$	$\overline{X_{i_1}} + X_{i_2}$
		$\overline{X_{i_1}}X_{i_2}$
		$X_{i_1} + \overline{X_{i_2}}$

Table 3.2: Logic gates for candidate model sets, static case, $l = 3$.

F_3^1	F_3^2	F_3^3
$\overline{X_{i_1}}X_{i_3} + X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}}X_{i_3} + X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}}X_{i_3} + X_{i_2} \oplus X_{i_3}$
$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$	$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$	$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$
	$\overline{X_{i_1}}(X_{i_2} \oplus X_{i_3}) + X_{i_1}\overline{X_{i_2}}$	$\overline{X_{i_1}}(X_{i_2} \oplus X_{i_3}) + X_{i_1}\overline{X_{i_2}}$
	$\overline{X_{i_1}}\,\overline{X_{i_2}} + X_{i_1}X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}}\,\overline{X_{i_2}} + X_{i_1}X_{i_2} \oplus X_{i_3}$
		$\overline{X_{i_1}}X_{i_2} + X_{i_1}\overline{X_{i_2} \oplus X_{i_3}}$
		$\overline{X_{i_1}}X_{i_3} + \overline{X_{i_2} \oplus X_{i_3}}$
		$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_2}} + X_{i_1}X_{i_2}$
		$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_2}} + X_{i_1}(X_{i_2} \oplus X_{i_3})$
		$\overline{X_{i_1}}\,\overline{X_{i_3}} + X_{i_1}(X_{i_2} \oplus X_{i_3})$
		$\overline{X_{i_1}}\,\overline{X_{i_3}} + X_{i_1}\overline{X_{i_2} \oplus X_{i_3}}$

of predictive variables correctly recovered For the latter, we count the number of correct predictive variables, not correct predicitive variable sets; this assigns partial credit if 3 out of $l = 4$ input variables are recovered for the wiring of a given target, for example.

The nonparametric CoD estimators are also employed to recover the wiring, through the simple inference procedure: all possible $\binom{d}{l}$ wirings $\{i_1, \ldots, i_l\}$ are considered, and the one that produces the largest estimated CoD is selected. The same measure of average rate of predictive variables correctly recovered, described above, is used to assess performance. Notice that nonparametric CoD estimators *cannot*

Table 3.3: Logic gates for candidate model sets, static case, $l = 4$.

F_4^1
$\overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$X_{i_1} \oplus X_{i_2} \oplus X_{i_3} \oplus X_{i_4}$
F_4^2
$\overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$X_{i_1} \oplus X_{i_2} \oplus X_{i_3} \oplus X_{i_4}$
$\overline{X_{i_1}X_{i_2}}\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_4}} + X_{i_1}\overline{X_{i_2}}\,\overline{X_{i_3}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
F_4^3
$\overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$X_{i_1} \oplus X_{i_2} \oplus X_{i_3} \oplus X_{i_4}$
$\overline{X_{i_1}X_{i_2}}\,\overline{X_{i_3} \oplus X_{i_4}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_4}} + X_{i_1}\overline{X_{i_2}}\,\overline{X_{i_3}} + X_{i_1}X_{i_2}(X_{i_3} \oplus X_{i_4})$
$\overline{X_{i_1}}\,\overline{X_{i_4}} + X_{i_1}\overline{X_{i_2} \oplus X_{i_3} \oplus X_{i_4}}$
$\overline{X_{i_1} \oplus X_{i_4}}\,\overline{X_{i_2}} + \overline{X_{i_1} \oplus X_{i_3} \oplus X_{i_4}}\,X_{i_2}$
$\overline{X_{i_2}}\,\overline{X_{i_4}} + \overline{X_{i_1} \oplus X_{i_3} \oplus X_{i_4}}\,X_{i_2}$
$\overline{X_{i_1} \oplus X_{i_2}}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}}$
$\overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{X_{i_3}X_{i_4}} + X_{i_1}X_{i_2}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2})\,\overline{X_{i_3} \oplus X_{i_4}}$
$\overline{X_{i_1}}\,\overline{X_{i_2}}\,\overline{\overline{X_{i_3}}X_{i_4}} + X_{i_1}X_{i_2}\,\overline{X_{i_4}} + (X_{i_1} \oplus X_{i_2)})\,\overline{X_{i_3}} \oplus X_{i_4}$

be used, by themselves, to recover the logic gates, only the wiring. Therefore, a comparison between ML and nonparametric methods for logic gate recovery cannot be performed.

Figure 3.5 and 3.6 display the results as a function of sample size, corresponding to the three candidate model sets $F_l^1 \subset F_l^2 \subset F_l^3$, for $l = 2, 3, 4$ and $p = 0.65, 0.75, 0.85$. We can see that in each case the recovery rates converge to 100% as sample size increases in all cases, but that convergence is much slower in the case of small predictive power p, i.e., more noise (note that the sample size scale is different among the plots). We can see that the performance of the ML-based inference method improves as more prior knowledge is available. We can see in Figure 3.6 that, in all cases, even for very small sample sizes, parametric ML-based inference is superior

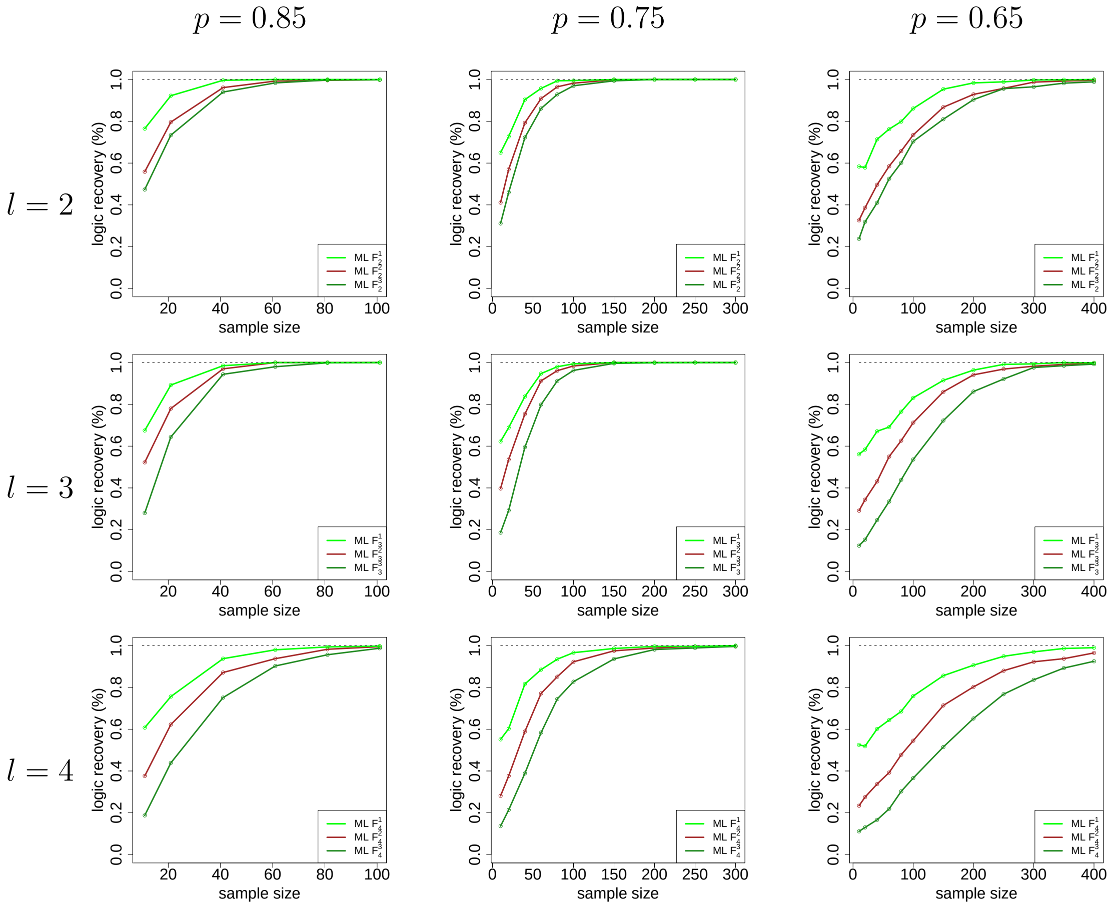


Figure 3.5: Average percentage of logic gates correctly recovered vs. sample size: static model.

to that of nonparametric methods. This is particularly true for $l = 3$ and $l = 4$ inputs, when the dimensionality and size of the search space becomes larger than for $l = 2$. Among the nonparametric methods, those based on the nonparametric ML (resubstitution) and bootstrap are the best, being nearly indistiguishable from each other (bootstrap in fact uses the nonparametric ML in its formulation), whereas cross-validation methods are the worst, with leave-one-out coming in last.

Notice that the performance of ML F_2^3 is very close to that of resubstitution and bootstrap. This is because in the $l = 2$ input case, there are only a total of

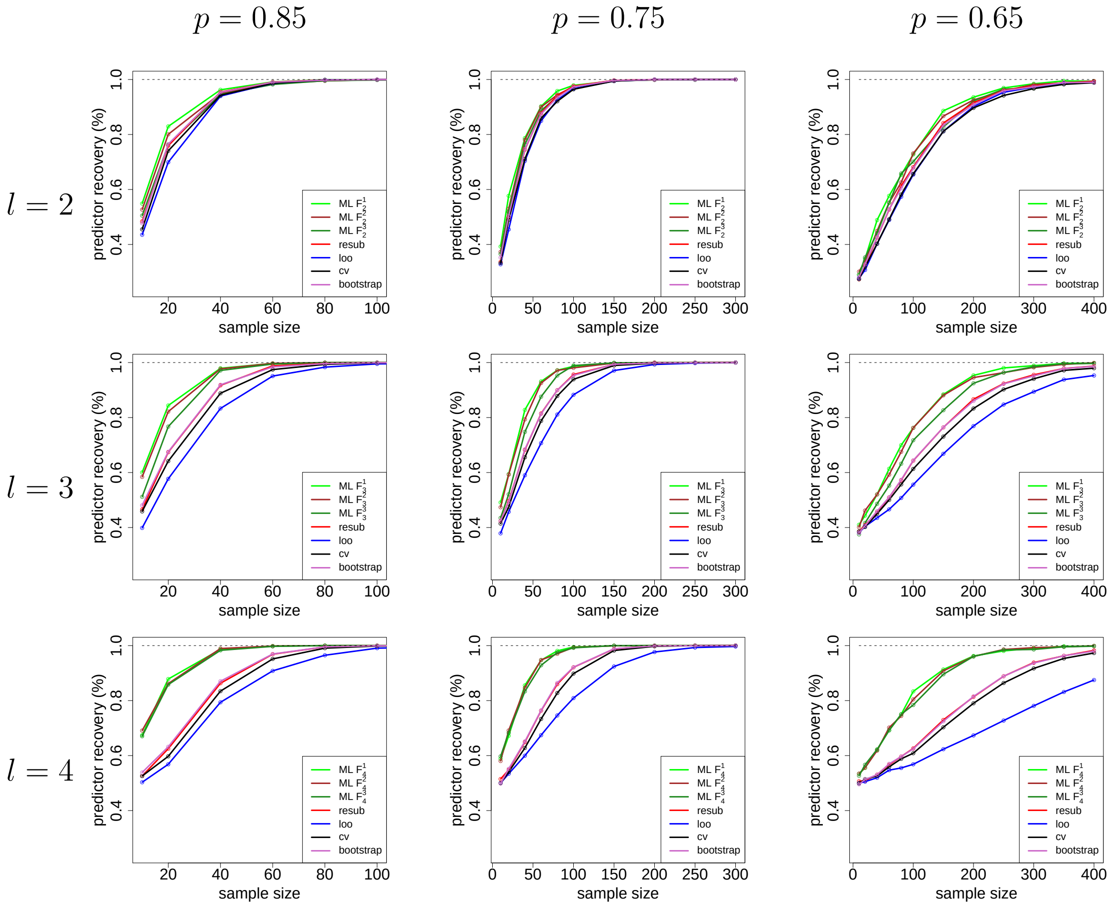


Figure 3.6: Average percentage of predictors correctly recovered vs. sample size: static model.

$2^4 = 16$ possible logics, among which only 10 are nontrivial 2-input logics, so that with a $c = 6$ candidate logic gates in model set F_2^3, there is little prior knowledge and performance of the parametric ML reduces to that of the nonparametric ML. Since the latter is less computationally expensive than the ML approach, and especially bootstrap, it would be the method of choice in the absence of any prior knowledge about the logic gates.

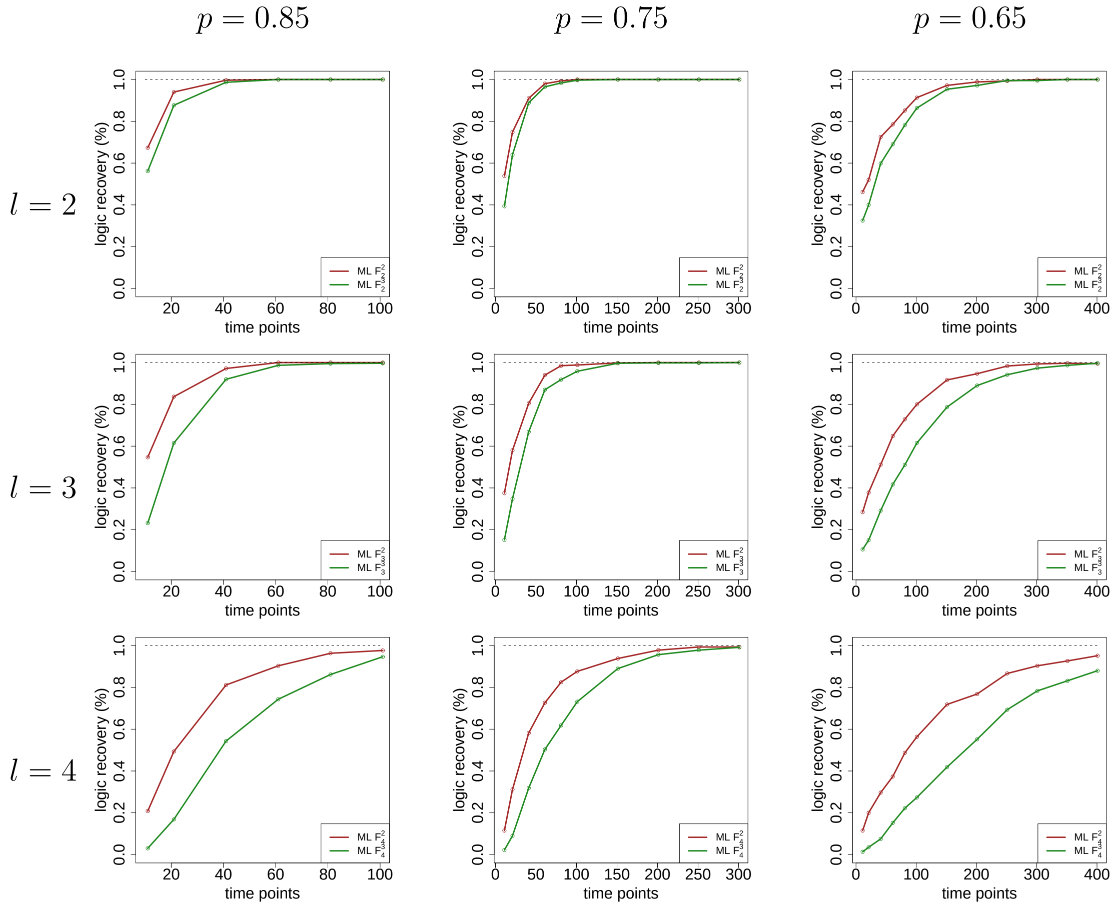


Figure 3.7: Average percentage of logics correctly recovered vs. time series length: dynamical model.

3.4.2 Dynamical Case: Network Inference

Here we address the inference of the network function $\mathbf{f}$, or simply network, in model (3.29). As in the static case, we assume that the unknown $\mathbf{f} = (f_1, \ldots, f_d)$ is a member of a candidate network set F. For simplicity, it is assumed here that the *connectivity* of the networks is fixed, i.e, the component Boolean functions f_i have the same number l of essential variables or inputs, for $i = 1, \ldots, d$. It has been suggested that low connectivity is a requirement for ordered system behavior [44] — accordingly, we consider here low connectivity values $l = 2, 3, 4$. Each network $\mathbf{f}$ is

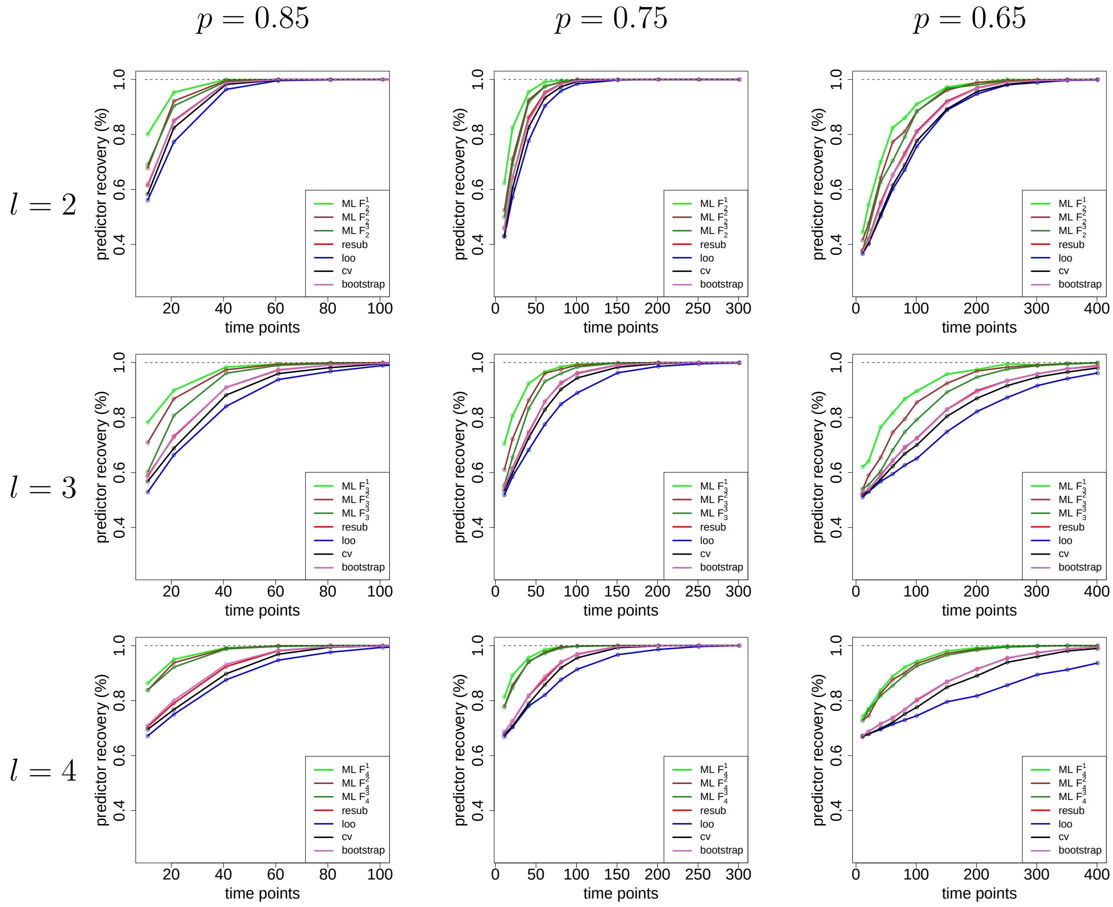


Figure 3.8: Average percentage of predictors correctly recovered vs. time series length: dynamical model

specified by the logic gates and wiring of its component Boolean functions (see the previous subsection). The total number of possible networks is thus $\left(2^l \times \binom{d}{l}\right)^d$, a very large number, even for modest values of d and l.

Here we assume that the model set F consists of a number c of possible logic gates and arbitrary wiring of connectivity l for each component Boolean function (the same set of c logic gates being considered for all components). This reduces the number of all possible networks to $\left(c \times \binom{d}{l}\right)^d$. As in the static case, c is inversely related to the amount of prior knowledge available. Notice that the number of networks is still very large – the inference procedure described below proposes a heuristic to reduce this search space to a manageable size. We remark that the number of networks cannot be reduced by considering the inference of the component Boolean functions separately, as the shared distribution of the noise $\mathbf{n}_k$ in (3.29) renders the inference problem irreducibly multivariate.

We propose the following network inference procedure to select a network from F.

1. For each of the d target variables, pick the two combinations of logic gate and wiring that present the largest predictive power estimate. Ties, if any, are broken randomly.

2. Compute the MLE $\widehat{\mathbf{CoD}}$ for each of the 2^d possible networks obtained form the previous step, and pick the one with the largest CoD in the L_1 sense, i.e. the one that maximizes $||\widehat{\mathbf{CoD}}||_1 = \sum_{i=1}^{d} \widehat{\mathbf{CoD}}(i)$. Ties, if any, are broken randomly.

The purpose of step 1. is to reduce the size of the search space in order to alleviate the computational complexity issue mentioned previously. After that, step 2. simply picks the network with the largest estimated CoD in the L_1 sense. The effectiveness of this procedure is assessed in the sequel by means of numerical experiments.

3.4.2.1 Numerical Experiments

We let $d = 6$, as opposed to $d = 8$ used in static case, for computational cost reasons. The network model consists of XOR logic gates regulating all targets and random wiring assignments corresponding to $l = 2, 3, 4$ connectivity. Furthermore, three different values of predictive power are considered, $p = 0.65$, $p = 0.75$, and $p = 0.85$.

The candidate model set F_l^1 consists of only the XOR gate with l inputs, and arbitrary wiring, for $l = 2, 3, 4$. This correspond to the situation where it is known that all logic gates in the network are XOR, but nothing is known about the wiring, which is to be inferred from the data. The candidate model sets F_l^2 and F_l^3, for $l = 2, 3, 4$, are the same as in the numerical experiments for the static case (see Tables 3.1–3.3), with the understanding that the logic gates in each model set apply to all the targets. The wiring for each target is entirely arbitrary, as before. Notice that

For each connectivity l and predictive power p, a total of $r = 100$ time series of length $n + 1$ (and thus n state transitions) are are drawn from each model in the steady-state regimen. The proposed inference procedure is applied to each sequence, and two performance measures are recorded for each of the three candidate model sets: the average rate of correct logic gates recovered and the average rate of predictive variables correctly recovered As before, we count the number of correct predictive variables recovered, as opposed to whether or not the entire wiring of the network is correctly recovered.

As in the static case, the nonparametric CoD estimators are also employed to recover the wiring, by simply picking, for each target, the wiring that produces the maximal CoD estimate.

Figure 3.7 and 3.6 display the results as a function of time series length. Note that in Figure 3.7, only two curves are plotted, for F_l^2 and F_l^3, since F_l^1 corresponds to full knowledge about the logic gates (XOR), for $l = 2, 3, 4$. Interestingly, the results are very similar to those obtained in the static case, and similar conclusions apply. For wiring recovery, the performance of parametric ML is superior to that of nonparametric methods. As the amount of prior knowledge is reduced, the performance of the parametric ML tends towards that of the nonparametric ML (resubstitution).

The latter is to be preferred in a situation where nothing is known about the network.

3.5 Summary

This chapter has presented a systematic theoretical framework for the inference of the CoD based upon a parametric maximum-likelihood approach, while highlighting its practical applications to estimation and system identification for static and dynamical Boolean models. Results reveal that the parametric ML CoD estimator outperforms the nonparametric alternatives provided that sufficient prior knowledge is available and the predictive power is not too small, i.e., the system noise level is not too high. The performance gap is larger for smaller sample sizes and larger dimensionality of the predictor vectors (i.e., larger connectivity of the regulatory network).

In fact, the parametric approach is especially suitable for small sample and large dimensionality situations, which can be ameliorated by the use of prior knowledge. Nonparametric approaches do not use prior knowledge and their performance thus degrades considerably with small sample sizes and large dimensionality. On the other hand, as less prior knowledge was available, the performance of the parametric and nonparametric ML CoD estimators were observed to equalize. This suggests that, in the no-information case, the NPML estimator would be preferable, due to its low computational complexity.

4. STATISTICAL DETECTION OF BOOLEAN REGULATORY RELATIONSHIPS*

DNA regulatory circuits can be often described by networks of Boolean logical gates updated and observed at discrete time intervals [2,9,37,38,44]. In a stochastic setting, the degree of association between Boolean predictors and targets can be quantified by means of the discrete Coefficient of Determination (CoD) [31], as discussed in previous chapters.

The CoD is often used in the inference of gene regulatory networks from gene-expression data [16,62,68]. However, applications of the CoD so far have been based on user-selected thresholds to decide on the presence of gene regulation between the given predictor and target genes. In this chapter, we will address this issue by providing a statistical test for a nonzero CoD between given Boolean predictors and a Boolean target in the context of a stochastic logic model that naturally allows the inclusion of prior knowledge if available. Rejection of the null hypothesis of zero CoD gives evidence for the presence of statistically-significant regulation. Even though the user still needs to choose the significance level, substituting this choice for the choice of an arbitrary CoD threshold has nevertheless advantages, beyond the fact that "standard" significance levels are available, such as $\alpha = 0.05$. The significance level can be interpreted as an upper bound on the false positive rate, whereas no such statistical interpretation can be attached to a user-selected CoD threshold.

Due to the multiple testing issue created by modern gene-expression experiments that monitor thousands of genes simultaneously, we furthermore propose multiple

*Parts of this section are reprinted with permission from "A Statistical Test for Intrinsically Multivariate Predictive Genes" by T. Chen and U.M. Braga-Neto, 2012, *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2012), Washington, DC, December 2012*, pp. 151–154, © 2012 IEEE.

testing procedures to control the overall Type I error rate, namely the single-step Bonferroni correction and the step-up Benjamini-Hochberg procedure, for controlling the family-wise error rate (FWER) and the false discovery rate (FDR), respectively [3,34]. We also discuss in this chapter the applications of the proposed methodology to real data sets for the detection of significant gene regulatory relationships.

4.1 Mathematical Preliminaries

After continuous measurements of gene expression have been binarized, a step that is not discussed here — for optimal methods to do this, see for example [60,71] — the sample data consist of a binary target random variable $Y \in \{0,1\}$ and a vector of binary predictor random variables $\mathbf{X} = (X_1, \ldots, X_d) \in \{0,1\}^d$. Due to uncertainty, noise affects the Boolean relationship between the predictors and the target, which is addressed here by a simple Boolean "additive-noise" model, that is, *stochastic logic model*, which has been discussed in our recent work [16,23]:

$$Y \;=\; f(\mathbf{X}) \oplus N, \tag{4.1}$$

where $f : \{0,1\}^d \to \{0,1\}$ is a Boolean logic function, the symbol "$\oplus$" indicates modulo-2 addition, and the noise N is a Bernoulli random variable that is independent of $\mathbf{X}$ and Y, such that $P(N = 1) = 1-p$, for $1/2 \leq p \leq 1$. Here, $1-p$ measures the amplitude of the noise. Please refer to Section 3.2 in Section 3 for more details.

We recall that the conditional distribution of the target given the predictor can be written entirely as a function of the logic function f and the parameter p:

$$\begin{aligned} &P(Y = 1 \mid \mathbf{X} = \mathbf{x}) \\ &\quad = \; I(f(\mathbf{x}) = 1)p \, + \, I(f(\mathbf{x}) = 0)(1-p)\,, \end{aligned} \tag{4.2}$$

where $I(A)$ is 1 when A is true, and 0, otherwise.

Let $\xi = P(f(\mathbf{X}) = 1)$; as we shall see, this distributional quantity plays a fundamental role in the sequel. In the context of model (4.1), it can be shown easily, by using (4.2), that the CoD is given by

$$\text{CoD} \;=\; 1 - \frac{1-p}{F\left[\xi p \,+\, (1-\xi)(1-p)\right]}\,, \tag{4.3}$$

where $F[u] = \min\{u, 1-u\}$, for $0 \leq u \leq 1$. The CoD is therefore a function of the distributional parameters $p \geq 1/2$ and $0 \leq \xi \leq 1$. Note that deterministic prediction is a function of p only: $\text{CoD} = 1 \Leftrightarrow \varepsilon_{\mathbf{X},Y} = 0 \Leftrightarrow p = 1$. The case $\text{CoD} = 0$ (i.e., no regulation) depends on both p and ξ, and is stated in the next proposition.

Proposition 1. *In the context of model (4.1), the following statements are equivalent:*

(i) $\text{CoD} = 0$.

(ii) $p = 1/2$ *or* $\xi \in \{0, 1\}$.

Proof. The result follows from equating the numerator and denominator in the ratio appearing in (4.3). Q.E.D.

For $0 < \xi < 1$, Proposition 1 assures us that $\text{CoD} = 0 \Leftrightarrow p = 1/2$, i.e., maximum noise. This would be the case, regardless of logic, if $P(\mathbf{X} = \mathbf{x}) > 0$ for all $x \in \{0,1\}^d$. Without distributional knowledge, one cannot however ignore the boundary condition $\xi \in \{0, 1\}$ when testing for null CoD.

As a concrete example, consider the case of $d = 2$ predictors, $\mathbf{X} = (X_1, X_2)$. In this case, there are a total of $2^{2^d} = 16$ possible prediction logics. Among those, six are either constant or depend only on one of the predictors, namely, 0, 1, X_1, X_2, $\overline{X}_1$, and $\overline{X}_2$. The remaining 10 logics are "true" 2-input logics, namely X_1X_2

(AND), $X_1 + X_2$ (OR), $X_1 \oplus X_2$ (XOR), $\overline{X}_1 X_2$, $X_1 \overline{X}_2$, and their negations. Logics can be represented by a bit string corresponding to the output column in its truth table; for example, 0001 (AND), 0111 (OR), 0110 (XOR), 0100 ($\overline{X}_1 X_2$), and 0010 ($X_1 \overline{X}_2$). The bit string representation is particularly convenient when checking the distributional constraint $\xi \in \{0,1\}$ in condition (ii) of Proposition 1. Now, note that if logic $\bar{f}$ is the negation of logic f, then $\bar{\xi} = 1 - \xi$, so that the constraint $\xi \in \{0,1\}$, and in fact the expression for the CoD in (4.3), are the same for f and $\bar{f}$, as can be easily checked. Among the 10 2-input logics, there are therefore a total of five cases to consider, which are listed in Table 4.1.

Similarly, for the case of $\text{CoD} = \theta \in [0,1)$, we can prove that

$$\text{CoD} = \theta \quad \Leftrightarrow \quad p = \frac{\delta + \min\{\xi, 1-\xi\}}{\delta + 2\min\{\xi, 1-\xi\}}, \tag{4.4}$$

where $\delta = \frac{\theta}{1-\theta}$. A small value of θ implies a loose regulation between a target and its predictors, whereas a large value implies a tight regulation.

Table 4.1: Distributional constraints for $\text{CoD} = 0:$ 2-input logic case

logic	bit string	constraint
OR / NOR	0111 / 1000	$p = 1/2$ or $P(0,0) \in \{0,1\}$
$\overline{X}_1 X_2$ / $X_1 + \overline{X}_2$	0100 / 1011	$p = 1/2$ or $P(0,1) \in \{0,1\}$
$X_1 \overline{X}_2$ / $\overline{X}_1 + X_2$	0010 / 1101	$p = 1/2$ or $P(1,0) \in \{0,1\}$
AND / NAND	0001 / 1110	$p = 1/2$ or $P(1,1) \in \{0,1\}$
XOR / NXOR	0110 / 1001	$p = 1/2$ or $P(0,0) + P(1,1) \in \{0,1\}$

4.2 CoD Hypothesis Test

The CoD is a function of the distribution parameters p and ξ of $(\mathbf{X}, Y)$, c.f. (4.3), and therefore statements about it can be statistically tested based on an i.i.d. sample $\mathbf{S}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ [14]. In particular, we are interested in the following hypothesis testing problem:

$$\begin{aligned} &\mathrm{H}_0 : \mathrm{CoD} = 0 \ \ (p = 1/2 \ \text{ or } \ \xi \in \{0, 1\}) \\ &\mathrm{H}_1 : \mathrm{CoD} > 0 \ \ (p > 1/2 \ \text{ and } \ 0 < \xi < 1)\,. \end{aligned} \tag{4.5}$$

The null hypothesis H_0 indicates the absence of useful prediction in $\mathbf{X}$ concerning the target Y, whereas the alternative hypothesis H_1 states that there is a degree of association between them.

This is a composite, multiparameter hypothesis testing problem. As the null parameter space is a union of two subsets $[p = 1/2]$ and $[\xi \in \{0, 1\}]$, the appropriate strategy to employ here is the *intersection-union test* (IUT) method; the individual tests for $p = 1/2$ and $\xi \in \{0, 1\}$ are level-α likelihood-ratio tests (LRTs), leading to an overall level-α IUT test [5,6]. This is summarized in the following result (details are found in the Appendix in the supplementary material).

