Condition-dependent coexpression in Arabidopsis and the bootstrap estimation of false discovery rate

by

Hui Lan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Condition-dependent coexpression in Arabidopsis
and the bootstrap estimation of false discovery rate

Hui Lan

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2015

Conventional coexpression analysis looks at gene expression levels over diverse conditions. Although such "condition independent" analyses provide useful generalized information, the coexpression of genes can depend on biological context. Looking at gene coexpression in a condition-dependent fashion identifies gene interactions relevant to more-specific biological questions. This thesis addresses condition-dependent coexpression in the context of seed germination in *Arabidopsis thaliana*. We study two broad aspects of the problem: (i) detecting pairwise interactions, and (ii) detecting three-way interactions. First, using gene expression data gathered exclusively from mature seeds, we develop a coexpression network, SeedNet. We analyze its statistical properties and find, for example, that SeedNet bifurcates into two previously undescribed components, one corresponding to germination and one to non-germination. Second, we develop a detector for three-way interactions in gene expression data and demonstrate its efficacy. In this approach, the coexpression of a pair of genes can depend on the expression level of a third gene. That is, coexpression is conditioned on the transcriptional environment. Because of the vast number of possible three-way interactions, there is an extremely large potential for false positives, so it is crucial to accurately estimate the False Discovery Rate (FDR). We show that extending approaches used for two-way interactions can significantly underestimate the FDR of three-way interactions. The thesis develops and evaluates a new approach for estimating FDR based on the bootstrap. We show that the bootstrap approach produces reasonable estimates over a wide range of statistical conditions, whereas conventional approaches rapidly break down. Finally, we use our detector and FDR estimates to add over 64,000 new, highly-confident edges to SeedNet.

Acknowledgements

I would like to thank my PhD supervisor Anthony Bonner for his invaluable guidance and crucial support during the years. He spent numerous hours with me on discussing the thesis work, through countless emails and meetings. I learned a great deal from him. Besides recommending highly technical books and papers on machine learning and statistics for me to read, he lent me the two little books that I particularly enjoyed, "The Elements of Style" and "The Cartoon Guide to Genetics". I also thank my PhD co-supervisor Nicholas Provart for providing important biological data and expert insights. We had several delightful meetings. Both of them proofread this thesis. I especially thank Igor Jurisica, Quaid Morris, George Bassel, Wei Xu, Wei Wang and Zhaolei Zhang for their valuable time and constructive inputs. I thank Wei Wang for flying from Los Angeles to Toronto to attend my defense. I thank Kamran Behdinan for chairing the defense. I thank Pingzhao Hu and Wei Xu for providing short-term research opportunities at SickKids and PMH, respectively. Thanks to the computing support team at the Computer Science Department for maintaining a strong computing environment.

Finally, I would like to thank my friends and parents for their long-term support.

Contents

1 Introduction **1**

2 Two-way interactions for seed germination **5**

2.1 Introduction 5

2.2 Biological background 6

2.3 Results 8

2.3.1 Two large clusters 8

2.3.2 Contributing factors to correlation 16

2.3.3 Changes in correlation and variance 21

2.3.4 Power-law analysis and cutoff estimation 28

2.3.5 Geometric random graphs 35

2.3.6 False discovery rate 35

2.4 Materials and Methods 38

2.4.1 Fitting a power law to a graph 38

2.4.2 Choosing a correlation threshold 39

2.4.3 Estimating the false discovery rate 41

2.5 Conclusions and biological interpretation 43

3 Three-way interactions and bootstrap estimation of FDR **46**

3.1 Introduction 46

3.1.1 Three-way interactions 46

3.1.2 False Discovery Rate 48

3.1.3 Validation 50

3.2 Background and related work . . . 51
3.2.1 Previous works and their limitations . . . 51
3.2.2 The bootstrap . . . 52
3.3 Methods . . . 57
3.3.1 Detecting three-way interactions in expression data using regression . . . 57
3.3.2 Estimating FDR . . . 60
3.4 Results and discussion . . . 65
3.4.1 Correlation change . . . 65
3.4.2 Extending SeedNet . . . 76
3.4.3 Enrichment tests . . . 78
3.4.4 False Discovery Rate . . . 87
3.4.5 Correlated predictors . . . 94

4 Validating FDR estimates **100**

4.1 Results on simulated data . . . 100
4.1.1 Quadratic data . . . 102
4.1.2 Cubic data . . . 106
4.1.3 Many predictor genes . . . 108
4.1.4 Dependent data . . . 115
4.1.5 Multi-modal data . . . 117
4.1.6 More-complex non-linearities . . . 120
4.2 Results on real/simulated data . . . 127
4.2.1 Cubic data . . . 127
4.2.2 Many predictor genes . . . 129
4.2.3 Max data . . . 131
4.2.4 Switch data . . . 133
4.3 Results on other non-Gaussian data . . . 135
4.3.1 Copulas . . . 135
4.3.2 Multivariate beta distributions . . . 139
4.3.3 Multivariate Laplacian distributions . . . 139

4.4 Results on mixtures of correlations 142

5 Summary and future work **150**

Appendices **157**

A Appendix for Chapter 2 **157**

A.1 Covariance decomposition 157

A.2 Preferential expression 160

A.3 Degree distributions in geometric random graphs 163

Bibliography **168**

List of Tables

3.1 Correlation properties of the top 50 discoveries on the seed germination/dormancy data from *Arabidopsis thaliana* (i.e., the 50 three-way interactions with the highest values of $|z|$). The FDR for these discoveries is 0.006 (as estimated by the bootstrap method). g_1, g_2 and g_3 form a gene triple, where g_3 is the target gene, and g_1 and g_2 are predictor genes. C_a and C_b are the Pearson correlation between the profiles of g_1 and g_3 on the 50 High Group samples and the 50 Low Group samples, respectively. C_1 and C_2 are the correlation on all samples between the profiles of g_1 and g_3 and between the profiles of g_2 and g_3, respectively. 68

3.2 Gene annotations for the genes in Table 3.1. 76

List of Figures

2.1 Histograms of raw and log-transformed gene expression levels for the 14,088 *Arabidopsis* genes on 138 seed samples. 9

2.2 A heatmap of a network constructed from gene pairs whose correlations exceed 0.6. Two large clusters emerge from the heatmap. 9

2.3 A heatmap of a gene network constructed from gene pairs whose correlations exceed 0.6. In addition, the gene pairs whose correlations below -0.6 are plotted on the heatmap. The two clusters along the diagonal line are the two clusters (as those in Figure 2.2), in which genes are correlated. The two clusters off the diagonal show gene pairs that are anti-correlated (correlation coefficients less than -0.6). They also indicate that genes between clusters tend to be anti-correlated. 10

2.4 The heatmaps of different networks that correspond to different thresholds, τ. The networks consistently bifurcate into two clusters. 11

2.5 Histograms of correlation coefficients for gene pairs (a) in which both genes are from cluster G, (b) in which both genes are from cluster D, and (c) in which one gene is from cluster G and the other is from cluster D. 12

2.6 Histograms of GS for gene profiles (red) and for randomly permuted gene profiles (green). 14

2.7 A heatmap of a network constructed from gene pairs whose correlations between the two genes are greater than 0.6. The yellow dots represent genes that are highly expressed in germinating seeds (GS $>$ 0.45), and the red dots represent genes that are highly expressed in non-germinating seeds (GS $<$ -0.45). 15

2.8 Histograms of GS values in cluster G and in cluster D. 16

2.9 An example illustrating the effect of preferential expression on correlation. (a) The green dots show one gene expression profile, and the red dots show another. The vertical heights of the first $n_1 = 65$ dots are $10 + e_i$, $i = 1, 2, ..., 65$. The vertical heights of the remaining $n_2 = 73$ dots are $2 + e_i$, $i = 1, 2, ..., 73$. Here, e_i is normally distributed with mean 0 and variance 1. The correlation coefficient between the green dots and the red dots is high (0.94) on all $n_1 + n_2 = 138$ samples, while it is low on the first $n_1 = 65$ samples (0.15) and on the remaining $n_2 = 73$ samples (0.04). In this case, preferential expression is the dominant cause of the high correlation over all samples. (b) The preferential expression of each gene is removed by making the first 65 expression levels have the same mean as the remaining 73 expression levels. The overall correlation now plummets from 0.94 to 0.08. 17

2.10 The decomposition of covariances for 100,000 randomly chosen gene pairs. Each gene pair is represented by a blue dot, a red dot and a black dot. For all three dots, the horizontal axis is the total covariance between the two genes within the pair, denoted Cov_{total}. The meaning of the vertical axis depends on dot colour. The y-coordinate of a blue dot represents their between-group covariance, $Cov_{between}$, i.e., the contribution to total covariance due to preferential expression. The y-coordinate of a red dot represents their within-group covariance, Cov_{within}, i.e., the contribution to total covariance due to coexpression during germination and during non-germination. The y-coordinate of a black dot, equal to its x-coordinate, represents their total covariance, Cov_{total}. For all gene pairs, $Cov_{total} = Cov_{between} + Cov_{within}$. 20

2.11 Correlation based on the adjusted gene profiles versus correlation based on the original (unadjusted) gene profiles. Each dot represents a gene pair. The horizontal axis, c, is the correlation computed on the original (unadjusted) gene profiles. The vertical axis, c', is the correlation computed on the adjusted gene profiles, after preferential expression has been removed. 22

2.12 (a) Histogram of correlation c_1 on germinating seeds for all gene pairs. It is roughly symmetric around 0. The distribution of c_2 for all gene pairs is similar. (b) Histogram of correlation change $c_1 - c_2$ from germination to non-germination for all gene pairs. 25

2.13 Histogram of correlation coefficients on germinating seeds for genes in set G (left), and histogram of correlation coefficients on non-germinating seeds for genes in set G (right). 25

2.14 Histogram of correlation change from germination to non-germination for genes in set G (left), and for genes in set D (right). Both histograms are roughly symmetric around 0. 26

2.15 Histograms of $\log_{10} v_1 - \log_{10} v_2$ for genes in set G, genes in set D, and all genes. 27

2.16 Empirical cumulative frequency versus node degree. A blue dot represent the number of nodes with degree at most d, where d is a node degree in the network. 29

2.17 Fitting different power laws, d^{-k}, to a network defined by correlation cutoff $\tau = 0.7$. Here, $k = 0.5$, 0.7, 0.9, 1.1, 1.3 and 1.5. Both visually and in terms of $nFit$, the best fit is achieved with $k = 0.9$. 31

2.18 A heatmap of a network constructed from gene pairs whose correlations exceed 0.7. 32

2.19 Empirical cumulative frequency versus expected cumulative frequency for node degree in a network with threshold $\tau = 0.7$. The red line shows the the power law fit for the cumulative node degree distribution. The distribution approximately follows $d^{-0.91}$, where d is node degree. $nFit = 0.00826538$. 33

2.20 Number of nodes versus node degree for the network of Figure 2.19. The red curve is the power law that best fits the network. 34

2.21 Power law fit for degree distribution in three networks with thresholds, $\tau = 0.5$, 0.6 and 0.8, respectively. The k value for d^{-k} is labelled on top of each plot of empirical cumulative frequency versus expected cumulative frequency. 36

2.22 False discovery rate (FDR) versus correlation threshold (τ). The right-hand curve is a close up of the tail of the left-hand curve. 37

2.23 A visualization of SeedNet (image courtesy of George W. Bassel), reproduced from [1]. The left part of the network represents cluster D, and the right part represents cluster G. The red and blue dots represent genes that are significantly associated with non-germination and germination, respectively [1]. 44

3.1 The distribution of 10,000 p-values for 10,000 randomly picked triples from the seed germination/dormancy data. This histogram shows a peaky uniform distribution, peaked near 0. The vertical yellow bar represents the value of λ. 65

3.2 The graphical representation of three-way interactions from Table 3.1. The dark grey nodes represent target genes (g_3). The light grey nodes represent predictor genes (g_1 or g_2). 77

3.3 Gene-function enrichment curve (X vs. N). X is the number of interesting triples among the top N detections. A triple is interesting if it is in S_I^1 (i.e., at least one of its predictor genes promotes germination/dormancy). The blue line is the plot for our detector. The red line is the plot for a random detector of three-way interactions. P-values at 0.1%, 1%, 10% and 50% of all triples are 1.097E-13, 0, 0 and 0, respectively. 81

3.4 Gene-function enrichment curve. A triple is interesting if it is in S_I^2 (i.e., both predictor genes in the triple promote germination/dormancy). P-values at 0.1%, 1%, 10% and 50% of all triples are 0.017, 2.506E-13, 0 and 0, respectively. 82

3.5 PPI enrichment curve. A triple is interesting if at least one pair of genes from the triple is a known protein-protein interaction. P-values at 1%, 10% and 50% of all triples are all 0. 84

3.6 TF-target enrichment curve. A triple (g_1, g_2, g_3) is interesting if both (g_1, g_3) and (g_2, g_3) are in `agris`. P-values at 1%, 10% and 50% of all triples are 0.0025, 0.000196 and 0.0002, respectively. 86

3.7 3-way-interaction enrichment curve. A triple (g_1, g_2, g_3) is interesting if it is in PTM-Switchboard. P-values at 10%, 20% and 50% of all triples are 0.0086, 1.06E-05 and 1.89E-14, respectively. 88

3.8 Histogram of z for the interesting triples, i.e., the triples in PTM-Switchboard. . 89

3.9 Generate an FDD curve. 91
3.10 Number of False Discoveries versus number of Discoveries for the FD estimation methods on the seed germination/dormancy data from *Arabidopsis thaliana*. The green curve is estimated by the bootstrap, the blue by total permutation, the pink by partial permutation, and the black by analytical t. 92
3.11 Number of False Discoveries versus number of Discoveries for the FD estimation methods on the yeast cogrim data. The green curve is estimated by the bootstrap, the blue by total permutation, the pink by partial permutation, and the black by analytical t. 93
3.12 Variance of the bootstrap FDD curve on the germination/dormancy data. The ten green curves are generated by repeating the bootstrap method ten times, each with a different randomization. The FDD curves for top two million discoveries are shown. 95
3.13 True FDD curves under different correlations, ρ, between two predictor genes. $\rho = 0, 0.3, 0.6$ and 0.9. Higher curves correspond to larger values of ρ. 96
3.14 Two distributions of z values of the interaction term in the second-order model, when the predictor genes are uncorrelated ($\rho = 0$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples. 97
3.15 Two distributions of z values of the interaction term in the second-order model, when the predictor genes are moderately correlated ($\rho = 0.8$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples. 98
3.16 Two distributions of z values of the interaction term in the second-order model, when the predictor genes are highly correlated ($\rho = 0.99$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples. 99

4.1 Generate data in which two predictor genes have a quadratic relationship with a target gene. Note that the coefficients of the quadratic (a through f) can be arbitrarily close to 0, thus allowing for arbitrarily-weak interactions. 103

4.2 FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0$, $nl_1 = 2$, $nl_2 = 0$ (no measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 104

4.3 FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0$, $nl_1 = 0.5$, $nl_2 = 0.5$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 105

4.4 FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0, nl_1 = 0.5, nl_2 = 0.5$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 106

4.5 Same as Figure 4.4 but showing only the top 2,000 discoveries. Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 107

4.6 Generate data in which two predictor genes have a cubic relationship with a target gene. 109

4.7 FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 110

4.8 FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0.3$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 111

4.9 FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0.6$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 112

4.10 Generate data in which each target gene has more than two predictor genes. Here, $P > 2$ is the number of predictor genes for each target gene, and λ is in $[0, 1]$ and determines the average correlation between the P predictor genes. This generates multi-way interactions between $P + 1$ genes. 114

4.11 FDD curves of the FD estimation methods on `DataMany`: $P = 3$, $\lambda = 0.2$, $nl_1 = 0.1$, $nl_2 = 0.1$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 115

4.12 Generate dependent data where triples can share genes and genes can be correlated across triples. Here, L is the size of the pool of shared genes, and λ is in $[0, 1]$ and determines the average correlation between genes in the pool. 116

4.13 FDD curves of the FD estimation methods on `DataDependent`: $L = 1000$, $\lambda = 0.2$, $nl_1 = 0.3$, $nl_2 = 0.3$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 118

4.14 Generate data in which the expression levels of predictor genes have a Gaussian mixture model. Here, K specifies the number of mixture components, and μ_{size} determines the average distance between the components. Since each target gene has two predictor genes, the mixture models are two-dimensional. 119

4.15 FDD curves of the FD estimation methods on `DataMixture`: $K = 3$, $\mu_{\text{size}} = 0.5$, $nl_1 = 0.1$, $nl_2 = 0.1$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 120

4.16 Representative multi-modal data for two predictor genes, x and y, generated using a Gaussian mixture model with three components. $N = 10,000, K = 3, \mu_{\text{size}} = 1$. 121

4.17 Generate data for which, in an interacting triple, the target gene is controlled by the stronger of the two predictor genes. 123

4.18 FDD curves of the FD estimation methods on `DataMax`: $\rho = 0$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 124

4.19 Generate data for which, in an interacting triple, the interaction between genes g_1 and g_3 switches between two modes, depending on the expression level of gene g_2. 125

4.20 FDD curves of the FD estimation methods on `DataSwitch`: $\rho = 0$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). . . . 126

4.21 Generate cubic data where two predictor genes have a cubic relationship with a target gene. The expression profiles for the two predictor genes are from real data. 128

4.22 FDD curves of the FD estimation methods on `RealCubic`: $nl_1 = 0.4, nl_2 = 0.4$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 129

4.23 FDD curves of the FD estimation methods on `RealCubic`: $nl_1 = 1, nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 130

4.24 FDD curves of the FD estimation methods on `RealMany` data: $nl_1 = 0.5, nl_2 = 0.5$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 131

4.25 FDD curves of the FD estimation methods on `RealMax`: $nl_1 = 0.1, nl_1 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 132

4.26 FDD curves of the FD estimation methods on `RealSwitch`: $nl_1 = 0.2, nl_2 = 0.2$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 133

4.27 FDD curves of the FD estimation methods on `RealSwitch`: $nl_1 = 0.5, nl_2 = 0.5$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 134

4.28 The scatter plot and marginal distribution of 10,000 2-dimensional normally distributed data points with mean $\mu = (0, 0)$ and covariance $\Sigma = [1.0, 0.8; 0.8, 1.0]$. Notice that the two variables, x and y, are strongly correlated, and each follows a Gaussian distribution. 136

4.29 The scatter plot and marginal distribution of 10,000 2-dimensional uniform data points from a Gaussian copula (constructed using the data points in Figure 4.28). Notice that the two variables, x and y, are still correlated, but now each has a uniform distribution. 137

4.30 The scatter plot and marginal distribution of 10,000 2-dimensional data points following a beta distribution. The data points are generated by applying the inverse CDF of a beta distribution (with parameters $A = 2$ and $B = 5$) to the data in Figure 4.29. Notice that the two variables, x and y, are still correlated, and each follows a beta distribution. 138

4.31 Generate bivariate beta-distributed data. 140

4.32 FDD curves of the FD estimation methods on `DataBetaRandomRho`: $M_1 = 15000, M_2 = 5000, N = 100, \sigma_u = 0.3, nl_1 = 0.1, nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 141

4.33 Generate bivariate Laplacian data. 142

4.34 The scatter plot and marginal distribution of 10,000 2-dimensional data points following a Laplacian distribution. The data points are generated by applying the inverse CDF of the Laplacian distribution to the data from Figure 4.29. The two variables, x and y, are correlated ($r = 0.57$), and each follows a Laplacian distribution. 143

4.35 FDD curves of the FD estimation methods on `DataLaplaceRandomRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $\sigma_u = 0.3$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 144

4.36 Distribution of 20,000 simulated correlation coefficients, ρ, where $\rho = (e^{2u} - 1)/(e^{2u} + 1)$, and $u \sim \mathcal{N}(0, \sigma_u)$ for $\sigma_u = 0.3$. 146

4.37 FDD curves of the FD estimation methods on `DataCubicRandomRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $\sigma_u = 0.3$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 147

4.38 Distribution of 20,000 real correlation coefficients in the seed germination/dormancy data for *Arabidopsis*. 148

4.39 FDD curves of the FD estimation methods on `DataCubicRealRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red). 149

A.1 Histograms of degree distribution in four geometric random graphs in 2-dimensional space, $G(14088, 0.161)$, $G(14088, 0.105)$, $G(14088, 0.058)$, and $G(14088, 0.024)$. The values of the second parameter in G, $r = 0.161, 0.105, 0.058$, and 0.024, are chosen such that the resulting graphs have roughly the same average node degree as the coexpression networks with correlation cutoffs $\tau = 0.5, 0.6, 0.7$, and 0.8, respectively. 165

A.2 Similar to Figure A.1, but the geometric random graphs are in 3-dimensional space. 166

A.3 Similar to Figure A.1, but the geometric random graphs are in 4-dimensional space. 167

Chapter 1

Introduction

Gene coexpression is a powerful bioinformatics tool [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Coexpressed genes may be involved in the same biological process, and thus coexpression networks are often used to investigate gene function [1, 2, 3, 5, 7, 8]. Methods for detecting, analyzing and clustering pairs of coexpressed genes are now well developed. However, conventional coexpression analysis looks at the expression levels of pairs of genes over a diverse set of conditions [2, 11]. Although such "condition independent" analyses provide useful generalized information, the coexpression of a pair of genes depends on the biological context, including environmental factors and the expression levels of other genes. Looking at gene coexpression in a "condition dependent" fashion should more precisely identify gene interactions relevant to answering more-specific biological questions [1].

This thesis addresses this issue in the specific context of seed germination in *Arabidopsis thaliana*, the model organism of plant biology. Intact viable seeds are in a state either of germination or of dormancy. Dormancy is an evolutionarily acquired trait that prevents germination when seeds are not under favourable conditions, and serves both to give sufficient time for seed dispersal and to distribute germination of a seed population over time. Germination is an irreversible developmental process that consists of complex events happening between imbibition (i.e., absorption of water) and emergence of the radicle, and has great agronomic and ecological significance [1, 12]. The full mechanism of seed germination and dormancy has attracted a lot of research but still remains unclear [1, 12]. This thesis aims to enhance our understanding of it with computational approaches applied to gene expression data. Our methods

are condition-dependent, taking both biological conditions and transcriptional conditions into consideration.

We study two aspects of the problem: (i) detecting pairwise interactions, and (ii) detecting three-way interactions. In the first case, we build and analyze SeedNet [1], a coexpression network based, not on a diverse set of data, but on data gathered exclusively from imbibed mature seeds of *Arabidopsis*. In SeedNet, the nodes are genes and an edge between two genes means that their expression profiles are highly correlated, i.e., the correlation coefficient exceeds some threshold, which is chosen so that the network fits a scale-free graph as closely as possible [13]. (Geometric random graphs [14] were also tried, but they do not fit the network at any threshold (Section 2.3.5).) We analyze the statistical properties of the network and find, for example, that the network consists of two main clusters, one corresponding to germination and one to non-germination. The analysis also reveals intriguing properties in the correlation and covariance structure of the two clusters. The thesis provides a detailed development of the methods used to construct and analyze SeedNet, including a robust algorithm for determining the optimum correlation threshold for approximating a scale-free graph. These methods and analysis techniques are not limited to seed germination in *Arabidopsis*, but are general and can be applied to other organisms and other condition-dependent data. SeedNet itself is available online as a community resource (http://vseed.nottingham.ac.uk). These results are described in Chapter 2. Material in this chapter forms the computational contribution of [1]. The chapter itself is also being submitted for publication as a separate paper.

In our second approach, detecting three-way interactions, the coexpression of a pair of genes can depend on the expression level of a third (unspecified) gene. That is, coexpression is conditioned on the transcriptional environment. Moreover, by searching for three-way interactions, the thesis addresses another limitation of conventional coexpression analysis: it focusses on *pairwise* relationships between genes. Pairwise coexpression is clearly too simplistic to describe the complex relationships between gene expression levels, since these relationships can involve multiple genes [15, 16, 17, 18, 19, 20, 21]. For example, the coexpression of two genes may be modulated by a third gene [16, 20], and in genetic regulatory networks, the expression level of a gene can be a combinatorial function of several transcription factors [15, 17, 18]. In general, pairwise coexpression does not capture higher-order statistical dependencies or the complex

biological relationships they reflect [19]. Methods are therefore needed for detecting three-way (and more generally, multi-way) interactions in gene expression data.

To this end, we develop a simple algorithm for detecting three-way interactions in gene expression data and demonstrate its efficacy on the seed data from *Arabidopsis* and on data from yeast. We also show that discoveries made by the detector exhibit the expected correspondence between three-way interaction and transcriptionally-dependent coexpression; that is, when three genes interact, the coexpression of two of them depends on the expression level of the third. Finally, we develop a bootstrap method for estimating the false discovery rate (FDR) of the discoveries made by our detector. Applying our detector to the seed data and estimating the FDR of the discoveries, we add over 64,000 new, highly-confident edges to SeedNet. These edges cannot be detected by traditional, pairwise coexpression analysis. This is one way in which three-way interaction analysis extends the usability of gene expression data. These results are described in Chapter 3.

Because of the vast number of possible three-way interactions, there is a large potential for false positives. Thus, a crucial step in the discovery process is the accurate estimation of the false discovery rate (FDR) [22], since only if the FDR is low can a discovery be confidently declared. Estimating the FDR of three-way interactions in gene expression data faces two main challenges: (i) the underlying distribution of the data is unknown, and (ii) estimating the null distribution is considerably more subtle than for two-way interactions.

The bootstrap is a well-known solution to the first problem [23], and this thesis explores its utility in addressing the second problem. In particular, we develop a method based on the bootstrap for estimating the FDR of our regression-based detector. We test the method and compare it to other methods used in the literature, including permutation tests and an analytical t-test [20, 23, 24]. Our tests show that these methods produce widely differing estimates on both real and simulated data, often differing by more than an order of magnitude. In particular, the bootstrap method consistently produces by far the largest estimates. This indicates that either the bootstrap method is overestimating or the other methods are underestimating the number of false discoveries.

It is impossible to determine which method is more accurate without knowing all the three-way interactions in a data set. Since this is unknown for large biological datasets, we test the

methods on simulated data, for which all interactions are known. While the simulations are a highly simplified approximation of biological reality, they do provide strong evidence that the bootstrap method is the most accurate over a wide range of statistical conditions. In particular, while all the methods give good estimates on idealized data, the bootstrap method consistently gives the best estimates on more complex data, while the other methods rapidly break down, consistently underestimating the true number of false discoveries, often by more than an order of magnitude. In sum, our bootstrap method gives the best available estimates of FDR for three-way interactions over a wide range of statistical conditions. These results are described in Chapter 4, which, together with Chapter 3, is being submitted for publication.

Chapter 2

Two-way interactions for seed germination

2.1 Introduction

Seed germination and dormancy are widely seen in flowering plants and have great ecological and agricultural importance. The internal mechanism by which genes work together to maintain dormancy and to facilitate germination, however, is still poorly understood, albeit actively sought. Two fundamental questions remain unanswered: how is germination completed, and how is dormancy maintained? A few genes involved in seed germination have been identified [25], but how they interact to regulate germination is still largely unknown.

Gene interaction and regulation is often investigated using coexpression networks [1, 2, 4, 7, 8], since genes that are coexpressed may be involved in the same biological process. However, context-independent coexpression analyses based on gene expression data from diverse sources, such as AraNet [11], are perhaps not useful in answering specific biological questions, such as questions about the regulation of seed germination. To address this specific question, we use gene expression data from *Arabidopsis thaliana* generated exclusively from samples of germinating and non-germinating seeds. More complicated methods such as supervised statistical learning do not produce useful results when classifying germinating genes and dormant genes. Instead we use a simple but robust correlation approach to build a coexpression network, called

SeedNet [1], that gives biological insight into seed germination and dormancy. Node degree in the network approximately obeys a power-law distribution. Three previously undescribed regions of interactions are clearly present in the network, corresponding to germination, dormancy and the transition between them [1]. The formation of these distinct regions is not due to preferential expression during germination or dormancy, but due to correlation during germination and during dormancy. Nevertheless, the region associated with germination is enriched with genes that are preferentially expressed during germination, while the region associated with dormancy is enriched with genes that are preferentially expressed during non-germination. The false discovery rates for edges in the network are estimated to be extremely low. Results based on our network also suggest that there is more complex transcriptional regulation during dormancy than during germination [1].

2.2 Biological background

A dry quiescent seed resumes its metabolic activities shortly after imbibing water. Imbibing water is crucial for either germination or dormancy. The uptake of water is rapid initially, followed by a plateau, and increases again after finishing germination. Imbibition enlarges the seed, weakens its coat, converts the membrane from gel state to sol state, and increases oxygen permeability. Meanwhile, respiration and protein synthesis commence, enzymes are activated, DNA damaged during maturation drying is repaired, and mitochondria is repaired or synthesized. Extant mRNAs and newly transcribed mRNAs encode proteins, supporting the germination process. Under favourable environmental conditions, the radicle (a part of seed embryo that later becomes the primary root) penetrates the seed coat, marking the completion of germination. If, for some unknown reasons, the radicle of an imbibed seed fails to protrude its surrounding structures, we say that the seed enters dormancy. There are two types of dormancy, coat-enhanced dormancy and embryo dormancy. Notice that dormancy does not mean quiescence. Conversely, dormancy is busily maintained through a complex sequence of enzymatic and metabolic activities. A dormant seed is internally busy, both in maintaining its dormancy state and in receiving external signals, such as chilling, to break dormancy. It germinates when properly stimulated by environmental and/or chemical stimuli.

Causes of dormancy. An imbibed seed becomes dormant if its radicle encounters a mechanical barrier (e.g., constraining endosperm and testa), or the seed embryo itself prevents the radicle from elongating (even in the absence of constraining structures). Dormancy is thought to be due to deficiency in required substance for completing germination, or due to blockage by the presence of certain substance. As an analogy, a car could not move forward if it is deficient in gasoline or blocked by some barrier. This suspension of growth not only has theoretical significance but also has practical benefits as it allows the seed to avoid bad seasons or premature germination.