Proposition 2. *For given $0 \le \alpha \le 1$, the test with rejection region*

$$\begin{aligned} \mathcal{R} \ = \ &\left\{ \mathbf{s}_n \ \middle| \ \sum_{i=1}^{n} I(f(\mathbf{x}_i) = y_i) \ge k \quad \textit{and} \right. \\ &\left. \exists\, 1 \le i, j \le n \ \textit{s.t.} \ f(\mathbf{x}_i) \ne f(\mathbf{x}_j) \right\} \,, \end{aligned} \tag{4.6}$$

where k is the $100(1-\alpha)\%$ percentile of a Binomial$(n, 1/2)$ distribution, i.e., k is the smallest integer such that

$$\sum_{l>k} \binom{n}{l} \left(\frac{1}{2}\right)^n \le \alpha\,, \tag{4.7}$$

is a level-α test for (5.7).
Proof. See Appendix D.

The following statements follow from Proposition 2.

(1) **Rejection region.** Notice that $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2$, where $\mathcal{R}_1 = \{\mathbf{s}_n \mid \sum_{i=1}^{n} I(f(\mathbf{x}_i) = y_i) \geq k\}$ is the rejection region for the $[p = 1/2]$ LRT, and expresses how tightly the data follows the proposed model, while $\mathcal{R}_2 = \{\mathbf{s}_n \mid \exists\, 1 \leq i, j \leq n \text{ s.t. } f(\mathbf{x}_i) \neq f(\mathbf{x}_j)\}$ is the rejection region for the $[\xi \in \{0, 1\}]$ LRT, and indicates that the null hypothesis cannot be rejected if $f(\mathbf{x}_i)$ is constant, for $i = 1, \ldots, n$. Notice that

$$\begin{aligned} P_{\mathcal{R}_2} \;&=\; P(\mathbf{S}_n \in \mathcal{R}_2) \;=\; 1 - P\left([f(\mathbf{X}_i) = 1, \forall i = 1, \ldots, n]\right. \\ &\cup\; \left.[f(\mathbf{X}_i) = 0, \forall i = 1, \ldots, n]\right) \\ &=\; 1 - \xi^n - (1 - \xi)^n \,. \end{aligned} \tag{4.8}$$

It follows that, unless $\xi \in \{0, 1\}$, in which case $\mathbf{S}_n \notin \mathcal{R}_2$ with probability 1, we have $P_{\mathcal{R}_2} \to 1$ as sample size increases to infinity. Therefore, the criterion for rejecting the null hypothesis will be, with probability approaching 1, whether or not $\mathbf{S}_n \in \mathcal{R}_1$, and the proposed test approaches an LRT for $p = 1/2$.

(2) **p-value.** The rejection regions for varying significance level α are *nested*, that is, $\mathcal{R}(\alpha_1) \subseteq \mathcal{R}(\alpha_2)$, whenever $\alpha_1 \leq \alpha_2$. This allows us to define a p-value for the proposed test as

$$\pi(\mathbf{s}_n) \;=\; \begin{cases} \displaystyle\sum_{l \,\geq\, \sum_{i=1}^{n} I(f(\mathbf{x}_i) = y_i)} \binom{n}{l} \left(\frac{1}{2}\right)^n, & \text{if } \mathbf{s}_n \in \mathcal{R}_2 \\ \qquad\qquad 1\,, & \text{otherwise.} \end{cases} \tag{4.9}$$

It is clear that $\pi(\mathbf{s}_n)$ is a *valid* p-value [14], i.e., under the null hypothesis, $P(\pi(\mathbf{s}_n) \leq u) \leq u$, for all $0 \leq u \leq 1$.

(3) **Statistical power.** The power function [14] of the proposed test can be

shown to be

$$\begin{aligned} \beta(p,\xi) \;=\; P(\mathbf{S}_n \in \mathcal{R}) \;= \\ \left(\sum_{l>k} \binom{n}{l} p^l (1-p)^{n-l}\right) \,\times\, (1-\xi^n \,-\, (1-\xi)^n)\,, \end{aligned} \tag{4.10}$$

for $p \geq 1/2$ and $0 \leq \xi \leq 1$, where k is given by (5.9). Note that, under the null hypothesis, either $\beta(p,\xi) = 0$, if $\xi \in \{0,1\}$, or $\beta(p,\xi) \leq \alpha$, if $p = 1/2$ and $0 < \xi < 1$ (by virtue of eq. 5.9). Therefore, $\sup \beta(p,\xi) \leq \alpha$ under the null hypothesis, so that this is indeed an α-level test. Under the alternative hypothesis, $\beta(p,\xi)$ gives the statistical power of the test. Notice from (4.10) that $\beta(p,\xi)$ not only on the distributional parameters p and ξ, but also on the level α sample size n, and logic function f (through ξ). Therefore, a power analysis has to take into account all of these factors. We consider below two important special cases for statistical power, where the analysis is facilitated.

Consider a uniform predictor distribution, $P(\mathbf{X} = \mathbf{x}) = 1/2^d$, for $\mathbf{x} \in \{0,1\}^d$. It is easy to see that this implies that the individual predictors $X_1, \ldots, X_d$ are independent. Clearly, $\xi = m/2^d$, where m is the number of minterms of logic f, i.e., the number of 1's in its bit string representation (c.f. Section 4.1). The cases $m = 0$ and $m = 2^d$ are uninteresting, since they correspond to the constant logics $f \equiv 0$ and $f \equiv 1$, respectively. In addition, m and $2^d - m$ lead to the same value for the CoD (c.f. equation 4.3), and hence for p and the power $\beta(p,\xi)$. It suffices thus to consider logics with $m = 1, \ldots, 2^{d-1}$ minterms, in which case it is possible to show that

$$\text{CoD} \;=\; \frac{m(2p-1)}{mp \,+\, (2^d - m)(1-p)} \;\;\Rightarrow\;\; p \;=\; \frac{1}{2}\,\frac{m + (2^d - m)\text{CoD}}{m + (2^{d-1} - m)\text{CoD}}\,. \tag{4.11}$$

Substituting this into (4.10) allows us to compute the power function in terms of the CoD (i.e., the "effect size"), and the number of minterms m and sample size n, which is displayed in Figure 4.1, in the case $d = 4$. A few values for the number of minterms are selected in the interval $m = 1, \ldots, 2^{d-1}$. We remark that for small values of m, the logic is of a *canalizing* type, which are very relevant in the investigation of gene regulatory relationships [44, 52]. Briefly, a canalizing logic is one where just one of the inputs alone can largely dictate the output, as in an AND logic ($m = 1$). We can see in Figure 4.1, that for large sample size, power increases to 1 very rapidly with effect size. In addition, power increases monotonically with a *decreasing* number of minterms, i.e., power is larger for canalizing logics. However, the behavior for small sample sizes is complex. Generally speaking, we can say that logics with fewer minterms lead to more powerful tests at small effect sizes, whereas logics with more minterms produce more power if the effect size is large. We can also see that the behavior of curves is qualitatively different at a severely small sample size, $n = 10$, where power is very small unless the CoD approaches 1.

Figure 4.2(a) on the other hand displays the minimum sample size necessary to achieve a standard power value of 80%, for $d = 4$ and a few values of number of minterms selected in the interval $m = 1, \ldots, 2^{d-1}$. The staircase pattern in the curves is due to the discrete nature of sample size. We can see that sample size requirement is monotonically decreasing with increasing CoD effect size, as expected. For small CoD effect size, the sample requirement is much larger for large values of m. For example, if CoD = 0.2, a 4-input AND logic ($m = 1$) would require a sample size of about $n = 40$, whereas the requirement for a 4-input XOR logic ($m = 8$) would be around $n = 180$ (for CoD = 0.1 the sample size for a XOR logic would be enormous). This shows the difficulty of detecting small CoDs, especially if the logic has many minterms. As for large CoDs, the situation improves considerably: we can see that sample requirement is low and essentially independent of m. In fact, the situation is reversed with respect to small CoDs, larger m here leads to slightly smaller required sample sizes.

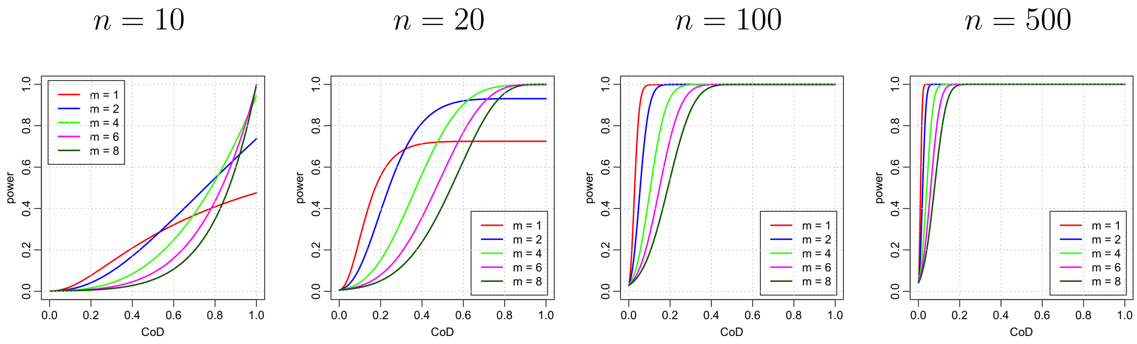
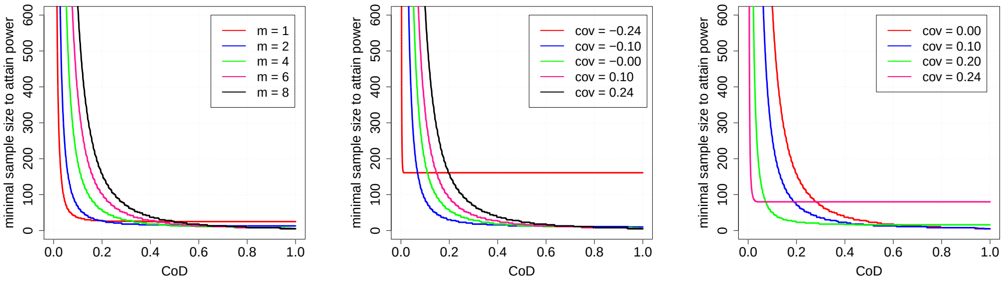


Figure 4.1: Statistical power vs. CoD for proposed test, in the uniform predictor case, with $d = 4$ and $\alpha = 0.05$, and varying sample size n and number of logic function minterms m.



(a) (b) (c)

Figure 4.2: Minimum sample size to achieve power = 0.8 vs. CoD for proposed test, with $\alpha = 0.05$. (a) Uniform predictor case, with $d = 4$ and varying number of logic function minterms m. (b-c) Correlated predictor case, with $d = 2$ and varying predictor covariance γ, for logic functions AND (b) and XOR (c).

Consider two predictors X_1 and X_2, such that $P(X_1) = P(X_2) = 1/2$; these are referred to as "unbiased" predictors in [16,52]. Let

$$\begin{aligned}\gamma &= \mathrm{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] \\ &= P(X_1 = 1, X_2 = 1) - \frac{1}{4}\,.\end{aligned} \tag{4.12}$$

From the constraint $P(X_1) = P(X_2) = 1/2$ it follows that $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$. When $\gamma = 0$ one obtains the case of uniform independent predictors previously considered, for $d = 2$.

With $d = 2$, there are only two families of useful logics to consider, according to number of minterms: the case $m = 1, 3$, represented here by the AND logic, and $m = 2$, represented here by the XOR logic. These cases correspond to the minimum (canalizing) and maximum (non-canalizing) number of minterms possible, respectively. For the AND logic, it is easy to see that $\xi = 1/4 + \gamma$. In addition, it can be shown that:

$$\mathrm{CoD} = \frac{(2p-1)(1+4\gamma)}{4(1-p)+(2p-1)(1+4\gamma)} \Rightarrow p = \frac{1}{2}\,\frac{(1+4\gamma)+(3-4\gamma)\mathrm{CoD}}{(1+4\gamma)+(1-4\gamma)\mathrm{CoD}}\,. \tag{4.13}$$

For the XOR logic, on the other hand, we have $\xi = 1/2 - 2\gamma$. Furthermore,

$$\mathrm{CoD} = \frac{(2p-1)(1-4|\gamma|)}{2(1-p)+(2p-1)(1-4|\gamma|)} \Rightarrow p = \frac{1}{2}\,\frac{(1-4|\gamma|)+(1+4|\gamma|)\mathrm{CoD}}{(1-4|\gamma|)+4|\gamma|\mathrm{CoD}}\,. \tag{4.14}$$

Substituting the expressions for p and ξ in each case above into (4.10) allows us to compute the power function in terms of the CoD effect size and the covariance parameter γ, which is displayed in Figure 4.3, for the AND and XOR logic cases. A few values of the covariance parameter are selected from the allowed interval $-1/4 \leq \gamma \leq 1/4$, but the case of perfectly negatively correlated predictors, $\gamma = -0.25$, is

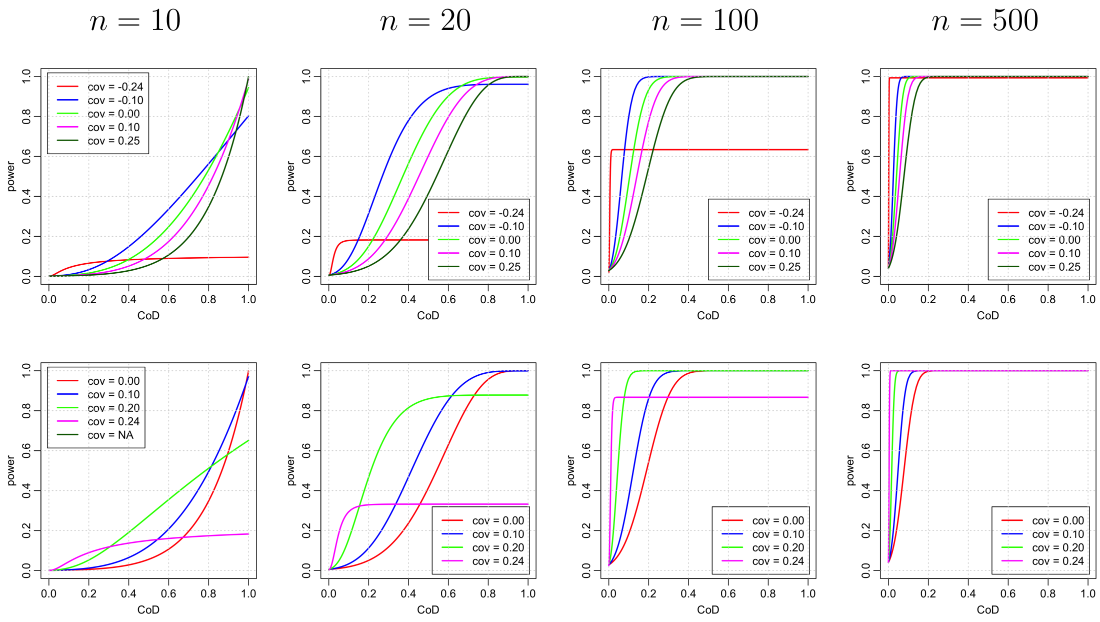


Figure 4.3: Statistical power vs. CoD for proposed test, in the correlated predictor case with $d = 2$, and varying sample size n and predictor covariance γ. Top row: AND logic. Bottom row: XOR logic.

excluded, as it corresponds to the null hypothesis CoD = 0, in both AND and XOR cases. In addition, power is a function of $|\gamma|$ in the XOR case, so that only curves for $\gamma \geq 0$ are plotted (each of which give the cases of both positive γ and negative $-\gamma$ correlation, of course). As in the previous example of uncorrelated predictors, we can see that for large sample size, power increases to 1 very rapidly with effect size. For $n = 500$, power decreases monotonically with increasing predictor correlation in the AND case; while it monotonically *increases* with increasing *magnitude* of predictor correlation, in the XOR case. However, as before, the behavior for small sample sizes is complex. It can be said that in the AND case, power generally is larger for negatively correlated predictors if the effect size is small, while positively correlated predictors lead to more powerful tests at large effect sizes. For the XOR logic, highly correlated predictors (regardless of sign) lead to more powerful tests for small effect size, while weakly correlated predictors produce more power at large effect sizes. As before, the behavior of curves is qualitatively different at a severely small sample size, $n = 10$, where power is very small unless the CoD approaches 1.

Figure 4.2(b-c) displays the minimum sample size necessary to achieve a standard power value of 80%, for for the AND and XOR logic cases, respectively, and a few values of the covariance parameter in the allowed interval $-1/4 \leq \gamma \leq 1/4$. As in the previous example of uncorrelated predictors, we can see that the sample size requirement is monotonically decreasing with increasing CoD effect size. For small CoD effect size, the sample requirement is much larger for large values of covariance γ (in the case of XOR, large values in magnitude). For large CoD size, the situation is reversed, dramatically so in the case of predictors with large negative correlation in the AND case, and uncorrelated predictors in the XOR case.

We remark than an extension of these results to $d \geq 3$ predictors is possible using an appropriate parametrization for the covariance structure of the predictor vector; such a parametrization is given in [16].

The results of the power analysis for the proposed test, displayed in Figures 4.1 and 4.3, may be summarized as follows. If a small CoD effect size is expected, then sample sizes in the neighborhood of $n = 100$ or larger are required for effective statistical power; in this case, small number of minterms (canalizing logics) lead to larger statistical power, while uncorrelated predictors lead to smaller power. If large CoD values, i.e., a tightly regulated target, is expected, then smaller sample sizes may be employed, as long as the logic of prediction contains a sufficiently large number of minterms and the predictors are weakly correlated, or, if the logic is closer to a canalizing type, the predictors are sufficiently positively correlated.

(4) **Confidence Interval.** A confidence interval for the CoD can be derived by

considering a test of

$$\mathrm{H}_0 : \mathrm{CoD} = \theta \text{ vs. } \mathrm{H}_1 : \mathrm{CoD} \neq \theta, \tag{4.15}$$

where $\theta \in (0, 1)$. The likelihood ratio test statistic is given by $(\{i_1, \ldots, i_d\} \subseteq \{0, 1\}^d)$

$$\begin{aligned} \lambda(\mathbf{s}_n; \theta) \; &= \; \frac{\sup_{\mathrm{CoD}=\theta} P(\mathbf{S}_n = \mathbf{s}_n)}{\sup P(\mathbf{S}_n = \mathbf{s}_n)} \; = \\ &\frac{\sup p^{n_f}(1-p)^{n_f} \prod_{\{i_1,\ldots,i_d\}} P(\mathbf{X} = \{i_1, \ldots, i_b\})^{n_{i_1 \ldots i_b}}}{(n_f/n)^{n_f}(1-n_f/n)^{n-n_f}(n_{i_1 \ldots i_b}/n)^{n_{i_1 \ldots i_b}}}, \end{aligned} \tag{4.16}$$

where $n_f = \sum_{i=1}^n I(f(\mathbf{X}_i) = Y_i)$, $n_{i_1 \ldots i_b} = \sum_{i=1}^n I(\mathbf{X}_i = \{i_1, \ldots, i_b\})$, and p is expressed by eq. (4.4). Note that the optimization problem of the numerator in eq. (4.16) can be solved by the method of gradient descent when there are multiple parameters [55]. Under regularity conditions, the LRT statistic follows an asymptotic distribution, that is, under the H_0, as $n \to \infty$, $-2 \log \lambda(\mathbf{S}_n; \theta) \to \chi_1^2$ [14]. Hence, given some θ, the rejection region of such an asymptotic size α test is formulated by

$$\mathcal{R} \; = \; \left\{ \mathbf{s}_n \;\middle|\; -2 \log \lambda(\mathbf{s}_n; \theta) \geq \chi_1^2(\alpha) \right\}, \tag{4.17}$$

where $\lambda(\mathbf{S}_n)$ is shown in eq. (4.16).

By inverting the LRT [14], the approximate $1 - \alpha$ confidence interval of the CoD, the set with plausible values of θ, is given by

$$\mathcal{C}(\mathbf{s}_n) \; = \; \left\{ \theta \;\middle|\; -2 \log \lambda(\mathbf{s}_n; \theta) \leq \chi_1^2(\alpha) \right\}, \tag{4.18}$$

which can be numerically solved by the bisection method [12].

In the following, we consider again two important special cases (i.e., uniform and correlated predictors) for estimation of the confidence interval.

In the uniform predictor case, CoD is a function of only p (c.f. eq. 4.11). Since

$n_f = \sum_{i=1}^{n} I(f(\mathbf{x}_i) = y_i)$ Binomial(n, p), the Clopper-Pearson interval is employed to calculate the $1-\alpha$ binomial confidence interval [25]. By substituting this confidence interval for p into eq. (4.11), we can obtain the confidence interval for the CoD, that is,

$$\left[\frac{m(2p_L - 1)}{mp_L \,+\, (2^d - m)(1 - p_L)}, \frac{m(2p_U - 1)}{mp_L \,+\, (2^d - m)(1 - p_U)}\right], \tag{4.19}$$

where $p_L = \mathrm{Beta}\left(\alpha/2; \mathrm{n_f}, \mathrm{n} - \mathrm{n_f} + 1\right)$ and $p_U = \mathrm{Beta}\left(1 - \alpha/2; \mathrm{n_f} + 1, \mathrm{n} - \mathrm{n_f}\right)$. Note that $\mathrm{Beta(t; a, b)}$ is the t-th quantile from a beta distribution with parameters a and b.

In the correlated predictor case, the confidence interval is approximated by the asymptotic distribution, that is, χ_1^2 distribution, as discussed in the general case. Table 4.2 shows the confidence interval estimate of the CoD based on random sample with $n = 100$ generated by a 2-input AND logic model in the general, uniform, and correlated predictor cases, respectively. We observe that the true values of θ lie in the corresponding confidence intervals in all cases. Note that the approximation works better for a larger sample size.

Table 4.2: 95% Confidence interval (CI) for the CoD based on one random sample generated from a 2-input AND logic model ($n = 100$): (a) in the general case ($P_1 = 0.8$, $P_2 = 0.6$, $\gamma = 0.05$, $d = 2$); (b) in the uniform predictor case ($m = 1, d = 2$); (c) correlated predictor case ($\gamma = 0.05$, $d = 2$)

θ	CI (General)	CI (Uniform)	CI (Correlated)
0.0	$[0.0000, 0.2153]$	$[-0.1229, 0.0668]$	$[0.0000, 0.1303]$
0.1	$[0.0000,\ 0.2369]$	$[0.0178, 0.2743]$	$[0.0096, 0.2506]$
0.2	$[0.0925,\ 0.4488]$	$[0.1113, 0.4037]$	$[0.0854, 0.3026]$
0.3	$[0.1023,\ 0.4738]$	$[0.2286, 0.5545]$	$[0.2917, 0.5201]$
0.4	$[0.1486,\ 0.4927]$	$[0.1813, 0.4953]$	$[0.3192, 0.4166]$

4.3 Multiple Testing Procedure

For a given target Y, the proposed test for multivariate Boolean relationships presupposes the model (4.1), which in turn depends on the choice of logic function f and predictor vector $\mathbf{X}$. Assuming dimensionality d and a number of genes G in the original gene-expression dataset, the total number of possible logic functions is 2^d and the number of distinct predictors is $\binom{G}{d}$. This creates a multiple testing issue; the total number of tests to be carried out would be, in this case, $M = 2^d \times \binom{G}{d}$. In typical gene-expression microarray or RNA-seq studies, G tends to be very large (in the order of thousands or more) so that, even if d is kept small, the number of tests may be very large indeed. In this section, we address the multiple testing problem in the context of the proposed detection method. We also comment on how to reduce the number of tests by use of prior knowledge.

4.3.1 Type-I Error Rates and Power

In a multiple testing procedure (MTP), there is a total of M null hypotheses to be simultaneously tested, $\{H_0(m) \mid m = 1, \ldots, M\}$. While there is no ambiguity in defining a type-I error for a single test, in the case of MTPs the situation is less clear [33]. Let $0 \leq R \leq M$ be the number of hypotheses rejected by the test, and let $0 \leq V \leq R$ be the number of hypotheses falsely rejected (i.e., "false positives"). We consider in this paper two specific definitions of type-I error rates for MTPs:

- *The family-wise error rate* [54] is defined as FWER $= P(V \geq 1)$.
- *The false discovery rate* [3] is defined as

$$\begin{aligned} \text{FDR} &= E\left[\frac{V}{R}\, I(R > 0)\right] \\ &= E\left[\frac{V}{R} \mid R > 0\right] P(R > 0)\,. \end{aligned} \tag{4.20}$$

The FWER gives the probability of at least one false positive, whereas the FDR essentially gives the average, or expected, proportion of false positives in the list of rejected hypotheses (with the proviso that, if no hypotheses are rejected, i.e., $R = 0$, then FDR = 0). It can be shown quite easily that the FDR is always smaller or equal than the FWER, with strict equality holding in the case where all the null hypotheses are true [3].

In the multiple testing procedures that control the Type-I error rate at a given level α, one also expects to maximize power. We consider here the definition of the power for MTPs as given by:

$$\mathrm{PWR} = \frac{\mathrm{E[S]}}{\mathrm{h_1}}, \tag{4.21}$$

where S is the true positives and h_1 is the number of false null hypotheses [34]. Obviously, The power gives the expected value of the proportion of true positives among the false null hypotheses. Note that the power estimate is mathematically equal to the true positive rate, that is, S/h_1.

4.3.2 Control of the Type-I Error Rate

For a given $0 < \alpha < 1$, an MTP is said to control the FWER at level α if FWER $\leq \alpha$. Similarly, an MTP is said to control the FDR at level α if FDR $\leq \alpha$. Notice that, since FDR $\leq$ FWER, any FWER-controlling procedure is also FDR-controlling, but the converse is not true in general, unless all null hypotheses are true, in which case FDR = FWER, as mentioned previously.

Suppose that individual tests of the hypotheses $\{H_0(m) \mid m = 1, \ldots, M\}$ are performed, producing a set of (valid) *unadjusted p-values* $\{\pi_1, \ldots, \pi_M\}$. Let

$$\pi'_m \;=\; \min\{M\pi_m, 1\}\,, \quad m = 1, \ldots, M \tag{4.22}$$

be the set of *adjusted p-values.* Then it can be shown, by an application of Boole's inequality, that rejection of $H_0(m)$ if $\pi'_m \leq \alpha$, for $m = 1, \ldots, m$, is an MTP that controls the FWER at level α [33]. This is the well-known *Bonferroni Correction* method [54].

Similarly, let $\{\pi^*_1, \ldots, \pi^*_M\}$ be the list of unadjusted p-values sorted in increasing order, and define the set of adjusted p-values by

$$\pi''_m = \min_{h=m,\ldots,M} \left\{ \min \left\{ \frac{M}{h} \pi^*(h), 1 \right\} \right\} , \quad m = 1, \ldots, M \, . \tag{4.23}$$

Then it can be shown that rejection of $H_0(m)$ if $\pi''_m \leq \alpha$, for $m = 1, \ldots, M$, is an MTP that controls the FDR at level α, under the assumption of independence of the p-values for the true null hypotheses [3, Thm. 1] or for certain dependence structures among the p-values [4, Thm. 1.2]. If the p-values have an arbitrary dependence structure, the previous procedure will only control the FDR approximately. Here we utilize this FDR-controlling procedure, and assess its efficacy by means of simulation (see the next subsection).

As pointed out in [3], the power of the FWER- and FDR-controlling procedures described previously decreases as the number of tests M increases. In practice, to have a useful MTP with reasonable power, the number of tests has to be reduced by using *prior knowledge.* In our case, let the true predictor set belong to a set L, and assume that it is related to the target via a logic function f in a set K. The total number of tests is thus $M = |L| \times |K|$. Provided that $|L| \ll \binom{G}{d}$ and $|K| \ll 2^d$, which are the prior knowledge constraints of the problem, then the number of tests M may be kept reasonably small.

From the previous considerations, we arrive at the following MTP.

Coefficient of Determination MTP.

(1) Set the significance level α, and model sets L and K. The total number of tests is $M = |L| \times |K|$.

(2) For the given data set $\mathbf{S}_n = \mathbf{s}_n$, compute the unadjusted p-values $\{\pi_1(\mathbf{s}_n), \ldots, \pi_M(\mathbf{s}_n)\}$ for the tests $H_0(m)$: CoD $= 0$ vs. $H_1(m)$: CoD > 0, for $m = 1, \ldots, M$, using Eq. (4.9).

(3-a) **FWER-controlling step.** Compute the adjusted p-values $\{\pi'_1(\mathbf{s}_n), \ldots, \pi'_M(\mathbf{s}_n)\}$ according to Eq. (4.22). Reject those hypotheses $H_0(m)$ such that $\pi'_m \leq \alpha$, for $m = 1, \ldots, M$.

(3-b) **FDR-controlling step.** Compute the adjusted p-values $\{\pi''_1(\mathbf{s}_n), \ldots, \pi''_M(\mathbf{s}_n)\}$ according to Eq. (4.23). Reject those hypotheses $H_0(m)$ such that $\pi''_m \leq \alpha$, for $m = 1, \ldots, M$.

It can be shown that the FDR-controlling step can be equivalently implemented by the following more efficient procedure [3]:

(3-b)' **FDR-controlling step.** Find the list of increasing unadjusted p-values $\{\pi^*_1(\mathbf{s}_n), \ldots, \pi^*_M(\mathbf{s}_n)\}$ and let $H^*_0(m)$ be the null hypothesis corresponding to $\pi^*_m(\mathbf{s}_n)$, for $m = 1, \ldots, M$. Let m^* be the largest m such that $\pi^*_m(\mathbf{s}_n) \leq \frac{m}{M}\alpha$. Reject all $H^*_0(m)$ for $m = 1, \ldots, m^*$. If $\pi^*_m(\mathbf{s}_n) > \frac{m}{M}\alpha$ for all $m = 1, \ldots, M$, then reject none of the hypotheses.

4.3.3 Performance of Multiple Testing Procedures

In this section, we assess the effectiveness of the previous CoD MTP by means of simulation experiments. For the first experiment, we assume that each target Y is regulated by predictors X_1 and X_2 among a set of possible predictors $X_1, \ldots, X_G$, such that $Y = X_1 \mathrm{XOR} \mathrm{X}_2 \oplus \mathrm{N}$, where $N \sim$ Bernoulli(1−p), for $1/2 \leq p \leq 1$, as before. Furthermore, we assume that the distribution of the random vector $(X_1, \ldots, X_G)$ is uniform. This specifies the stochastic model. Notice that here $L = \binom{G}{2}$. Provided that G is not too large, this does not create a serious multiplicity issue; in our simulation, G ranges from 4 to 24. In addition, we consider a number of targets D varying from 1 to 8. As for the logic model set K, we consider three scenarios: (1) the prediction logic is known, $K_1 = \{\mathrm{XOR}\}$; (2) $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and (3) $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \bar{\mathrm{X}}_1\bar{\mathrm{X}}_2, \bar{\mathrm{X}}_1 + \mathrm{X}_2\}$. The total number of tests is given by $M_i = D \times \binom{G}{2} \times K_i$, under each of the prior-knowledge scenarios $i = 1, 2, 3$ described previously. Hence, the MTP increases in difficulty as the number of predictors and targets increase, and as less prior knowledge is available. We draw 5000 samples of varying size n and form averages of FWER, FDR, and power estimates under FWER-controlling and FDR-controlling procedures.

In the first set of results, we fix $D = 1$, $G = 24$, and plot the results as a function of the sample size n. The total number of tests is $M_1 = 276$, $M_2 = 552$, and $M_3 = 1104$, under each logic model set. Note that, as there is only one target under consideration, the number of false null hypotheses is one, whereas the number of true null hypotheses is the total of tests in each case minus one. The results are displayed in Figure 4.4. We can observe that the FWER- and FDR-controlling procedures are able to control the FWER and FDR, respectively, at all sample sizes. In addition, as expected from the theory, FWER estimates are always larger than FDR estimates (in particular, FDR is controlled by the FWER-controlling procedure, but *not* vice-versa). We can also see that the FWER-controlling procedure is more conservative, producing smaller FWER and FDR estimates than the FDR-controlling procedure. As for power, there is little difference between the two procedures in this case, with a very small advantage for the FDR-controlling procedure, as expected. We can see, for small sample sizes, that there is a loss of power as less prior knowledge is available.

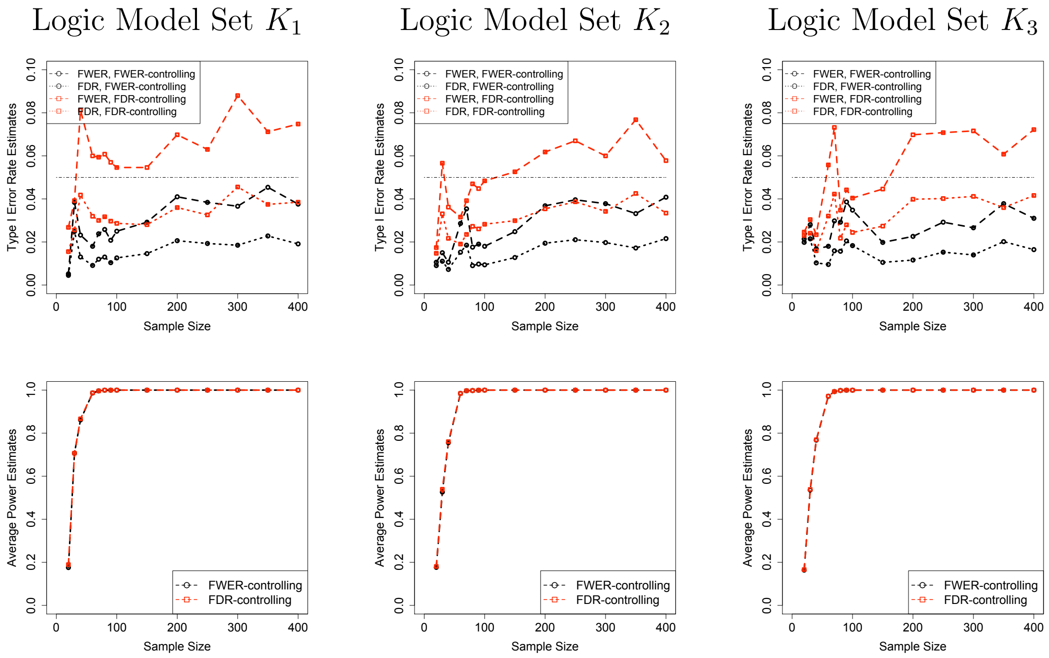


Figure 4.4: Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\text{XOR}\}$; $K_2 = \{\text{AND}, \text{XOR}\}$; and $K_3 = \{\text{AND}, \text{XOR}, \bar{\text{X}}_1\bar{\text{X}}_2, \bar{\text{X}}_1 + \text{X}_2\}$, and predictive power $p = 0.85$. There is a single target to be predicted by two among $G = 24$ genes.