The effects of hormones in germination and dormancy. Two plant hormones, abscisic acid (ABA) and gibberellins (GAs), play an important role in promoting dormancy and promoting germination, respectively. ABA can prevent the embryo radicle from elongating, possibly by preventing the cell wall from loosening. Endogenous ABA is prevalent in many dormant seeds. For instance, sunflower seeds require continuous synthesis of ABA to maintain dormancy. Developing seeds on a parent plant deficient in and insensitive to ABA would germinate precociously. In contrast, GA promotes and maintains germination, for example, by activating endosperm-weakening enzymes to reduce the resistance to radicle growth from the surrounding structures, or by neutralizing the effects of ABA .

Release from dormancy usually consists of two major steps. First, the dormant seed receives dormancy-breaking signals and forms a signal transduction chain. The chain perhaps initiates the synthesis of GAs, or causes the seed to become sensitive to GAs, or both. Second, dormancy-maintaining mode is switched to dormancy-breaking mode, and structures surrounding the embryo are weakened (by enzymes) to facilitate radicle emergence.

Seed dormancy has been studied for many years but still remains mysterious. The difficulty mainly comes from the fact that the metabolism and respiration rate of imbibed dormant seeds are virtually indistinguishable (sometimes only subtly different) from that of germinating seeds, and that no cause-and-effect relationship can be established between particular proteins and dormancy/germination. Many so-called germination-specific mRNAs are actually only related to post-germination. In addition, the decisive events for dormancy or germination perhaps only happen in a few embryonic cells, and thus could be easily overlooked in the presence of data from other cells. It has been suggested that studying the changes of embryo transcripts during

dormancy and germination may unravel this mystery [12].

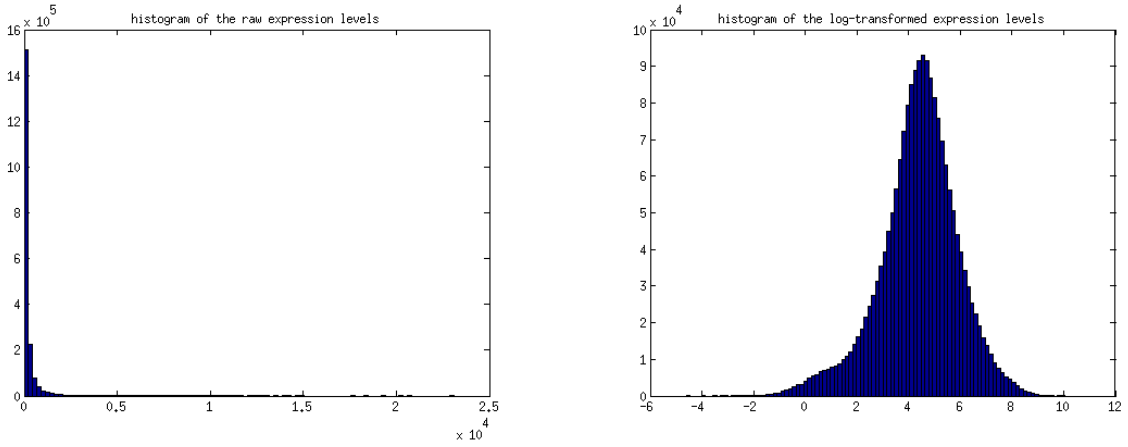
2.3 Results

As described in [1], the data we use contain gene expression levels for 14,088 *Arabidopsis* genes (after filtering) on 138 seed samples, of which 73 are non-germinating seeds and 65 are germinating seeds. The seeds were maintained under diverse physiological and environmental conditions (e.g., wild-type, mutation, low temperature, nitrate, far-red light, white and red light, abscisic acid treatment, and gibberellin acid treatment), and represent a wide range of developmental stages (e.g., after-ripened, primary dormant, and secondary dormant). The data forms a 14,088 by 138 matrix of gene expression levels. Figure 2.1 shows the data's histograms. We use log-transformed expression data as it is much more Gaussian-like than the raw data, as seen in Figure 2.1, and is therefore more appropriate for Pearson correlation.

2.3.1 Two large clusters

Using the gene expression profiles on the 65 germinating seeds and 73 non-germinating seeds, a simple gene regulatory network, pertaining to seed germination and non-germination, was constructed by putting a threshold (cutoff) on the correlation coefficients between genes. Specifically, the network has an edge between two genes if and only if $r >= \tau$, where r is the Pearson correlation coefficient of the two gene expression profiles and τ is a given threshold. Different values of τ define different networks, and we chose a value for τ that makes the network as close to scale-free as possible (Section 2.3.4.2).

Agglomerative hierarchical clustering was performed on the above network using the R function `flashClust` in the WGCNA package [7]. Clusters were merged based on their average dissimilarity, where two genes have a dissimilarity of 0 if they are connected by an edge in the network, and a dissimilarity of 1 otherwise. The `flashClust` function returns an ordering of the genes, which was used in generating heatmaps of the network. Figure 2.2 gives an example of such a heatmap of a network with $\tau = 0.6$. Each column and each row represents a gene, and a dark blue dot means that the two genes in that particular row and column are highly correlated (i.e., have a correlation coefficient at least 0.6).



(a) Raw expression levels

(b) Log-transformed expression levels

Figure 2.1: Histograms of raw and log-transformed gene expression levels for the 14,088 *Arabidopsis* genes on 138 seed samples.

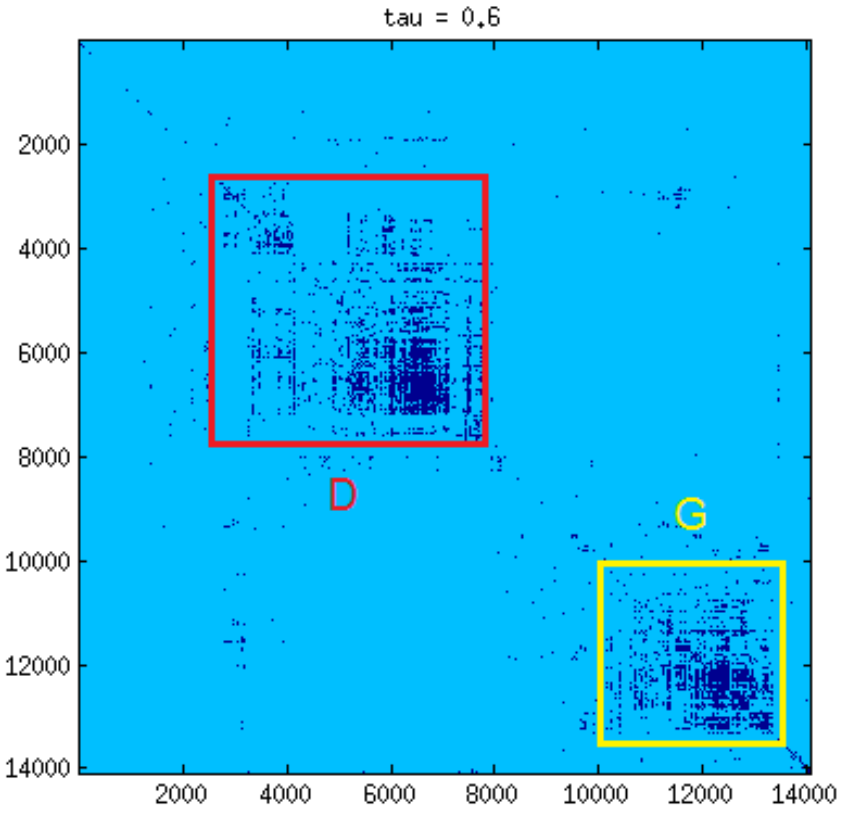


Figure 2.2: A heatmap of a network constructed from gene pairs whose correlations exceed 0.6. Two large clusters emerge from the heatmap.

The heatmap shows two large clusters. One cluster, extending roughly from genes 2,500 to 7,800, we call cluster D (for dormancy). The other cluster, extending roughly from genes 10,000 to 13,500, we call cluster G (for germination). It is worth noting that it is *not* the case that one cluster is due to correlation, while the other is due to anti-correlation, since the edges in our network (and the dark blue dots in the heatmap) only represent positive correlations. If we also display anti-correlated gene pairs, then the anti-correlations give rise to two new distinct clusters, as seen in Figure 2.3, where the upper-right and lower-left clusters show anti-correlated gene pairs.

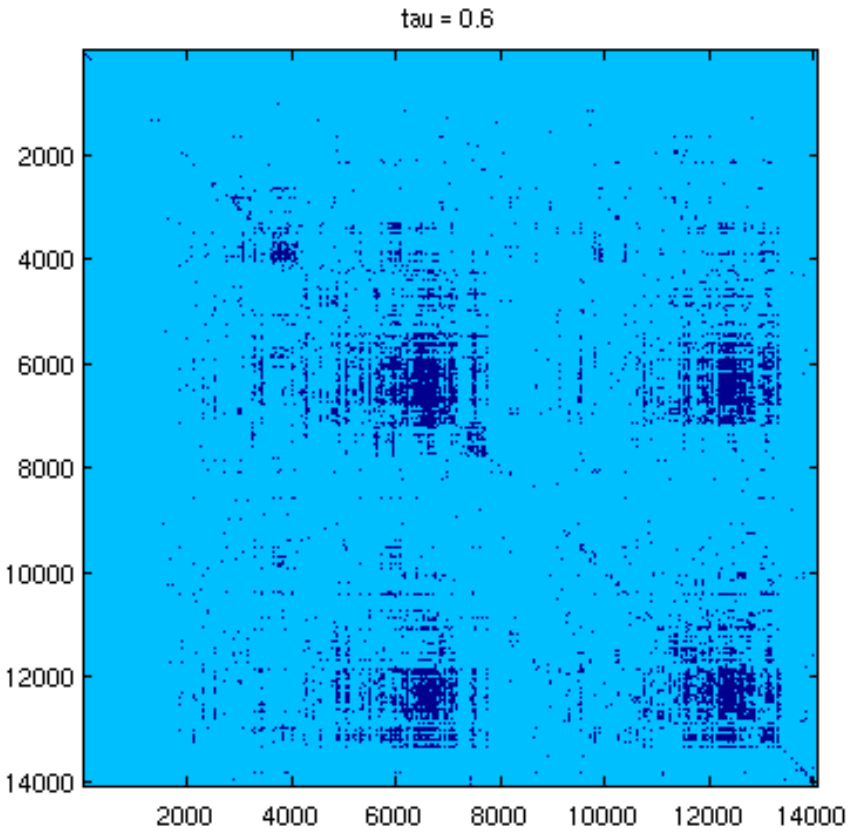
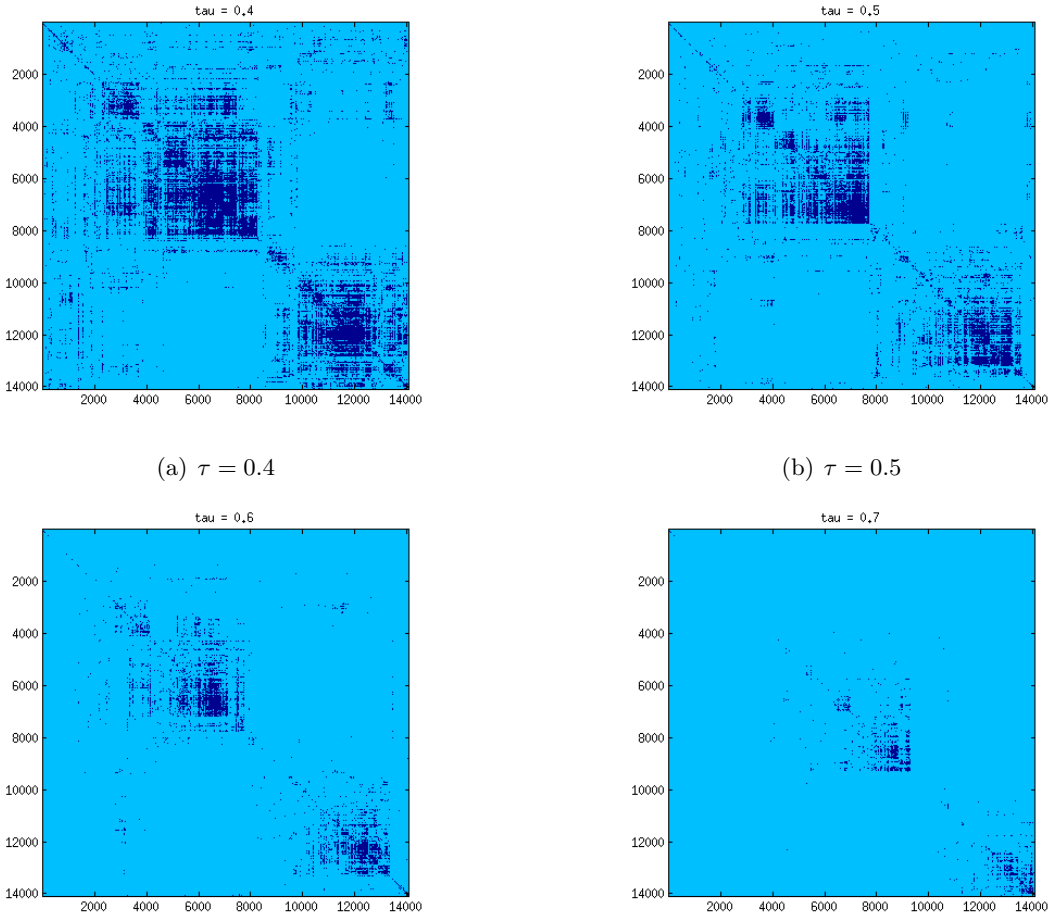


Figure 2.3: A heatmap of a gene network constructed from gene pairs whose correlations exceed 0.6. In addition, the gene pairs whose correlations below -0.6 are plotted on the heatmap. The two clusters along the diagonal line are the two clusters (as those in Figure 2.2), in which genes are correlated. The two clusters off the diagonal show gene pairs that are anti-correlated (correlation coefficients less than -0.6). They also indicate that genes between clusters tend to be anti-correlated.

The two large clusters are robust. That is, they continue to emerge over a wide range of

thresholds, τ, not just a single τ. We tested several values of τ, namely, $\tau = 0.40, 0.50, 0.60, 0.65, 0.70, 0.75$ and 0.80, and found that each version of the network bifurcated into two large clusters. Figure 2.4 shows the heatmaps of four of these networks. Different thresholds produce different networks with different sizes: the higher the threshold, the smaller the network. Clearly, each network has two large clusters. It would be suspicious if this bifurcation did not happen, as that would suggest the two clusters were an artefact of a particular parameter setting rather than a true biological phenomenon.



(a) $\tau = 0.4$

(b) $\tau = 0.5$

(c) $\tau = 0.6$

(d) $\tau = 0.7$

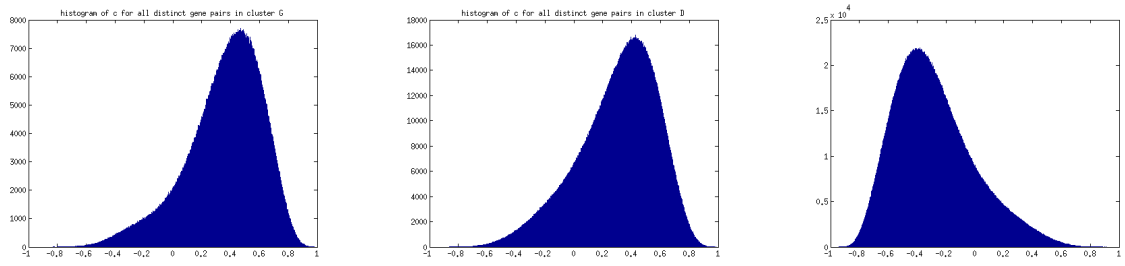
Figure 2.4: The heatmaps of different networks that correspond to different thresholds, τ. The networks consistently bifurcate into two clusters.