For the next group of experiments, we fix $D = 1$, $n = 40$ and plot the results as a function of the initial number of genes G. The total number of tests varies from a minimum of 6 in the case of $G = 4$ and complete knowledge about the prediction logic to a maximum of 1104, for $G = 24$ and logic model set K_3. The results are displayed in Figure 4.5. The previous observations regarding the FWER and FDR estimates are still valid in this case. As for power, we can again observe little difference between the FWER- and FDR-controlling procedures, but it is possible to observe a clear and accentuated decrease in power as G increases. This indicates that in experiments with more than a few dozen initial genes and small sample sizes (here, $n = 40$), one can expect to face the issue of lack of power, in case of a very small number of false

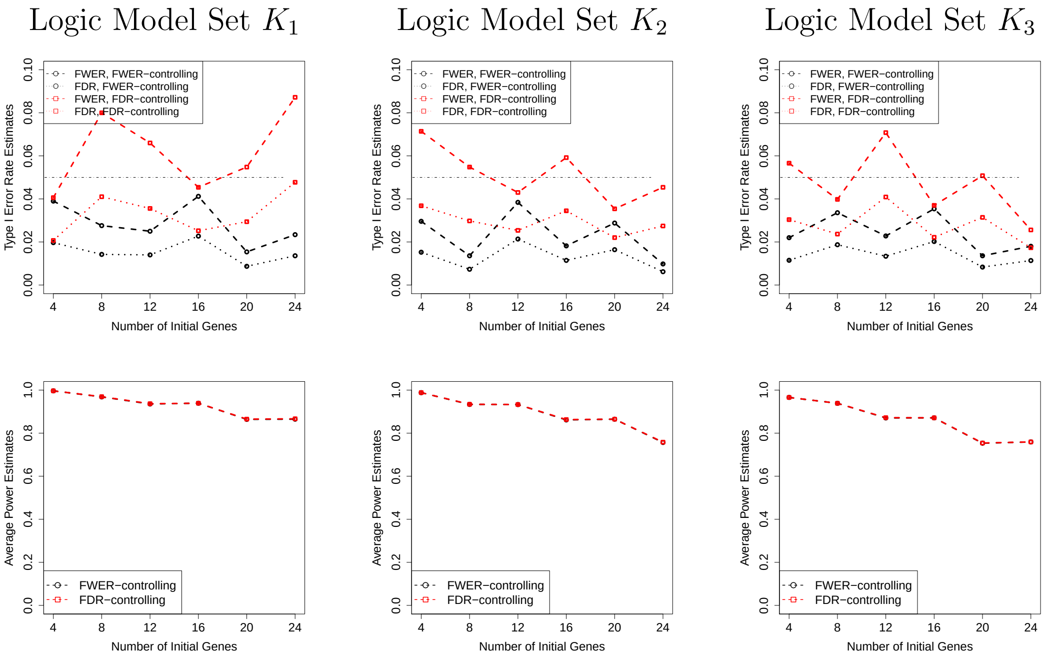


Figure 4.5: Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of initial number of genes under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\text{XOR}\}$; $K_2 = \{\text{AND}, \text{XOR}\}$; and $K_3 = \{\text{AND}, \text{XOR}, \bar{\text{X}}_1\bar{\text{X}}_2, \bar{\text{X}}_1 + \text{X}_2\}$, and predictive power $p = 0.85$. There is a single target to be predicted by two among a varying number of initial genes. Sample size is fixed at $n = 40$.

null hypotheses. Finally, it is again possible to see a decrease in power as less prior knowledge is available.

For the final group of experiments, we investigate how the number of targets to be tested can affect the FWER- and FDR-controlling procedures. We fix $n = 40$ and $G = 24$, and plot the results as a function of the number of targets D. The total number of tests varies from a minimum of 276 in the case of $D = 1$ and complete knowledge about the prediction logic to a maximum of 8832, for $D = 8$ and logic model set K_3. Note that here the number of false null hypotheses is D, whereas the number of true null hypotheses is obviously the total of tests in each case minus D. The results are displayed in Figure 4.6. The previous observations regarding the

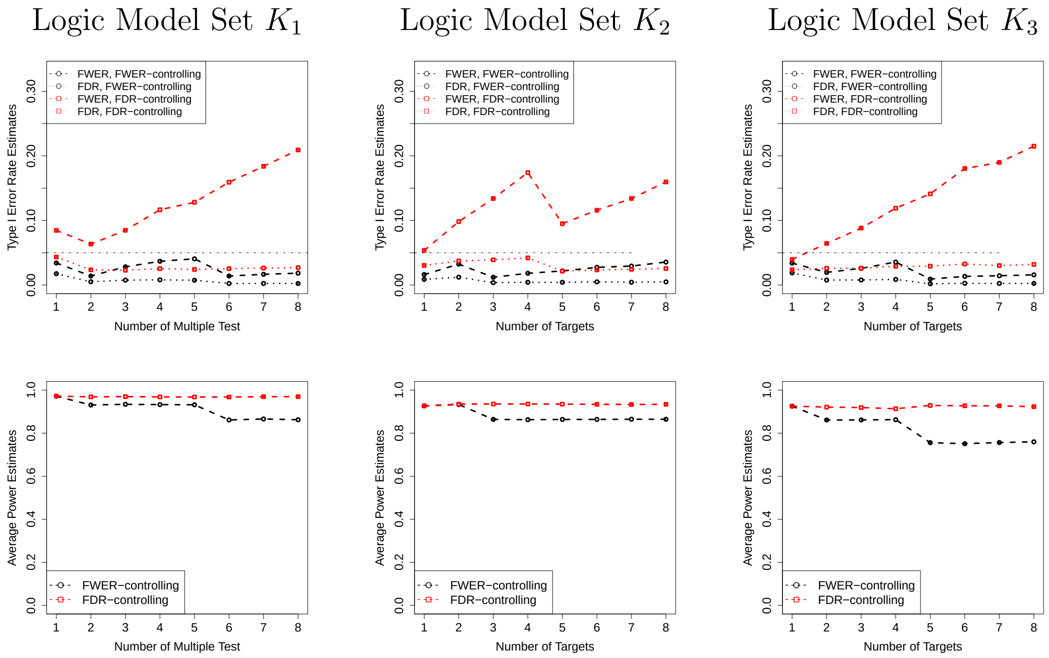


Figure 4.6: Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of number of targets under FWER- and FDR-controlling procedures, for three logic model sets, $K_1 = \{\mathrm{XOR}\}$; $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \bar{\mathrm{X}}_1\bar{\mathrm{X}}_2, \bar{\mathrm{X}}_1 + \mathrm{X}_2\}$, and predictive power $p = 0.85$. There is a varying number of targets D to be predicted by two among $G = 24$ genes. Sample size is fixed at $n = 40$.

FWER and FDR estimates are still valid in this case. As for power, however, we can observe a clear superiority of the FDR- over the FWER-controlling procedure. This is of course related to the presence of a larger number of true alternative hypotheses in this case. We can see that the power of the FDR-controlling procedure, besides being excellent, is also robust to the increase in number of tests, in contrast to the FWER-controlling procedure.

We have selected to run the previous two simulation experiments with $n = 40$, small sample settings, due to limited availability of sample gene-expression data in practice. To investigate the appropriateness of this choice, we have re-run these simulations with $n = 20$ and $n = 60$ —results are shown in Figures 1-4 in the supplementary material. We observed that the general conclusions from the $n = 40$ case were still valid. With a smaller sample size $n = 20$, the FDR-controlling procedure has a very clear superiority over the FWER-controlling one, as was already observed with $n = 40$. With $n = 60$, the performance of the FWER- and FDR-controlling procedures become very close due to the fact that larger sample size leads to stronger power.

The overall conclusion on the comparison between FWER- and FDR-controlling procedures is that in application with multiple targets, the FDR-controlling procedure is to be preferred due its superior power, whereas the FWER-controlling procedure is to be preferred in applications with very small number of targets since there is no appreciable difference in power, while the FWER and FDR rates are smaller.

4.4 Case Study: Genotoxic Stress Responsive Genes

In this section, we illustrate the application of the proposed multivariate Boolean detection methodology based on the CoD to real gene expression data, from a study on ionizing radiation (IR) responsive genes in [46]. This data set consists of 12 genes under 3 conditions (i.e., IR, MMS, UV) in 30 cell lines of both *p53* proficient and *p53* deficient cells. The data is ternary, indicating up-regulated (+1), down-regulated (-1), or no-change (0) status. Here we map this to binary expression using the following code: change (1), for either up-regulated or down-regulated genes, and no-change (0), as before. Additionally, we consider the three binary conditions (IR, MMS, and UV) as possible predictive factors, for a total of 15 Boolean variables in the data set.

4.4.1 Detection of Significant Regulatory Relationships

In the first group of experiments, we use the proposed approach to find significant regulatory relationships between two predictors and a target. We assume no prior knowledge, and thus make no constraints on the allowed regulatory relationships, other than a gene does not predict itself. Hence, all $\binom{14}{2}$ two-predictor sets and 10 possible "true" 2-predictor logic candidates are considered for each target, for a total of $\binom{14}{2} \times 10 = 910$ possible models; note that each gene can appear in multiple models, both as a member of different pairs and under different logic relationships. In addition, we consider each of the 12 genes in the data set as a possible target, so that the number of multiple tests performed is $M = \binom{14}{2} \times 10 \times 12 = 10,920$. We apply both the FWER- and the FDR-controlling procedures outlined in the previous section with a significance level $\alpha = 0.05$. Figure 4.7 displays the gene targets possessing significant predictors and the number of significant predictive relationships (out of the maximum of 910) detected, under each of the two approaches.

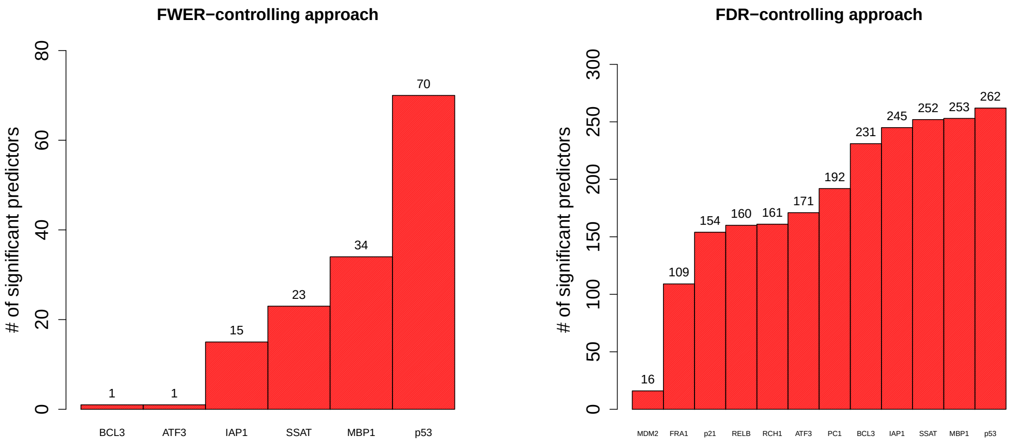


Figure 4.7: Significant predictive relationships detected in the IR-response stress gene-expression data of [46], under the FWER- and FDR-controlling approaches.

Table 4.3: Examples of detected relationships that are consistent with known biological groundtruth

Target	Pred. 1	Pred. 2	Controlling	Logic	Adjusted P-value
p53	p21	MDM2	FDR	OR	7.1025×10^{-3}
p21	MDM2	ATF3	FDR	OR	6.1272×10^{-4}
p21	MDM2	ATF3	FDR	$X_1 + \bar{X}_2$	3.9910×10^{-2}

Interestingly, *p53* turns out to possess the largest number of significant predictive relationships, under both approaches. This is in accordance with the known fact that *p53* is a significantly active gene involved in various pathways associated with stress responses. Notice that the FWER-controlling approach is more conservative and thus produces fewer significant predictive relationships than the FDR-controlling approach, for each of the targets. Table 4.3 provides examples of detected regulatory relationships that are consistent with well-known biological groundtruth. All of these relationships are detected under the FDR-controlling approach. As is known in the biological literature, p53 is found to be expressed when at least one of p21 and MDM2 is expressed, while p21 is found to be regulated in two ways: is is expressed when MDM2 is expressed or ATF3 is expressed, or when MDM2 is expressed or ATF3 is not expressed — the adjusted p-value for the former result is smaller than that for the latter, which may be evidence that the OR logic can provide a better model for this regulatory relationship. Table 4.4 lists top 20 significant regulatory relationships under FDR- and FWER-controlling approaches. These results could serve as candidate regulatory relationships for further experimental verification.

Notice that the adjusted p-values in Table 4.4 are identical for FWER- and FDR-controlling procedures, respectively. This is due to the discrete nature of the problem. For instance, considering the FWER-controlling procedure, all the 20 detections with their predicted logics share the same $k = 28$ in eq. (5.9) in form of the rejection

Table 4.4: A list of top 20 significant regulatory relationships detected in the IR-response stress gene-expression data of [46] under the FWER- and FDR-controlling approaches

Target	Pred. 1	Pred. 2	Logic	π' (FWER)	π'' (FDR)
MBP1	RCH1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	BCL3	IAP1	$\bar{X}_1 X_2$	3.153×10^{-4}	7.006×10^{-6}
MBP1	FRA1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	FRA1	SSAT	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	ATF3	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	SSAT	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	MDM2	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	p21	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	SSAT	MDM2	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	BCL3	MBP1	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	BCL3	p21	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	FRA1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	FRA1	MBP1	AND	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	BCL3	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	IAP1	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	MMS	NAND	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	UV	NAND	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	ATF3	$\bar{X}_1 + X_2$	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	IAP1	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	MBP1	NAND	3.153×10^{-4}	7.006×10^{-6}

region, which naturally leads to the same adjusted p-values according to eqs. (4.9) and (4.22).

4.4.2 Detection of Synthetic Target Genes

Following [47], we further examine the properties of the proposed methodology by generating 8 synthetic target genes, SYN1, SYN2, . . . SYN8, which are assumed to be predicted by two of 12 genes in the IR-response stress gene-expression data of [46]. Hence, each new data set consists of 23 genes (with 3 conditions included). The synthetic relationships are shown in Table 4.5, where the noise $N \sim \text{Bernoulli}(1-p)$. A total of $M = 100$ realizations are generated for the eight synthetic genes, based on the relationships in Table 4.5. As for the logic model set K, we consider three cases: (1) the logic is known, $K_1 = \{\text{XOR}\}$; (2)$K_2 = \{\text{XOR, AND, NAND}\}$; and $K_3 = \{10\ 2 - \text{predictorlogics}\}$. We assume here that a gene cannot predict itself. Hence, with the addition of the 8 synthetic target genes, the total number of multiple tests is $M_1 = \binom{22}{2} \times 1 \times 8 = 1848$ for the set K_1, $M_2 = \binom{22}{2} \times 3 \times 8 = 5,544$ for the set K_2, and $M_3 = \binom{22}{2} \times 10 \times 8 = 18,480$ for the set K_3. We apply the FWER- and FDR-controlling procedures with a significance level $\alpha = 0.05$. Figure 4.8 shows the power estimates as a function of the predictive power under each of the two procedures. It is observed that the FDR-controlling approach achieves larger power than the FWER-controlling one as expected. As the predictive power increases, the power increases to 1 for both approaches. When we have less prior knowledge about logic models, the power tends to be smaller.

Table 4.5: Synthetic Relationships based on the IR-response stress gene-expression data of [46]

Target	Synthetic Relationship
1	SYN1 = PC1 XOR MDM2 $\oplus$ N
2	SYN2 = IAP1 XOR SSAT $\oplus$ N
3	SYN3 = PC1 XOR MMS $\oplus$ N
4	SYN4 = ATF3 XOR p53 $\oplus$ N
5	SYN5 = RCH1 XOR FRA1 $\oplus$ N
6	SYN6 = RELB XOR MMS $\oplus$ N
7	SYN7 = p53 XOR IR $\oplus$ N
8	SYN8 = BCL3 XOR IAP1 $\oplus$ N

4.5 Summary

We have described in this paper a rigorous statistical testing framework to investigate regulatory relationships among genes, by using the discrete CoD. This marks

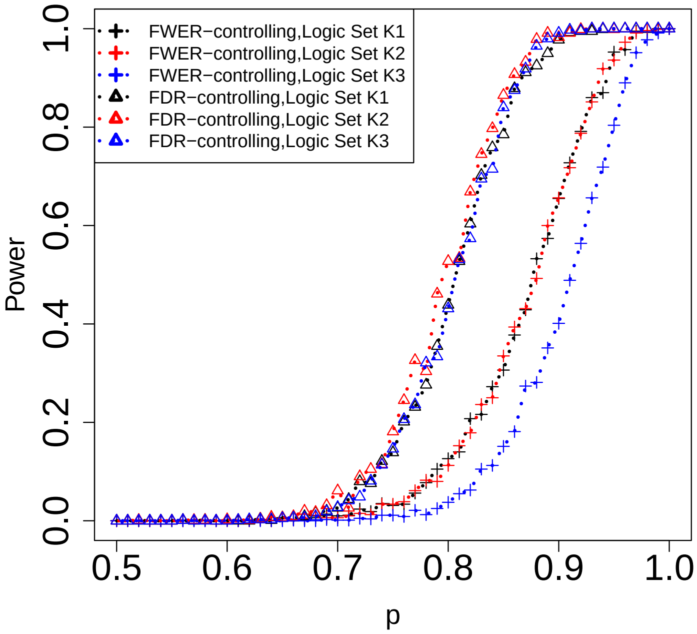


Figure 4.8: Power estimates as a function of predictive power for 8 synthetic targets using both FWER- and FDR-controlling procedures, for three logic candidate sets, $K_1 = \{\text{XOR}\}$, $K_2 = \{\text{XOR, AND, NAND}\}$ and $K_3 = \{10 \text{ meaningful logics}\}$.

a significant change in the application of the CoD to such problems, since thus far its use depended on user-selected thresholds to characterize the presence of significant relationships. Multiple-testing procedures are also described, which make the methodology applicable to large data sets. Furthermore, software that implements the COD test is made available to the scientific community as an R *codtest* package through our website (http://gsp.tamu.edu/Publications/ supplementary/ting13a). It is expected that this methodology will be a useful practical tool for the inference of gene regulatory relationships and networks from gene-expression data.

5. STATISTICAL DETECTION OF INTRINSICALLY MULTIVARIATE PREDICTIVE GENES*

Canalization, i.e. buffering or robustness, of genotypes plays an important role in the developmental processes of organisms, which suppresses phenotypic variation. Back in 1942, Waddington proposed the existence of canalizing genes that can constrain a biological system to acquired characters in the face of environmental stimuli [69]. Canalizing genes make adaptive and optimal reactions to environmental perturbations, and can produce reliable developmental effects against genetic mutations or environmental changes during evolution [49, 70]. In one word, canalization preserves biological systems with characteristics born from natural selection. However, this significant property of biological systems during the course of evolution is not well understood and verified since then. Until recently, Lehner has studied global quantitative gene datasets in yeast to investigate Waddington's intuition, and confirmed that canalizing genes, also known as "hub" genes, present similar robustness when faced with environmental, stochastic and genetic perturbations [49].

Canalizing genes are frequently found in signalling pathways, which deliver information from a variety of sources to the machinery that enacts central cellular functions such as cell-cycle, survival, apoptosis and metabolism [52]. For example, DUSP1 antagonizes the activity of the p38 mitogen activated kinase, MAPK1 (ERK), which is known to be a central component that assists extracellular signal-regulated kinases to send mitogenic signals [15]. Hence, the gene DUSP1 canalyzes when it dephosphorylates MAPK1. DUSP1 provides a complicated transcriptional

*Parts of this section are reprinted with permission from "A Statistical Test for Intrinsically Multivariate Predictive Genes" by T. Chen and U.M. Braga-Neto, 2012, *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2012), Washington, DC, December 2012*, pp. 151–154, © 2012 IEEE.

mechanism for dephosphorylating MAPK1, and the expression of DUSP1 is induced strongly by growth factors and cellular stresses [13, 56]. Since the function of DUSP1 might lead to abnormal MAPK1 signalling, this will have negative impact both on processes like proliferation and apoptosis critical to the development of human cancer and on the active response of tumour cells to conventional cancer therapies [45, 64]. Canalyzing behavior is often observed in signal transducing pathways. For instance, canalyzation was associated with the behavior of RAS gene family in the mitogenic pathway [67]. In addition, the p53 (TP53) gene is also well known to be a canalyzing gene for signal integration under stresses, which exerts strong control with cellular stress responses [39].

Martins and collaborators [52] defined the concept of *intrinsically multivariate prediction*, in which case when the controlling gene is active, it cannot be well-predicted by subsets of its predictor genes, but it can be predicted by the full set with great accuracy. Such a set of predictor genes is called Intrinsically Multivariate Predictive (IMP) set for the target gene [52]. The IMP characterizes the property of a canalyzing gene that it can be able to exert overriding control. Based on the notion of IMP, they proposed a very nice mathematical expression of IMP in the context of the binary Coefficient of Determination (CoD), the IMP score being used to measure how closely a series of slave genes coordinate with their master gene [31]. As such, IMP depends on the probability model connecting one controlling gene and its slave genes, which, however, is usually unknown, or only partially known in practice.

Their work showed that DUSP1 had the largest number of IMP gene sets in related pathways, thereby providing evidence that the IMP criterion could be used as a practical tool for discovery of canalyzing genes [52]. However, applications of the IMP criterion so far have been based on user-selected thresholds to decide on the presence of gene multivariate prediction between the given predictor and target genes. In this chapter, we describe a multiple testing framework for the detection of significant intrinsically multivariate predictive genes, by providing a statistical test for a nonzero IMP score between given a Boolean target and Boolean predictors [18, 24]. Our proposed multiple testing procedures are validated by using both synthetic and real data sets.

5.1 Intrinsically Multivariate Prediction

We first review the concept of intrinsically multivariate prediction in the context of the CoD, based on a proposed stochastic logic model [16] that mimics the behavior of stochastic biological systems in practice.

The concept of intrinsically multivariate prediction (IMP) was first introduced by [52] for the investigation of canalyzing genes. A predictor set $\mathbf{X}$ is said to be intrinsically multivariate predictive (IMP) of the target Y if $\mathbf{X}$ predicts Y accurately, but Y cannot be predicted accurately by any subset of $\mathbf{X}$. Mathematically, this can be expressed by the *IMP score* of the pair $(\mathbf{X}, Y)$ [52]

$$\mathrm{IMP_Y}(\mathbf{X}) = \mathrm{CoD_Y}(\mathbf{X}) - \max_{\mathbf{Z} \subsetneq \mathbf{X}} \mathrm{CoD_Y}(\mathbf{Z}), \tag{5.1}$$

where $\mathbf{Z} \neq \emptyset$. In the two-predictor case, the IMP score is given by

$$\mathrm{IMP_Y}(\mathrm{X_1}, \mathrm{X_2}) = \mathrm{CoD_Y}(\mathrm{X_1}, \mathrm{X_2}) - \max_{\mathrm{i}=1,2} \mathrm{CoD_Y}(\mathrm{X_i})\,. \tag{5.2}$$

Clearly, $\mathrm{IMP_Y}(\mathbf{X}) = 0$ implies that $\mathbf{X}$ is not IMP of Y. The larger the IMP score is, the stronger the IMP effect is. Note that, since $\mathbf{Z}$ cannot be either $\emptyset$ or full set $\mathbf{X}$, there are totally $2^d - 2$ subsets $\mathbf{Z}$'s in the set $\mathbf{X}$.

What of our interest is the case of IMP $= 0$ (i.e., no imp affect). This can furthermore be written into the equivalent statement via the definition of IMP, that is, $\varepsilon_Y(\mathbf{X}) = \min_{\mathbf{Z} \subsetneqq \mathbf{X}} \varepsilon_Y(\mathbf{Z})$ by assuming that $\varepsilon_Y \neq 0$. Since predictor $\mathbf{X}$ is the perfect predictor of target Y, $\varepsilon_Y(\mathbf{Z}) \geq \varepsilon_Y(\mathbf{X})$, for any $\mathbf{Z} \subsetneqq \mathbf{X}$. Suppose the predictive power of $\mathbf{X}$ over Y is p, and then we have the optimal error $\varepsilon_Y(\mathbf{X}) = 1 - p$. Hence, if $\varepsilon_Y(\mathbf{T}) = \varepsilon_Y(\mathbf{X}) = 1 - p$ for some $\mathbf{T} \subsetneqq \mathbf{X}$, then $\varepsilon_Y(\mathbf{T})$ is clearly the minimum of $\varepsilon_Y(\mathbf{Z})$ for all $\mathbf{Z} \subsetneqq \mathbf{X}$. Let $\mathcal{V}(\mathbf{X}) := \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{2^d-2}\} = \mathcal{P}(\mathbf{X}) \backslash \{\{\emptyset\}, \{\mathbf{X}\}\}$, that is, the power set of $\mathbf{X}$ excluding empty set and $\mathbf{X}$ itself. We give next a result that relates $\mathrm{IMP_Y}\mathbf{X} = 0$ (i.e., no IMP) with parameter p and the joint probability distribution of predictor $\mathbf{X}$ under the d-predictor logic model (3.12).

Proposition 3. *Under a d-predictor stochastic logic model, the following statements are equivalent:*

(i) $\mathrm{IMP_Y}(\mathbf{X}) = 0$.

(ii) $p = 1/2$ *or any of statement* W_i $(i = 1, 2, \ldots, 2^d - 2)$ *works, where*

$$W_i = \left\{ \sum_{\mathbf{x}_i^{(1)} \in \{0,1\}^{|\mathbf{x}_i^{(1)}|}} P\left(\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}, \mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}\right) \mathbf{1}(f(\mathbf{x}) = 1) = 0 \text{ or } \sum_{\mathbf{x}_i^{(1)} \in \{0,1\}^{|\mathbf{x}_i^{(1)}|}} P\left(\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}, \mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}\right) \mathbf{1}(f(\mathbf{x}) = 0) = 0, \text{ for all } \mathbf{x}_i^{(2)} \in \{0,1\}^{|\mathbf{x}_i^{(2)}|} \,|\, \mathbf{X}_i^{(2)} = \mathbf{V}_i \in \mathcal{V}(\mathbf{X}) \right\}, \tag{5.3}$$

for $i = 1, 2, \ldots, 2^d - 2$ *and* $\mathbf{X} = \mathbf{X}_i^{(1)} \cup \mathbf{X}_i^{(2)}$.

Proof. See Appendix E. Q.E.D.

It is easy to check that a logic f and its negated logic share the same W_i, for $i = 1, \ldots, 2^d - 2$. It should be noted that, for some fixed $\mathbf{X}_i^{(2)}$, the corresponding statement W_i contains $s_i = 2^{2^{|\mathbf{x}_i^{(2)}|}}$ sub-statements, namely, $\{W_i^{(l)}\}_{l=1}^{s_i}$. When statement W_i holds, $W_i^{(1)}$ works, $\ldots$, or $W_i^{(s_i)}$ works. Furthermore, IMP $= 0$ is equivalent to the statement that at least one of $W_i^{(l)}(l = 1, \ldots, s_i, i = 1, \ldots, 2^d - 2)$ holds, where

$$W_i^{(l)} = \left\{ \sum_{\mathbf{x}_i^{(1)} \in \{0,1\}^{|\mathbf{x}_i^{(1)}|}} P(\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}, \mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}) \times \right. \\ \left. \mathbf{1}(f(\mathbf{x}) = z) = 0, \text{ for all } z \in \mathbf{a}_l \right\} , \tag{5.4}$$

where $\mathbf{a}_l$ is the l-th element of $2^{|\mathbf{x}_i^{(2)}|}$-ary Cartesian product over $2^{|\mathbf{x}_i^{(2)}|}$ equivalent sets of $\{0, 1\}$.

For conciseness, we further explain IMP $= 0$ with an equivalent expression by eliminating repeated results from all $W_i^{(l)}$'s, which is formulated by

$$\begin{aligned} &\sum_{\mathbf{x} \in \mathcal{D}_i} P(\mathbf{X} = \mathbf{x}) = 0, \ldots, \ or \sum_{\mathbf{x} \in \mathcal{D}_{d^*}} P(\mathbf{X} = \mathbf{x}) = 0, \ or \\ &\sum_{\mathbf{x} \in \overline{\mathcal{D}}_i} P(\mathbf{X} = \mathbf{x}) = 0, \ldots, \ or \sum_{\mathbf{x} \in \overline{\mathcal{D}}_{d^*}} P(\mathbf{X} = \mathbf{x}) = 0 \, , \end{aligned} \tag{5.5}$$

where $\overline{\mathcal{D}}_i$ is the complementary set of $\mathcal{D}_i$, for $i = 1, \ldots, d^*$. For example, under a 2-predictor stochastic AND logic model, we have $D_1, D_2, \ldots, D_3$ expressed by

$$\mathcal{D}_1 = \{(0, 1)\}, \ \mathcal{D}_2 = \{(1, 0)\}, \ \mathcal{D}_3 = \{(1, 1)\}. \tag{5.6}$$

Thus, we give next a proposition relating $\mathrm{IMP}_{\mathrm{Y}}(\mathbf{X}) = 0$ with the model information in the model (3.12) in the 2-predictor case.

Proposition 4. *Given a 2-input stochastic logic model, the following statements are equivalent:*

(i) $\mathrm{IMP}_{\mathrm{Y}}(\mathbf{X}) = 0$ *.*

(ii) $p = 1/2$ *or* $\sum_{\mathbf{x}\in\mathcal{D}_i} P(\mathbf{X}=\mathbf{x}) = 0$*, or* $\sum_{\mathbf{x}\in\overline{\mathcal{D}}_i} P(\mathbf{X}=\mathbf{x}) = 1$*, for* $i = 1, 2, ..., d^*$*, where* $\mathcal{D}_i$*'s regarding all 10 meaningful logics are shown in Table 5.1.*

Table 5.1: $\mathcal{D}_i$'s in $\mathrm{IMP}_{\mathrm{Y}}(\mathbf{X}) = 0$ for 2-Input Logics

logic	bit string	constraint
AND / NAND	0001 / 1110	$\{(0,1)\}$, $\{(1,0)\}$, $\{(1,1)\}$
XOR / NXOR	0110 / 1001	$\{(0,0),(1,1)\}$, $\{(0,0),(0,1)\}$, $\{(0,0),(1,0)\}$
$\overline{X}_1X_2$ / $X_1+\overline{X}_2$	0100 / 1011	$\{(0,1)\}$, $\{(1,0)\}$, $\{(1,1)\}$
$X_1\overline{X}_2$ / $\overline{X}_1+X_2$	0010 / 1101	$\{(0,0)\}$, $\{(0,1)\}$, $\{(1,0)\}$
OR / NOR	0111 / 1000	$\{(0,0)\}$, $\{(0,1)\}$, $\{(1,0)\}$

. It is easy to check that CoD = 0 implies that IMP = 0. For example, if the logic f is AND, then $\mathrm{IMP}_{\mathrm{Y}}(\mathbf{X}) = 0$ if and only if $p = 1/2$ or $P(0,1) = 0$ or 1 or $P(1,0) = 0$ or 1 or $P(1,1) = 0$ or 1, from which it follows that $\mathrm{CoD}_Y(\mathbf{X}) = 0 \rightarrow \mathrm{IMP}_{\mathrm{Y}}(\mathbf{X}) = 0$ (it can be shown that this is a general fact). Notice that only 10 out of $2^4 = 16$ possible logics are shown in Table 5.1, since the remaining 6 logics are either constant or depend on only one of the predictors. Note also that there are only 5 rows in Table 5.1, since a logic f and its negated logic share the same $\mathcal{D}$, as they share the same expression for their full CoD with respect to the full set $\mathbf{X}$ and individual CoD's with respect to any subset of $\mathbf{X}$.

For conciseness, we will denote in the sequel $\mathrm{CoD}_Y(\mathbf{X})$ and $\mathrm{IMP}_{\mathrm{Y}}(\mathbf{X})$ by CoD and IMP, respectively.