However, this raises the question of which value of the threshold, τ, best defines the seed-germination network. Too small a value would include many genes and edges that have nothing to do with germination, while too large a value would exclude many genes and edges that are

important for germination. In Section 2.3.4.2, we show that $\tau = 0.7$ defines a network that best-fits a scale-free graph; and in [1], we provide experimental evidence that at this threshold, the network explains many of the properties of seed germination in *Arabidopsis* and has predictive value.

In addition to robustness, the two large clusters also have a suggestive correlation structure. Within each cluster, genes tend to be correlated (as one might expect), but between clusters, genes tend to be anti-correlated (and not merely uncorrelated). For example, Figure 2.5(a) shows a histogram of the correlation coefficients of all gene pairs within cluster G. Notice that the vast majority of gene pairs within cluster G are positively correlated, and the histogram is strongly biased towards positive values, with a peak near 0.5. Although some genes within cluster G are anti-correlated, since the left-hand tail of the histogram includes negative values, some as low as -0.6, the great majority of correlations are positive. Correlation coefficients for gene pairs within cluster D have a very similar histogram (Figure 2.5(b)).

In contrast, Figure 2.5(c) shows a histogram of correlation coefficients of all gene pairs in which one gene comes from cluster G and the other from cluster D. Note that the histogram is biased towards negative values, with a peak near -0.4. That the two clusters tend to be anti-correlated, instead of merely uncorrelated, suggests that there is an antagonising relationship between them: the genes of one cluster tend to be upregulated while those of the other cluster are downregulated. This seems consistent with the hypotheses that cluster G is related to germination and cluster D to non-germination, a hypothesis that we examine more closely next.



(a) (b) (c)

Figure 2.5: Histograms of correlation coefficients for gene pairs (a) in which both genes are from cluster G, (b) in which both genes are from cluster D, and (c) in which one gene is from cluster G and the other is from cluster D.

2.3.1.1 The clusters are enriched with preferentially expressed genes

A gene is said to be preferentially expressed if its expression levels tend to be high in germinating seeds and low in non-germinating seeds. We use gene significance [7] (abbreviated as GS) to quantify the preferential expression of a gene.

Formally, GS is the correlation coefficient between gene expression levels and binary sample traits, so its value is between -1 and +1. Specifically, let $x = [x_1, x_2, ..., x_N]$ denote a gene's expression profile over N samples, and let $y = [y_1, y_2, ..., y_N]$ denote sample traits, where y_i is either 1 or -1. In our case, the samples are seeds, and $y_i = 1$ if seed i is germinating, and $y_i = -1$ otherwise. We call x an expression vector and y a trait vector. The Pearson correlation coefficient between x and y is the gene significance, GS. In Appendix A.2 (Corollary 4), we show that GS is proportional to the difference between the mean expression level on germinating seeds and the mean expression level on non-germinating seeds. In fact, GS is equivalent to a t statistic that measures the significance of this difference (Corollary 5 and [26]).

Intuitively, a GS value is significant if it rarely arises by chance. Figure 2.6 shows a histogram (in red) of GS values from gene profiles and a histogram (in green) of GS values from randomly permuted gene profiles.[1] The green histogram shows that a GS value between -0.1 and +0.1 could easily arise by chance, and is therefore insignificant. In contrast, a GS values smaller than -0.3 or larger than +0.3 is unlikely to arise by chance, and is therefore significant. This amounts to a permutation test for statistical significance. We therefore define a gene to be preferentially expressed during germination if its GS $>$ +0.3, and to be preferentially expressed during non-germination if its GS $<$ -0.3.

The two clusters in Figure 2.2 are each enriched with preferentially expressed genes. Figure 2.7 shows a heatmap of a simple correlation network with $\tau = 0.6$. Genes that are strongly preferentially expressed ($|\text{GS}| > 0.45$) are highlighted on the diagonal.[2] Yellow dots represent genes that have high preferential expression during germination (GS $>$ 0.45); red dots represent genes that have high preferential expression during non-germination (GS $<$ -0.45). Notice that yellow dots are concentrated in and around one cluster (cluster G), and red dots

[1] Because of the large number of genes involved, the shape of the green histogram is quite stable and changes very little if it is regenerated using different random permutations [data not shown].

[2] 0.45 was chosen because it is larger than the significance level 0.3 and because it gave enough diagonal dots in Figure 2.7 to illustrate the trend, but not so many as to saturate the figure.

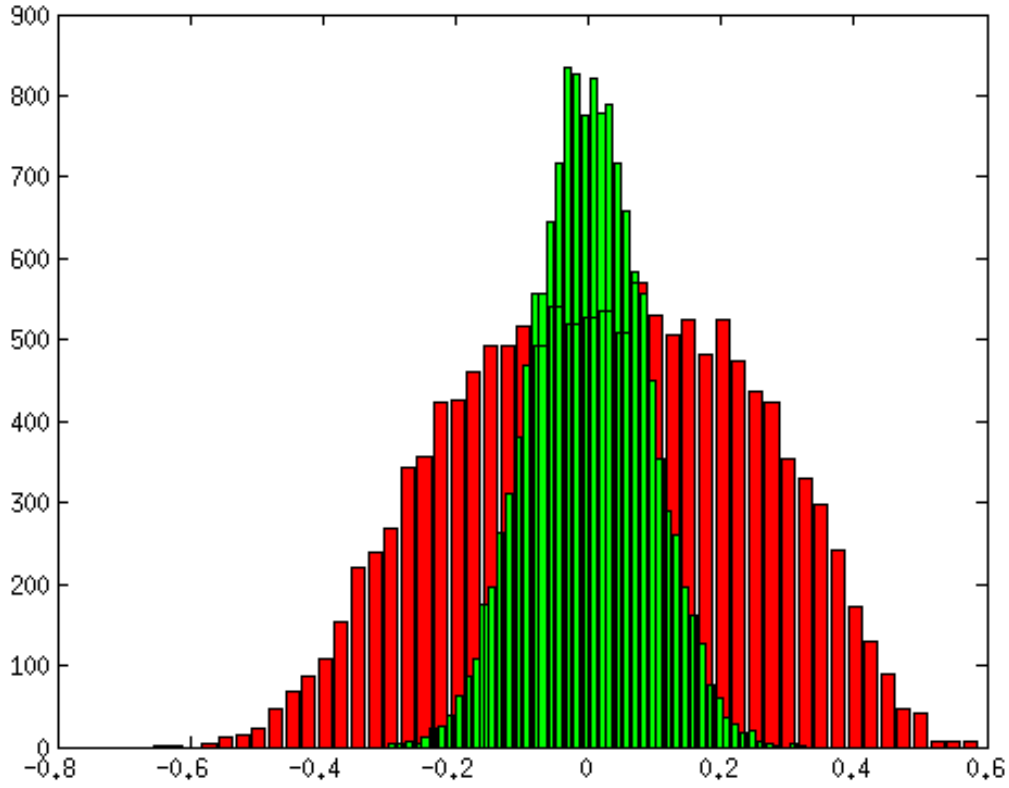


Figure 2.6: Histograms of GS for gene profiles (red) and for randomly permuted gene profiles (green).

are concentrated in and around the other cluster (cluster D). This figure shows the preferential enrichment of the two clusters: cluster G is enriched with genes that have high preferential expression during germination, and cluster D is enriched with genes that have high preferential expression during non-germination.

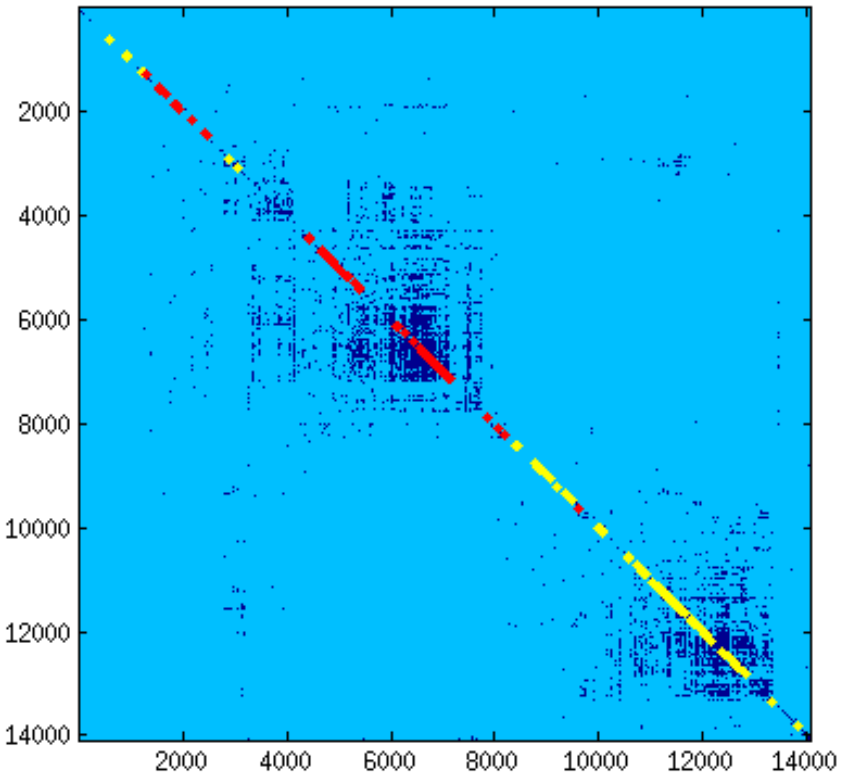
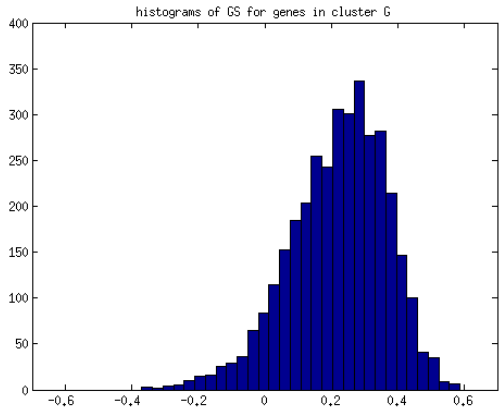
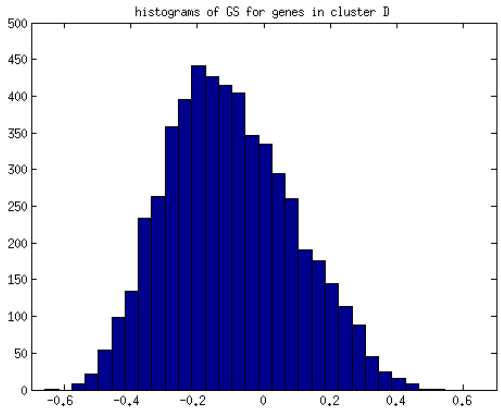


Figure 2.7: A heatmap of a network constructed from gene pairs whose correlations between the two genes are greater than 0.6. The yellow dots represent genes that are highly expressed in germinating seeds (GS $>$ 0.45), and the red dots represent genes that are highly expressed in non-germinating seeds (GS $<$ -0.45).

Figure 2.8 provides additional evidence for this preferential enrichment. It shows two histograms of GS for each of the two clusters. Notice that the histogram for cluster G is biased towards positive values, while the histogram for cluster D is biased towards negative values, showing the tendency for genes in cluster G to be preferentially expressed during germination, and the tendency for genes in cluster D to be preferentially expressed during non-germination.



(a) histogram of GS values for cluster G



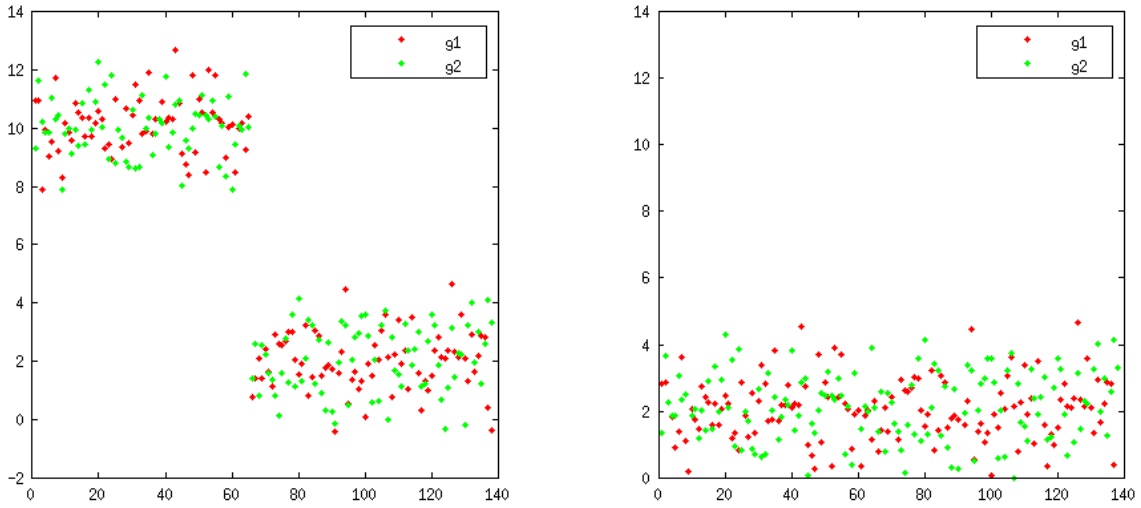
(b) histogram of GS values for cluster D

Figure 2.8: Histograms of GS values in cluster G and in cluster D.

2.3.2 Contributing factors to correlation

As described in Section 2.3.1, agglomerative hierarchical clustering based on correlation produces two major clusters, one closely related to genes that are preferentially expressed during germination, and the other to genes that are preferentially expressed during non-germination. Preferential expression itself is a contributing factor to correlation. Coexpression or interaction during germination and during dormancy is another contributing factor.

Since the two clusters are closely associated with preferential expression, preferential expression itself might be the dominant cause of correlation and cluster formation. After all, genes that are high during germination and low during non-germination seem, on average, to be high at the same time and low at the same time, and therefore correlated, even though they might not be correlated during germination or during dormancy (Figure 2.9). Similarly, they would be anti-correlated with the genes that are low during germination and high during non-germination, which would form a different cluster. If this were the case, then none of the results we have presented so far would be surprising. However, this section shows that preferential expression contributes little to correlation and cluster formation. Instead, correlation during germination and during non-germination is the dominant factor. High correlation between genes (*i.e.*, edges in our network) therefore reflects how genes work together during germination and during dormancy.