5.2 IMP Hypothesis Test

The IMP is a function of the logic f, the distribution parameters p and the joint probability distribution of $\mathbf{X}$, and therefore statements about it can be statistically tested based on an i.i.d. sample. Following the CoD hypthesis test in [17], we are particularly interested in the following statistical hypothesis problem:

$$\begin{aligned} &\mathrm{H_0} : \mathrm{IMP} = 0 \ \ (\mathrm{p} = 1/2 \ \text{ or } \ \lambda_1 = 0 \text{ or } 1, \text{or} \ldots \ \ \lambda_{\mathrm{d}*} = 0 \text{ or } 1) \\ &\mathrm{H_1} : \mathrm{IMP} > 0 \ \ (\mathrm{p} > 1/2 \ \text{ and } \ 0 < \lambda_1 < 1 \ \text{ and} \ldots \ \ 0 < \lambda_{\mathrm{d}*} < 1)\,, \end{aligned} \tag{5.7}$$

where $\lambda_i = \sum_{\mathbf{x} \in \mathcal{D}_i} P(\mathbf{X} = \mathbf{x})$. The null hypothesis $\mathrm{H_0}$ indicates the absence of IMP affect of $\mathbf{X}$ concerning the target Y, whereas the alternative hypothesis $\mathrm{H_1}$ states that there is a degree of IMP effect between them.

This is a composite, multiparameter hypothesis testing problem. As the null parameter space is a union of $2d^* + 1$ subsets $[p = 1/2]$, $[\lambda_1 = 0]$, $[\lambda_1 = 1]$,. . ., $[\lambda_{d^*} = 0]$, and $[\lambda_{d^*} = 1]$, the appropriate strategy to employ here is the *intersection-union test* (IUT) method; the individual tests for $p = 1/2$, $\lambda_1 = 0$, $\lambda_1 = 1$, . . ., $\lambda_{d^*} = 0$ and $\lambda_{d^*} = 1$ are level-α likelihood-ratio tests (LRTs), leading to an overall level-α IUT test [5,6]. It is proven that, when using the IUT method, the composite test is a level α test if each test divided from the composite test is a level α test [5,6]. Let $\mathbf{S}_n = \{X_{i1}, \ldots, X_{id}, Y_i\}_{i=1}^n$ be a random vector of i.i.d. observations whose distribution follows the stochastic model in (3.12). One observation, or sample, of $\mathbf{S}_n$ is denoted by $\mathbf{s}_n$. For simplicity we will introduce the notation $\mathbf{X}_i$ to denote the random vector with the components $X_{i1}, \ldots, X_{id}$, and $\mathbf{x}_i$ is one observation of this random vector. This is summarized in the following proposition for the 2-predictor case. Details can be found in the supplementary information.

Proposition 5. *Under a 2-predictor stochastic logic model, a level α IUT test of H_0 : $\mathrm{IMP} = 0$ vs. $\mathrm{H_1}$: $\mathrm{IMP} > 0$ is given by $\Phi = \mathbf{1}(\mathbf{s}_n \in \mathcal{R})$, where $\mathcal{R} = \mathcal{R}_1 \cap$*

$\mathcal{R}_{21} \cdots \cap \mathcal{R}_{2d^*} \cap \mathcal{R}_{31} \cdots \cap \mathcal{R}_{3d^*}$ *with* $(j = 1, 2, ..., d^*)$

$$\begin{aligned}
\mathcal{R}_1 &= \{\mathbf{s}_n \mid \sum_{i=1}^{n} \mathbf{1}(f(\mathbf{x}_i) = y_i) \geq K\}, \\
\mathcal{R}_{2j} &= \{\mathbf{s}_n \mid \mathbf{x}_i \in \mathcal{D}_j \text{ for some } i \in \{1, \ldots, n\}\} \\
\mathcal{R}_{3j} &= \{\mathbf{s}_n \mid \mathbf{x}_i \in \overline{\mathcal{D}}_j \text{ for some } i \in \{1, \ldots, n\}\},
\end{aligned} \tag{5.8}$$

where D_j $(j = 1, \ldots, d^*)$ *is formed in Table 5.1,* $\overline{\mathcal{D}}_j$ *is a complementary set of* $\mathcal{D}_j$*, for* $j = 1, \ldots, d^*$*, and* k *is the* $100(1 - \alpha)\%$ *percentile of a Binomial(n,1/2) distribution, i.e.,* k *is the smallest integer such that*

$$\sum_{t \geq k} \binom{n}{t} \left(\frac{1}{2}\right)^n \leq \alpha\ , \tag{5.9}$$

is a level-α *test for (5.7).*

Proof. See Appendix F. Q.E.D.

The following statements are made from Proposition 5.

(1) **Rejection region.** Notice that $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_{21} \cdots \cap \mathcal{R}_{2d^*} \cap \mathcal{R}_{31} \cdots \cap \mathcal{R}_{3d^*}$, where $\mathcal{R}_1 = \{\mathbf{s}_n \mid \sum_{i=1}^{n} I(f(\mathbf{x}_i) = y_i) \geq k\}$ is the rejection region for the $[p = 1/2]$ LRT, and expresses how tightly the data follows the proposed model, while $\mathcal{R}_{2i} = \{\mathbf{s}_n \mid \mathbf{x}_i \in \mathcal{D}_j \text{ for some } i \in \{1, \ldots, n\}\}$ is the rejection region for the $[\sum_{\mathbf{x} \in \mathcal{D}_i} P(\mathbf{X} = \mathbf{x}) = 0]$ LRT, and $\mathcal{R}_{3i} = \{\mathbf{s}_n \mid \mathbf{x}_i \in \overline{\mathcal{D}}_j \text{ for some } i \in \{1, \ldots, n\}\}$ is the rejection region for the $[\sum_{\mathbf{x} \in \mathcal{D}_i} P(\mathbf{X} = \mathbf{x}) = 1]$ LRT, for $i = 1, \ldots, d$, and indicates that the null hypothesis cannot be rejected if these constraints on the sample of predicor $\mathbf{X}$ are not satisfied. For the simplification of the following formulation, let $\tilde{\mathcal{R}} = \mathcal{R}_{21} \cdots \cap \mathcal{R}_{2d^*} \cap \mathcal{R}_{31} \cdots \cap \mathcal{R}_{3d^*} = \tilde{\mathcal{R}}_1 \cap \tilde{\mathcal{R}}_2 \cdots \cap \tilde{\mathcal{R}}_{2d^*}$ and $(\tilde{\mathcal{D}}_1, \ldots, \tilde{\mathcal{D}}_{2d^*}) = (\mathcal{D}_1, \ldots, \mathcal{D}_{d^*}, \overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_{2d^*})$. Notice

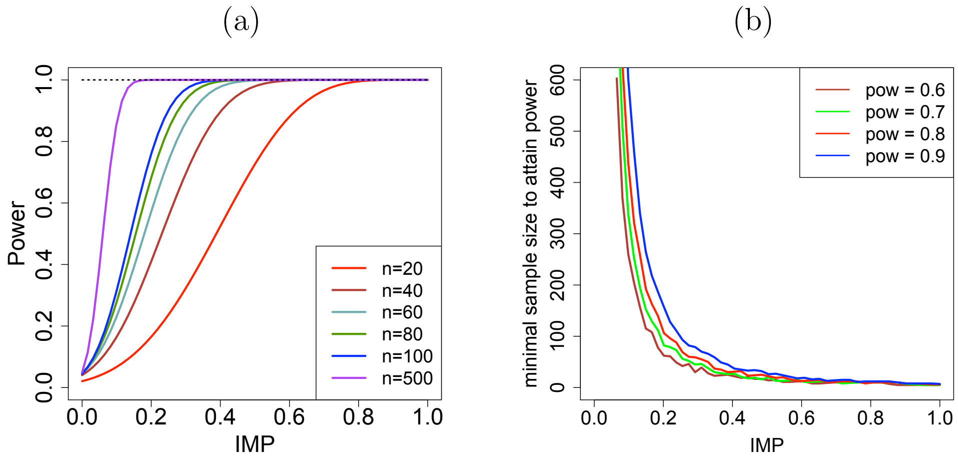


Figure 5.1: Relationship among power function, sample size and the IMP value given a 2-input stochastic XOR model with $P_1 = P_2 = 0.05, \gamma = 0.05$, for the proposed IMP test with $\alpha = 0.05$. (a) Statistical power vs. IMP over varying sample size n. (b) Minimum sample size to achieve varying power vs. IMP.

that

$$
\begin{aligned}
&P_{\tilde{\mathcal{R}}} = P(\mathbf{S}_n \in \tilde{\mathcal{R}}) = 1 - P\left(\tilde{\mathcal{R}}_1^c \cup \cdots \cup \tilde{\mathcal{R}}_{2d^*}^c\right) = 1 - \sum_{i=1}^{2d^*} P(\tilde{\mathcal{R}}_i^c) \\
&+ \sum_{1\le i<j\le 2d^*} P(\tilde{\mathcal{R}}_i^c \cap \tilde{\mathcal{R}}_j^c) - \cdots + (-1)^{2d^*} P(\tilde{\mathcal{R}}_1^c \cap \cdots \cap \tilde{\mathcal{R}}_{2d^*}^c) = \\
&1 - \sum_{i=1}^{2d^*} \left(\sum_{\mathbf{x}\in\tilde{\mathcal{D}}_i^c} P(\mathbf{X} = \mathbf{x}) \right)^n + \sum_{1\le i<j\le 2d^*} \left(\sum_{\mathbf{x}\in\tilde{\mathcal{D}}_i^c \cap \tilde{\mathcal{D}}_j^c} P(\mathbf{X} = \mathbf{x}) \right)^n \qquad (5.10) \\
&- \ldots + (-1)^{2d^*} \left(\sum_{\mathbf{x}\in\tilde{\mathcal{D}}_1^c \cap\cdots\cap \tilde{\mathcal{D}}_{2d^*}^c} \right)^n .
\end{aligned}
$$

When the joint probability of predictors satisfies eq. (5.5), $P_{\tilde{\mathcal{R}}}$ is always zero; Otherwise, $P_{\tilde{\mathcal{R}}} \to 0$ as $n \to \infty$.

(2) **p-value.** The rejection regions for varying significance level α is *nested*, that is, $\mathcal{R}_n(\alpha_1) \subseteq \mathcal{R}_n(\alpha_2)$, whenever $\alpha_1 \le \alpha_2$.. This allows us to define a p-value for the

proposed test as

$$\pi(\mathbf{s}_n) \;=\; \begin{cases} \displaystyle\sum_{t \geq \sum_{i=1}^{n} \mathbf{1}(f(\mathbf{x}_i)=y_i)} \binom{n}{t} \left(\frac{1}{2}\right)^n, & \text{for} \quad \mathbf{s}_n \in \mathcal{R}_2 \\ 1. & \text{otherwise} \end{cases} \tag{5.11}$$

(3) **Statistical power.** The power function [14] of the proposed test can be shown to be

$$\beta(p, f, P(\mathbf{X})) \;=\; P(\mathbf{S}_n \in \mathcal{R}) \;=\; \left(\sum_{t=k}^{n} \binom{n}{t} p^t (1-p)^{n-t} \right) \;\times\; P_{\mathcal{R}_{23}}, \tag{5.12}$$

for $p \geq 1/2$, where k us given by (5.9). For instance, we can see in Fig. 5.1(a), that for large sample size, power increases to 1 very rapidly. In addition, power increases monotonically with increasing IMP. Fig. 5.1(b) displays the minimum sample size necessary to achieve varying standard power value. As expected, the larger the power value is, the less sample size is needed for a fixed IMP effect size. We may summerize that, if a small IMP effect size is expected, then sample sizes in the neighborhood of $n = 100$ or larger are required for effective statistical power.

5.3 Multiple Testing Procedures

For a given target Y, the proposed test for IMP effect presupposes the model (3.12), which in turn depends on the choice of logic function f and predictor vector $\mathbf{X}$. Assuming dimensionality d (that is, d-predictor per target) and a number of genes G in the original gene-expression dataset, the total number of possible logic functions is 2^d and the number of distinct predictors is $\binom{G}{d}$. This creates a multiple testing issue with the total number of tests to be carried out being, in this case, $M = 2^d \times \binom{G}{d}$. In typical gene-expression microarray or RNA-seq studies, G tends to be very large (in the order of thousands or more). Therefore, even if d is kept small, the number of tests may be very large indeed, which may lead to no rejections of the null hypotheses (and no significant results can be concluded). In this section, we address the multiple testing problem in the context of the proposed detection method. We also comment on how to reduce the number of tests by use of prior knowledge.

In a multiple testing procedure (MTP), there is a total number of M null hypotheses to be simultaneously tested, $\{H_0(m) \mid m = 1, \dots, M\}$. The basic rationale behind a MTP is that, when there are $M \geq 1$ parallel null hypotheses, we need to provide rejection regions for each null hypothesis $H_0(m)(m = 1, \dots, M)$, and then to decide which of the M hypotheses should be rejected with a controlled Type-I error rate.

We recall that, for a given $0 < \alpha < 1$, an MTP is said to control the FWER at level α if FWER $\leq \alpha$. Similarly, an MTP is said to control the FDR at level α if FDR $\leq \alpha$. Notice that, since FDR $\leq$ FWER, any FWER-controlling procedure is also FDR-controlling, but the converse is not true in general, unless all null hypotheses are true, in which case FDR = FWER, as mentioned previously.

Formally, we develop in the following a statistical multiple testing framework for the identification of significant IMP pairs of predictors and targets. Suppose that a given target may be predicted by a predictor set $\mathbf{X}^i$ among a possible number L of predictor sets. Suppose that only partial knowledge about the logical regulations is known, that is, a number K_i of candidate logics for each predictor set $\mathbf{X}_j$, and the total number of tests to be performed is therefore $M = \sum_{i=1}^{L} K_i$. Given a significance level α, the significant IMP gene sets for the target can be found by the following procedures:

(1) We compute the unadjusted p-values $\{\pi(1), \dots, \pi(M)\}$ for tests $H_0(m)$: IMP $= 0$ vs. $\mathrm{H_1(m)}$: IMP > 0, for $m = 1, \dots, M$.

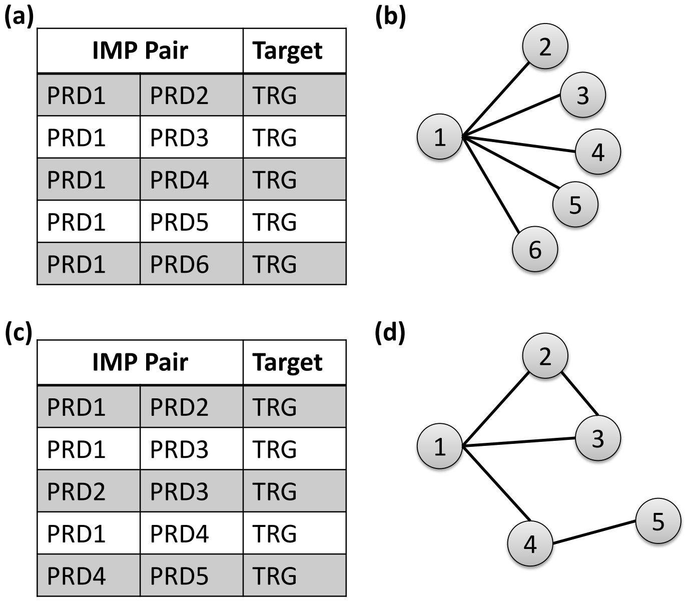
(2-a) **FWER-controlling approach:** Reject those null hypotheses $H_0(m)$ such that the corresponding adjusted p-value $\pi'(m) \leq \alpha$, and the corresponding predictor sets are regarded as the significant IMP sets of the given target.

(2-b) **FDR-controlling approach:** Reject those null hypotheses $H_0(m)$ such that the corresponding adjusted p-value $\pi''(m) \leq \alpha$. This can be realized in an equivalent way: order the unadjusted p-values to obtain the vector $\{\pi^*(1), \ldots, \pi^*(M)\}$ such that $\pi^*(1) \leq \pi^*(2) \leq \ldots \leq \pi^*(M)$. Let m^* be the largest m such that $\pi^*(m) \leq \frac{m}{M}\alpha$. Then reject the null hypotheses $H^*(m)(m = 1, \ldots, m^*)$ associated with the p-vlaues $\{\pi^*(1), \ldots, \pi^*(m^*)\}$. Hence, the corresponding predictor sets are IMP sets of statistical significance.

Note that, given D multiple targets, two approaches, depending on the largeness of M, can be employed with the above procedures. Suppose that, each target shares the same lists of candidate predictor sets and candidate logic sets. The proposed multiple testing procedures can be applied to mulitple D targets in parallel tests with the numbe of tests $M = \sum_{i=1}^{L} k_i \times D$ to be performed being reasonably large. Otherwise, the proposed procedures are used for each target (with $M = \sum_{i=1}^{L} k_i$), respectively. Details will be discussed in the applications to real data sets in the next Section.

5.4 Results and Discussion

In this section, we illustrate the application of the proposed multivariate Boolean detection methodology based on the IMP in a number of experiments using both synthetic and real data. The performance of effective recovery of canalyzing genes is investigated.



(a)

IMP Pair		Target
PRD1	PRD2	TRG
PRD1	PRD3	TRG
PRD1	PRD4	TRG
PRD1	PRD5	TRG
PRD1	PRD6	TRG

(c)

IMP Pair		Target
PRD1	PRD2	TRG
PRD1	PRD3	TRG
PRD2	PRD3	TRG
PRD1	PRD4	TRG
PRD4	PRD5	TRG

Figure 5.2: Examples of IMP pairs for a target and graphs of IMP pairs. (a) An example of 5 IMP pairs out of 6 predictor genes for one target; (b) Graph for IMP pairs in (a) withouth cycle (A line means the two connected genes function as an IMP pair for the target); (c) An example of 5 IMP pairs out of 5 predictor genes for one target ; (d) Graph for IMP pairs in (c) with cycle.

5.4.1 Synthetic Data

Consider a target gene TRG as variable Y and a set of G predictors PRD1, ..., PRDG as variable vector $X_1, \ldots, X_G$. We assume that the target TRG is predicted by T IMP pairs with corresponding logic functions LGC1, ..., LGCT. Suppose that the T IMP pairs include U unrepeated predictors PRD1, ..., PRDU, and then the remaining $G - U$ predictors can be regarded regarded as "noises" that do not influence the expression of TRG. Fig. 5.2 gives an example of TRGT and its IMP pairs. For example, TRG (Y) is regulated by the IMP pair (PRD1(X_1), PRD2(X_2)) through a stochastic XOR logic function, such that $Y = X_1 \text{XOR} X_2 \oplus N$, where $N \sim$ Bernoulli$(1 - p)$, for $1/2 \leq p \leq 1$, as before. We generate i.i.d. sample data of size n for TRG, PRD1, ..., PRDG with following steps.

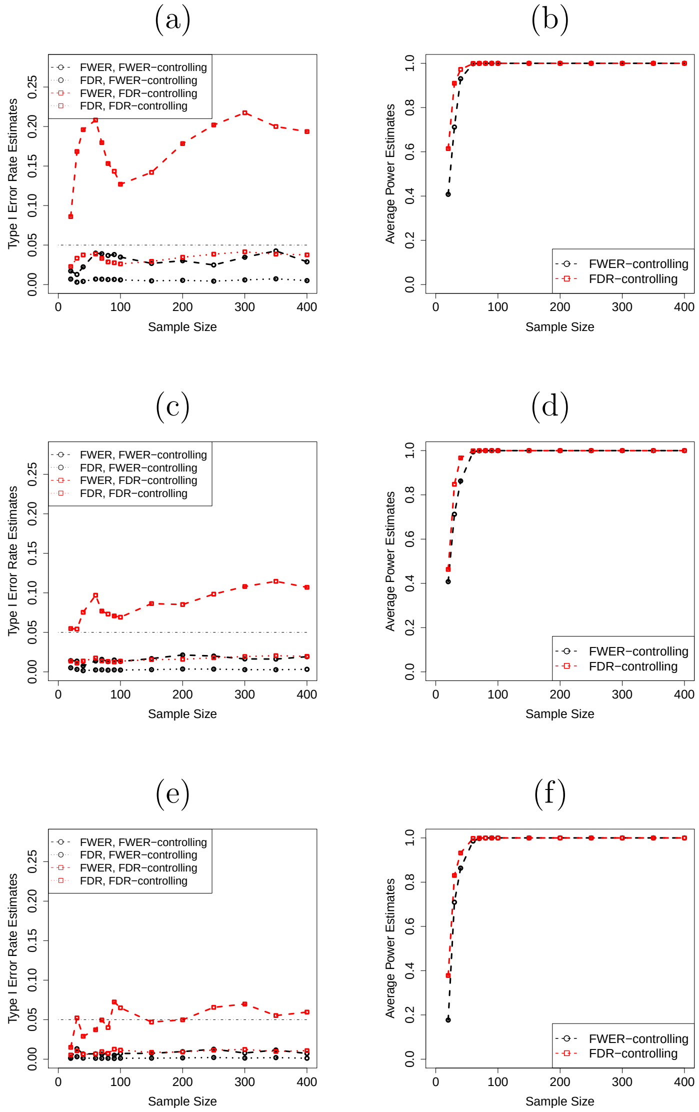


Figure 5.3: Average FWER and FDR estimates (left column) and power estimates (right column) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets and preditive power $p = 0.85$. (a) FWER and FDR estimates for model set $K_1 = \{\text{AND}\}$. (b) Power estimates for model set $K_1 = \{\text{AND}\}$. (c) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (d) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (e) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. (f) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. There is a single target to be predicted by 5 IMP pairs (as shown in Figs. 5.2) among $G = 24$ genes.

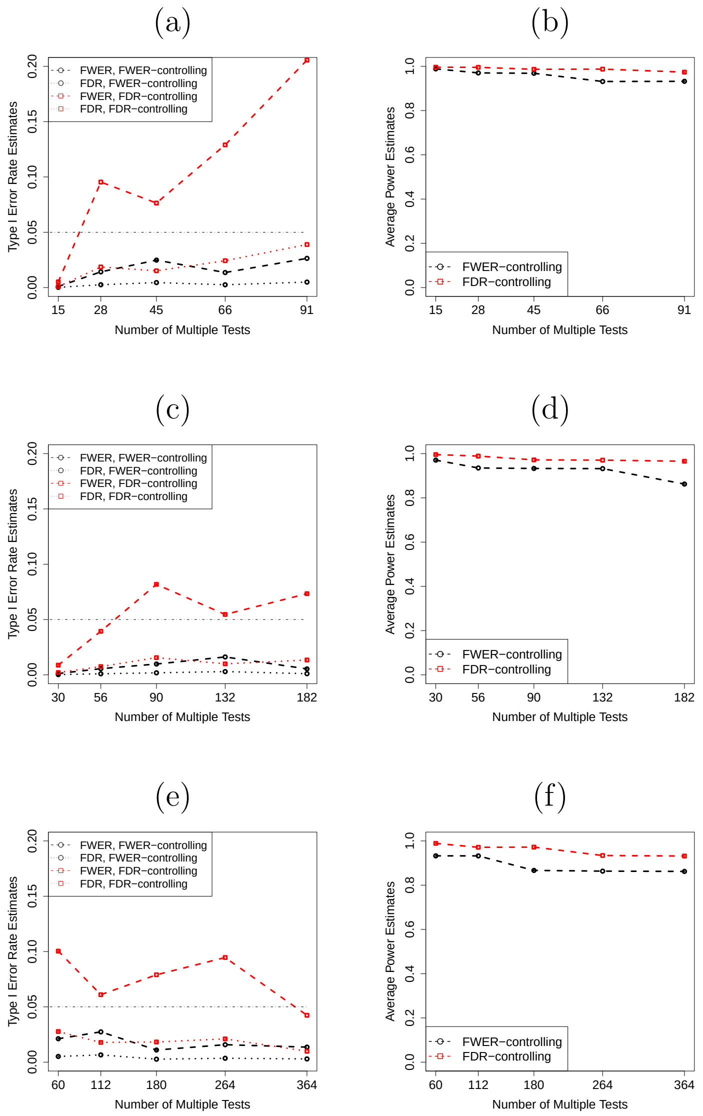


Figure 5.4: (a) Average FWER and FDR estimates (left column) and power estimates (right column) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets and preditive power $p = 0.85$. (a) FWER and FDR estimates for model set $K_1 = \{\text{AND}\}$. (b) Power estimates for model set $K_1 = \{\text{AND}\}$. (c) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (d) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}\}$. (e) FWER and FDR estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. (f) Power estimates for model set $K_1 = \{\text{AND}, \text{XOR}, \overline{X_1 X_2}\}$. There is a single test to be predicted by 5 IMP pairs (as shown in Figs. 5.2) among a vaying number of initial genes $K = 6, 8, 10, 12, 14$. Sample size is fixed at $n = 40$.

Step 1: Generate i.i.d. sample binary data of size n, $y_1, \ldots, y_n$, for gene TRG following $P(Y = 0) = c \leq 0.5$, for a given c.

Do **Step 2** and **Step 3** from the 1st IMP pair to the t-th pair:

Step 2: Generate n i.i.d. nosie samples $n_1, n_2, \ldots, n_n$ that satisfies $P(N = 1) = 1 - p$. Next, we obtain a new sequence $y_i^* = n_i \oplus y_i$, for $i = 1, \ldots, n$, which is the true sample of Y before it is contaminated by noise.

Step 3: For the i-th IMP pair, if the sample data of the 1st predictor have not been generated yet, we generate n i.i.d. samples for this predictor by following a uniform distribution. By knowing the logic function LGCi associated with output y^* and input of the 1st predictor, sample data for the 2nd predictor can be generated deterministically. If there are more than 1 possible solutions, just randomly pick one.

Step 4: Generate i.i.d. sample binary data of size n for PRD$U + 1$, ..., PRDG following a uniform distribution.

Note that the proposed data-generating procedure has its limitations on the relationships among predictors in the IMP pairs. To put this more clear, examples of graphs (with cycle or without cycle) of IMP pairs are shown in Fig. 5.2. In the graphs, all unrepeated predictors in the IMP pairs are considered as vertices of the graphs, and two vertices are connected if the corresponding two predictors consist of an IMP pair for the target. Obviously, the proposed procedure can only be employed to the case with no cycle in the graph. This is because the existence of cycles in the graph of IMP pairs will result in conflicts in the generation of samples for predictors, which, however, can be avoided in the case of no cycles using the above data-generating procedure.

Here the number of candidate predictor set for a given target is $L = \binom{G}{2}$. Provided that G is not too large, this does not create a serious multiplicity issue; in our simulation, G ranges from 6 to 14. As for the logic model set K, we consider three scenarios: (1) the prediction logic is known, $K_1 = \{\mathrm{XOR}\}$; (2) $K_2 = \{\mathrm{AND}, \mathrm{XOR}\}$; and (3) $K_3 = \{\mathrm{AND}, \mathrm{XOR}, \overline{\mathrm{X}_1\mathrm{X}_2}, \overline{\mathrm{X}_1}\mathrm{X}_2\}$. The total number of tests to be performed is given by $M_i = \binom{G}{2} \times K_i$, under each of the prio-knowledge scenarios $i = 1, 2, 3$ described previously. Hence, the MTP increases in difficulty as the number of predictors and targets increase, and as less prior knowledge is available. We draw 5000 samples of varying size n and form averages of FWER, FDR, and power estimates under FWER-controlling and FDR-controlling procedures.

Figures 5.3 and 5.4 show the performance of FWER and FDR estimates and power estimates for varying sample size and number of multiple tests, where FWER- and FDR-controlling procedures are employed, respectively. In Fig. 5.3, we fix $G = 24$, and we plot the results as a function of sample size n. The total number of tests is $M_1 = 276$, $M_2 = 553$ and $M_3 = 828$, under each logic model set. In Fig. 5.4, we fix $n = 40$ and vary G from 6 to 14. The total number of tests varies from a minimum of 15 in the case of $G = 6$ and complete knowledge (K_1) about the prediction logic to a maxim of 273, for $G = 14$ and logic model set K_3. Results are plotted as a function of M. Several observations are made in the following from these figures.

- The FWER- and FDR-controlling procedures are able to control the FWER and FDR, respectively. In addition, FWER estimates are always larger than FDR estimates, as expected from theroy, and thus the FDR is always controlled under the FWER-controlling procedure, which is more conservative for the smaller FWER and FDR estimates than the FDR-controlling procedure, as we observe.

- As for the power, we can see a clear superiority of the FDR- over the FWER-controlling procedure. Moreover, the power of the FDR-controlling procedure, besides being excellent, is also robust to the increase in number of tests (M), in comparison the FWER-controlling procedure.

- As the sample size increases, the power estimates increase for both procedures, whereas, as the number of multiple tests increases, the power estimates decrease, as expected.

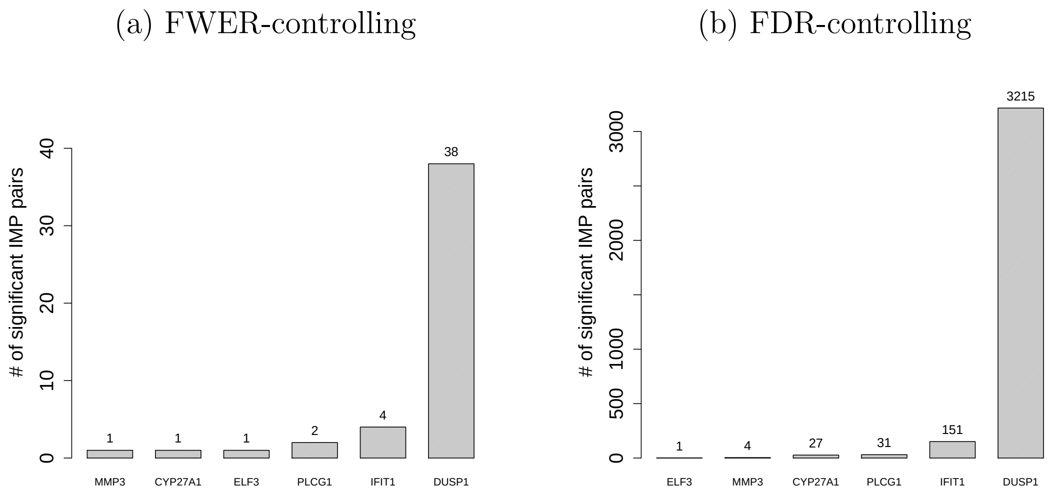
These results indicate that the FDR-controlling procedure is preferred in real applications due to its superior power over the FWER-controlling procedure.

5.4.2 Real Data

In this section, the proposed multiple testing procedures are applied to real data sets for the identification of canalyzing genes and their IMP sets of statistical significance.

5.4.2.1 Case study I: melanoma and gene DUSP1

One data set of interest consists of 31 samples with 587 gene expressions. 19 sample out of the 31 samples are normal tissues and the remaining 12 samples are tissues with melanoma. All the gene expressions are binarized into 0 or 1, where 0 indicates no significant expression whereas 1 represents significant expression (either over- or under-expression). We eliminate 469 genes out of 587 genes in the original dataset by following the criterion that there should be enough variability in the data. As a comparison, we preserve the gene DUSP1 of our particular interest. Hence, we have 119 genes left for analysis.


(a) FWER-controlling

(b) FDR-controlling

Figure 5.5: Number of significant IMP pairs versus target gene discovered from melanoma data set (a) using the FWER-controlling approach; (b) using the FDR-controlling approach.

We fix the significance level α to be 0.05. We assume no prior knowledge, and thus make no constraints on the allowed regulatory relationships, other than a gene does not predict itself. Suppose that each target is predicted by $d = 2$ predictors, and there are 10 possible 2-predictor candidate logics for each target gene. Hence, the total number of multiple tests to be performed is $M = \binom{118}{2} * 10 = 69,030$ for each target gene. Note that, we conduct here multiple testing procedures target by target as proposed in Section 5.3 due to the large number of genes in the processed data.

Figure 5.5 shows the number of significant IMP pairs for six targets (*CYP27A1*, *ELF3*, *MMP3*, *PLCG1*, *IFIT1* and *DUSP1*) by using FWER- and FDR-controlling procedures, respectively. It is observed that, *DUSP1*, a hypothesized canalizing gene, has the largest number of significant IMP pairs for both approaches. This is consistent with the fact that the gene *DUSP1* plays an active role in regulating central and process-integrating signaling pathways. By using the FWER-controlling approach, there are 38 significant IMP pairs for *DUSP1*, whereas, 3215 significant IMP pairs for *DUSP1* by the FDR-controlling approach, since the latter approach is less conservative than the former one. We present in Table 5.2 in the supplementary information the top 20 significant IMP sets for target genes under both FDR- and FWER-controlling approaches, which gives the potential multivariate predictions of statistical significance for the guidance of further biological experimental studies.