(a) preferential expression

(b) preferential expression removed

Figure 2.9: An example illustrating the effect of preferential expression on correlation. (a) The green dots show one gene expression profile, and the red dots show another. The vertical heights of the first $n_1 = 65$ dots are $10 + e_i$, $i = 1, 2, ..., 65$. The vertical heights of the remaining $n_2 = 73$ dots are $2 + e_i$, $i = 1, 2, ..., 73$. Here, e_i is normally distributed with mean 0 and variance 1. The correlation coefficient between the green dots and the red dots is high (0.94) on all $n_1 + n_2 = 138$ samples, while it is low on the first $n_1 = 65$ samples (0.15) and on the remaining $n_2 = 73$ samples (0.04). In this case, preferential expression is the dominant cause of the high correlation over all samples. (b) The preferential expression of each gene is removed by making the first 65 expression levels have the same mean as the remaining 73 expression levels. The overall correlation now plummets from 0.94 to 0.08.

2.3.2.1 Covariance decomposition

This section shows that in our gene expression data, preferential expression is a minor contributing factor to covariance. We focus on covariance here, rather than correlation, because covariance can be decomposed into a sum of contributions from independent parts: the contribution due to preferential expression, and the contribution due to coexpression during germination and during non-germination. (Section 2.3.2.2 focusses on correlation.)

Our decomposition of covariance uses an ANOVA-like approach. First, we divide the seeds into two groups, with group 1 consisting of germinating seeds, and group 2 consisting of non-germinating seeds. The seeds in group 1 are labelled with $1, 2, ..., n_1$, and the seeds in group 2 are labelled with $1, 2, ..., n_2$. Letting x_{ij} be the expression level of gene x in seed j of group i (and likewise for y_{ij}), the following equation gives the total covariance of the profiles of genes x and y over all seeds:

$$Cov_{total} = \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}), \tag{2.1}$$

where $N = n_1 + n_2$ is the total number of seeds, and $\bar{x}_{\bullet\bullet} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} x_{ij}/N$ is the mean gene expression level of gene x over all seeds (and likewise for $\bar{y}_{\bullet\bullet}$). Cov_{total} can be decomposed into within-group covariance and between-group covariance, which are given by the following equations:

$$Cov_{within} = \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \tag{2.2}$$

$$Cov_{between} = \frac{1}{N} \sum_{i=1}^{2} n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \tag{2.3}$$

where $\bar{x}_{i\bullet} = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the mean gene expression level of gene x over seeds in group i (and likewise for $\bar{y}_{i\bullet}$). That is, $\bar{x}_{1\bullet}$ is the mean gene expression level over germinating seeds, and $\bar{x}_{2\bullet}$ is the mean gene expression level over non-germinating seeds.

It can be shown that $Cov_{total} = Cov_{within} + Cov_{between}$ (Appendix A.1). Intuitively, $Cov_{between}$ is the contribution to total covariance from preferential expression, i.e., from the

change in mean gene expression level between germinating seeds and non-germinating seeds. In contrast, Cov_{within}, is the contribution to total covariance from correlations during germination and correlations during non-germination. More specifically,

$$Cov_{within} = \alpha_1 Cov_1 + \alpha_2 Cov_2, \tag{2.4}$$

where $\alpha_i = n_i/N$, $i = 1, 2$, and

$$Cov_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \tag{2.5}$$

is the covariance over seeds in group i. Note that $\alpha_1 + \alpha_2 = 1$. Thus, the within-group covariance is a weighted average of the covariance over germinating seeds, Cov_1, and the covariance over non-germinating seeds, Cov_2.

Figure 2.10 shows the decomposition of total covariance, Cov_{total} (in black), into Cov_{within} (in red) and $Cov_{between}$ (in blue). Each dot represents a randomly chosen gene pair. The horizontal axis is their total covariance. The meaning of the vertical axis depends on dot colour: a blue dot represents their between-group covariance, a red dot represent their within-group covariance, and a black dot represent their total covariance. The red dots are clustered near the black diagonal line, while the blue dots are clustered near the horizontal axis, meaning that preferential expression accounts for almost none of the total covariance between genes, while coexpression during germination and during non-germination accounts for almost all of the total covariance. This is the case for both positive and negative covariance (i.e., for correlated and anti-correlated gene pairs). In other words, preferential expression is a minor contributing factor to correlation.

2.3.2.2 Removing preferential expression

This section provides additional evidence that preferential expression contributes little to correlation. We first remove preferential expression from each gene's profile. We then show that the correlation of the gene profiles with preferential expression removed is almost identical to that of the original gene profiles. Preferential expression is thus a minor contributing factor to

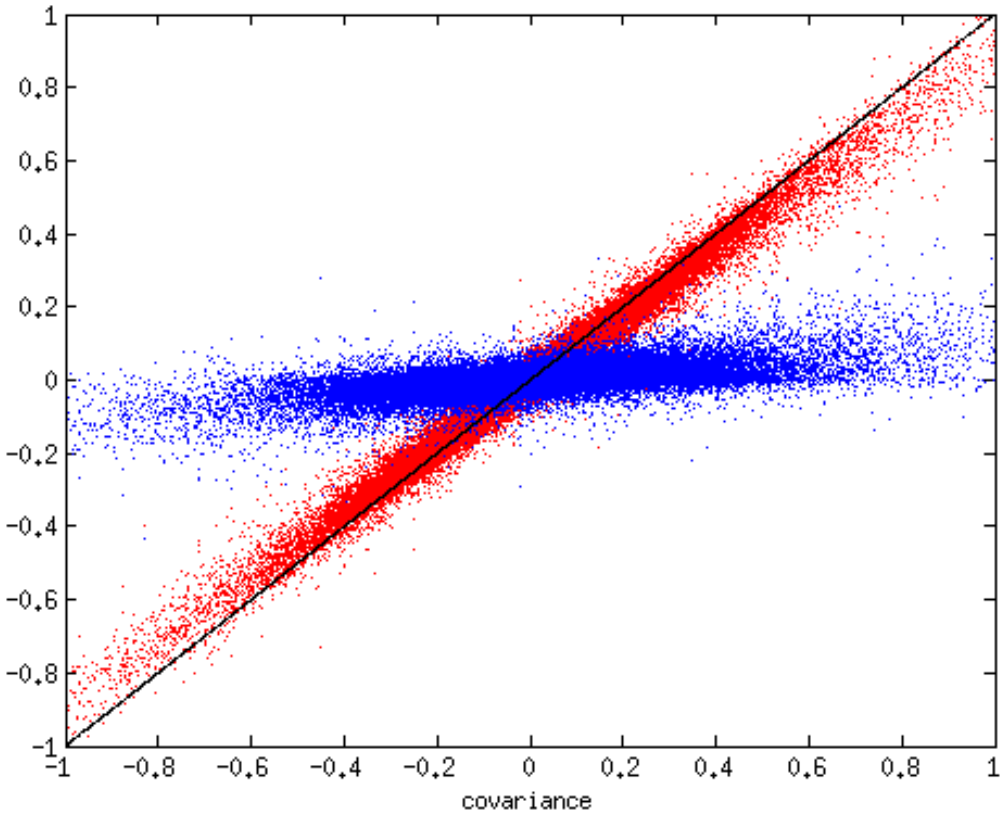


Figure 2.10: The decomposition of covariances for 100,000 randomly chosen gene pairs. Each gene pair is represented by a blue dot, a red dot and a black dot. For all three dots, the horizontal axis is the total covariance between the two genes within the pair, denoted Cov_{total}. The meaning of the vertical axis depends on dot colour. The y-coordinate of a blue dot represents their between-group covariance, $Cov_{between}$, i.e., the contribution to total covariance due to preferential expression. The y-coordinate of a red dot represents their within-group covariance, Cov_{within}, i.e., the contribution to total covariance due to coexpression during germination and during non-germination. The y-coordinate of a black dot, equal to its x-coordinate, represents their total covariance, Cov_{total}. For all gene pairs, $Cov_{total} = Cov_{between} + Cov_{within}$.

correlation, and therefore to cluster formation.

To remove preferential expression from a gene's profile, we adjust the expression levels so that the mean expression levels on the germinating and non-germinating seeds are the same (as in Figure 2.9(b)). Specifically, using the notation of Section 2.3.2.1, we let $x'_{ij} = x_{ij} - \bar{x}_{i\bullet}$. Then $\bar{x}'_{i\bullet} = 0$. That is, in the adjusted profiles, the mean expression level on the germinating seeds, $\bar{x}'_{1\bullet}$, and the mean expression level on the non-germinating seeds, $\bar{x}'_{2\bullet}$, are both 0. There is therefore no preferential expression in the adjusted profiles (GS = 0, by Corollary 4 in Appendix A.2); however, all other sources of variation remain (as in Figure 2.9).

Let c' denote the correlation coefficient between two adjusted gene profiles, and c denote the correlation coefficient between the two original (unadjusted) profiles. Figure 2.11 shows c' versus c, where each dot represents a gene pair, the vertical axis represents c', and the horizontal axis represents c. The linear shape of the plot with a 45 degree slope means that c' is very close in value to c, i.e., removing preferential expression from gene profiles results in little change to the correlations between genes. Preferential expression is therefore a minor contributing factor to correlation and cluster formation.

2.3.3 Changes in correlation and variance

In this section, we look for changes in the behaviour of genes between germination and non-germination. We consider two simple measures of behaviour: variance and correlation. Perhaps surprisingly, we find little overall change in these behaviours. For instance, for genes that are preferentially expressed during germination, there is little tendency for correlations to go up or down when going from germination to non-germination. Individual gene pairs may change their correlation, but the distribution of correlations does not change. In fact, in this regard, genes that are preferentially expressed during germination are indistinguishable from randomly chosen genes. Likewise for genes that are preferentially expressed during non-germination.

We let set G be the set of genes that are preferentially expressed during germination (i.e., for which $GS > 0.3$), and let set D be the set of genes that are preferentially expressed during non-germination (i.e., for which $GS < -0.3$). The results below describe the behaviour of genes in these two sets. Similar results hold for genes in the two large clusters (clusters G and D) [data not shown].

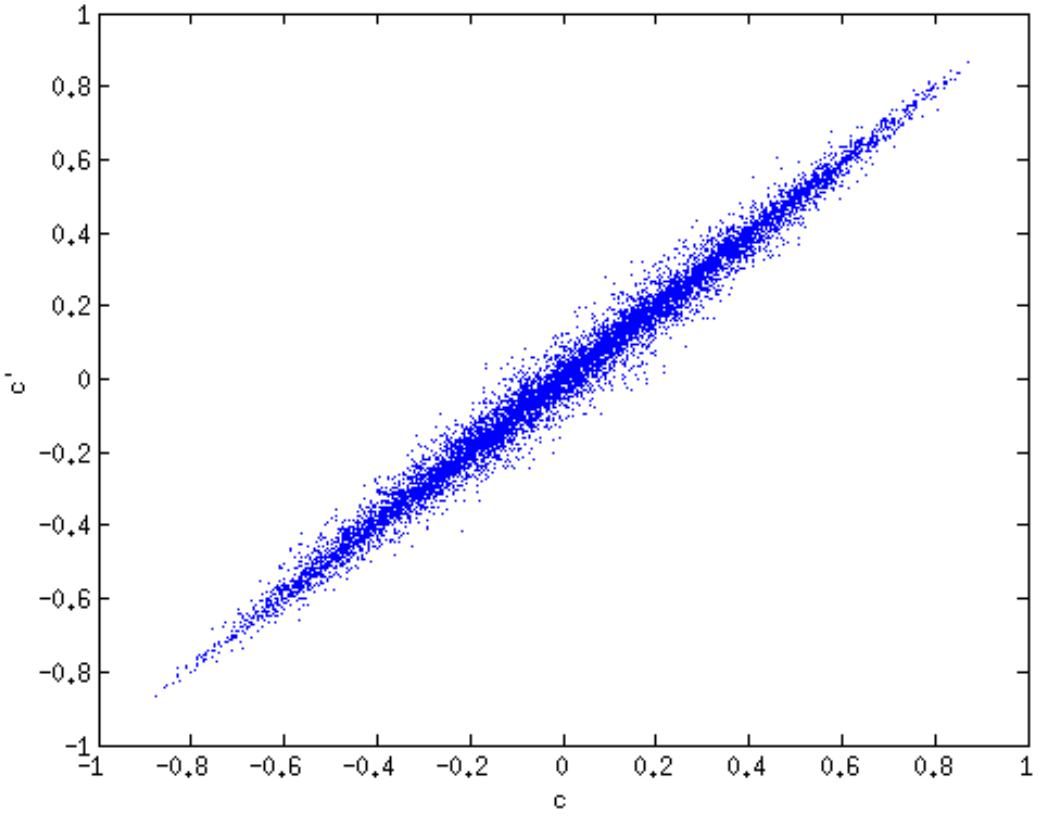


Figure 2.11: Correlation based on the adjusted gene profiles versus correlation based on the original (unadjusted) gene profiles. Each dot represents a gene pair. The horizontal axis, c, is the correlation computed on the original (unadjusted) gene profiles. The vertical axis, $c^{'}$, is the correlation computed on the adjusted gene profiles, after preferential expression has been removed.

2.3.3.1 Preferential correlation

Until now, we have looked at how genes are correlated overall. Now, we look at how they are correlated during germination, and how they are correlated during non-germination, and we look for differences. For instance, we might expect genes that are preferentially expressed during germination to also be preferentially correlated during germination, that is, to be more correlated during germination than during non-germination. In the simplest case, a cluster of genes that work together to promote germination might turn on and off together during germination, while they might simply remain off during non-germination. More generally, the activity levels of such genes might be high and fluctuate coordinatedly during germination, but remain low and fluctuate randomly during non-germination. Such genes would be both preferentially expressed and preferentially correlated during germination.

Interestingly, our results show that such genes are not the norm. Instead, genes that are preferentially expressed during germination tend to be equally correlated during germination and non-germination. Some gene pairs may change their correlation, but there is no overall tendency for correlation to increase or decrease. In fact, for genes that are preferentially expressed during germination, the distribution of correlation change is almost identical to that of randomly chosen genes. This leads to the following rather unintuitive conclusion: the tendency for such genes to work together does not seem to depend in any special way on whether a seed is germinating or dormant. This can perhaps be interpreted as follows: these genes tend to work together like a machine, both during germination and during non-germination (and perhaps at all times), but the machine is turned up during germination and turned down during non-germination.[3]

The rest of this section provides evidence to support this conclusion. To this end, we let c_1 denote the correlation coefficient of two gene expression profiles on the germinating seeds, and let c_2 denote their correlation coefficient on the non-germinating seeds. We shall present histograms of c_1, c_2 and $c_1 - c_2$ for various sets of genes.

Figure 2.12(a) shows a histogram of c_1 for all gene pairs. Notice that the histogram is roughly symmetric with a mean of 0. The distribution of c_2 for all gene pairs is similar [data

[3]This interpretation applies to seed samples and may not extend, e.g., to AraNet [11].

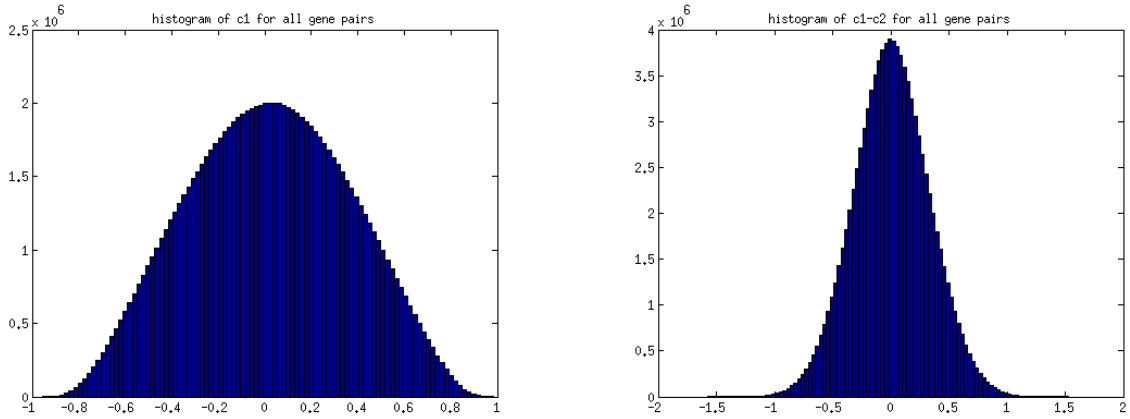
not shown]. Intuitively, this means that randomly chosen gene pairs have no tendency to be either correlated or anti-correlated during germination. Likewise during non-germination [data not shown]. Thus there is no bias in the background distributions of correlation. In contrast, Figure 2.13(a) shows a histogram of c_1 for gene pairs in which both genes are preferentially expressed during germination (i.e., both genes are in set G). This histogram is clearly biased towards positive values, with a peak near $c_1 = 0.4$. These gene pairs therefore have a strong bias towards correlation during germination, as one might expect. Figure 2.13(b) shows a histogram of c_2 for the same gene pairs. This histogram is similar to the previous one. It is clearly biased towards positive values, with a peak just above $c_2 = 0.4$. Thus, genes that are preferentially expressed during germination tend to be just as correlated during non-germination as during germination. If anything, they tend to be slightly more correlated during non-germination.

Figure 2.14(a) verifies this. It shows a histogram of $c_1 - c_2$ for the same genes pairs (i.e., for genes in set G). The histogram is roughly symmetric with a peak near 0, demonstrating that there is no tendency for c_1 to be higher than c_2. If anything, there is a very slight tendency for c_1 to be lower than c_2 (i.e., for $c_1 - c_2$ to be negative). Figure 2.12(b) shows that genes in set G are very much like other genes in this respect. The figure shows a histogram of $c_1 - c_2$ for all gene pairs. Notice that the mean, variance and shape of the two histograms are very similar. Thus, there appears to be nothing special about the change in correlation between germination and non-germination for genes that are preferentially expressed during germination. Thus, even though their expression levels decrease significantly during non-germination, these genes are co-expressed just as much during non-germination as during germination. If anything, they are co-expressed slightly more during non-germination.

Similar results hold for genes that are preferentially expressed during non-germination. This is illustrated in Figure 2.14(b), which shows a histogram of $c_1 - c_2$ for such genes (i.e., genes in set D). Again, the histogram is roughly symmetric with a peak near zero. In this case, however, the peak is slightly more negative than for genes that are preferentially expressed during germination. Thus, genes in set D tend to be slightly more correlated during non-germination.

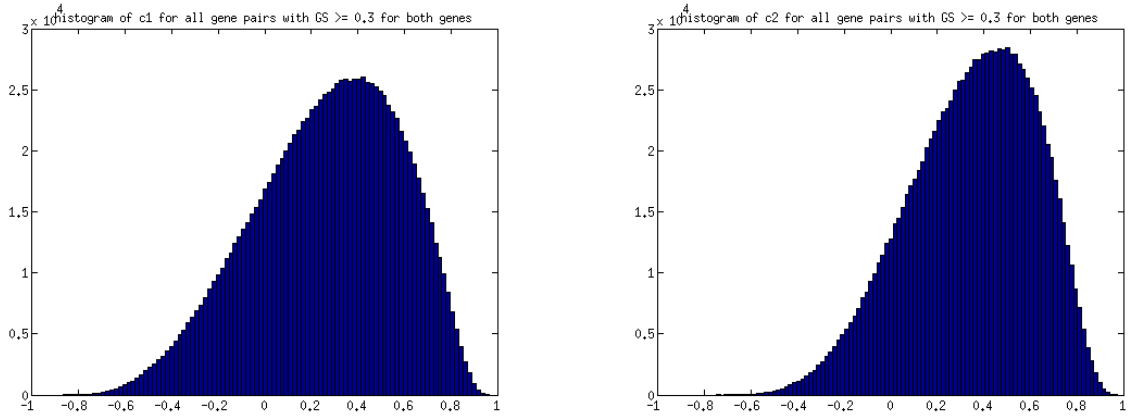
In sum, for genes that are preferentially expressed during germination or non-germination, the histograms of correlation are very different from the background, while the histograms of

correlation change are not.



(a) (b)

Figure 2.12: (a) Histogram of correlation c_1 on germinating seeds for all gene pairs. It is roughly symmetric around 0. The distribution of c_2 for all gene pairs is similar. (b) Histogram of correlation change $c_1 - c_2$ from germination to non-germination for all gene pairs.

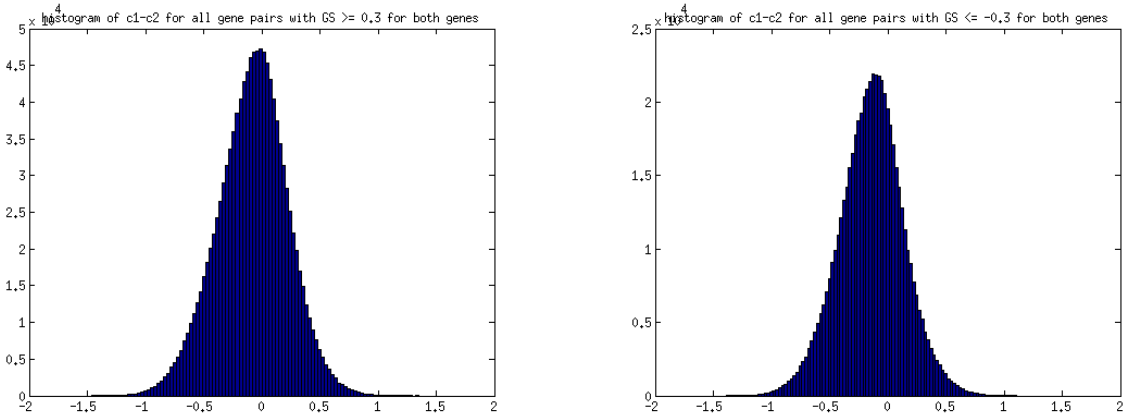


(a) (b)

Figure 2.13: Histogram of correlation coefficients on germinating seeds for genes in set G (left), and histogram of correlation coefficients on non-germinating seeds for genes in set G (right).

2.3.3.2 Preferential variance

In this section, we look at the variance of a gene's expression levels during germination, and its variance during non-germination, and we look for differences. For instance, we might expect that genes that are preferentially expressed during germination to also have preferential variance during germination, that is, to have higher variance during germination than during



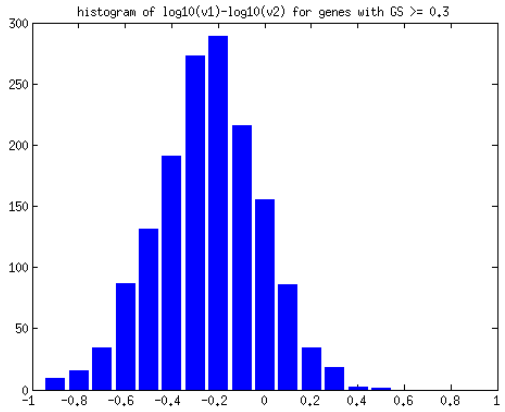
(a) (b)

Figure 2.14: Histogram of correlation change from germination to non-germination for genes in set G (left), and for genes in set D (right). Both histograms are roughly symmetric around 0.

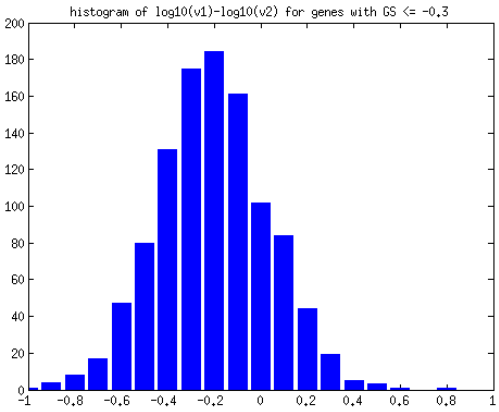
non-germination. As in the case of preferential correlation, a group of genes that promote germination might turn on and off together during germination, and simply remain off during non-germination. More generally, the expression levels of such genes might be high and fluctuate widely (and coordinatedly) during germination, but remain largely quiescent during non-germination. Such genes would be both preferentially expressed and have preferential variance during germination.

Our results, however, do not support this expectation. In fact, genes of all kinds show a slightly higher variance during non-germination than during germination.[4] To show this, let v_1 be the variance of a gene expression profile on the germinating seeds, and let v_2 be its variance on the non-germinating seeds. Figure 2.15 shows histograms of $\log v_1 - \log v_2$. Figure 2.15(c) shows that randomly chosen genes tend to have slightly higher variance during non-germination. Figures 2.15(a) and 2.15(b) show the same result for genes that are preferentially expressed during germination and non-germination, respectively. In fact, all three histograms have roughly the same shape, with a bias towards negative values and a peak at -0.2. (v_2 therefore tends to be slightly greater than v_1.) Thus, in terms of change of variance (as for change of correlation), genes that are preferentially expressed during germination are indistinguishable from randomly chosen genes. Likewise for genes that are preferentially expressed during non-germination.

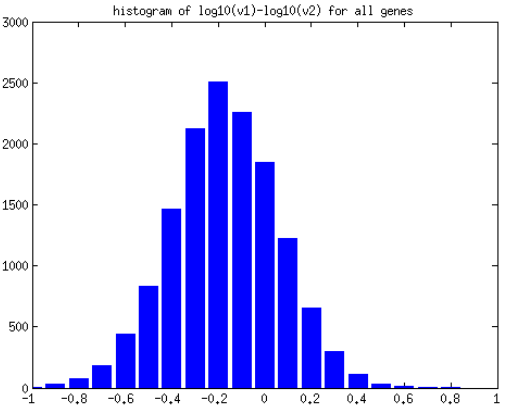
[4]We are speaking here of the variance not of the raw expression data, but of the log-transformed data, which is used throughout this thesis. This variance thus represents not the absolute change in raw expression levels, but their proportionate change.



(a) Genes in set G



(b) Genes in set D



(c) All genes

Figure 2.15: Histograms of $\log_{10} v_1 - \log_{10} v_2$ for genes in set G, genes in set D, and all genes.

2.3.4 Power-law analysis and cutoff estimation

As described in Section 2.3.1, SeedNet [1] was constructed by putting a threshold (cutoff) on the correlation coefficients between genes. Specifically, the network has an edge between two genes if and only if $r >= \tau$, where r is the Pearson correlation coefficient of the two gene expression profiles and τ is a given threshold. Different values of τ define different networks. For SeedNet, a threshold of $\tau = 0.75$ was used. This threshold was chosen because it gives a network that closely fits a scale-free graph [13]. This section describes our method for fitting a correlation network to a scale-free graph and for choosing the optimum threshold. A graph is scale-free if its node degree has a power-law distribution, that is, if the probability that a node has degree d is proportional to d^{-k}, for some k.

2.3.4.1 Fitting a power law with cumulative frequencies

One way to fit a power law distribution to a graph involves using a histogram to approximate the probability distribution of node degree, and then fit the histogram to a power law [26]. Unfortunately, methods based on this idea depend on the arbitrary choice of number of histogram bins or bin width. Furthermore, grouping data points into bins can cause information loss since points within each bin are rendered indistinguishable. Moreover, histograms can also be quite noisy, since bin height can have a high variance. The noise can be reduced by using wider bins, but doing so increases information loss.

To avoid these problems, the method used to construct SeedNet is based on cumulative frequencies. Given a graph, we define the empirical cumulative frequency of degree d, denoted empCF(d), to be the number of nodes in the network with degree at most d. Given a power law, we define the expected cumulative frequency of degree d, denoted expCF(d), to be the expected number of nodes with degree at most d. Empirical cumulative frequency is the cumulative analogue of a histogram. Figure 2.16 shows a plot of empirical cumulative frequency for SeedNet. Notice that it is a smooth curve that increases from left to right. It has more points than a histogram, because it has one point per degree, instead of one per histogram bin. Also, there is no information loss since nodes of different degree are not grouped into a single bin.

Our method fits empCF(d) to expCF(d) over all values of d in the network (Section 2.4.1).

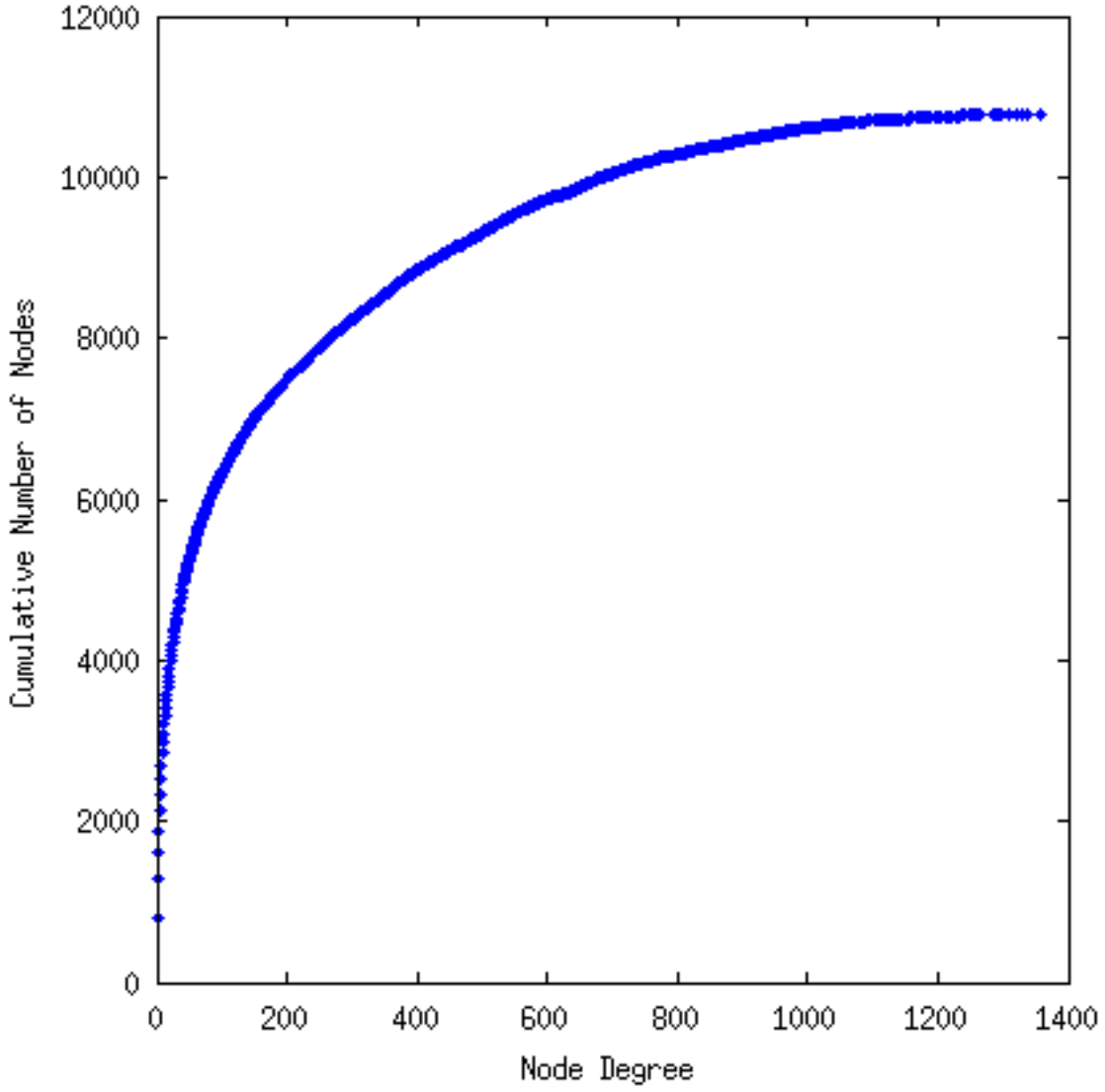


Figure 2.16: Empirical cumulative frequency versus node degree. A blue dot represent the number of nodes with degree at most d, where d is a node degree in the network.