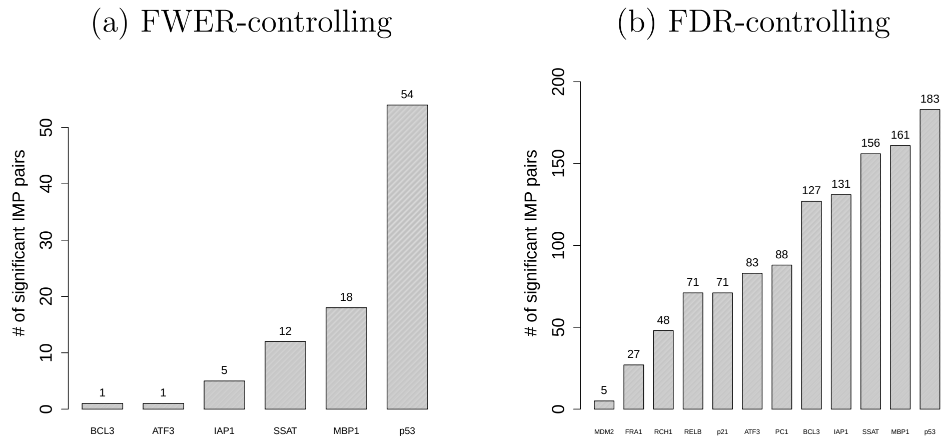
5.4.2.2 Case study II: genotoxic stresses and gene p53

This data set consists of 12 genes under 3 conditions (i.e., IR, MMS, UV) in 30 cell lines of both *p53* proficient and *p53* deficient cells. The data is ternary, indicating up-regulated (+1), down-regulated (-1), or no-change (0) status. Here we map this to binary expression using the following code: change (1), for either up-regulated or down-regulated genes, and no-change (0), as before. Additionally, we consider the three binary conditions (IR, MMS, and UV) as possible predictive factors, for a total of 15 Boolean variables in the data set.

We employ the proposed multiple testing procedures to find significant IMP sets of target genes. We again assume no prior knowledge about regulatory relationships and that a gene does not predict itself. Hence, all $\binom{14}{2}$ two-predictor sets and 10 possible "true" 2-predictor logic candidates are considered for each target, for a total of $\binom{14}{2} \times 10 = 910$ possible models; note that each gene can appear in multiple models, both as a member of different pairs and under different logic relationships. In addition, we consider each of the 12 genes in the data set as a possible target, so that the number of multiple tests performed is $M = \binom{14}{2} \times 10 \times 12 = 10,920$. We apply both the FWER- and the FDR-controlling procedures outlined in the previous section with a significance level $\alpha = 0.05$. Figure 5.6 displays the number of significant IMP set for corresponding gene targets, under each of the two approaches.

Interestingly, *p53* turns out to possess the largest number of significant IMP sets, under both approaches. This is in accordance with the known fact that *p53* is a significantly active gene involved in various pathways associated with stress responses. *p53* plays a crucial role in arresting the cell cycle, inhibiting angiogenesis, activating DNA repair and conserving genome stability. In unstressed cells, *p53* is kept in a low level through a continuous degradation of itself. However, it becomes activated in reponse to environmental stresses like UV, IR and oxidative stress, gaining a quick accumulation of *p53* in stressed cells and acting as a transcriptional regulator in cells.

Notice that the FWER-controlling approach is more conservative and thus produces fewer significant IMP sets than the FDR-controlling approach, for each of the targets. By using the FDR-controlling approach, one detection is consistent with biological groundtruth that *p21* is found to be expressed when *MDM2* is expressed or *ATF3* is expressed. Table 5.3 lists top 20 significant IMP pairs of target genes under FDR- and FWER-controlling approaches. These results could serve as candidate regulatory relationships for further experimental verification.


(a) FWER-controlling

(b) FDR-controlling

Figure 5.6: Number of significant IMP pairs versus target gene discovered from genotoxic stress-responsive data set (a) using the FWER-controlling approach; (b) using the FDR-controlling approach.

5.5 Summary

We have presented a rigorous statistical testing framework to detect canalyzing genes, by using the intrinsically multivariate predictive (IMP) criterion in the context of discrete CoD. Multiple-testing procedures are also proposed by taking advantage of a-priori knowledge about logical predictions if available, thus making the methodology applicable to large data sets. Furthermore, an R *imptest* package is developed for the implementation of the IMP hypothesis test, which is available to the scientific community through our website (http://gsp.tamu.edu/Publications/supplementary/ting13c). It is expected that this methodology will serve as a potential tool for the inference of canalyzing genes from discrete gene-expression data.

Table 5.2: A list of top 20 significant IMP pairs detected in the melanoma data

Target	Predictor 1	Predictor 2	Logic	π' (FWER)	π'' (FDR)
IFIT1	MMP3	TNF1F7	OR	1.209×10^{-3}	NA
DUSP1	MMP3	TNF1C2	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	UG5F5	LO1D6	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	TNF1F7	TNF1C2	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	HV2h5	HV70c10	$\bar{X}_1\bar{X}_2$	1.209×10^{-3}	1.286×10^{-4}
DUSP1	IFIT1	TNF1C2	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	CYP27A1	TNF1C2	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	CYP27A1	HV48d10	$X_1\bar{X}_2$	1.209×10^{-3}	1.286×10^{-4}
DUSP1	HV5c12	LO1D6	AND	1.209×10^{-3}	1.286×10^{-4}
DUSP1	HV25e5	PLCG1	AND	1.598×10^{-2}	4.204×10^{-4}
DUSP1	HV14e11	CYP27A1	$\bar{X}_1X_2$	1.598×10^{-2}	4.204×10^{-4}
DUSP1	MMP3	UG3G1	AND	1.598×10^{-2}	4.204×10^{-4}
MMP3	IFIT1	HV70c10	$X_1\bar{X}_2$	1.598×10^{-2}	NA
PLCG1	CYP27A1	ELF3	OR	1.598×10^{-2}	NA
PLCG1	ELF3	DUSP1	OR	1.598×10^{-2}	NA
IFIT1	MMP3	HV23e2	OR	1.598×10^{-2}	NA
IFIT1	MMP3	ELF3	OR	1.598×10^{-2}	NA
IFIT1	TNF1F7	DUSP1	OR	1.598×10^{-2}	NA
CYP27A1	PLCG1	HV12d1	AND	1.598×10^{-2}	NA
ELF3	PLCG1	HV24f12	AND	1.598×10^{-2}	NA
DUSP1	MMP3	HV5d9	$X_1\bar{X}_2$	NA	4.204×10^{-4}
DUSP1	MMP3	HV2h5	$X_1\bar{X}_2$	NA	4.204×10^{-4}
DUSP1	MMP3	ESTs	AND	NA	4.204×10^{-4}
DUSP1	MMP3	HV5c12	AND	NA	4.204×10^{-4}
DUSP1	PLCG1	HV2h5	$X_1\bar{X}_2$	NA	4.204×10^{-4}
DUSP1	PLCG1	TNF1C2	AND	NA	4.204×10^{-4}
DUSP1	PLCG1	HV5c12	AND	NA	4.204×10^{-4}
DUSP1	PLCG1	HV48d10	$X_1\bar{X}_2$	NA	4.204×10^{-4}
DUSP1	UG5F5	HV48d10	$X_1\bar{X}_2$	NA	4.204×10^{-4}

Table 5.3: A list of top 20 significant IMP pairs detected in IR-response stress data under FWER- and FDR-controlling approaches

Target	Predictor 1	Predictor 2	Logic	π' (FWER)	π'' (FDR)
MBP1	RCH1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	BCL3	IAP1	$\bar{X}_1 X_2$	3.153×10^{-4}	7.006×10^{-6}
MBP1	FRA1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	FRA1	SSAT	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	ATF3	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	SSAT	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	MDM2	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	IAP1	p21	AND	3.153×10^{-4}	7.006×10^{-6}
MBP1	SSAT	MDM2	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	BCL3	MBP1	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	BCL3	p21	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	FRA1	IAP1	AND	3.153×10^{-4}	7.006×10^{-6}
SSAT	FRA1	MBP1	AND	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	BCL3	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	IAP1	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	MMS	NAND	3.153×10^{-4}	7.006×10^{-6}
p53	RCH1	UV	NAND	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	ATF3	$\bar{X}_1 + X_2$	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	IAP1	$X_1 + \bar{X}_2$	3.153×10^{-4}	7.006×10^{-6}
p53	BCL3	MBP1	NAND	3.153×10^{-4}	7.006×10^{-6}

6. BAYESIAN COD ESTIMATION

As discussed in Sections 2 and 3, the CoD was estimated through nonparametric and nonparametric methods from a frequentist perspective, respectively [16, 21]. We investigated the performance of four nonparametric CoD estimators, based on the resubstitution, leave-one-out, bootstrap and cross-validation error estimators and that of parametric maximum-likelihood (ML) CoD estimator, based on parametric models for gene regulatory relationships. It was observed that, with the availability of prior knowledge about logic predictions, the ML CoD estimator is preferred for its best performance, whereas, one, without any prior knowledge, should use the resubstitution CoD estimator, provided one has evidence of moderate to tight regulation between predictors and target, and the number of predictors is not too large.

The nonparametric CoD estimators are defined by the discrete histogram prediction rule, while ML model-based CoD estimators are defined with respect to a parametric model. However, none of these CoD estimators are optimized based on statistical inference across a family of possible joint distributions between target and predictors, where the mass of the random parameter is concentrated around true parameter values for the true target-predictor distribution. This leads to a Bayesian approach to CoD estimation based on a parametrized family of target-predictor distributions as a function of random parameters characterized by assumed prior distributions. Such an idea was first introduced in the study of Bayesian error estimation for classification, which optimizes sample-based error estimation relative to mean-square error (MSE) between the error estimator and true error across a family of feature-label distributions [26, 27].

Following the Bayesian idea, we first introduce in this Chapter the exact formulation of the Bayesian CoD estimator in a minimum mean-square error (MMSE) sense, and the Bayesian CoD estimator based on the optimal Bayesian classifier [19]. Next, we employ Monte Carlo sampling experiments to assess the performance of the Bayesian CoD estimator against that of resubstitution, leave-one-out, bootstrap and cross-validation CoD estimators. Finally, Bayesian inference algorithms are developed with comparison to frequentist inference algorithms in Section 3. We also provide examples of their practical applications to gene-expression data sets.

6.1 Discrete Model

We define in this Section the discrete prediction setting. Let $\mathbf{X} = (X_1, X_2, \ldots, X_d) \in \{0,1\}^d$ be a predictor random vector and $Y \in \{0,1\}$ be a target random variable in our discrete prediction problem. The predictors as a group can take on values in a finite space with $b = 2^d$ possible states. For analysis purposes, we establish a bijection between this finite state space and a single predictor variable X taking values in the set $X \in \{1, 2, \ldots, b\}$. One specific value of X corresponds to a specific combination of the values of the original predictors, *i.e.*, a "bin" into which the data is categorized. The value b is the number of bins, which provides a direct measure of predictor complexity.

The probability distribution of the pair (X, Y) is specified by target prior probabilities: $c = P(Y = 0), 1 - c = P(Y = 1)$, and probabilities $p_i = P(X = i \mid Y = 0)$ and $q_i = P(X = i \mid Y = 1)$, for $i = 1, \ldots, b$. Notice that $\sum_{i=1}^{b} p_i = 1$ and $\sum_{i=1}^{b} q_i = 1$. Let the vector $\mathbf{p}$ denote $(p_1, \ldots, p_{b-1})$, $\mathbf{q}$ denote $(q_1, \ldots, q_{b-1})$ and θ be the parameter vector $(c, \mathbf{p}, \mathbf{q})$. Given sample data, define U_i as the number of samples with $Y = 0$ in bin $X = i$, and V_i as the number of samples with $Y = 1$ in bin $X = i$, for $i = 1, \ldots, b$. Define also the sample sizes $N_0 = \sum_{i=1}^{b} U_i$ and $N_1 = \sum_{i=1}^{b} V_i$. In what follows, realizations of the random variables N_0, N_1, U_i, V_i will be denoted by the respective small letters.

In the discrete prediction setting formulated previously, the CoD in eq. (2.5) can clearly be formulated as

$$\text{CoD} \;=\; \frac{1-c}{g(c)} + \sum_{i=1}^{b} \left(\frac{c}{g(c)} p_i - \frac{1-c}{g(c)} q_i \right) I_{p_i < \frac{1-c}{c} q_i} \,, \tag{6.1}$$

where $g(x) = \min(x, 1-x)$, for $x \in [0,1]$, and I_A is an indicator function giving 1 if condition A is satisfied; otherwise 0.

6.2 Bayesian CoD Estimators

We present in this Section the formulation of two well-defined Bayesian MMSE estimators for the CoD in eq. (6.1). One approach is analogous to that followed by [26] in defining the Bayesian MMSE classification error estimator, whereas the other one makes use of the optimal Bayesian classifier in [28].

In the Bayesian setting, our model set is indexed by the parameter vector $\theta = (c, \mathbf{p}, \mathbf{q})$, defined previously. The appropriate definitions of the priors for these parameters could take advantage of prior knowledge about the biological problem. For simplicity, here we will consider as priors the Dirichlet and Beta distributions [14]:

$$c \sim \text{Beta}(\alpha, \beta)\,,\;\; \mathbf{p} \sim \text{Dirichlet}(\alpha_1^0, \ldots, \alpha_b^0)\,,\;\; \mathbf{q} \sim \text{Dirichlet}(\alpha_1^1, \ldots, \alpha_b^1)\,, \tag{6.2}$$

where these hyperparameters α, β, α_i^0, α_i^1, $i = 1, \ldots, b$, are positive numbers. The case $\alpha_i^0 = \alpha_i^1 = 1$, for all $i = 1, \ldots, b$, corresponds to uninformative uniform priors. It is well-known that these are conjugate priors that take the same form as the corresponding posteriors, which are shown in [26, 27] to be:

$$\begin{aligned} &c \mid \mathbf{S}_n \sim \text{Beta}(n_0+\alpha, n_1+\beta)\,,\ \mathbf{p} \mid \mathbf{S}_n \sim \text{Dirichlet}(u_i+\alpha_1^0, \ldots, u_b+\alpha_b^0)\,, \\ &\mathbf{q} \mid \mathbf{S}_n \sim \text{Dirichlet}(v_i+\alpha_1^1, \ldots, v_b+\alpha_b^1)\,. \end{aligned} \tag{6.3}$$

Furthermore, it is also known that each element in $\mathbf{p}$ and $\mathbf{q}$ is beta-distributed: $p_i \sim \text{Beta}(t_{ap}^i, t_{bp}^i)$ and $q_i \sim \text{Beta}(t_{aq}^i, t_{bq}^i)$, where $t_{ap}^i = u_i+\alpha_i^0$, $t_{bp}^i = n_0+\alpha_0-(u_i+\alpha_i^0)$, $t_{aq}^i = u_i+\alpha_i^1$, and $t_{bq}^i = n_1+\alpha_1-(v_i+\alpha_i^1)$, for $i = 1, \ldots, b$.

6.2.1 The Bayesian MMSE CoD Estimator

We are interested in finding a sample-based estimator $\widehat{\text{CoD}}$ that minimizes $E_{\theta,\mathbf{S}_n}[|\widehat{\text{CoD}}-\text{CoD}|^2]$. The solution is the *Bayesian MMSE CoD estimator* $\widehat{\text{CoD}}^*$, which can be shown to be given by:

$$\widehat{\text{CoD}}^* = E_\theta[\text{CoD} \mid \mathbf{S}_n], \tag{6.4}$$

where the CoD is expressed in eq. (6.1). Notice that $\widehat{\text{CoD}}^*$ is an unbiased estimator and displays the least root mean-square error (RMS) over the distribution of $(\theta, \mathbf{S}_n)$. However, for a specific model with fixed θ, $\widehat{\text{CoD}}^*$ might not be unbiased or have the least RMS.

An interesting and useful fact proved in [26] is that c, $\mathbf{p}$ and $\mathbf{q}$ are independent given the sample data. Starting from (6.1), we can exploit this independence to write the Bayesian MMSE CoD estimator as

$$\begin{aligned} E_\theta[\text{CoD} \mid \mathbf{S}_n] = 1 - &\underbrace{E_{c|\mathbf{S}_n}\left[\frac{1-c}{g(c)}\right]}_{A} - \sum_{i=1}^{b}\{\underbrace{E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[\frac{c}{g(c)}p_i I_{p_i<\frac{1-c}{c}q_i}\right]\right]\right]}_{B_i} \\ &+ \underbrace{E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[\frac{1-c}{g(c)}q_i I_{p_i<\frac{1-c}{c}q_i}\right]\right]\right]}_{C_i}\}. \end{aligned} \tag{6.5}$$

In what follows, we give expressions for A, B_i, and C_i.

(1) Term A is given by

$$\begin{aligned} A = E_{c|\mathbf{S}_n}\left[\frac{1-c}{c}I_{c<1/2} + I_{c\geq 1/2}\right] = 1 + \\ \frac{1}{\mathrm{B}(n_0+\alpha, n_1+\beta)} \times \{\mathrm{IB}(1/2; n_0+\alpha-1, n_1+\beta+1) \\ -\mathrm{IB}(1/2; n_0+\alpha, n_1+\beta)\}\,, \end{aligned} \tag{6.6}$$

where B is the Beta function and IB is the *incomplete* Beta function:

$$\begin{aligned} \mathrm{IB}(k; a, b) = \int_0^k x^{a-1}(1-x)^{b-1}dx = \\ \begin{cases} \sum_{i=0}^{b-1} \frac{(-1)^i k^{a+i}}{a+i}\binom{b-1}{i}, & b \text{ is an integer,} \\ \sum_{i=0}^{\infty} \frac{(-1)^i k^{a+i}}{a+i}\binom{b-1}{i}, & \text{o.w.} \end{cases} . \end{aligned} \tag{6.7}$$

for $a, b > 0$ and $0 \leq k \leq 1$.

Before we proceed, we mention a useful fact concerning a Beta random variable.

Proposition 6. *Given $X \sim Beta(\alpha, \beta)$, we have*

$$E[XI_{X\leq k}] = \frac{IB(k; \alpha+1, \beta)}{B(\alpha, \beta)}I_{k<1} + \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)}I_{k\geq 1}\,. \tag{6.8}$$

Proof. This is obvious due to the fact that $0 \leq X \leq 1$. □

(2) By taking first the expectation over $\mathbf{p} \mid \mathbf{S}_n$ and using the definition of function

IB in eq. (6.7), we have that

$$
\begin{aligned}
B_i = \frac{1}{\mathrm{B}(t^i_{ap}, t^i_{bp})} \sum_{i=1}^{b} \Bigg(\sum_{j=0}^{P^i} \frac{(-1)^j \binom{t^i_{bp}-1}{j}}{t^i_{ap}+1+j} \times \\
\underbrace{E_{c|\mathbf{S}_n}\left[\frac{1}{g(c)} \frac{(1-c)^{t^i_{ap}+1+j}}{c^{t^i_{ap}+j}} E_{\mathbf{q}|\mathbf{S}_n}\left[q_i^{t^i_{ap}+1+j} I_{q_i<\frac{c}{1-c}} \right] \right]}_{B^1_i} \\
+ \mathrm{B}(t^i_{ap}+1, t^i_{bp}) \underbrace{E_{c|\mathbf{S}_n}\left[\frac{c}{g(c)} E_{\mathbf{q}|\mathbf{S}_n}\left[I_{q_i \geq \frac{c}{1-c}} \right] \right]}_{B^2_i} \Bigg) \quad ,
\end{aligned}
\tag{6.9}
$$

where $P^i = t^i_{bp} - 1$ if t^i_{bp} is an integer; otherwise $P^i = \infty$, and B^1_i and B^2_i can be obtained by taking expectation over $\mathbf{q} \mid \mathbf{S}_n$, using the definition of function IB, and applying Proposition A:

$$
\begin{aligned}
B^1_i = \frac{1}{\mathrm{B}(t^i_{aq}, t^i_{bq})} \Bigg(\sum_{k=0}^{Q^i} \frac{(-1)^k \binom{t^i_{bq}-1}{k}}{t^i_{aq}+t^i_{ap}+1+j+k} \\
\underbrace{E_{c|\mathbf{S}_n}\left[\frac{c}{g(c)} \left(\frac{c}{1-c} \right)^{t^i_{aq}+k} I_{c<1/2} \right]}_{B^3_i} + \mathrm{B}(t^i_{aq}+t^i_{ap}+1+j, t^i_{bq}) \\
\underbrace{E_{c|\mathbf{S}_n}\left[\frac{c}{g(c)} \left(\frac{1-c}{c} \right)^{t^i_{ap}+1+j} I_{c\geq 1/2} \right]}_{B^4_i} \Bigg) \quad ,
\end{aligned}
\tag{6.10}
$$

$$
\begin{aligned}
B^2_i = \underbrace{E_{c|\mathbf{S}_n}\left[\frac{c}{g(c)} I_{c\leq 1/2} \right]}_{B^5} - \frac{1}{\mathrm{B}(t^i_{aq}, t^i_{bq})} \sum_{k=0}^{Q^i} \frac{(-1)^k \binom{t^i_{bq}-1}{k}}{t^i_{aq}+k} \times \\
\underbrace{E_{c|\mathbf{S}_n}\left[\frac{c}{g(c)} \left(\frac{c}{1-c} \right)^{t^i_{aq}+k} I_{c<1/2} \right]}_{B^6_i} ,
\end{aligned}
$$

where $Q^i = t^i_{bq} - 1$ if t^i_{bq} is an integer; otherwise $Q^i = \infty$, while

$$\begin{aligned}
B^i_3 &= \mathrm{IB}(1/2; n_0 + \alpha + t^i_{aq} + k, n_1 + \beta - t^i_{aq} - k)\,, \\
B^i_4 &= \frac{\mathrm{B}(n_0 + \alpha - t^i_{ap} - j, n_1 + \beta + t^i_{ap} + j)}{\mathrm{B}(n_0 + \alpha, n_1 + \beta)} - \\
&\quad \frac{\mathrm{IB}(1/2; n_0 + \alpha - t^i_{ap} - j, n_1 + \beta + t^i_{ap} + j)}{\mathrm{B}(n_0 + \alpha, n_1 + \beta)}\,, \\
B_5 &= \mathrm{IB}(1/2; n_0 + \alpha, n_1 + \beta)\,, \\
B^i_6 &= \mathrm{IB}(1/2; n_0 + \alpha + t^i_{aq} + k, n_1 + \beta - t^i_{aq} - k)\,.
\end{aligned} \tag{6.11}$$

(3) Similarly as in item (2), we have that:

$$\begin{aligned}
C_i = \frac{1}{\mathrm{B}(t^i_{ap}, t^i_{bp})} \sum_{i=1}^{b} \Bigg(& \sum_{j=0}^{P^i} \frac{(-1)^j \binom{t^i_{bp}-1}{j}}{t^i_{ap} + j} \times \\
& \underbrace{E_{c|\mathbf{S}_n}\left[\frac{1}{g(c)} \frac{(1-c)^{t^i_{ap}+j+1}}{c^{t^i_{ap}+j}} E_{\mathbf{q}|\mathbf{S}_n}\left[q_i^{t^i_{ap}+1+j} I_{q_i < \frac{c}{1-c}}\right]\right]}_{C^1_i} + \\
& \mathrm{B}(t^i_{ap}, t^i_{bp}) \underbrace{E_{c|\mathbf{S}_n}\left[\frac{1-c}{g(c)} E_{\mathbf{q}|\mathbf{S}_n}\left[q_i I_{q_i \geq \frac{c}{1-c}}\right]\right]}_{C^2_i} \Bigg)\,,
\end{aligned} \tag{6.12}$$

with $C^1_i = B^1_i$ and C^2_i being given by:

$$\begin{aligned}
C^2_i = & \frac{\mathrm{B}(t^i_{aq} + 1, t^i_{bq})}{\mathrm{B}(t^i_{aq}, t^i_{bq})} \underbrace{E_{c|\mathbf{S}_n}\left[\frac{1-c}{g(c)} I_{c \leq 1/2}\right]}_{C^5} - \frac{1}{\mathrm{B}(t^i_{aq}, t^i_{bq})} \\
& \sum_{k=0}^{Q^i} \frac{(-1)^k \binom{t^i_{bq}-1}{k}}{t^i_{aq} + k + 1} \underbrace{E_{c|\mathbf{S}_n}\left[\frac{1-c}{g(c)} \left(\frac{c}{1-c}\right)^{t^i_{aq}+k+1} I_{c<1/2}\right]}_{C^6_i}\,,
\end{aligned} \tag{6.13}$$

where

$$C^5 = \mathrm{IB}(1/2; n_0 + \alpha - 1, n_1 + \beta + 1)\,, \tag{6.14}$$

and $C_i^6 = B_i^6$.

Finally, in order to get positive a and b in eq. (6.7), the hyperparameters for c, $\mathbf{p}$, $\mathbf{q}$ must satisfy the following conditions:

$$\alpha > \sum_{i=1}^{b} \alpha_i^0 - 1, \quad \beta > \sum_{i=1}^{b} \alpha_i^1 - 1\,. \tag{6.15}$$

Hence, if we choose uniform priors for $\mathbf{p}$ and $\mathbf{q}$, it is clear that the prior for c cannot be uniform.

6.2.2 The Bayesian CoD Estimator Based on the Optimal Bayesian Classifier

In Section 2, we have discussed several nonparametric CoD estimators based on the resubstitution, leave-one-out, bootstrap and cross-validation error estimators. Likewise, we will investigate another Bayesian CoD estimator in terms of Bayesian error estimators, in which case the Bayesian error estimator is minimized over some optimal Bayesian classifier [28, 29]. Such a Bayesian CoD estimator is quite similar to the nonparametric CoD estimator as a function of corresponding nonparametric error estimators in Section 2.

Let us first recall the concepts of Bayesian error estimation and optimal Bayesian classification. The optimization of error estimation is addressed in a Bayesian modelling framework throughout a family of distributions between target and predictors [26–29]. For an arbitrary classifier ψ, the Bayesian MMSE error estimator based on given information of X is expressed as [26, 27]:

$$\hat{\varepsilon} = \sum_{j=1}^{b} \left\{ \frac{n_0 + \alpha}{n + \alpha + \beta} \times \frac{U_j + \alpha_j^0}{n_0 + \alpha^0} I(\psi(j) = 1) + \frac{n_1 + \beta}{n + \alpha + \beta} \times \frac{V_j + \alpha_j^1}{n_1 + \alpha^1} I(\psi(j) = 0) \right\}\,. \tag{6.16}$$

To optimize classifier design, an optimal Bayesian classifier, ψ_{OBC}, is defined as

$$E_{\mathbf{p}|\mathbf{S}_n,\mathbf{q}|\mathbf{S}_n,c|\mathbf{S}_n}[\varepsilon(\theta,\psi_{\text{OBC}})] \leq E_{\mathbf{p}|\mathbf{S}_n,\mathbf{q}|\mathbf{S}_n,c|\mathbf{S}_n}[\varepsilon(\theta,\psi)]\,, \tag{6.17}$$

for all $\psi \in \mathcal{C}$, where $\mathcal{C}$ is an arbitrary family of classifier [28,29]. It has been shown that the optimal Bayesian classifier in the discrete model with $(\mathbf{p},\mathbf{q},c)$ is formed as

$$\psi_{\text{OBC}} = \begin{cases} 1, & \frac{n_0+\alpha}{n+\alpha+\beta}\frac{U_j+\alpha_j^0}{n_0+\alpha^0} < \frac{n_1+\beta}{n+\alpha+\beta}\frac{V_j+\alpha_j^1}{n_1+\alpha^1} \\ 0, & \text{o.w.} \end{cases}\,. \tag{6.18}$$

By substituting ψ_{OBC} for ψ in eq. (6.16), we have the Bayesian error estimator based on the optimal Bayesian classifier:

$$\hat{\varepsilon}_{\text{OBC}} = \sum_{i=1}^{b} \min\left\{\frac{n_0+\alpha}{n+\alpha+\beta}\frac{U_j+\alpha_j^0}{n_0+\alpha^0}, \frac{n_1+\beta}{n+\alpha+\beta}\frac{V_j+\alpha_j^1}{n_1+\alpha^1}\,,\right\} \tag{6.19}$$

which is shown to minimize Bayesian error estimator over all possible $\psi \in \mathcal{C}$.

Similarly, given no information about predictor X, its corresponding minimum Bayesian error estimator of Y is formed as (in terms of the optimal Bayesian classifier):

$$\hat{\varepsilon}_{0,\text{OBC}} = \min\left\{\frac{n_0+\alpha}{n+\alpha+\beta}, \frac{n_1+\beta}{n+\alpha+\beta}\right\}\,. \tag{6.20}$$

In terms of $\hat{\varepsilon}_{\text{OBC}}$ in eq. (6.19) and $\hat{\varepsilon}_{0,\text{OBC}}$ in eq. (6.20), the Bayesian CoD estimator based on the optimal Bayesian classifier, $\widehat{\text{CoD}}_{\text{OBC}}$, is given by:

$$\widehat{\text{CoD}}_{\text{OBC}} = 1 - \frac{\hat{\varepsilon}_{\text{OBC}}}{\hat{\varepsilon}_{0,\text{OBC}}}\,. \tag{6.21}$$

It is easy to show that $0 < \hat{\varepsilon}_{\text{OBC}} < \hat{\varepsilon}_{0,\text{OBC}}$, and thus $\widehat{\text{CoD}}_{\text{OBC}} \in (0,1)$.

6.3 Exact Moments of Bayesian CoD Estimator Based on Optimal Bayesian Classifier

As noted in Section 2, the performance metrics for an CoD estimator $\widehat{\text{CoD}}$ are its bias,

$$\text{Bias}\left[\widehat{\text{CoD}}\right] \;=\; E\left[\widehat{\text{CoD}}\right] - \text{CoD}\,, \tag{6.22}$$

the deviation variance,

$$\text{Var}_\text{d}\left[\widehat{\text{CoD}}\right] \;=\; \text{Var}\left(\widehat{\text{CoD}} - \text{CoD}\right) \;=\; \text{Var}\left(\widehat{\text{CoD}}\right)\,, \tag{6.23}$$

and the root mean-square (RMS) error,

$$\text{RMS}\left[\widehat{\text{CoD}}\right] \;=\; \sqrt{\text{Var}\left[\widehat{\text{CoD}}\right] \;+\; \text{Bias}\left[\widehat{\text{CoD}}\right]^2} \tag{6.24}$$

According to eqs. (2.21), (2.22) and (2.23), the peformance metrics (i.e., bias, deviation variance and RMS) for the Bayesian CoD estimator based on the optimal Bayesian classifier, $\widehat{\text{CoD}}_{\text{OBC}}$, can be obtained from the first and second momemts of $\hat{\varepsilon}_{\text{OBC}}/\hat{\varepsilon}_{0,\text{OBC}}$, namesly, $E\left[\frac{\hat{\varepsilon}_{\text{OBC}}}{\hat{\varepsilon}_{0,\text{OBC}}}\right]$ and $E\left[\frac{\hat{\varepsilon}^2_{\text{OBC}}}{\hat{\varepsilon}^2_{0,\text{OBC}}}\right]$.

The first moment of $\hat{\varepsilon}_{\text{OBC}}/\hat{\varepsilon}_{0,\text{OBC}}$ is given by

$$E\left[\frac{\hat{\varepsilon}_{\text{OBC}}}{\hat{\varepsilon}_{0,\text{OBC}}}\right] \;=\; \sum_{m\in U} E\left[\frac{\hat{\varepsilon}_{\text{OBC}}}{m/n+\alpha+\beta} \mid M=m\right] P(M=m), \tag{6.25}$$

where $U = \left\{\alpha, \alpha+1, \ldots, \lfloor\frac{n+\beta-\alpha}{2}\rfloor + \alpha, \beta, \beta+1, \ldots, \lfloor\frac{n+\alpha-\beta}{2}\rfloor + \beta\right\}$ and $M = (n + \alpha + \beta)\hat{\varepsilon}_{0,\text{OBC}}$. Since $\hat{\varepsilon}_{0,\text{OBC}} = \frac{1}{n+\alpha+\beta}\min(N_0+\alpha, N_1+\beta)$, we have $M = \min(N_0 + \alpha, n - N_0 + \beta)$. Notice that $\lfloor A\rfloor$ denote that the largest integer that is not greater than A. Let $I_0 = \left\{\alpha, \alpha+1, \ldots, \lfloor\frac{n+\beta-\alpha}{2}\rfloor + \alpha\right\}$, $I_1 = \left\{\beta, \beta+1, \ldots, \lfloor\frac{n+\alpha-\beta}{2}\rfloor + \beta\right\}$ and $n' = n + \alpha + \beta$. Suppose $\lfloor\alpha\rfloor \neq \lfloor\beta\rfloor$, and it follows that the event $[M = m]$ is equal to the union of the disjoint events $[N_0 = m - \alpha]$, for $m \in I_0$, and $[N_0 = n - m + \beta]$, for $m \in I_1$. By using Proposition 7 in the Appendix A, we can write $E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{\hat{\varepsilon}_{0,\mathrm{OBC}}}\right]$ as:

$$
\begin{aligned}
E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{\hat{\varepsilon}_{0,\mathrm{OBC}}}\right] &= \sum_{m \in I_0} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{m/n'} \mid N_0 = m - \alpha\right] P(N_0 = m - \alpha) + \\
&\quad \sum_{m \in I_1} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{m/n'} \mid N_0 = n - m + \beta\right] P(N_0 = n - m + \beta), \\
&= \sum_{n_{r_1}=0}^{\lfloor \frac{n+\beta-\alpha}{2} \rfloor} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{(n_{r_1} + \alpha)/n'} \mid N_0 = n_{r_1}\right] P(N_0 = n_{r_1}) + \qquad (6.26) \\
&\quad \sum_{n_{r_2}=0}^{\lfloor \frac{n+\alpha-\beta}{2} \rfloor} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{(n_{r_2} + \beta)/n'} \mid N_0 = n - n_{r_2}\right] P(N_0 = n - n_{r_2}) \\
&\quad (n_{r_1}, n_{r_2} \text{ } are \text{ } integers),
\end{aligned}
$$

where

$$
\begin{aligned}
E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{(n_{r_1} + \alpha)/n'} \mid N_0 = t\right] &= \frac{1}{n_{r_1} + \alpha} \sum_{i=1}^{b} \left\{ \sum_{\substack{\frac{(t+\alpha)(k+\alpha_i^0)}{t+\alpha^0} < \frac{(n-t+\beta)(l+\alpha_i^1)}{n-t+\alpha^1} \\ k \le t,\ k+l \le n}} \frac{(t+\alpha)(k+\alpha_i^0)}{t+\alpha^0} + \right. \\
&\left. \sum_{\substack{\frac{(t+\alpha)(k+\alpha_i^0)}{t+\alpha^0} \ge \frac{(n-t+\beta)(l+\alpha_i^1)}{n-t+\alpha^1} \\ k \le t,\ k+l \le n}} \frac{(n-t+\beta)(l+\alpha_i^1)}{n-t+\alpha^1} \right\} P(U_i = k, V_i = l \mid N_0 = t)\,,
\end{aligned}
\qquad (6.27)
$$

with $P(U_i = k, V_i = l \mid N_0 = t)$ expressed in eq. (2.30), for $t = n_{r_1}$, and $E\left[\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{(n_{r_2}+\beta)/n'} \mid N_0 = t\right]$ is formed as the one in eq. (6.27) with $n_{r_1} + \alpha$ replaced with $n_{r_2} + \beta$, for $t = n - n_{r_2}$. It is easy to show that eqs. (6.26) and (6.27) can be applied to the general case associated with α and β.

The second moment of $\hat{\varepsilon}^2_{\mathrm{OBC}}/\hat{\varepsilon}^2_{0,\mathrm{OBC}}$ is given by

$$E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{\hat{\varepsilon}^2_{0,\mathrm{OBC}}}\right] \;=\; \sum_{m\in U} E\left[\left(\frac{\hat{\varepsilon}_{\mathrm{OBC}}}{m/n'}\right)^2 \mid M=m\right] P(M=m)\,, \tag{6.28}$$

where $M = n'\hat{\varepsilon}_{0,\mathrm{OBC}}$, as before. By using Proposition 7 in the Appendix A, and the same reasoning applied previously in the case of the first moment, we further have

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{\hat{\varepsilon}^2_{0,\mathrm{OBC}}}\right] \;&=\; \sum_{m\in I_0} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{m^2/n'^2} \mid N_0 = m-\alpha\right] P(N_0 = m-\alpha)\;+ \\ &\quad\; \sum_{m\in I_1} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{m^2/n'^2} \mid N_0 = n{-}m+\beta\right] P(N_0 = n{-}m+\beta), \\ &=\; \sum_{n_{r_1}=0}^{\lfloor\frac{n+\beta-\alpha}{2}\rfloor} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{(n_{r_1}+\alpha)^2/n'^2} \mid N_0 = n_{r_1}\right] P(N_0 = n_{r_1})\;+ \\ &\quad\; \sum_{n_{r_2}=0}^{\lfloor\frac{n+\alpha-\beta}{2}\rfloor} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{(n_{r_2}+\beta)^2/n'^2} \mid N_0 = n{-}n_{r_2}\right] P(N_0 = n{-}n_{r_2})\,, \end{aligned} \tag{6.29}$$

$$\begin{aligned} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBC}}}{(n_{r_1}+\alpha)^2/n'^2} \mid N_0 = t\right] \;&=\; \frac{1}{(n_{r_1}+\alpha)^2}\times \\ \sum_{i=1}^{b} \Bigg\{ \sum_{l'_i>k'_i} k'^2_j P(U_i = k'_i, V_i = l'_i \mid N_0 = t) \;+\; &\sum_{k\geq l} l'^2_i P(U_i = k'_i, V_i = l'_i \mid N_0 = t) \Bigg\} \;+ \\ \frac{1}{(n_{r_1}+\alpha)^2} \sum_{\substack{i,j=1\\ i\neq j}}^{b} \Bigg\{ &\sum_{l'_i>k'_i}\sum_{s'_j>r'_j} k'_i r'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, V_j = s'_j \mid N_0 = t) + \\ &\sum_{l'_i>k'_i}\sum_{r'_j\geq s'_j} k'_i s'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, V_j = s'_j \mid N_0 = t) + \\ &\sum_{k'_i\geq l'_i}\sum_{s'_j>r'_j} l'_i r'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, V_j = s'_j \mid N_0 = t) + \\ &\sum_{k'_i\geq l'_i}\sum_{r'_j\geq s'_j} l'_i s'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, V_j = s'_j \mid N_0 = t) \Bigg\}\,, \end{aligned} \tag{6.30}$$

with $P(U_i = k, V_i = l \mid N_0 = t)$ as in (2.30) and $P(U_i = k, V_i = l, U_j = r, V_j = s \mid N_0 = t)$ expressed in eq. (2.35), for $t = n_{r_1}$, and $E\left[\frac{\hat{\varepsilon}^2_{\text{OBC}}}{(n_{r_2}+\beta)^2/n'^2} \mid N_0 = t\right]$ is formed as the one in eq. (6.30) with $n_{r_1} + \alpha$ replaced with $n_{r_2} + \beta$, for $t = n - n_{r_2}$.

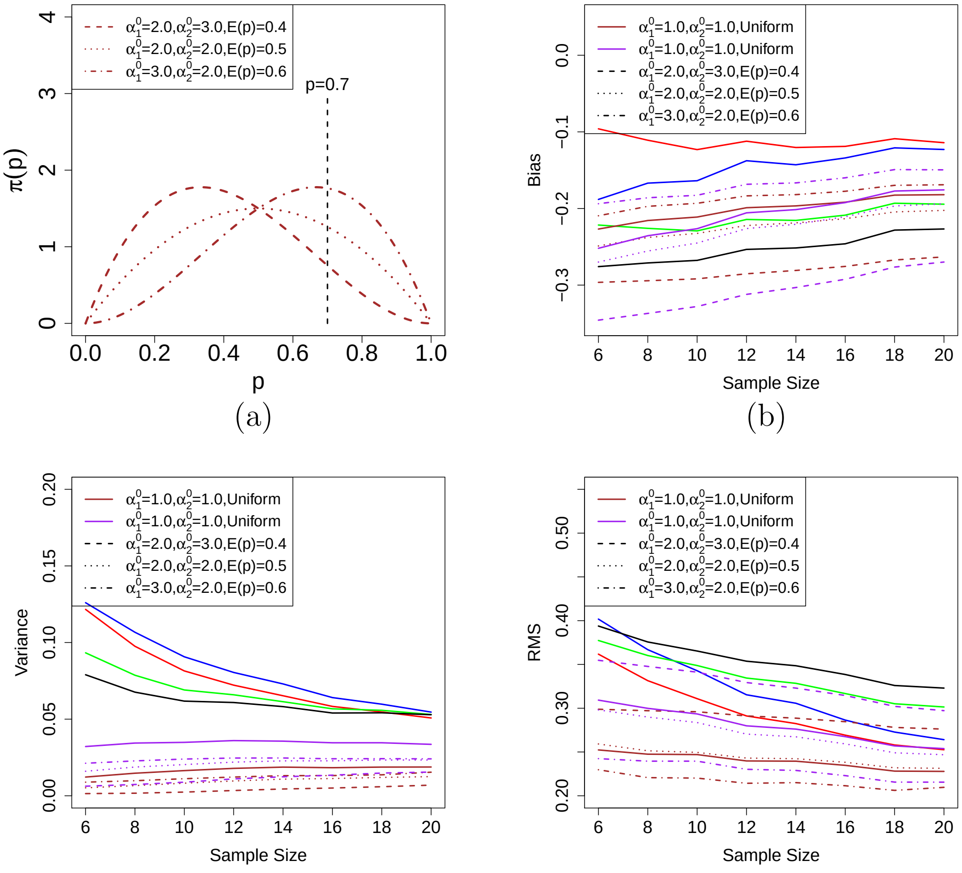
6.4 Performance of Bayesian CoD Estimators

In this Section, we study the performance of two well-defined Bayesian CoD estimators in Section 6.2 in two simulation studies. One study investigates how noninformative and informative priors can affect the performance of Bayesian CoD estimators by considering a discrete distribution with one single predictor (i.e., b =2) and its target, whereas the other study discusses their performance averaged over all the distributions and observes the optimality of the Bayesian MMSE CoD estimation. All the results are compared with the performance of nonparametric CoD estimators like resubstitution, leave-one-out, cross-validation and bootstrap.

6.4.1 Performance Over One Specific Distribution

In this Section, we consider a binary problem with $b = 2$. Let p be the probability for bin 1 with $Y = 0$ and q be the probability for bin 1 with $Y = 1$, that is, $p = p_1 = 1 - p_2$ and $q = q_1 = 1 - q_2$. We assume beta priors for p with hyperparameters α_1^0 and α_2^0. As to the priors for q, we set $\alpha_1^1 = \alpha_2^0$ and $\alpha_2^1 = \alpha_1^0$, and thus $E[p] = 1 - E[q]$.

In our simulations, we fix $p = 0.7$, $q = 0.3$ and $c = 0.5$. We first generate a random non-stratified sample for the sample size of data with $Y = 0$ (n_0) by following the fact that $n_0 \sim Binomial(n, c)$. Then the sample point of each bin ($u_1, \ldots, u_b$ and $v_1, \ldots, v_b$) is assigned by using the binomial or multinomial distribution associated. For each sample, we calculate the Bayesian MMSE CoD estimate, Bayesian CoD estimate based on the optimal Bayesian classifier and all the nonparameteric CoD estimates based on the discrete histogram rule. Finally, we generate 5000 Monte Carlo samples to obtain approximations for the bias, variance and RMS of all the



(a)

(b)

(c)

(d)

Figure 6.1: Bias, variance, and RMS for several CoD estimators vs. sample size over one distribution in the 2-predictor case. Fix $p = 0.7$, $q = 1 - p = 0.3$ and $c = 0.5$. Plot key: bayesian (brown), obc (purple), resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). As comparison, we assume noninformative uniform priors for p, q and c as shown in solid brown and purple lines. In dashed lines, a beta prior for c with $\alpha = \beta = 6.0$ is specified. All results are approximated by Monte Carlo sampling method. Note that computations of the Bayesian MMSE CoD estimates associated with beta priors (in dashed lines) are exact, whereas the Bayesian MMSE CoD estimate with uniform priors for true distributions is approximated with Monte Carlo sampling method.

Bayesian and non-Bayesian CoD estimators. In order to examine how different priors affect the results of Bayesian estimation, both non-informative priors (uniform priors for p, q and c) and informative priors (beta priors for p, q and c) are discussed in our

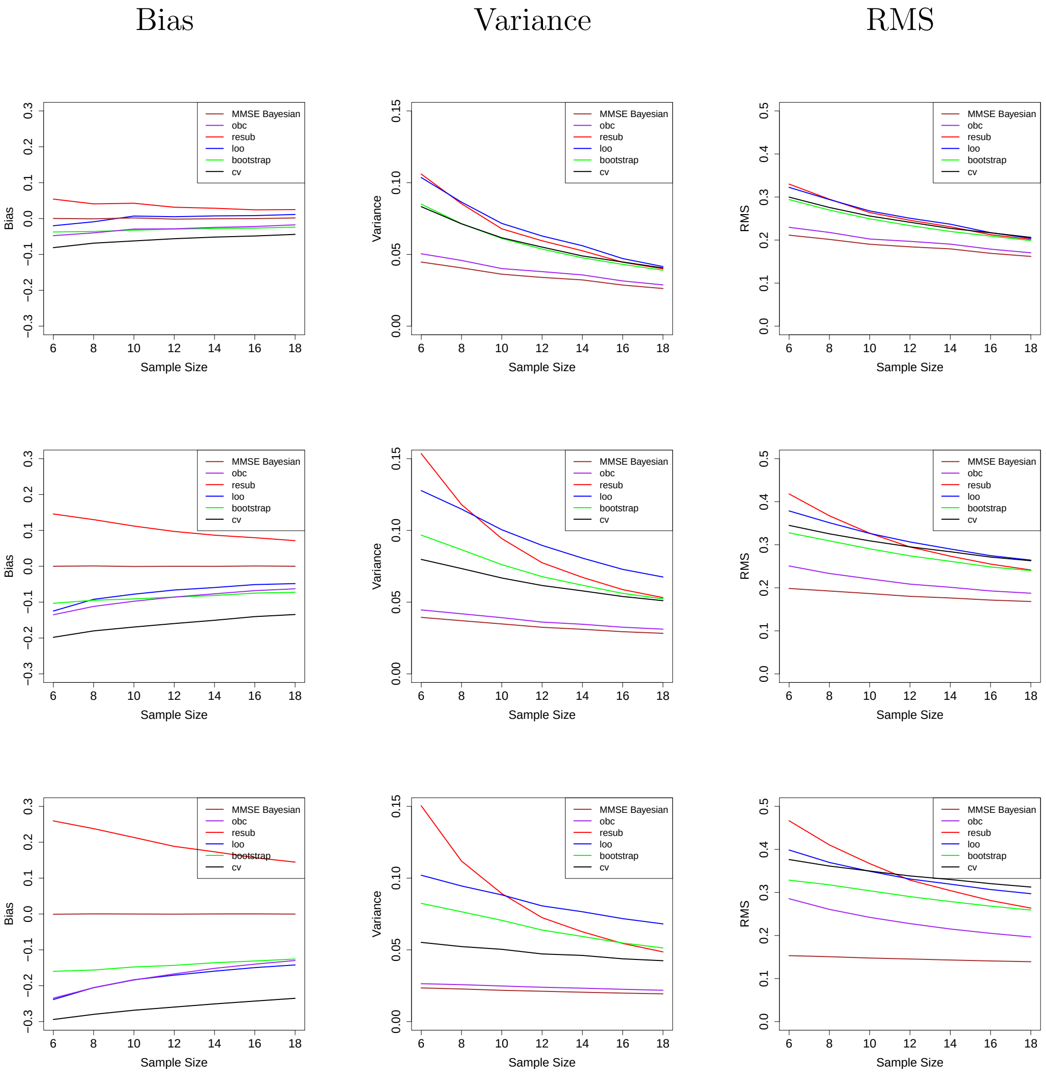


Figure 6.2: Bias, variance, and RMS for several CoD estimators vs. sample size over all distributions. Top row: $b = 2$; Middle row: $b = 4$; Bottom row: $b = 8$ Plot key: bayesian (brown), obc (purple), resubstitution (red), leave-one-out (blue), 0.632 bootstrap (green), 10-repeated 2-fold cross-validation (black). We assume uniform priors for all bin probabilities and a beta distribution $B(\alpha, \beta)$ for c, with $\alpha = b + 1$ and $\beta = b + 1$. All results are approximated by the Monte Carlo sampling method, and computations of the Bayesian MMSE CoD estimates are exact.

studies.

Figure 6.1 shows the bias, variance and RMS of Bayesian MMSE CoD estimator and Bayesian CoD estimator based on the optimal Bayesian classifier associated with various priors (i.e., noninformative uniform priors and informative beta priors) for p and q and their comparison to the performance of nonparametric CoD estimators. Figure 6.1(a) shows the beta priors for p we use in the Bayesian case. It is observed that the prior distribution with $E(p) = 0.8$ has the highest density at the true value of $p = 0.70$. The closer one prior centers at the true distribution, the better estimation it is expected to achieve. As a result, we observe that the prior with a higher density (e.g., the prior with $E[p] = 0.8$ in our simulations) at the true distributions tends to give better performance small(i.e. smaller RMS) than Bayesian CoD estimators with other priors. In addition, when the prior distribution has a smaller density around the true value of p, the performance of Bayesian estimators can be even worse than the resubstitution and leave-one-out. For instance, assuming the prior with $E[p] = 0.4$ as shown in the Figure 6.1, we can see that the resubstitution and leave-one-out converge to the optimal CoD much faster than the Bayesian ones regarding the RMS. Among the Bayesian CoD estimators associated with various priors, the one based on the prior with $E[p] = 0.6$ has the highest bias in amplitude and the least variance, whereas the Bayesian CoD estimator with uniform priors has the largest variance. As a summary, we can forecast that, with available knowledge about true distributions $\mathbf{p}, \mathbf{q}$ and c in the d-predictor case (with $b = 2^d$), the priors with higher densities around these true distributions are preferred for better estimation of the CoD.

6.4.2 Performance Over All Distributions

Following the simulation studies in [26], we compute the performance metrics of the Bayesian CoD estimator, for a given sample size, over all distributions in the probability model, with a beta prior for target probability c and uniform priors for the bin probabilities $(\mathbf{p}, \mathbf{q})$. This is done by the Monte Carlo sampling method drawing $M = 10000$ simulated training data sets of the required sample size from the probability model in two steps. In the first step, we randomly generate the true distributions of c and $(\mathbf{p}, \mathbf{q})$ based on the assumptions of priors, and then, in the second step, collect samples that are randomly generated according to the current distributions. Given sample data, we can compute the exact Bayesian MMSE CoD estimate as expressed in Section 6.2, as well as obtain Monte Carlo approximations of nonparametric CoD estimators such as resubstitution, leave-one-out, bootstrap and cross-validation. Based on a large number of simulated experiments, sample means and sample variances are employed to approximate the performance metrics.

Figure 6.2 shows the comparison results between the performance of the Bayesian CoD estimator and that of the other four nonparametric CoD estimators, as a function of varying sample size, for difference bin sizes $b = 2$, $b = 4$ and $b = 8$. Several observations are made in what follows. First, as expected, the Bayesian CoD estimator is observed to perform the best, given its unbiasedness and least RMS, when averaged over all distributions. Secondly, the leave-one-out CoD estimator has the second-best performance according to RMS when averaged over all the distributions, whereas we know from a previous publication that the resubstitution performs best among the nonparametric list for a fixed model [21]. Last but not least, as the sample size or bin size increases, the performance of the Bayesian CoD estimator has obvious improvement over the others.

6.5 Applications to System Identification Problems

By following the problems of system identification in Section 3, we consider in this section the inference of gene regulatory relationships with partial knowledge about the logic gates regulating each target variable but no knowledge about the wiring associated with each logic gate. We propose inference procedures based on the two proposed Bayesian estimators (i.e., Bayesian MMSE CoD estimator and Bayesian CoD estimator based on the optimal Bayesian classifier) to recover both wiring and logic information. In the case of wiring recovery, we compare the performance of two Bayesian approaches against the use of the parametric ML approach and nonparametric approaches, whereas, in the case of logic gate recovery, we compare the performance of Bayesian approaches with the ML one and the resubstitution among nonparametric approaches. Notice that the nonparametric CoD estimators are not capable of taking advantage of the available incomplete knowledge, which only depend on the discrete histogram rule to decide on the logic prediction.

We consider the static case only. Like what is described in Section 3, we also consider nested sets of candidate models, from more (smaller set) to less (larger set) informative, in the simulated numerical examples in this section, which allows us to investigate how the amount of prior knowledge can affect inference accuracy.

We consider here inference of the Boolean function f, or predictor, in the static model (3.12). It is assumed that the true predictor f is unknown but is a member of a candidate model set F containing several Boolean functions, as mentioned in Section 3. Again we assume that each predictor f in F depends on the same number l of essential predictive variables, or inputs. It is assumed that the model set F consists of a number c of possible logic gates and arbitrary wiring of connectivity l. The larger c is, the less is known about the system.

We propose the following Bayesian predictor inference procedure to select a predictor from F based on Bayesian approaches. For each target and its d-predictor set, we assume Dirichlet distributions for class-conditional probabilities $(p_1, \ldots, p_{2^d-1})$, $(q_1, \ldots, q_{2^d-1})$ and a beta distribution of class 0 probability c, as mentioned in eqs. (6.2).

1. For each logic gate, specify the hyperparameters of priors in eqs. (6.2) and then pick the wiring that produces the largest MMSE Bayesian CoD estimate / Bayesian CoD estimate based on the optimal Bayesian classifier. Ties, if any, are broken randomly.

2. Among the c candidate predictors obtained from the previous step, select the one that presents the largest predictive power estimate. Ties, if any, are broken randomly.

Notice that the specification of the hyperparameters of priors is very important since an informative prior will probably lead to a good Bayesian CoD estimator that better recovers the regulatory relationship between one target and its predictors. A detailed discussion of initiation of those hyperparameters will be given regarding the numerical experiments in the following section. Moreover, we will make assessment of the effectiveness of our propose inference procedures based on the Bayesian approaches by means of numerical experiments.

6.5.1 Numerical Experiments

In this section, we follow the numerical experiment settings as discussed in Section 3.4.1.1, where the static model in eq. (3.12) is employed.

6.5.1.1 Experimental Settings

We let $d = 8$ and set up two groups of experiments, corresponding to $l = 2, 3$. A set of $k = 8$ models are considered in each case, each model being obtained by a random wiring assignment $\{i_1, \dots, i_l\}$ and a choice of a logic gate:

- $l = 2$: $g(X_{i_1}, X_{i_2}) = X_{i_1} \oplus X_{i_2}$.

- $l = 3$: $g(X_{i_1}, X_{i_2}, X_{i_3}) = X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$.

- $l = 4$: $g(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}) = X_{i_1} \oplus X_{i_2} \oplus X_{i_3} \oplus X_{i_4}$.

In addition, two different values of predictive power ($p = 0.75$ and $p = 0.85$) are considered. For each value of l, three nested candidate model sets $F_l^1 \subset F_l^2 \subset F_l^3$ are employed, each containing all $\binom{8}{l}$ possible predictor variable assignments $\{i_1, \ldots, i_l\}$, for $l = 2, 3, 4$, and the logic gates depicted in Tables 6.1 and 6.2.

Table 6.1: Logic gates for candidate model sets, static case, $l = 2$.

F_2^1	F_2^2	F_2^3
$X_{i_1} \oplus X_{i_2}$	$X_{i_1} X_{i_2}$	$X_{i_1} X_{i_2}$
	$X_{i_1} \oplus X_{i_2}$	$X_{i_1} \oplus X_{i_2}$
	$X_{i_1} \overline{X_{i_2}}$	$X_{i_1} \overline{X_{i_2}}$
	$\overline{X_{i_1}} + X_{i_2}$	$\overline{X_{i_1}} + X_{i_2}$
		$\overline{X_{i_1}} X_{i_2}$
		$X_{i_1} + \overline{X_{i_2}}$

6.5.1.2 *Specification of Hyperparameters of Priors*

Given a $l-$input stochastic logic model, class condidtional probabilities $\mathbf{p}$ and $\mathbf{q}$ and class 0 probability c are functions of predictive power p and joint distributions of predictors. Assuming the uniformity of predictors, we can easily show that:

$$\begin{aligned} p_i &= \frac{pI_{f(X=i)=0} + (1-p)I_{f(X=i)=1}}{\sum_{i=1}^{2^l} pI_{f(X=i)=0} + (1-p)I_{f(X=i)=1}} \; , i = 1, \ldots, 2^l \\ q_i &= \frac{pI_{f(X=i)=1} + (1-p)I_{f(X=i)=0}}{\sum_{i=1}^{2^l} pI_{f(X=i)=1} + (1-p)I_{f(X=i)=0}} \; , i = 1, \ldots, 2^l \\ c &= \frac{1}{2^l} \sum_{i=1}^{2^l} pI_{f(X=i)=0} + (1-p)I_{f(X=i)=1} \, . \end{aligned} \tag{6.31}$$

For improved Bayesian estimation, the choice of priors for $\mathbf{p}$, $\mathbf{q}$ and c is desired to concentrate their densities at the true values of $\mathbf{p}$, $\mathbf{q}$ and c in eqs. (6.31), as concluded in Section sec:pm-fix. In practice, the model parameter p is not known,

Table 6.2: Logic gates for candidate model sets, static case, $l = 3$.

F_3^1	F_3^2	F_3^3
$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}} X_{i_3} + X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}} X_{i_3} + X_{i_2} \oplus X_{i_3}$
	$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$	$X_{i_1} \oplus X_{i_2} \oplus X_{i_3}$
	$\overline{X_{i_1}}(X_{i_2} \oplus X_{i_3}) + X_{i_1}\overline{X_{i_2}}$	$\overline{X_{i_1}}(X_{i_2} \oplus X_{i_3}) + X_{i_1}\overline{X_{i_2}}$
	$\overline{X_{i_1}}\,\overline{X_{i_2}} + X_{i_1}X_{i_2} \oplus X_{i_3}$	$\overline{X_{i_1}}\,\overline{X_{i_2}} + X_{i_1}X_{i_2} \oplus X_{i_3}$
		$\overline{X_{i_1}} X_{i_2} + X_{i_1}\overline{X_{i_2} \oplus X_{i_3}}$
		$\overline{X_{i_1}} X_{i_3} + \overline{X_{i_2} \oplus X_{i_3}}$
		$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_2}} + X_{i_1}X_{i_2}$
		$\overline{X_{i_1}}\,\overline{X_{i_2} \oplus X_{i_2}} + X_{i_1}(X_{i_2} \oplus X_{i_3})$
		$\overline{X_{i_1}}\,\overline{X_{i_3}} + X_{i_1}(X_{i_2} \oplus X_{i_3})$
		$\overline{X_{i_1}}\,\overline{X_{i_3}} + X_{i_1}\overline{X_{i_2} \oplus X_{i_3}}$

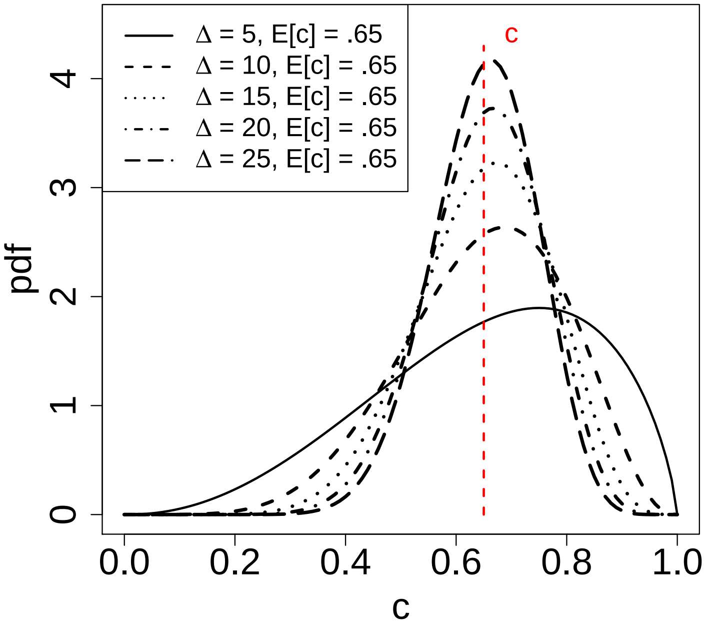


Figure 6.3: An example of probability distribution functions of beta priors for the class 0 probability c in a 2-input AND logic model for varying Δ. Set $p = 0.8$ and thus $c = \frac{2p+1}{4} = 0.65$.

Table 6.3: Specification of hyperparameters of priors for $\mathbf{p}$ using sample data drawn from the static model in the 2-predictor case

Logic	$(\alpha_1^0, \ldots, \alpha_4^0)$
AND	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil\right)$
NAND	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil\right)$
XOR	$\left(\lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil\right)$
NXOR	$\left(\lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil\right)$
$X_1 + \bar{X}_2$	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$
$\bar{X}_1 X_2$	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$
$X_1 \bar{X}_2$	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$
$X_2 + \bar{X}_1$	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$
OR	$\left(\lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$
NOR	$\left(\lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$

which, however, can be estimated from sample data drawn from the given logic model. By using the maximum-likelihood estimation approach, p can be estimated as a function of sample data, that is, $\hat{p}$, as shown in eq. (3.21). By substituting $\hat{p}$ into eqs. (6.31), we can obtain $\hat{\mathbf{p}}, \hat{\mathbf{q}}$ and $\hat{c}$ as a function of $\hat{p}$. To adjust the shape of concentration, we multiply $\hat{\mathbf{p}}, \hat{\mathbf{q}}$ and $\hat{c}$ with a factor Δ, and take $\lceil\hat{p}_1\Delta\rceil, \ldots, \lceil\hat{p}_{2^l}\Delta\rceil$, $\lceil\hat{q}_1\Delta\rceil, \ldots, \lceil\hat{q}_{2^l}\Delta\rceil$ and $\lceil\hat{c}\Delta\rceil$ as the hyperparameter values of these priors. Note that $\lceil x\rceil$ gives the smallest integer that is not less than x. Here we Tables 6.3–6.5 presents the specification of hyperparameters of priors for $\mathbf{p}, \mathbf{q}$ and c based on sample data drawn from the 2-input stochastic logic model. To examine how Δ affects the concentration of priors, Figure 6.3 shows the probability distribution of the beta prior for the class 0 probability c for a varying factor Δ by considering a 2-input AND logic model. It is observed that, as the factor Δ increases, the distribution tends to center at the true value of $c = 0.65$. The larger the Δ is, the lower variance

Table 6.4: Specification of hyperparameters of priors for $\mathbf{q}$ using sample data drawn from the static model in the 2-predictor case

Logic	$(\alpha_1^1, \ldots, \alpha_4^1)$
AND	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil\right)$
NAND	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil\right)$
XOR	$\left(\lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil\right)$
NXOR	$\left(\lceil\frac{\Delta\hat{p}}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2}\rceil, \lceil\frac{\Delta\hat{p}}{2}\rceil\right)$
$\bar{X}_1 X_2$	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$
$X_1 + \bar{X}_2$	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$
$X_1 \bar{X}_2$	$\left(\lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$
$X_2 + \bar{X}_1$	$\left(\lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$
OR	$\left(\lceil\frac{\Delta(1-\hat{p})}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil, \lceil\frac{\Delta\hat{p}}{2\hat{p}+1}\rceil\right)$
NOR	$\left(\lceil\frac{\Delta\hat{p}}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil, \lceil\frac{\Delta(1-\hat{p})}{3-2\hat{p}}\rceil\right)$

the prior presents. In our simulations, we set $\Delta = 10$.

6.5.1.3 Simulation Results

For each number of inputs l, predictive power p, and sample size n, a total of $r = 50$ datasets are drawn from each model. After applying the proposed Bayesian inference procedures, we record the average percentage of correctly-recovered logic gates and the average percentage of correct predictive variables for each of the three candidate model sets, as shown in Section 3.4.1.1. Moreover, we compare these results with those of using the nonparametric and parametric CoD estimators in the inference procedures in Section 3.4.1.1. Notice that, for the quickness in producing results, we employ the Monte Carlo sampling method to obtain the Bayesian MMSE CoD estimates throughout all the simulation studies in this Section and such approximations have been checked to guarantee good accuracy.

Table 6.5: Specification of hyperparameters of priors for c using sample data drawn from the static model in the 2-predictor case

Logic	(α, β)
AND	$\left(\lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil, \lceil\frac{\Delta(3-2\hat{p})}{4}\rceil\right)$
NAND	$\left(\lceil\frac{\Delta(3-2\hat{p})}{4}\rceil, \lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil\right)$
XOR	$\left(\lceil\frac{\Delta}{2}\rceil, \lceil\frac{\Delta}{2}\rceil\right)$
NXOR	$\left(\lceil\frac{\Delta}{2}\rceil, \lceil\frac{\Delta}{2}\rceil\right)$
$\bar{X}_1X_2$	$\left(\lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil, \lceil\frac{\Delta(3-2\hat{p})}{4}\rceil\right)$
$X_1 + \bar{X}_2$	$\left(\lceil\frac{\Delta(3-2\hat{p})}{4}\rceil, \lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil\right)$
$X_1\bar{X}_2$	$\left(\lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil, \lceil\frac{\Delta(3-2\hat{p})}{4}\rceil\right)$
$X_2 + \bar{X}_1$	$\left(\lceil\frac{\Delta(3-2\hat{p})}{4}\rceil, \lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil\right)$
OR	$\left(\lceil\frac{\Delta(3-2\hat{p})}{4}\rceil, \lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil\right)$
NOR	$\left(\lceil\frac{\Delta(2\hat{p}+1)}{4}\rceil, \lceil\frac{\Delta(3-2\hat{p})}{4}\rceil\right)$

Figure 6.4 – 6.7 display the results as a function of sample size, corresponding to the three candidate model sets $F_l^1 \subset F_l^2 \subset F_l^3$, for $l = 2, 3$ and $p = 0.75, 0.85$. Several observations are made in the following.

- As the sample size increases, the performance of the two Bayesian methods increases accordingly. Obviously, the more prior knowledge we know, the more quickly their performance converges to 100%. The same results apply to the other methods.

- It is observed that, in the 2-predictor case, the performance of the Bayesian-based inference methods is very close to the ML-based one, and they all beat the nonparametric methods. As the number of predictors (l) increases (e.g. $l = 3$), the performance of the ML-based inference method performs better than the two Bayesian methods over the sample size.