If the fit is perfect, a plot of empCF(d) versus expCF(d) should be a straight line passing through the origin. The worse the fit, the more the plot will deviate from this straight line. This is illustrated in Figure 2.17, where empCF(d) is plotted against expCF(d) for several different power laws, d^{-k}. In each plot, the blue curve is a plot of empCF(d) versus expCF(d) for SeedNet. The diagonal red line is defined by the equation empCF(d) = expCF(d), which represents a perfect fit. The closer the blue curve is to the red line, the better the fit between the power law and SeedNet. Each plot in Figure 2.17 also gives a value for $nFit$, a quantitative measure of how well the graph fits the power law based on total squared error (Sections 2.4.1 and 2.4.2). The smaller the value of $nFit$, the better the fit.

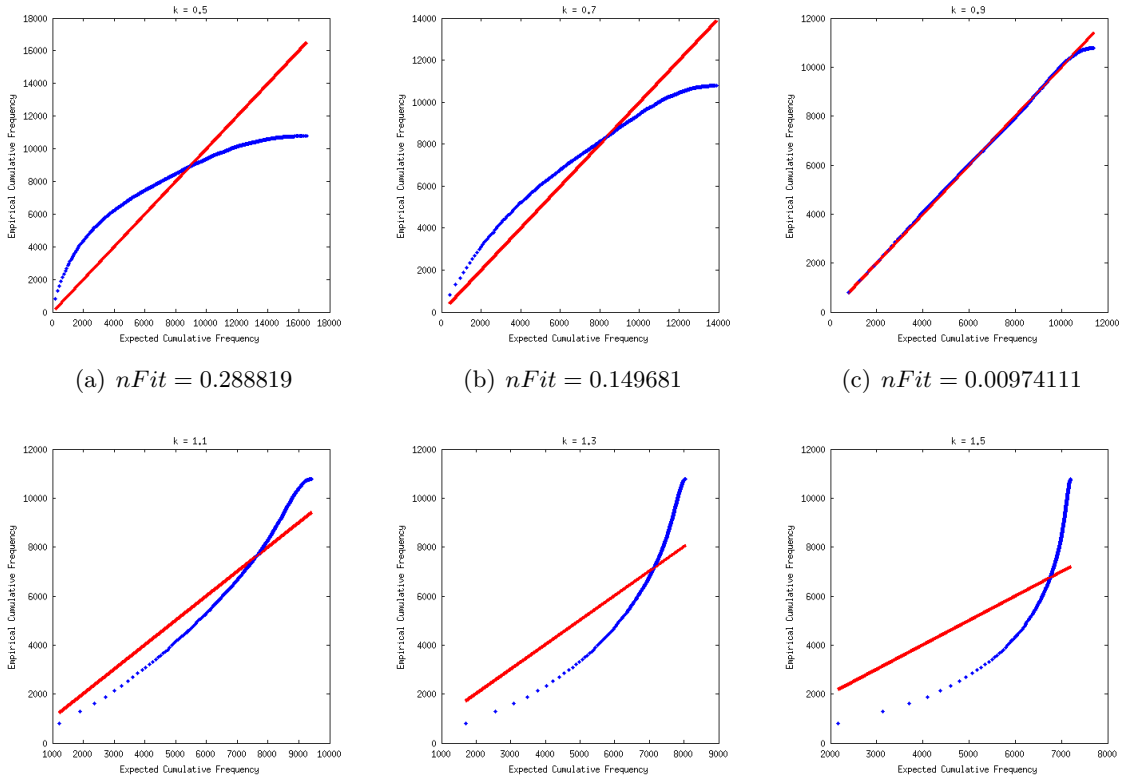
In Figure 2.17, the best fit occurs at $k = 0.9$. A more careful analysis (below) shows that the optimal value of k is actually 0.91, which is illustrated in Figure 2.19. As k increases above 0.91, the fit worsens: the blue curve deviates from the red line and becomes more and more concave. As k decreases below 0.91, the fit also worsens: the blue curve deviates from the red line and becomes more and more convex. Section 2.4.1 describes a process for finding the optimal value of k.

2.3.4.2 The optimal network

We analyzed many different networks, each corresponding to a different correlation threshold. For each network, we found a power law that best fits the network. For some networks, a power law could be found that fit the network well. For other networks, no power had a good fit. We found that a network with threshold $\tau = 0.695$ has the best fit to a power law (Section 2.4.2). For simplicity, we round τ to two decimal places, giving $\tau = 0.70$. Figure 2.18 shows a heatmap of this network.

Figure 2.19 shows the empirical CF of this network versus the expected CF of the power law with $k = 0.91$. This is the power law that fits the network best. Among the 10,782 nodes (shown as blue dots) with degree greater than 0, the first 10,500 (those with lowest degree) closely obey a power law with $k = 0.91$, while the last 282 nodes (those with highest degree) disobey the power law.

Figure 2.20 gives a different view of the fit between the network and the power law. The blue dots represent the height of a histogram bin of node degree (with bin width 20), and the



(a) $nFit = 0.288819$ (b) $nFit = 0.149681$ (c) $nFit = 0.00974111$

(d) $nFit = 0.122717$ (e) $nFit = 0.221711$ (f) $nFit = 0.292304$

Figure 2.17: Fitting different power laws, d^{-k}, to a network defined by correlation cutoff $\tau = 0.7$. Here, $k = 0.5$, 0.7, 0.9, 1.1, 1.3 and 1.5. Both visually and in terms of $nFit$, the best fit is achieved with $k = 0.9$.

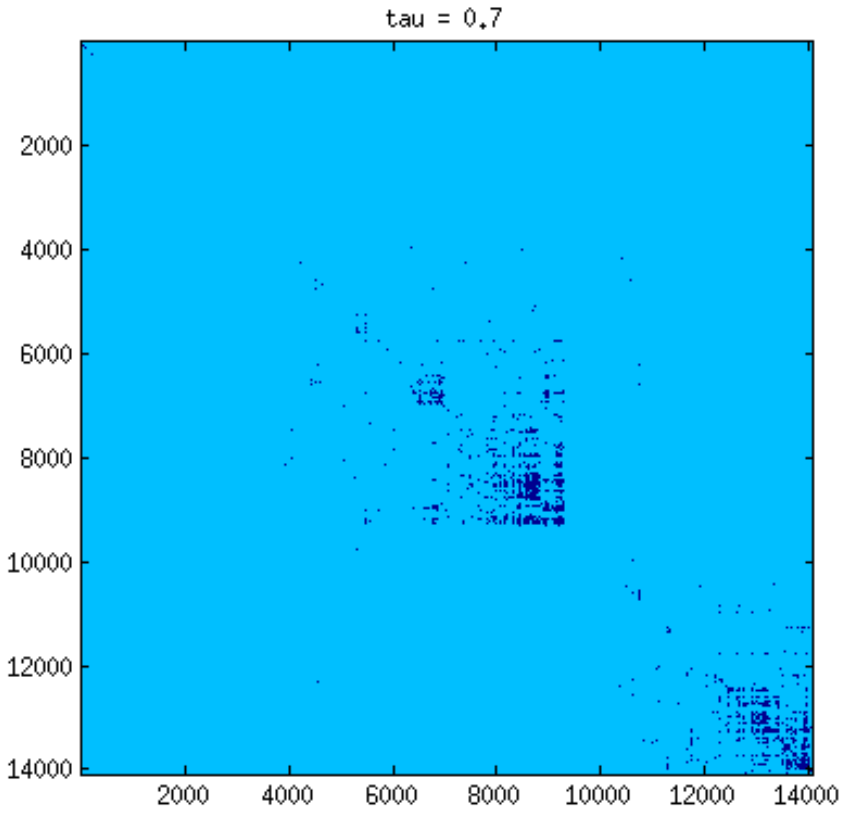


Figure 2.18: A heatmap of a network constructed from gene pairs whose correlations exceed 0.7.

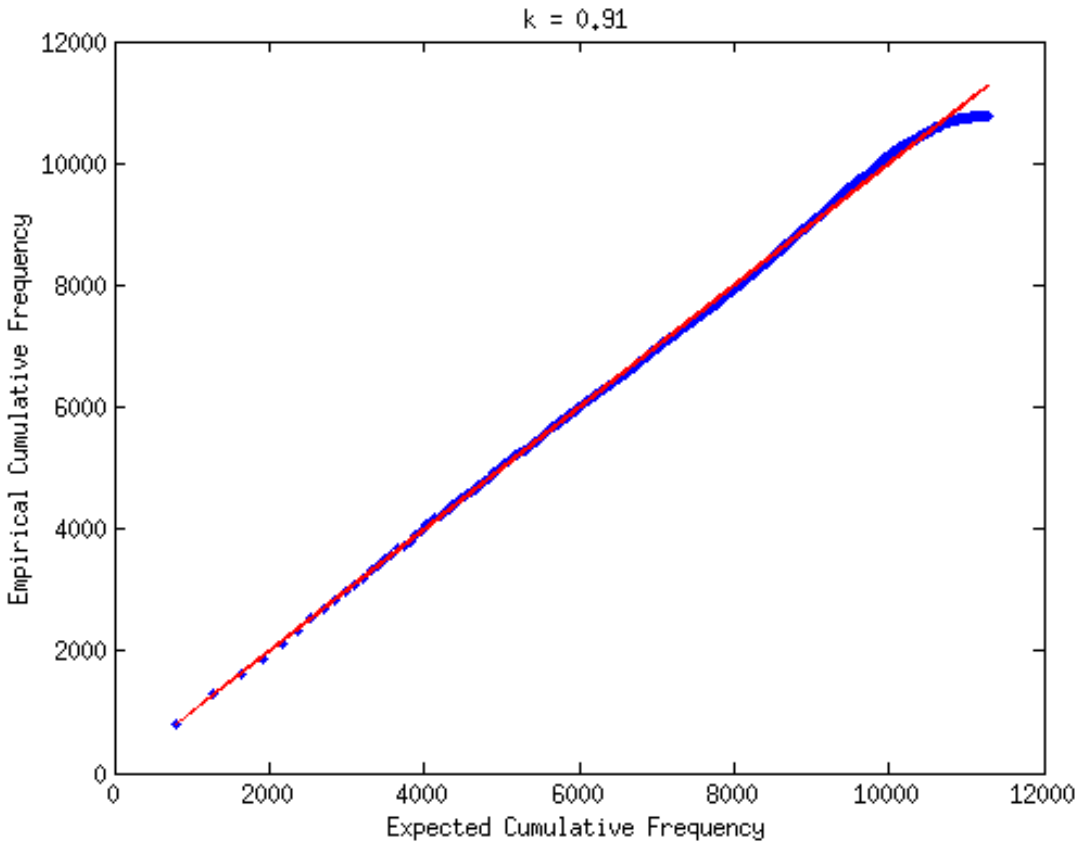


Figure 2.19: Empirical cumulative frequency versus expected cumulative frequency for node degree in a network with threshold $\tau = 0.7$. The red line shows the the power law fit for the cumulative node degree distribution. The distribution approximately follows $d^{-0.91}$, where d is node degree. $nFit = 0.00826538$.

red curve represents the best fitting power law (proportional to d^{-k}, where $k = 0.91$). Note that the network closely obeys the power law for nodes with degree ranging from 1 to over 1000, that is, over more than three orders of magnitude.

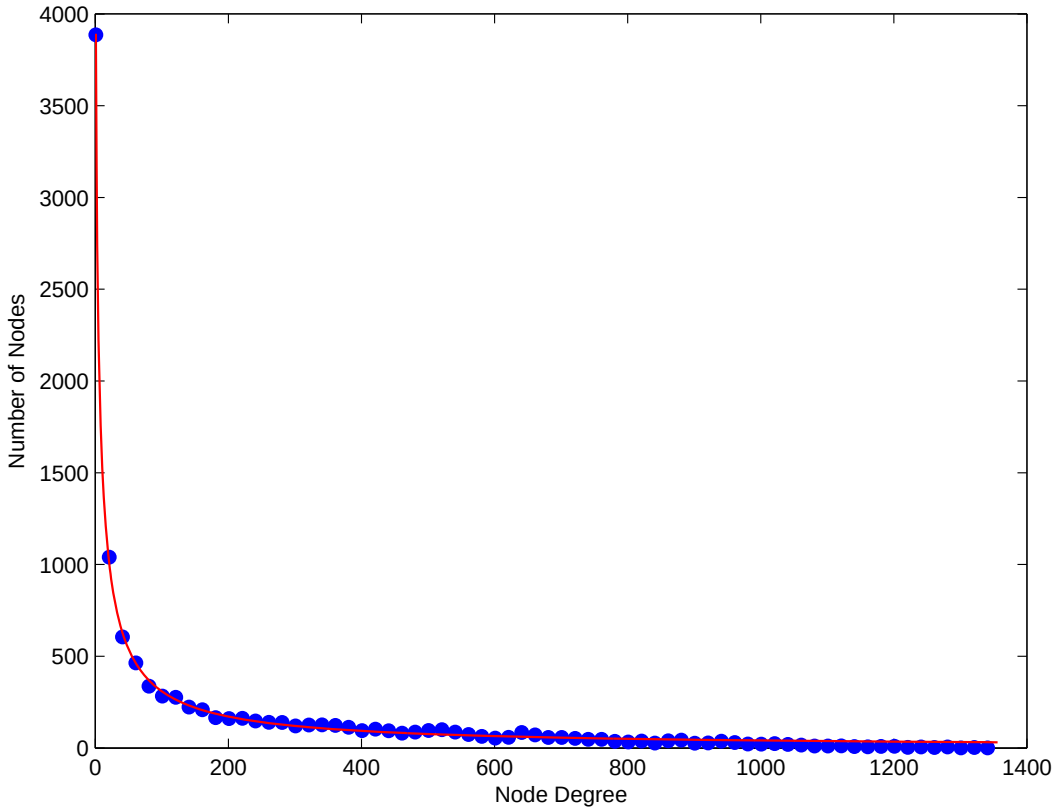


Figure 2.20: Number of nodes versus node degree for the network of Figure 2.19. The red curve is the power law that best fits the network.