- We can see in Figures 6.6 and 6.7 that, when $l = 2$, the parametric ML-based

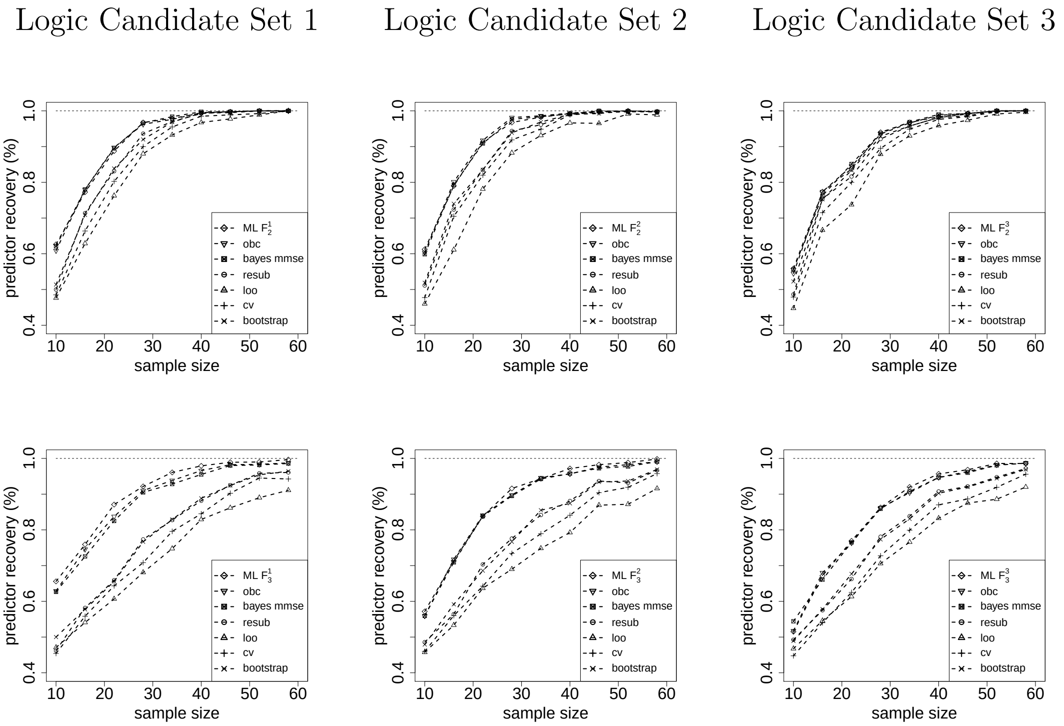


Figure 6.4: Percentage of predictor recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.85. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method.

inference is superior to that of the Bayesian methods for very small sample size (e.g., $n = 10$). As the sample sizes increases, the Bayesian methods start to outperform the ML-based one only by very little improvement. When $l = 3$, the ML approach performs better than the Bayesian approaches, which is more obvious for a smaller predictive power value $p = 0.75$.

- We can see that the performance of the Bayesian-based inference methods improve as more prior knowledge is available since the specification of hyperparameters of priors can take more advantage of the prior knowledge by allowing the prior distributions to center at true distributions $(\mathbf{p}, \mathbf{q}, c)$.

- In the case of larger dimensionality of the predictor vector (e.g., $l = 3$), it

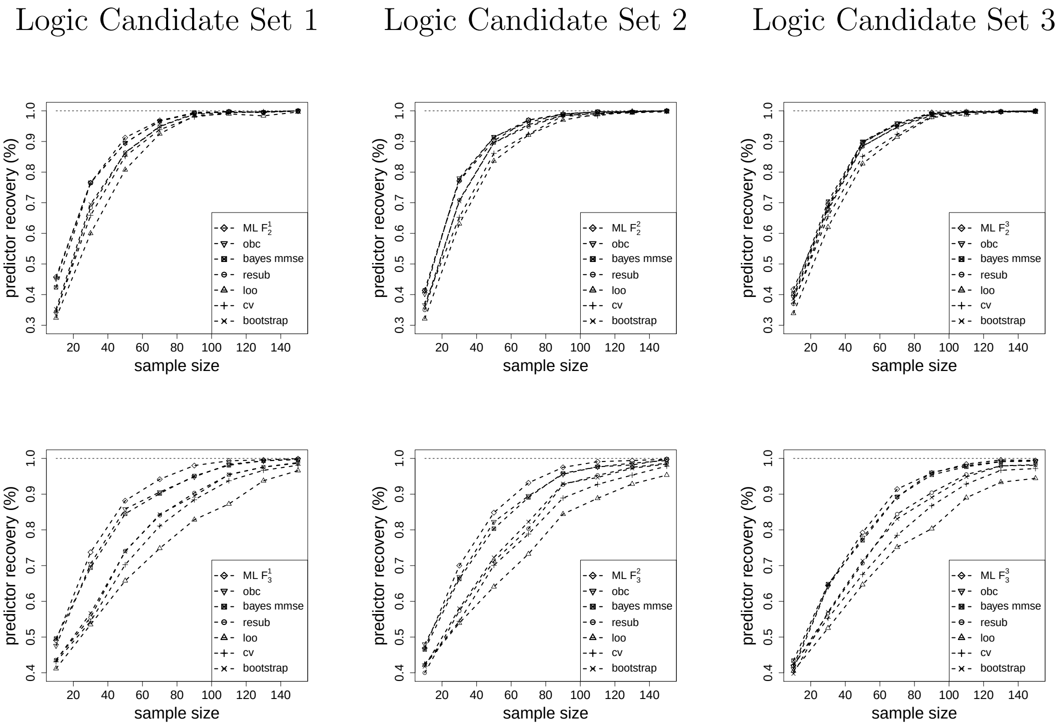


Figure 6.5: Percentage of predictor recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.75. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method.

is more obvious that both Bayesian and ML-based approaches are superior to nonparametric approaches, since the former both take advantage of prior knowledge about gene regulation.

6.6 Summary

In this paper, we have introduced a Bayesian framework to estimate the CoD in discrete prediction settings and its applications to inference problems in Genomics. We have defined two Bayesian CoD estimators, one from a MMSE perspective and the other based on the optimal Bayesian classifier. We have derived exact analytical expressions of the Bayesian MMSE CoD estimator that optimizes CoD estimation with respect to MSE, across a family of target-predictor distributions, and exact

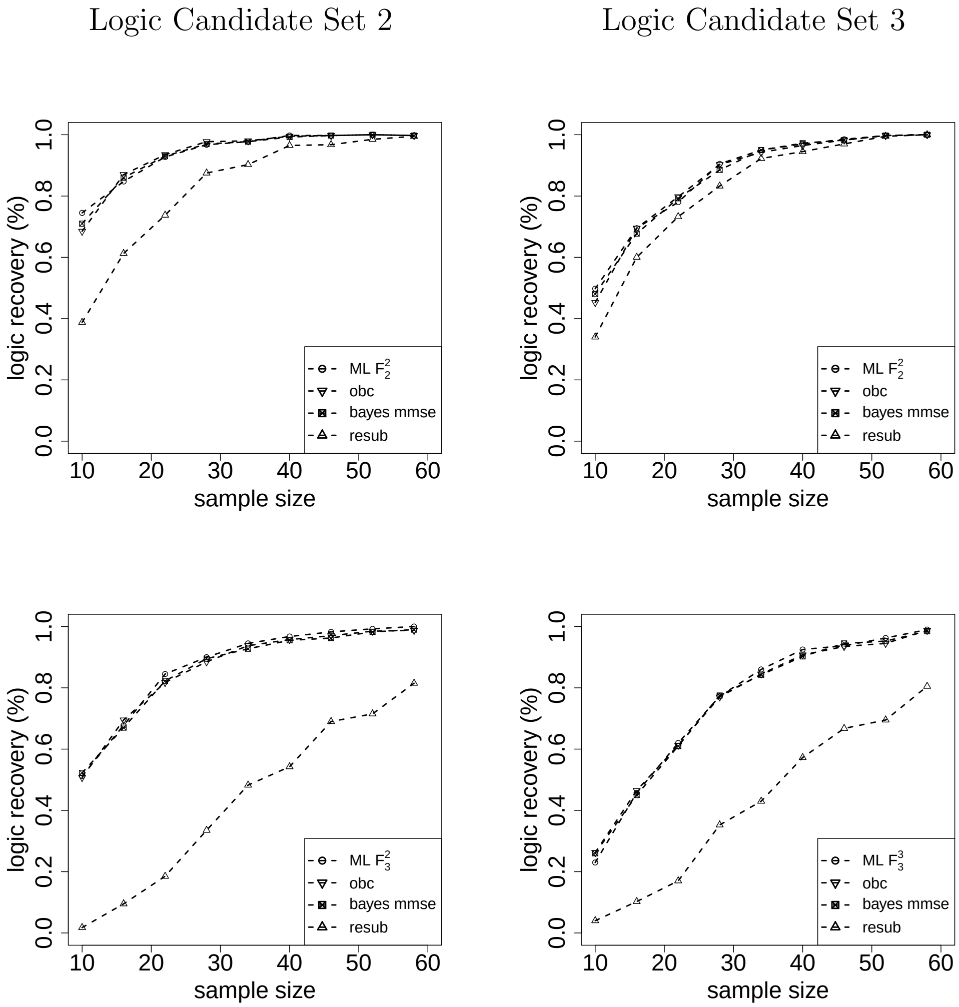


Figure 6.6: Percentage of logic recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.85. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method.

formulas for the performance metrics (i.e., bias, variance and RMS) of the Bayesian CoD estimator based on the optimal Bayesian classifier. We have compared the performance metrics of the two Bayesian CoD estimators against those of resubstitution, leave-one-out, bootstrap and cross-validation CoD estimators over all the distributions and over one specific distribution, by means of Monte Carlo sampling experiments. Our results demonstrate that the Bayesian MMSE CoD estimator has

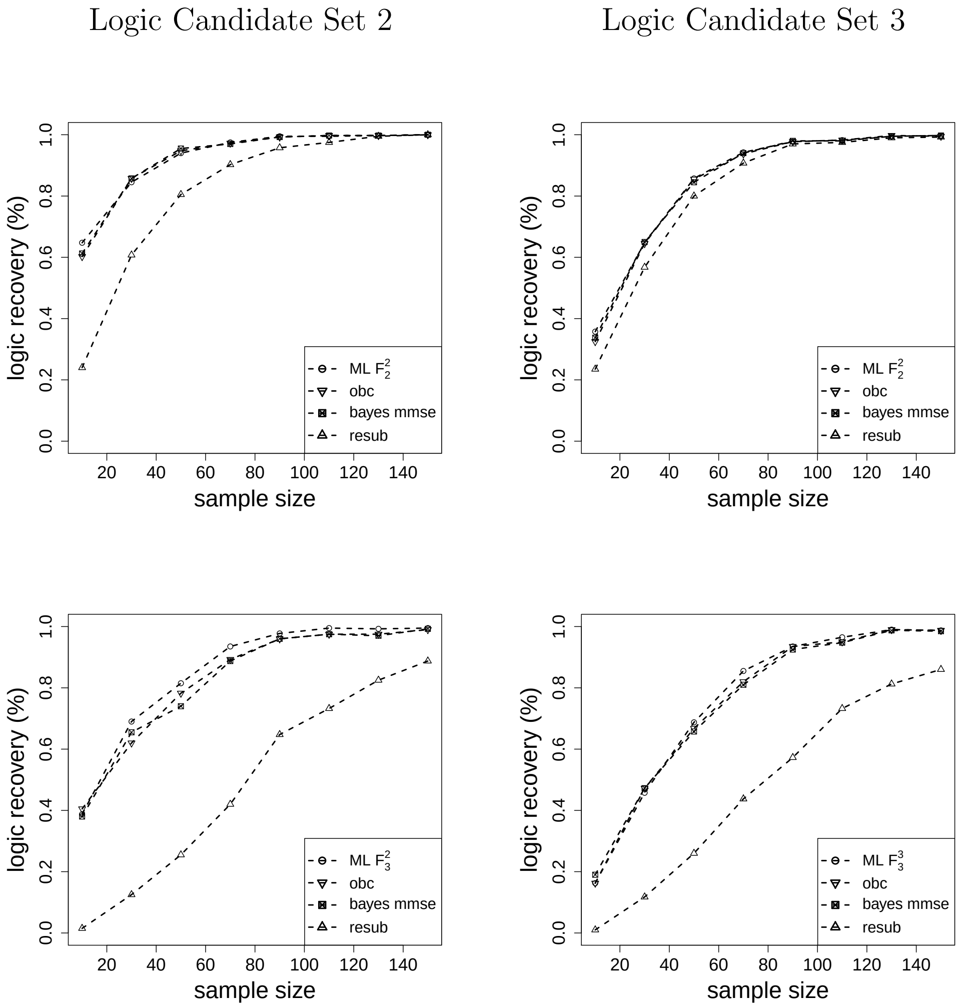


Figure 6.7: Percentage of logic recovery vs. sample size. Top row: $b = 2$; Bottom row: $b = 3$. Predictive power p is set to be 0.75. The Bayesian MMSE CoD estimator is approximated by the Monte Carlo sampling method.

excellent performance with zero bias and least RMS, when averaged over all distributions and sample data. According to results with respect to one specific distribution, we conclude that priors with higher densities around true distributions present better performance with less RMS.

We have studied the applications of CoD estimation to the inference of gene regulatory relationships based on sample microarray data, from a frequentist viewpoint [16]. Likewise, we have proposed predictor inference procedures based on Bayesian CoD estimators for the recovery of both wiring and logic gates of target and predictor genes of interest. We address the issue of incorporation of prior knowledge in the Bayesian setting by specifying the hyperparameters of priors from sample data with a possible list of candidate models. Therefore, we have made the unsurprising observation that the proposed Bayesian procedures give better prediction than the ones using nonparametric CoD estimators such as resubstitution, leave-one-out, cross-validation and bootstrap, and present very close results to the ML-based inference procedures that also allow the inclusion of prior knowledge.

7. CONCLUSION

In this dissertation, we have presented a comprehensive study of the inference of the discrete CoD from both frequentist and Bayesian perspectives, with the applications to the system identification problems in Genomics. In addition, we develop two promising statistics tools for the detection of multivariate gene regulatory relationships and canalyzing genes of statistical significance, respectively. We make significant contributions in this dissertation by not only enriching the theoretical understanding of inference problems of the discrete CoD but also improving the applications of the CoD to the inference of multivariate gene regulatory relationships in practice.

First, we define the sample-based nonparametric CoD estimators from a frequentist perspective, and derive exact analytical expressions of performance metrics of the resubstitution and leave-one-out CoD estimators. Using a parametric Zipf model, we have compared the exact performance metrics of resubstitution and leave-one-out between each other and against approximate performance metrics of cross-validation and bootstrap CoD estimators. Our results indicate that, provided one has evidence of moderate to tight regulation between the genes, and the number of predictors is not too large, one should use the CoD estimator based on resubstitution.

Secondly, we have presented a systematic theoretical framework for the inference of the CoD based upon a parametric maximum-likelihood approach, with its applications to estimation and system identification for static and dynamical Boolean models. Inference algorithms are proposed for both static and dynamic cases to recover gene regulatory relationships (i.e., wiring and logic gates). Analytical and numerical results show that the parametric ML CoD estimator outperforms the nonparametric alternatives when sufficient prior knowledge is available and the system noise level is not too high. The performance gap is larger for smaller sample sizes and larger dimensionality of the predictor vectors, in which situations the estimation via the parametric approach can be ameliorated by the use of prior knowledge. In addition, as less prior knowledge was available, the performance of the parametric and nonparametric ML CoD estimators were observed to equalize. This suggests that, in the no-information case, the NPML estimator (i.e. resubstitution estimator) would be preferred, due to its low computational complexity.

Thirdly, we have described a rigorous statistical testing framework to investigate regulatory relationships among genes, by using the discrete Coefficient of Determination (CoD), and to discover canalyzing genes by using the intrinsically multivariate prediction (IMP). This marks a significant change in the application of the CoD to such problems, since thus far its use depended on user-selected thresholds to characterize the presence of significant relationships or canalyzing genes. Multiple-testing procedures are also described, which make the methodology applicable to large data sets. Furthermore, software that implements the CoD test is made available to the scientific community as an R *codtest* package through our website (http://gsp.tamu.edu/Publications/supplementary/ting13a), and the R *imptest* package for the IMP test is available at our website (http://gsp.tamu.edu/Publications/supplementary/ting13c). It is expected that this methodology will be a useful practical tool for the inference of gene regulatory relationships and canalyzing genes from gene-expression data.

Finally, we have proposed a Bayesian estimation framework for the inference of CoD across a parametrized family of joint distributions between target and predictors, where the prior distribution of the parameters are desired to concentrate around the true distributions. We have shown that the Bayesian CoD estimator that achieves minimum mean-square error between one CoD estimator and the optimal CoD possesses the best performance when averaged over a given family of distributions and sample data. We also define another Bayesian CoD estimator based on the optimal Bayesian classifier, which performs better than the four nonparametric CoD estimators but worse than the Bayesian MMSE one. Moreover, inference algorithms based on these Bayesian CoD estimators have been developed to recover the gene regulatory relationships (i.e., wiring and logic gates) by using the discrete gene-expression data. Results show that the Bayesian inference algorithms are very comparable to the ML-based algorithms that could take advantage of available prior knowledge.

In conclusion, this dissertation is intended to serve as foundation for a detailed study of the application of CoD estimation in Genomics and related fields. An obvious application is the inference of genomic regulatory networks from sample microarray data, as discussed here. In addition to that, there are several issues related to nonlinear prediction in the discrete domain, which can benefit from the work presented here. Still there are several important problems to be investigated, as summarized in the following:

- Regarding the maximum-likelihood inference of the discrete CoD in dynamical systems, future investigations should include the extension to suitably-constrained nonstationary dynamical systems, as well as the comparison to alternative approaches for small-sample inference of discrete systems, such as discrete Bayesian networks [40].

- The Bayesian approach to hypothesis testing of the discrete CoD should be studied to take model uncertainty into account [7, 8, 42]. What also deserves careful investigation is the parametric model we could use, appropriate priors of parameters we could specify for possible closed-form solutions and the calculation of the Bayes factor for the formulation of one Bayesian test with its applications to detection of significant gene regulatory relationships in Genomics problems.

REFERENCES

[1] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.

[2] R. Albert and H. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 1–18, 2003.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–230, 2001.

[4] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[5] R. BERGER, "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, vol. 24, no. 1, pp. 295–300, 1982.

[6] R. L. Berger, "Likelihood ratio tests and intersection-union tests," in *Advances in statistical decision theory and applications.* New York, NY: Springer, 1997, pp. 225–237.

[7] J. M. Bernardo and J. M. Ramón, "An introduction to bayesian reference analysis: inference on the ratio of multinomial parameters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 1, pp. 101–135, 1998.

[8] J. M. Bernardo and R. Rueda, "Bayesian hypothesis testing: A reference approach," *International Statistical Review*, vol. 70, no. 3, pp. 351–372, 2002.

[9] S. Bornholdt, "Boolean network models of cellular regulation: prospects and limitations," *J. R. Soc. Interface*, vol. 5, no. 1, pp. S85–S94, 2008.

[10] U. Braga-Neto and E. Dougherty, "Exact performance of error estimators for discrete classifiers," *Pattern recognition*, vol. 38, no. 11, pp. 1799–1814, 2005.

[11] U. M. Braga-Neto, "Classification and error estimation for discrete data," *Current genomics*, vol. 10, no. 7, p. 446, 2009.

[12] R. L. Burden and J. D. Faires, "Numerical analysis," *Pacific Grove, CA: Brooks Cole*, 2001.

[13] M. Camps, A. Nichols, and S. Arkinstall, "Dual specificity phosphatases: a gene family for control of map kinase function," *The FASEB Journal*, vol. 14, no. 1, pp. 6–16, 2000.

[14] G. Casella and R. L. Berger, *Statistical inference*. Pacific Grove, CA: Duxbury, 2002.

[15] L. Chang and M. Karin, "Mammalian map kinase signalling cascades," *Nature*, vol. 410, no. 6824, pp. 37–40, 2001.

[16] T. Chen and U. Braga-Neto, "Maximum-likelihood estimation of the discrete coefficient of determination in stochastic boolean systems," *IEEE Transactions on Signal Processing*, vol. 61, no. 15, pp. 3880–3894, 2013, doi:10.1109/TSP.2013.2264054.

[17] ——, "Statistical detection of boolean regulatory relationships," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, 2013, submitted.

[18] ——, “Statistical detection of intrinsically multivariate predictive genes,” *Bioinformatics*, vol. PP, 2013, submitted.

[19] ——, “Optimal bayesian mmse estimation of the coefficient of determination in discrete prediction,” in *proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS’2013), Houston, TX, November 2013.*

[20] ——, “Approximate expressions for the variances of non-randomized error estimators and cod estimators for the discrete histogram rule,” in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS’2010), Cold Spring Harbor, NY, November 2010.* IEEE, 2010, pp. 1–4.

[21] ——, “Exact performance of cod estimators in discrete prediction,” *EURASIP Journal on Advances in Signal Processing (JASP), Special Issue on Genomic Signal Processing*, vol. 2010, pp. 1–13, 2010, doi:10.1155/2010/487893.

[22] ——, “Maximum likelihood estimation of the binary coefficient of determination,” in *proceedings of 45th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA.* IEEE, 2011, pp. 1012–1016.

[23] ——, “Sample-based estimators for the instrinsically multivariate prediction score,” in *proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS’2011), San Antonio, TX.* IEEE, 2011, pp. 139–142.

[24] ——, “A statistical test for intrinsically multivariate genes,” in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS’2012), Washington, DC, December 2012.* IEEE, 2012, pp. 151–154.

[25] C. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial,” *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[26] L. A. Dalton and E. R. Dougherty, “Bayesian minimum mean-square error estimation for classification error part i: Definition and the bayesian mmse error estimator for discrete classification,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 1, pp. 115–129, 2011.

[27] ——, “Bayesian minimum mean-square error estimation for classification error part ii: Linear classification of gaussian models,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 1, pp. 130–144, 2011.

[28] ——, “Optimal classifiers with minimum expected error within a bayesian framework–part i: Discrete and gaussian models,” *Pattern Recognition*, vol. 46, no. 5, pp. 1301–1314, 2012.

[29] ——, “Optimal classifiers with minimum expected error within a bayesian framework–part ii: Properties and performance analysis,” *Pattern Recognition*, vol. 46, no. 5, pp. 1288–1300, 2012.

[30] L. Devroye, *A probabilistic theory of pattern recognition.* New York, NY: Springer, 1996, vol. 31.

[31] E. R. Dougherty, S. Kim, and Y. Chen, “Coefficient of determination in nonlinear signal processing,” *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.

[32] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* New York, NY: John Wiley & Sons, 2012.

[33] S. Dudoit, J. Shaffer, and J. Boldrick, “Multiple hypothesis testing in microarray experiments,” *Statistical Science*, vol. 18, no. 1, pp. 71–103, 2003.

[34] S. Dudoit and M. van der Laan, *Multiple testing procedures with applications to genomics.* New York, NY: Springer, 2008.

[35] B. Efron, “Bootstrap methods: another look at the jackknife,” *The annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[36] ——, “Estimating the error rate of a prediction rule: improvement on cross-validation,” *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[37] Y. L. Q. O. F. Li, T. Long and C. Tang, “The yeast cell-cycle network is robustly designed,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 14, pp. 4781–4876, 2004.

[38] A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry, “Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle,” *Bionformatics*, vol. 22, no. 14, pp. 124–131, 2006.

[39] M. Gomez-Lazaro, F. Fernandez-Gomez, and J. Jordan, “p53: twenty five years understanding the mechanism of genome protection,” *Journal of physiology and biochemistry*, vol. 60, no. 4, pp. 287–307, 2004.

[40] D. Heckerman and D. Geiger, “Learning bayesian networks: a unification for discrete and gaussian domains,” in *Proceedings of the eleventh conference on uncertainty in artificial intelligence.* Montreal, Quebec: Morgan Kaufmann, 1995, pp. 274–284.

[41] J. J. Hunter, “A survey of generalized inverses and their use in stochastic modelling,” *Research Letters in the Information and Mathematical Sciences*, vol. 1, no. 1, pp. 25–33, 2000.

[42] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[43] S. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets." *Journal of theoretical biology*, vol. 22, no. 3, pp. 437–467, 1969.

[44] ——, *The origins of order: Self organization and selection in evolution.* New York, NY: Oxford University Press, 1993.

[45] S. Keyse, "Dual-specificity map kinase phosphatases (mkps) and cancer," *Cancer and Metastasis Reviews*, vol. 27, no. 2, pp. 253–261, 2008.

[46] S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, and J. M. Trent, "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," *Journal of biomedical optics*, vol. 5, no. 4, pp. 411–424, 2000.

[47] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner, "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, no. 2, pp. 201–209, 2000.

[48] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[49] B. Lehner, "Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast," *PLoS One*, vol. 5, no. 2, pp. 9035–9035, 2010.

[50] L. Ljung, *System identification.* New York, NY: Prentice Hall, 1999.

[51] S. Marshall, L. Yu, Y. Xiao, and E. R. Dougherty, "Inference of a probabilistic boolean network from a single observed temporal sequence," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, no. 1, pp. 1–15, 2007.

[52] D. C. Martins, U. M. Braga-Neto, R. F. Hashimoto, M. L. Bittner, and E. R. Dougherty, "Intrinsically multivariate predictive genes," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, no. 3, pp. 424–439, 2008.

[53] P. McCullagh, *Tensor methods in statistics.* London: Chapman and Hall, 1987, vol. 161.

[54] R. Miller, *Simultaneous Statistical Inference.* New York, NY: Sprnger Series in Statistics.

[55] A. Mordecai, *Nonlinear programming: Analysis and methods*, 2003.

[56] T. Noguchi, R. Metz, L. Chen, M. Mattei, D. Carrasco, and R. Bravo, "Structure, mapping, and expression of erp, a growth factor-inducible gene encoding a nontransmembrane protein tyrosine phosphatase, and effect of erp on cell growth." *Molecular and cellular biology*, vol. 13, no. 9, pp. 5195–5205, 1993.

[57] A. B. Owen, *Empirical likelihood.* Boca Raton, FL: Chapman and Hall/CRC, 2001.

[58] X. Qian, J. Hua, U. M. Braga-Neto, Z. Xiong, E. Suh, and E. R. Dougherty, "Confidence intervals for the true classification error conditioned on the estimated error," *Technology in cancer research & treatment*, vol. 5, no. 6, pp. 579–589, 2006.

[59] S. Ross, *Stochastic Processes.* New York, NY: Wiley, 1995.

[60] I. Shmulevich and W. Zhang, "Binary analysis and optimization-based normalization of gene expression data," *Bioinformatics*, vol. 18, no. 4, pp. 555–565, 2002.

[61] I. Shmulevich and E. R. Dougherty, *Genomic Signal Processing.* Princeton, NJ: Princeton University Press, 2007.

[62] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, “Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks,” *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[63] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches.* New York, NY: Wiley-Interscience, 2006.

[64] K. Smalley, “A pivotal role for erk in the oncogenic behaviour of malignant melanoma?” *International journal of cancer*, vol. 104, no. 5, pp. 527–532, 2003.

[65] C. A. Smith, “Some examples of discrimination,” *Annals of Eugenics*, vol. 13, no. 1, pp. 272–282, 1946.

[66] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 1, pp. 111–147, 1974.

[67] C. J. Tabin and R. A. Weinberg, “Analysis of viral and somatic activations of the cha-ras gene.” *Journal of virology*, vol. 53, no. 1, pp. 260–265, 1985.

[68] G. Vahedi, I. Ivanov, and E. Dougherty, “Inference of boolean networks under constraint on bidirectional gene relationships,” *Systems Biology, IET*, vol. 3, no. 3, pp. 191–202, 2009.

[69] C. Waddington, “Canalization of development and the inheritance of acquired characters,” *Nature*, vol. 150, no. 3811, pp. 563–565, 1942.

[70] A. Wagner, *Robustness and evolvability in living systems.* New Jersey: Princeton University Press Princeton, 2005.

[71] X. Zhou, X. Wang, and E. R. Dougherty, “Binarization of microarray data on the basis of a mixture model1,” *Molecular cancer therapeutics*, vol. 2, no. 7, pp. 679–684, 2003.

[72] G. K. Zipf, *The psycho-biology of language.* Oxford, England: Houghton, Mifflin, 1935.

APPENDIX A

RESULTS ON CONDITIONAL EXPECTATION GIVEN DISJOINT EVENTS

Proposition 7. *For a discrete random variable X and disjoint events A and B, we have*

$$E[X \mid A \cup B] \;=\; \frac{P(A)}{P(A) \;+\; P(B)}\, E[X \mid A] \;+\; \frac{P(B)}{P(A) + P(B)}\, E[X \mid B]\,. \tag{A.1}$$

Proof.

$$\begin{aligned} E[X \mid A \cup B] \;&=\; \sum_x x\, P(X = x \mid A \cup B) \\ &=\; \sum_x x\, \frac{P(A \cup B \mid X = x) P(X = x)}{P(A \cup B)} \\ &=\; \sum_x x\, \frac{[P(A \mid X = x) + P(B \mid X = x)] P(X = x)}{P(A) + P(B)} \\ &=\; \sum_x x\, \frac{P(X = x \mid A) P(A) + P(X = x \mid B) P(B)}{P(A) + P(B)} \\ &=\; \frac{P(A)}{P(A) + P(B)} \sum_x x\, P(X = x \mid A) \;+\; \frac{P(B)}{P(A) + P(B)} \sum_x x\, P(X = x \mid B) \\ &=\; \frac{P(A)}{P(A) + P(B)}\, E[X \mid A] \;+\; \frac{P(B)}{P(A) + P(B)}\, E[X \mid B]\,. \end{aligned} \tag{A.2}$$

Q.E.D.

APPENDIX B

EXPRESSIONS OF BIAS AND VARIANCE OF THE ML COD ESTIMATOR IN 3-INPUT AND LOGIC MODEL

The bias is expressed in the form of

$$\mathrm{Bias}\left[\widehat{\mathrm{CoD}}^{\mathrm{ML}}_{\mathrm{AND}^3}\right] \approx \left\{ \begin{array}{l} -\dfrac{(1-p)(1-2p)\gamma\left[-\frac{\gamma}{n^2}+\frac{n-1}{n^2}(P_1\gamma_{23}+P_2\gamma_{13}+P_3\gamma_{12})\right]}{[1-P_1P_2P_3-\gamma-(1-2P_1P_2P_3-2\gamma)p]^2}, \\ \qquad\qquad\qquad\qquad\qquad\qquad P_1P_2P_3+\gamma<\frac{1}{2} \\ \dfrac{(1-p)(1-2p)\gamma\left[-\frac{\gamma}{n^2}+\frac{n-1}{n^2}(P_1\gamma_{23}+P_2\gamma_{13}+P_3\gamma_{12})\right]}{[P_1P_2P_3+\gamma+(1-2P_1P_2P_3-2\gamma)p]^2}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \mathrm{P_1P_2P_3}+\gamma>\frac{1}{2} \end{array} \right. \tag{B.1}$$

and the variance is given by

$$\mathrm{Var}[\widehat{\mathrm{CoD}}^{\mathrm{ML}}_{\mathrm{AND}^3}] \approx \left\{ \begin{array}{l} \frac{(P_1P_2P_3+\gamma)^2\mathrm{Var}(\hat{p})}{[1-P_1P_2P_3-\gamma-(1-2P_1P_2P_3-2\gamma)p]^4} \ + \ (1-p)^2(1-2p)^2\times \\ \Big[\frac{2P_1P_2P_3^2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)+2P_1P_2^2P_3\mathrm{Cov}(\hat{P}_1,\hat{P}_3)+2P_1^2P_2P_3\mathrm{Cov}(\hat{P}_2,\hat{P}_3)}{[1-P_1P_2P_3-\gamma-(1-2P_1P_2P_3-2\gamma)p]^4} + \\ \frac{P_2^2P_3^2\mathrm{Var}(\hat{P}_1)+P_1^2P_3^2\mathrm{Var}(\hat{P}_2)+P_1^2P_2^2\mathrm{Var}(\hat{P}_3)+2P_1P_2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[1-P_1P_2P_3-\gamma-(1-2P_1P_2P_3-2\gamma)p]^4} + \\ \frac{\mathrm{Var}(\hat{\gamma})+2P_2P_3\mathrm{Cov}(\hat{P}_1,\hat{\gamma})+2P_1P_3\mathrm{Cov}(\hat{P}_2,\hat{\gamma})+2P_1P_2\mathrm{Cov}(\hat{P}_3,\hat{\gamma})}{[1-P_1P_2P_3-\gamma-(1-2P_1P_2P_3-2\gamma)p]^4}\Big] \ , \ P_1P_2P_3+\gamma<\frac{1}{2} \\ \frac{(P_1P_2P_3+\gamma-1)^2\mathrm{Var}(\hat{p})}{[P_1P_2P_3+\gamma+(1-2P_1P_2P_3-2\gamma)p]^4} \ + \ (1-p)^2(1-2p)^2\times \\ \Big[\frac{2P_1P_2P_3^2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)+2P_1P_2^2P_3\mathrm{Cov}(\hat{P}_1,\hat{P}_3)+2P_1^2P_2P_3\mathrm{Cov}(\hat{P}_2,\hat{P}_3)}{[P_1P_2P_3+\gamma+(1-2P_1P_2P_3-2\gamma)p]^4} + \\ \frac{P_2^2P_3^2\mathrm{Var}(\hat{P}_1)+P_1^2P_3^2\mathrm{Var}(\hat{P}_2)+P_1^2P_2^2\mathrm{Var}(\hat{P}_3)+2P_1P_2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[P_1P_2P_3+\gamma+(1-2P_1P_2P_3-2\gamma)p]^4} \\ \frac{\mathrm{Var}(\hat{\gamma})+2P_2P_3\mathrm{Cov}(\hat{P}_1,\hat{\gamma})+2P_1P_3\mathrm{Cov}(\hat{P}_2,\hat{\gamma})+2P_1P_2\mathrm{Cov}(\hat{P}_3,\hat{\gamma})}{[P_1P_2P_3+\gamma+(1-2P_1P_2P_3-2\gamma)p]^4}\Big] \ , \ P_1P_2P_3+\gamma>\frac{1}{2} \end{array} \right. \tag{B.2}$$

where the ML estimators for the three-input logic model parameters satisfy:

$$\begin{aligned}\mathrm{Var}[\hat{\gamma}_{ij}] &= \frac{n-1}{n^2} P_i P_j (1-P_i)(1-P_j) + \\ &\frac{(n-1)^2}{n^3}(1-2P_i)(1-2P_j)\gamma - \frac{(n-1)(n-2)}{n^3}\gamma_{ij}^2,\end{aligned} \tag{B.3}$$

for $i, j \in 1, 2, 3$ and $i < j$,

$$\begin{aligned}\mathrm{Cov}(\hat{P}_i, \hat{\gamma}) \; = \; &\frac{n-1}{n^3}\left[(n+1)\gamma - (n+2)P_i\gamma - (n-2)(P_iP_k\gamma_{ij} + P_iP_j\gamma_{ik}) - \right. \\ &\left. (P_i\gamma_{jk} + P_j\gamma_{ik} + P_k\gamma_{ij}) - 2\gamma_{ij}\gamma_{ik} + 2P_i^2\gamma_{jk}\right],\end{aligned}$$

$$\begin{aligned}&\mathrm{Cov}(\hat{\gamma}_{ij}, \hat{\gamma}) \; = \\ &\frac{n-1}{n^2}(P_iP_jP_k + P_i^2P_j^2P_k - P_i^2P_jP_k - P_iP_j^2P_k) - \frac{(n-1)^2(n+2)}{n^4}(P_i\gamma + P_j\gamma) + \\ &\frac{n-1}{n^4}(P_i\gamma_{jk} + P_j\gamma_{ik} - (n-1)P_k\gamma_{ij}) + \frac{(n-1)(n-2))}{n^4}(P_i^2\gamma_{jk} + P_j^2\gamma_{ik}) + \\ &\frac{(n-1)(3n-4)}{n^4}P_k^2\gamma_{ij} + \frac{(n-1)(5n-8)}{n^4}(P_i\gamma_{ij}\gamma_{jk} + P_j\gamma_{ij}\gamma_{ik}) + \\ &\frac{(n-1)(n-2)}{n^4}(P_iP_j\gamma_{jk} + P_iP_j\gamma_{ik}) - \frac{(n-1)^2(n-2)}{n^4}(P_jP_k\gamma_{ij} + P_iP_k\gamma_{ij}) + \\ &\frac{(n-1)(n-2)^2}{n^4}(P_iP_j^2\gamma_{ik} + P_i^2P_j\gamma_{jk}) + \frac{4(n-1)^2}{n^4}P_iP_j\gamma + \frac{4(n-1)^3}{n^4}P_iP_jP_k\gamma_{ij} + \\ &\frac{(n-1)^2(n+1)}{n^4}\gamma - \frac{(n-1)(n^2+n-4)}{n^4}\gamma_{ij}\gamma - \frac{(n-1)(n-2)}{n^4}(\gamma_{ij}\gamma_{ik} + \gamma_{ij}\gamma_{jk}),\end{aligned} \tag{B.4}$$

for $i, j, k \in \{1, 2, 3\}$ and $i < j$ $(i \neq k, j \neq k)$ with $\gamma_{ij} = \gamma_{ji}$,

$$\begin{aligned}&\mathrm{Cov}(\hat{p}, \hat{P}_i) \; = \; 0, \text{ for } i = 1, 2, 3\,, \; \mathrm{Cov}(\hat{p}, \hat{\gamma}) \; = \; 0, \\ &\mathrm{Cov}(\hat{P}_i, \hat{P}_j) \; = \; \gamma_{ij}/n, \quad \text{for } i, j \in \{1, 2, 3\} \text{ and } i < j \text{ with } \gamma_{ij} = \gamma_{ji}, \\ &\mathrm{Cov}(\hat{P}_i, \hat{\gamma}_{ij}) \; = \; \frac{n-1}{n^2}(1-2P_i)\gamma_{ij}, \; for \, 1 \leq i < j \leq 3, \\ &\mathrm{Cov}(\hat{P}_1, \hat{\gamma}_{23}) \; = \; \mathrm{Cov}(\hat{P}_2, \hat{\gamma}_{13}) \; = \; \mathrm{Cov}(\hat{P}_3, \hat{\gamma}_{12}) \; = \; \frac{n-1}{n^2}(\gamma - P_1\gamma_{23} - P_2\gamma_{13} - P_3\gamma_{12}),\end{aligned} \tag{B.5}$$

and we have the variance for $\hat{\gamma}$ as given by:

$$
\begin{aligned}
&\mathrm{Var}[\hat{\gamma}] \;=\; \frac{n^2-1}{n^3}P_1P_2P_3 - \frac{n-1}{n^2}(P_1^2P_2^2P_3 + P_1^2P_2P_3^2 + P_1P_2^2P_3^2) - \\
&\frac{n-1}{n^3}(P_1^2P_2P_3 + P_1P_2^2P_3 + P_1P_2P_3^2) - \\
&\frac{(n-1)(2n-1)}{n^3}P_1^2P_2^2P_3^2 + \frac{(n-1)(4n^3+4n^2-16n+8)}{n^5}P_1P_2P_3\gamma + \\
&\frac{(n-1)(2n^3-12n^2+20n-8)}{n^5}(P_1^2P_2P_3\gamma_{23} + P_1P_2^2P_3\gamma_{13} + P_1P_2P_3^2\gamma_{12}) + \\
&\frac{(n-1)(2n^2-8n+4)}{n^5}(P_1P_2P_3\gamma_{12} + P_1P_2P_3\gamma_{13} + P_1P_2P_3\gamma_{23}) + \\
&\frac{(n-1)^2(4-2n^2)}{n^5}(P_1P_2 + P_1P_3 + P_2P_3)\gamma - \frac{2(n-1)^2(n+1)}{n^5}(P_1 + P_2 + P_3)\gamma \\
&\frac{2(n-1)^2(n-2)}{n^5}(P_1P_2^2\gamma_{13} + P_1^2P_2\gamma_{23} + P_1P_3^2\gamma_{12} + P_1^2P_3\gamma_{23} + P_2P_3^2\gamma_{12} + P_2^2P_3\gamma_{13}) + \\
&\frac{(n-1)(6n^2-30n+32)}{n^5}(P_1P_2\gamma_{13}\gamma_{23} + P_1P_3\gamma_{12}\gamma_{23} + P_2P_3\gamma_{12}\gamma_{13}) + \\
&\frac{2(n-1)^2}{n^5}(P_1P_2\gamma_{13} + P_1P_2\gamma_{23} + P_2P_3\gamma_{12} + P_2P_3\gamma_{13} + P_1P_3\gamma_{12} + P_1P_3\gamma_{23}) + \\
&\frac{4(n-1)(n-2)}{n^5}(P_1\gamma_{13}\gamma_{23} + P_1\gamma_{12}\gamma_{23} + P_2\gamma_{13}\gamma_{23} + P_2\gamma_{12}\gamma_{13} + P_3\gamma_{12}\gamma_{23} + P_3\gamma_{12}\gamma_{13}) + \\
&\frac{(n-1)(n^2-9n+12)}{n^5}(P_1^2\gamma_{23}^2 + P_2^2\gamma_{13}^2 + P_3^2\gamma_{12}^2) + \\
&\frac{(n-1)(4n^2+6n-16)}{n^5}(P_1\gamma_{23} + P_2\gamma_{13} + P_3\gamma_{12})\gamma + \\
&\frac{3(n-1)(n-2)}{n^5}(P_1\gamma_{23}^2 + P_2\gamma_{13}^2 + P_3\gamma_{12}^2) + \frac{(n-1)}{n^5}(P_1\gamma_{23} + P_2\gamma_{13} + P_3\gamma_{12}) + \\
&\frac{8(n-1)(n-2)}{n^5}\gamma_{12}\gamma_{13}\gamma_{23} + \frac{2(n-1)}{n^5}(\gamma_{12}\gamma_{13} + \gamma_{12}\gamma_{23} + \gamma_{13}\gamma_{23}) - \\
&\frac{2(n-1)(n^2-2)}{n^5}(\gamma_{12} + \gamma_{13} + \gamma_{23})\gamma - \frac{(n-1)(n^3+n^2-n-4)}{n^5}\gamma^2 + \\
&\frac{(n-1)^2(n+1)^2}{n^5}\gamma.
\end{aligned}
\tag{B.6}
$$

Note that, when $P_1P_2P_3 + \gamma = \frac{1}{2}$, the CoD is not differentiable, and thus the asymptotic approximation cannot be made as mentioned in the paper. However, in this case we could obtain $\mathrm{CoD}_{\mathrm{AND}^3} = 2p - 1$, which then gives $\widehat{\mathrm{CoD}}^{\mathrm{ML}} = 2\hat{p} - 1$.

Hence, it produces in this case, the exact bias with $\text{Bias}\left[\widehat{\text{CoD}}^{\text{ML}}_{\text{AND}^3}\right] = 0$, for all n, and the exact variance with $\text{Var}\left[\widehat{\text{CoD}}^{\text{ML}}_{\text{AND}^3}\right] = \frac{4}{n}p(1-p)$.

APPENDIX C

ASYMPTOTIC EXPRESSIONS OF BIAS AND VARIANCE OF THE ML COD ESTIMATOR FOR 10 2-PREDICTOR LOGICS

Table C.1: Formulas for ML CoD estimator and its bias asymptotic approximations for the five representative two-predictor logic models.

Logic	ML CoD Estimator	Bias
AND	$1 - \frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$	$\frac{(2\mathbf{1}_{1-2A>0}-1)(1-p)(1-2p)\gamma}{n[1-A+(1-2A)p-\mathbf{1}_{1-2A<0}]^2}$
	$\hat{A} = \hat{P}_1\hat{P}_2 + \hat{\gamma}$	$A = P_1P_2 + \gamma$
XOR	$1 - \frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$	$\frac{2(1-2\mathbf{1}_{1-2A>0}(1-p)(1-2p)\gamma)}{n[1-A+(1-2A)p-\mathbf{1}_{1-2A<0}]^2}$
	$\hat{A} = \hat{P}_1 + \hat{P}_2 - 2\hat{P}_1\hat{P}_2 - 2\hat{\gamma}$	$A = P_1 + P_2 - 2P_1P_2 - 2\gamma$
OR	$1 - \frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$	$\frac{(1-2\mathbf{1}_{1-2A>0})(1-p)(1-2p)\gamma}{n[1-A+(1-2A)p-\mathbf{1}_{1-2A<0}]^2}$
	$\hat{A} = \hat{P}_1 + \hat{P}_2 - \hat{P}_1\hat{P}_2 - \hat{\gamma}$	$A = P_1 + P_2 - P_1P_2 - \gamma$
$X_1\bar{X}_2$	$1 - \frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$	$\frac{(1-2\mathbf{1}_{1-2A>0})(1-p)(1-2p)\gamma}{n[1-A+(1-2A)p-\mathbf{1}_{1-2A<0}]^2}$
	$\hat{A} = \hat{P}_1 - \hat{P}_1\hat{P}_2 - \hat{\gamma}$	$A = P_1 - P_1P_2 - \gamma$
$\bar{X}_1X_2$	$1 - \frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$	$\frac{(1-2\mathbf{1}_{1-2A>0})(1-p)(1-2p)\gamma}{n[1-A+(1-2A)p-\mathbf{1}_{1-2A<0}]^2}$
	$\hat{A} = \hat{P}_2 - \hat{P}_1\hat{P}_2 - \hat{\gamma}$	$A = P_2 - P_1P_2 - \gamma$

Table C.2: Formulas for ML CoD estimator and its variance asymptotic approximations for the five representative two-predictor logic models.

Logic	ML CoD Estimator	Variance
AND	$1-\frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$ $\hat{A}=\hat{P}_1\hat{P}_2+\hat{\gamma}$	$\frac{(P_1P_2+\gamma-\mathbf{1}_{1-2A<0})^2\mathrm{Var}(\hat{p})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}+(1-p)^2(1-2p)^2\times$ $\frac{P_2^2\mathrm{Var}(\hat{P}_1)+P_1^2\mathrm{Var}(\hat{P}_2)+2P_1P_2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)+\mathrm{Var}(\hat{\gamma})+2P_2\mathrm{Cov}(\hat{P}_1,\hat{\gamma})+2P_1\mathrm{Cov}(\hat{P}_2,\hat{\gamma})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}$
XOR	$1-\frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$ $\hat{A}=\hat{P}_1+\hat{P}_2-2\hat{P}_1\hat{P}_2-2\hat{\gamma}$	$\frac{(P_1+P_2-2P_1P_2-2\gamma-\mathbf{1}_{1-2A<0})^2\mathrm{Var}(\hat{p})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}+(1-p)^2(1-2p)^2\times$ $\Big[\frac{(1-2P_2)^2\mathrm{Var}(\hat{P}_1)+(1-2P_1)^2\mathrm{Var}(\hat{P}_2)+2(1-2P_1)(1-2P_2)\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}$ $\frac{4\mathrm{Var}(\hat{\gamma})-4(1-2P_2)\mathrm{Cov}(\hat{P}_1,\hat{\gamma})-4(1-2P_1)\mathrm{Cov}(\hat{P}_2,\hat{\gamma})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}\Big]$
OR	$1-\frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$ $\hat{A}=\hat{P}_1+\hat{P}_2-\hat{P}_1\hat{P}_2-\hat{\gamma}$	$\frac{(P_1+P_2-P_1P_2-\gamma-\mathbf{1}_{1-2A<0})^2\mathrm{Var}(\hat{p})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}+(1-p)^2(1-2p)^2\times$ $\Big[\frac{(1-P_2)^2\mathrm{Var}(\hat{P}_1)+(1-P_1)^2\mathrm{Var}(\hat{P}_2)+2(1-P_1)(1-P_2)\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}$ $\frac{\mathrm{Var}(\hat{\gamma})-2(1-P_2)\mathrm{Cov}(\hat{P}_1,\hat{\gamma})-2(1-P_1)\mathrm{Cov}(\hat{P}_2,\hat{\gamma})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}\Big]$
$X_1\bar{X}_2$	$1-\frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$ $\hat{A}=\hat{P}_1-\hat{P}_1\hat{P}_2-\hat{\gamma}$	$\frac{(P_1-P_1P_2-\gamma-\mathbf{1}_{1-2A<0})^2\mathrm{Var}(\hat{p})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A\geq 0}]^4}+(1-p)^2(1-2p)^2\times$ $\Big[\frac{(1-P_2)^2\mathrm{Var}(\hat{P}_1)+P_1^2\mathrm{Var}(\hat{P}_2)-2P_1(1-P_2)\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}$ $\frac{\mathrm{Var}(\hat{\gamma})-2(1-P_2)\mathrm{Cov}(\hat{P}_1,\hat{\gamma})+2P_1\mathrm{Cov}(\hat{P}_2,\hat{\gamma})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}\Big]$
$\bar{X}_1X_2$	$1-\frac{1-\hat{p}}{F[\hat{A}+(1-2\hat{A})\hat{p}]}$ $A=P_2-P_1P_2-\gamma$	$\frac{(P_2-P_1P_2-\gamma-\mathbf{1}_{1-2A<0})^2\mathrm{Var}(\hat{p})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}+(1-p)^2(1-2p)^2\times$ $\Big[\frac{P_2^2\mathrm{Var}(\hat{P}_1)+(1-P_1)^2\mathrm{Var}(\hat{P}_2)-2(1-P_1)P_2\mathrm{Cov}(\hat{P}_1,\hat{P}_2)}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}$ $\frac{\mathrm{Var}(\hat{\gamma})+2P_2\mathrm{Cov}(\hat{P}_1,\hat{\gamma})-2(1-P_1)\mathrm{Cov}(\hat{P}_2,\hat{\gamma})}{[1-A-(1-2A)p-\mathbf{1}_{1-2A<0}]^4}\Big]$

APPENDIX D

PROOF OF PROPOSITION 2

Proof of Proposition 2. Using Proposition 1, we know that to test H_0 : CoD $= 0$ vs. H_1 : CoD > 0, is equivalent to test

$$\begin{aligned} &H_0 : p = 1/2 \text{ or } \xi \in \{0, 1\} \text{ vs.} \\ &H_1 : p \neq 1/2 \text{ and } \xi \neq 0 \text{ and } \xi \neq 1\,, \end{aligned} \tag{D.1}$$

where $\xi = P(f(\mathbf{X}) = 1)$.

The IUT method is applied here. First we derive a LRT of $H_{01} : p = 1/2$ vs. H$_{11}$: p $> 1/2$. Assuming a stochastic logic model in eq. (1), a level α LRT of $H_0 : p = 1/2$ versus $H_1 : p > 1/2$ can be based on the test statistic

$$\lambda(\mathbf{s}_n) = \begin{cases} \left[\dfrac{(1 - z_n)^{z_n}}{2(1 - z_n)z_n^{z_n}}\right]^n =: g(z_n) & \mathbf{s}_n \in \mathcal{R}_2 \\ \\ \qquad 1, & \text{otherwise} \end{cases} \tag{D.2}$$

where $z_n = \sum_{i=1}^n \mathbf{1}(f(\mathbf{x}_i = y_i))$. When $\mathbf{s}_n \in \mathcal{R}_2$, $g(z_n)$ is decreasing in $z_n \in [0, 1]$, and so that $\lambda(\mathbf{s}_n) \leq c$ is equivalent to $z_n \geq k$. Since $\sum_{i=1}^n \mathbf{1}(f(\mathbf{X}_i = Y_i))$ follows a Binomial(n,p) distribution, k is the $100(1 - \alpha)\%$ percentile of a Binomial(n,1/2) distribution, i.e., k is the smallest integer such that $\sum_{l>k} \binom{n}{l}(1/2)^n \leq \alpha$.

Secondly, we need to test $H_{02} : \xi = 0$ vs. H$_{12}$: $\xi \neq 0$. Note that ξ is a function of $P_1, \ldots, P_d$ and γ's to the order d. The maximum-likelihood estimator of ξ, denoted as $\hat{\xi}$, is the function ξ with $P_1 \ldots, P_d$ and γ's to the order d replaced by their corresponding ML estimators as given in [20]. We denote $\hat{\xi} = \sum_{f(\mathbf{x})=1} \hat{P}(\mathbf{X} = \mathbf{x})$, where $\hat{P}(\mathbf{X} = \mathbf{x})$ is also the sample proportion of samples of $(\mathbf{X} = \mathbf{x})$. Furthermore, we can prove the equivalence between $\hat{\xi} = 0$ and $f(\mathbf{x}_i) = 0$ for all $i \in \{1, \ldots, n\}$ as shown by:

(a) $\hat{\xi} = 0 \Longrightarrow \sum_{f(\mathbf{x})=1} \hat{P}(\mathbf{X} = \mathbf{x}) = 0 \Longrightarrow f(\mathbf{x}_i) = 0$ for all $i \in \{1, \ldots, n\}$;

(b) $f(\mathbf{x}_i) = 0$ for all $i \Longrightarrow \hat{P}(\mathbf{X} = \mathbf{x}) = 0$ for any $\mathbf{x}_i$ satisfying $f(\mathbf{x}_i) = 1 \Longrightarrow \hat{\xi} = 0$.

Thus, we have the LRT statistic formed by

$$\begin{aligned} \lambda(\mathbf{s}_n) &= \frac{\sup_{\xi=0} L(\boldsymbol{\theta}|\mathbf{s}_n)}{\sup L(\boldsymbol{\theta}|\mathbf{s}_n)} \\ &= \begin{cases} v < 1, & f(\mathbf{x}_i) = 1 \text{ for some } i \\ 1, & f(\mathbf{x}_i) = 0 \text{ for all } i \end{cases}, \end{aligned} \tag{D.3}$$

Let us choose $c = v$, the rejection region $\mathcal{R}_2 = \{\lambda(\mathbf{s}_n) \leq c\}$ is equivalent to $\mathcal{R}_2 = \{\mathbf{s}_n | f(\mathbf{x}_i) = 1 \text{ for some } i\}$. The type-I error can be computed by:

$$\begin{aligned} \beta_{\xi=0}(\phi) &= P_{\xi=0}(f(\mathbf{x}_i) = 1 \text{ for some } i) \\ &= 1 - (1-\xi)^n|_{\xi=0} = 0 < \alpha. \end{aligned} \tag{D.4}$$

Therefore, the test function $\phi = \mathbf{1}_{S_n \in \mathcal{R}_2}$ is a level α test here.

Thirdly, we could prove that the test function $\phi = \mathbf{1}_{S_n \in \mathcal{R}_3}$ is a level α test to test $H_{03} : \xi = 1$ vs. $\mathrm{H}_{13} : \xi \neq 1$, where $\mathcal{R}_3 = \{\mathbf{s}_n | f(\mathbf{x}_i) = 0 \text{ for some } i\}$.

Lastly, we get the rejection region $\mathcal{R}$ of testing $H_0 : \text{CoD} = 0$ vs. $H_1 : \text{CoD} > 0$ as formed by $R = \mathcal{R}_1 \cap \mathcal{R}_2 \cap \mathcal{R}_3$ according to the IUT theorem, where $\mathcal{R}_2 \cap \mathcal{R}_3$ is equivalent to

$$\left\{ \mathbf{s}_n \,\middle|\, \exists\, 1 \leq i, j \leq n \text{ s.t. } f(\mathbf{x}_i) \neq f(\mathbf{x}_j) \right\}$$

. Q.E.D.

APPENDIX E

PROOF OF PROPOSITION 4

Proof of Proposition 4. We are concerned with testing H_0 : IMP $= 0$ against H_1 : IMP > 0. The null hypothesis can furthermore be written into the equivalent statement via definition of IMP, that is, $\varepsilon_Y(\mathbf{X}) = \min_{\mathbf{Z} \subsetneqq \mathbf{X}} \varepsilon_Y(\mathbf{Z})$. Since predictor $\mathbf{X}$ is the perfect predictor of target Y, $\varepsilon_Y(\mathbf{Z}) \geq \varepsilon_Y(\mathbf{X})$, for any $\mathbf{Z} \subsetneqq \mathbf{X}$. Suppose the predictive power of $\mathbf{X}$ over Y is p, and then we have $\varepsilon_Y(\mathbf{X}) = 1 - p$. Hence, for some $\mathbf{T} \subsetneqq \mathbf{X}$, if $\varepsilon_Y(\mathbf{T}) = \varepsilon_Y(\mathbf{X}) = 1 - p$, then $\varepsilon_Y(\mathbf{X}^{(2)})$ is clearly the minimum of $\varepsilon_Y(\mathbf{Z})$ for all $\mathbf{Z} \subsetneqq \mathbf{X}$. Therefore, H_0 : IMP $= 0$ is equivalent to

$$H_0 : \varepsilon_Y(\mathbf{X}) = \varepsilon_Y(\mathbf{V}_1) \text{ or } \varepsilon_Y(\mathbf{X}) = \varepsilon_{\mathrm{Y}}(\mathbf{V}_2) \ \ldots \ \text{ or } \varepsilon_{\mathrm{Y}}(\mathbf{X}) = \varepsilon_{\mathrm{Y}}(\mathbf{V}_{2^{\mathrm{d}}-2}), \tag{E.1}$$

where $\mathcal{V}(\mathbf{X}) := \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{2^d-2}\} = \mathcal{P}(\mathbf{X}) \backslash \{\{\emptyset\}, \{\mathbf{X}\}\}$, that is, the power set of $\mathbf{X}$ excluding empty set and set $\mathbf{X}$.

Let $\mathbf{X}^{(2)}$ be an element in $\mathcal{V}(\mathbf{X})$ and $\mathbf{X}^{(1)} = \mathbf{X} \backslash \mathbf{X}^{(2)}$. Assume $P(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}) > 0$ for any $\mathbf{x}^{(2)}$, and we have

$$\begin{gathered} p \geq P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \ = \\ \frac{\sum_{\mathbf{x}^{(1)} \in \{0,1\}^{|\mathbf{x}^{(1)}|}} P(\mathbf{X} = \mathbf{x}) \left[p \cdot \mathbf{1}(f(\mathbf{x}) = 1) + (1 - p) \cdot \mathbf{1}(f(\mathbf{x}) = 0)\right]}{P(\mathbf{X}^{(2)} = \mathbf{x}^{(2)})} \geq 1 - p. \end{gathered} \tag{E.2}$$

Since $F(x) = \min(x, 1 - x)$ is strictly increasing in $x \in [1 - p, 1/2]$ and decreasing in $x \in [1/2, p]$,

$$F\left[P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)})\right] \in [1 - p, 1/2] \tag{E.3}$$

.

Now consider $\varepsilon_Y(\mathbf{X}) = \varepsilon_Y(\mathbf{X}^{(2)})$, for any $\mathbf{X}^{(2)} \in \mathcal{V}$, and we have

$$
\begin{aligned}
1 - p = & \sum_{\mathbf{X}^{(2)}=\mathbf{x}^{(2)}} F\left[P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)})\right] P(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \\
\Leftrightarrow & \sum_{\mathbf{X}^{(2)}=\mathbf{x}^{(2)}} \big(F\left[P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)})\right] - (1-p)\big) P(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \;=\; 0 \\
\Leftrightarrow & \, F\left[P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)})\right] = 1 - p, \quad \text{for all } \mathbf{x}^{(2)} \in \{0,1\}^{|\mathbf{x}^{(2)}|} \\
\Leftrightarrow & \, P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) = 1 - p, \text{ or } P(Y = 1 \,|\, \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) = p, \\
& \qquad\qquad \text{for all } \mathbf{x}^{(2)} \in \{0,1\}^{|\mathbf{x}^{(2)}|} \\
\Leftrightarrow & \, p = 1/2, \text{ or } \sum_{\mathbf{x}^{(1)} \in \{0,1\}^{|\mathbf{x}^{(1)}|}} P(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \mathbf{1}(f(\mathbf{x}) = 1) = 0, \text{ or} \\
& \sum_{\mathbf{x}^{(1)} \in \{0,1\}^{|\mathbf{x}^{(1)}|}} P(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \mathbf{1}(f(\mathbf{x}) = 0) = 0\,, \\
& \qquad\qquad \text{for all } \mathbf{x}^{(2)} \in \{0,1\}^{|\mathbf{x}^{(2)}|}\,.
\end{aligned}
\tag{E.4}
$$

Note that the last third expression is derived using eq. (E.3). It is easy to check this includes results with $P(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}) = 0$ for some $\mathbf{x}^{(2)}$. Hence, the proposition holds. Q.E.D.

APPENDIX F

PROOF OF PROPOSITION 5

Proof of Proposition 5. Using Proposition 1, we know that to test H_0 : IMP $=$ 0 vs. H$_1$: IMP > 0, is equivalent to test

$$\begin{aligned} &H_0 : p = 1/2 \text{ or } P(\mathbf{X} \in \mathcal{D}_1) = 0, \ldots, \text{ or } P(\mathbf{X} \in \mathcal{D}_{d^*}) = 0 \text{ or} \\ &\qquad P(\mathbf{X} \in \mathcal{D}_1) = 1, \ldots, \text{ or } P(\mathbf{X} \in \mathcal{D}_{d^*}) = 1 \\ &\text{vs. } H_1 : 1 \geq p > 1/2 \text{ and } 1 > P(\mathbf{X} \in \mathcal{D}_1) > 0, \ldots, \text{ and } 1 > P(\mathbf{X} \in \mathcal{D}_{d^*}) > 0\,. \end{aligned} \tag{F.1}$$

The IUT method is applied here.

First we derive a LRT of $H_{01} : p = 1/2$ vs. H$_{11}$: p $> 1/2$. Assuming a stochastic logic model in eq. (1), a level α LRT of $H_0 : p = 1/2$ versus $H_1 : p > 1/2$ can be based on the test statistic

$$\lambda(\mathbf{s}_n) = \begin{cases} \left[\dfrac{(1-z_n)^{z_n}}{2(1-z_n)z_n^{z_n}}\right]^n =: g(z_n) & \mathbf{s}_n \in \mathcal{R}_2 \\ \\ \qquad\qquad 1, & \text{otherwise} \end{cases} \tag{F.2}$$

where $z_n = \sum_{i=1}^n \mathbf{1}(f(\mathbf{x}_i = y_i))$. When $\mathbf{s}_n \in \mathcal{R}_2$, $g(z_n)$ is decreasing in $z_n \in [0, 1]$, and so that $\lambda(\mathbf{s}_n) \leq c$ is equivalent to $z_n \geq K$.

Secondly, we need to test $H_{0j} : P(\mathbf{X} \in \mathcal{D}_j) = 0$ vs. $\mathrm{H_{1j}} : \mathrm{P}(\mathbf{X} \in \mathcal{D}_\mathrm{j}) > 0$, for any $j \in \{1, \ldots, d^*\}$. Note that $P(\mathbf{X} \in \mathcal{D}_j)$ is a function of $P_1, \ldots, P_d$ and γ's to the order d. The maximum-likelihood estimator of $P(\mathbf{X} \in \mathcal{D}_j)$, denoted as $\widehat{P_{\mathcal{D}_j}}$, is the function $P(\mathbf{X} \in \mathcal{D}_j)$ with $P_1 \ldots, P_d$ and γ's to the order d replaced by their corresponding ML estimators (that is, frequency estimators for probabilities), and thus we have

$\widehat{P_{\mathcal{D}_j}} = 1/n \sum_{i=1}^{n} \mathbf{1}(\mathbf{x}_i \in \mathcal{D}_j)$. Thus, we have the LRT statistic formed by

$$\lambda(\mathbf{s}_n) \;=\; \frac{\sup_{P(\mathbf{X}\in\mathcal{D}_j)=0} L(\boldsymbol{\theta}|\mathbf{s}_n)}{\sup L(\boldsymbol{\theta}|\mathbf{s}_n)} \;=\; \begin{cases} 0 < 1, & \mathbf{x}_i \in \mathcal{D}_j \text{ for some } i \\ 1, & \mathbf{x}_i \notin \mathcal{D}_j \text{ for all } i \end{cases}, \tag{F.3}$$

where $\lambda(\mathbf{s}_n) = 1$ under $\mathbf{x}_i \notin \mathcal{D}_j$ for all i holds since $\widehat{P_{\mathcal{D}_j}} = 0 \Leftrightarrow \mathbf{x}_i \notin \mathcal{D}_j$ for all i. Let us choose $c = 1/2$, the rejection region $\mathcal{R}_{2j} = \{\lambda(S_n) \le c\}$ is equivalent to $\mathcal{R}_{2j} = \{\mathbf{s}_n|\mathbf{x}_i \in \mathcal{D}_j \text{ for some } i\}$. The type-I error can be computed by:

$$\begin{aligned} \beta_{P(\mathbf{X}\in\mathcal{D}_j)=0}(\phi) &= P_{P(\mathbf{X}\in\mathcal{D}_j)=0}(\mathbf{X}_i \in \mathcal{D}_j \text{ for some } i) \\ &= 1 - (1 - P(\mathbf{X} \in \mathcal{D}_j))^n|_{P(\mathbf{X}\in\mathcal{D}_j)=0} = 0 < \alpha. \end{aligned} \tag{F.4}$$

Therefore, the test function $\phi_j(\mathbf{s}_n) = \mathbf{1}(\mathbf{s}_n \in \mathcal{R}_j)$ is a level α test here.

Next, similarly we can obtain the rejection region $\mathcal{R}_{3j} = \{\mathbf{s}_n|\mathbf{x}_i \in \overline{\mathcal{D}}_j \text{ for some } i\}(j = 1, \ldots, d^*)$ for testing $H_{0j} : P(\mathbf{X} \in \overline{\mathcal{D}}_j) = 0$ vs. $\mathrm{H_{1j} : P(\mathbf{X} \in \overline{\mathcal{D}}_j) > 0}$.

Lastly, we obtain the rejection region $\mathcal{R}$ of testing $H_0 : \mathrm{IMP} = 0$ vs. $\mathrm{H_1 : IMP > 0}$ as formed by $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_{21} \cap \cdots \cap \mathcal{R}_{2d^*} \cap \mathcal{R}_{31} \cap \cdots \cap \mathcal{R}_{3d^*}$. And the test function $\phi = \mathbf{1}_{\mathbf{s}_n \in \mathcal{R}}$ is also a level α test by the IUT theorem. Q.E.D.