The networks generated by other correlation cutoffs, $\tau = 0.5$, 0.6 and 0.8, are best fit by power laws with $k = 0.49$, 0.72 and 1.04, respectively. Figure 2.21 shows the fits. Although each of the blue curves represents a best fit, they are not equally good. Also, the network has an imperfect fit to a power law for nodes of low degree (Figure 2.21(a) and Figure 2.21(c)) and/or for nodes of high degree (all of the blue curves). Notice that for $\tau = 0.5$, the fit is particularly bad for nodes of high degree (Figure 2.21(b)).

Among the four networks defined by $\tau = 0.5, 0.6, 0.7$ and 0.8, the blue curve is visually closest to the red diagonal line in the network with threshold $\tau = 0.7$ (Figure 2.19). This

is also true in a least squares sense, as reflected by the fitness measure $nFit$ (Section 2.4.2). Moreover, $nFit$ is equivalent to non-centered correlation (Section 2.4.2). Thus, when $\tau = 0.7$ and $k = 0.91$, the empirical CF is more highly correlated with the expected CF than for any other combination of τ and k.

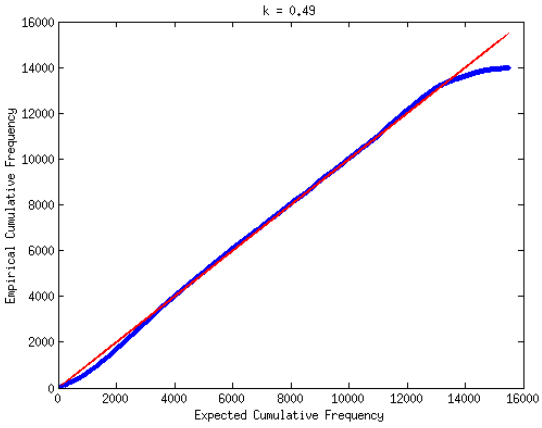
2.3.5 Geometric random graphs

There is evidence that, when fitting protein-protein networks, a geometric random graph is more appropriate than a scale-free model in terms of modelling graphlet frequency, network diameter and clustering coefficients [14]. However, we show that geometric random graphs are not appropriate for modelling node degree distribution in coexpression networks, such as SeedNet. (This is also true for protein-protein networks, as noted by the same authors [14].) In fact, the node degree distribution in geometric random graphs is very different from the node degree distribution in coexpression networks (Appendix A.3). We conjecture that one reason for this is that coexpression networks have a transitive property that protein-protein networks do not. Specifically, if gene g_1 is correlated with g_2, and g_2 is correlated with g_3, then in general, g_1 will also be correlated with g_3 (though to a lesser extent). This is not true of protein-protein networks. That is, if protein p_1 interacts with p_2, and p_2 interacts with p_3, then in general p_1 need not interact with p_3.

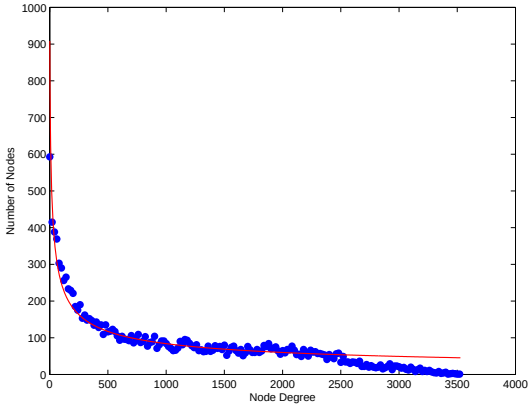
2.3.6 False discovery rate

This section describes how we estimate an upper bound on the false discovery rate (FDR) for each edge in our network.

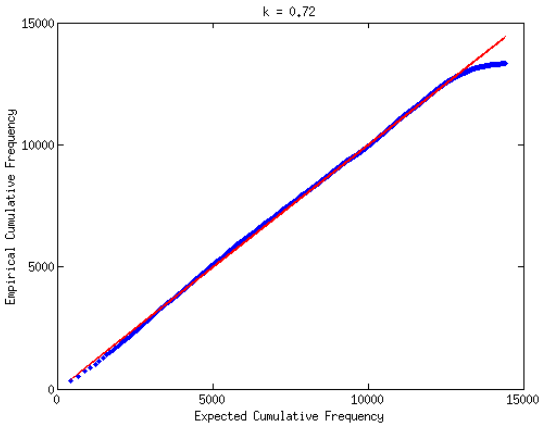
Given a threshold τ, a discovery is a pair of genes for which $\rho \geq \tau$, where ρ is the Pearson correlation coefficient of the two gene expression profiles in our data set. A discovery is intended to be a pair of genes whose expression levels are correlated. However, because any real data set is finite, the correlation coefficient estimated from the data will in general be different from the true correlation coefficient (*i.e.*, the correlation coefficient estimated from an infinite set of data). In particular, two genes that are independent (and therefore uncorrelated) may appear to be correlated and lead to false discoveries. More precisely, a false discovery is a pair of genes whose expression levels are independent, but whose correlation coefficient on our data set



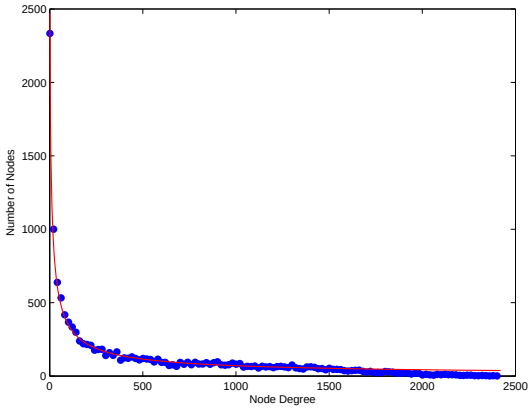
(a) $\tau = 0.5, \quad nFit = 0.0228881$



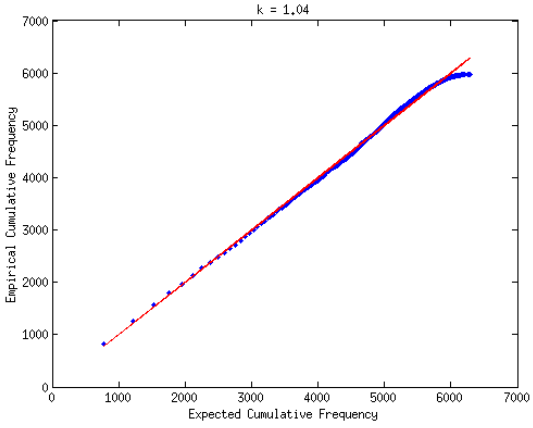
(b) $\tau = 0.5$



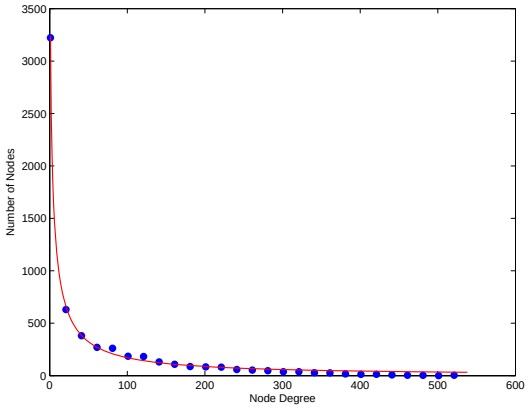
(c) $\tau = 0.6, \quad nFit = 0.0152530$



(d) $\tau = 0.6$



(e) $\tau = 0.8, \quad nFit = 0.0121066$



(f) $\tau = 0.8$

Figure 2.21: Power law fit for degree distribution in three networks with thresholds, $\tau = 0.5$, 0.6 and 0.8, respectively. The k value for d^{-k} is labelled on top of each plot of empirical cumulative frequency versus expected cumulative frequency.

exceeds τ. The false discovery rate (FDR) is the proportion of discoveries that are expected to be false [22]. That is, FDR $= m/M$, where M is the number of discoveries, and m is the expected number of false discoveries. In general, as the threshold τ increases, the number of discoveries and the FDR both decrease.

FDRs cannot be computed exactly, since although M is known, m is not, since we do not know which discoveries are true and which are false. However, it is possible to estimate an upper bound on the FDR, and a framework for doing so is developed in 2.4.3. Adapting this framework, we developed a method for estimating an upper bound on the FDR for the detection of coexpressed genes (*i.e.*, genes with correlated expression levels). The details are given in Algorithm 2.4.2 in Section 2.4.3.

As the threshold, τ, varies, the FDR also varies, tracing out a curve. Figure 2.22 shows such a curve for our *Arabidopsis* data. Using this curve, we can read off the FDR for any given threshold. For the optimal value of $\tau = 0.7$, the estimated FDR is 0. At this value, the variance of the estimated FDR is extremely small [data not shown].

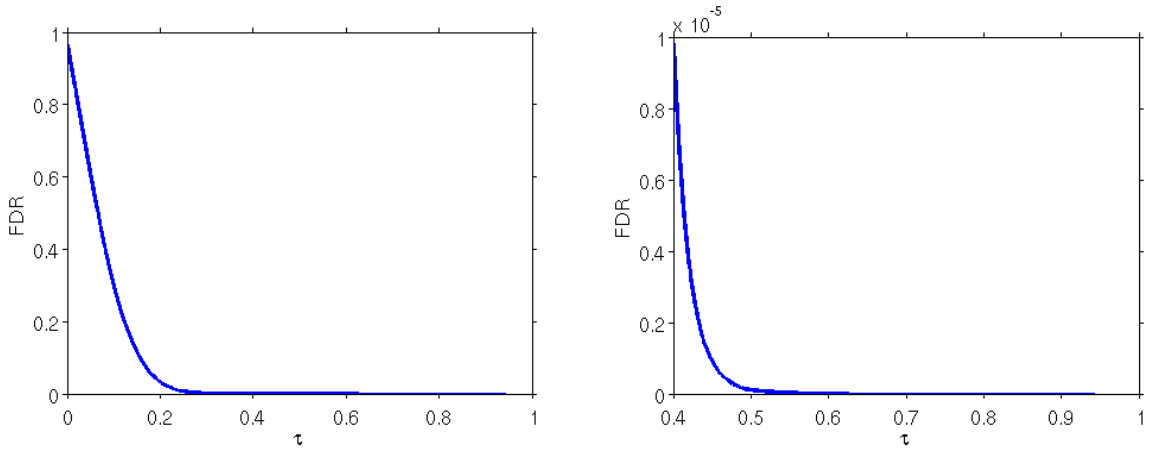


Figure 2.22: False discovery rate (FDR) versus correlation threshold (τ). The right-hand curve is a close up of the tail of the left-hand curve.

2.4 Materials and Methods

2.4.1 Fitting a power law to a graph

This section describes our algorithm for finding the power law, d^{-k}, that best fits a given graph, $\mathcal{G}$. We break the problem into two parts. First, we assume that k is given and measure how well graph $\mathcal{G}$ fits the power law d^{-k}. Second, we test many different values of k and choose the value that best fits the graph.

To measure how well a graph fits a given power law, recall from Section 2.3.4.1 that a good fit implies that for nodes of any degree, d, the empirical cumulative frequency, empCF(d), is approximately equal to the expected cumulative frequency, expCF(d). To make this idea precise, let $\mathcal{D} = \{d_1, ..., d_n\}$ be the set of node degrees in the graph. In addition, let y_i be the number of nodes with degree at most d_i in graph $\mathcal{G}$. That is, $y_i = \text{empCF}(d_i)$. If we let $z_i = \sum_{d=1}^{d=d_i} d^{-k}$, then for graphs that obey the power law d^{-k}, the expected number of nodes of degree at most d_i is proportional to z_i. That is, $\text{expCF}(d_i) = \alpha z_i$, for some $\alpha > 0$.

If graph $\mathcal{G}$ obeys the power law d^{-k}, then we should have that $\text{empCF}(d_i) \approx \text{expCF}(d_i)$ for $1 \leq i \leq n$. That is, $y_i \approx \alpha z_i$, for some $\alpha > 0$. One way to measure the accuracy of this approximation is with a least squares fit, that is, by finding the value of α that minimizes the total squared error, $\sum_i (y_i - \alpha z_i)^2$. This is a 1-dimensional regression problem and has the solution $\alpha = \sum_i y_i z_i / \sum_i z_i^2$ [27]. Using this value of α, the total squared error simplifies to

$$\sum_i (y_i - \alpha z_i)^2 \;=\; \frac{\sum y_i^2 \sum z_i^2 - (\sum y_i z_i)^2}{\sum z_i^2} \tag{2.6}$$

This equation tells us how well a graph fits a particular power law, d^{-k}. To find the power law that best fits the graph, we test many different values of k within a range, *e.g.*, from 0.5 to 1.5 in increments of 0.1. The values of z_i depend on k, so for each value of k, we recompute the z_i and then recompute the total squared error using Equation 2.6. Finally, we choose the value of k that gives the smallest total squared error. This process is illustrated in Figure 2.17.

2.4.2 Choosing a correlation threshold

Recall that SeedNet contains an edge between two genes if and only if the correlation between the two gene profiles exceeds a given threshold, τ. Different values of τ give different networks, and as described in Section 2.3.4.2, τ is chosen to yield a network that fits a power law as closely as possible. However, when judging how well different networks fit a power law, it is inappropriate to simply compare the total squared errors as given by Equation 2.6. This is because two networks can fit a power law equally well but have very different squared errors.

To see this, consider the three plots on the left side of Figure 2.21. Each plot represents a different network and has a different scale on the vertical axis. If two such plots looked exactly the same, then the two networks would fit a power law equally well, even if the vertical scales were different. However, changing the vertical scale changes the squared error. For instance, doubling the vertical scale would double all the errors and quadruple the total squared error.

To fix this problem, we normalize the total squared error to compensate for differences in vertical scale. In particular, we define the normalized fit, $nFit$, of a graph to a given power law as follows:

$$\begin{aligned} nFit^2 &= \text{Total Squared Error}/\textstyle\sum_i y_i^2 \\ &= \textstyle\sum_i (y_i - \alpha z_i)^2 / \sum_i y_i^2 \\ &= \frac{\sum y_i^2 \sum z_i^2 - (\sum y_i z_i)^2}{\sum y_i^2 \sum z_i^2} \\ &= 1 - (\textstyle\sum y_i z_i)^2 / (\sum y_i^2 \sum z_i^2) \end{aligned} \tag{2.7}$$

where the third equality is from Equation 2.6. Geometrically, $nFit$ is the sine of the angle between the vectors $y = (y_1, ..., y_n)$ and $z = (z_1, ..., z_n)$.

Note that since the y_i's do not depend on the power law, d^{-k}, the same power law that minimizes total squared error for a given graph will also minimize $nFit$. We can therefore use $nFit$ to determine which power law best fits a graph (Figure 2.17). More importantly, we can use $nFit$ to compare how well *different* graphs fit a power law (Figure 2.21).

$nFit$ is also closely related to correlation. In particular, the last line of Equation 2.7 can be rewritten as follows:

$$nFit^2 = 1 - corr(y, z)^2$$

where $corr(y, z) = \sum y_i z_i / \sqrt{\sum y_i^2 \sum z_i^2}$. Note that $corr(y, z)$ can be viewed as an uncentered

correlation coefficient of y and z, *i.e.*, one in which the means are not subtracted from y_i and z_i. Thus, $nFit$ is minimized when uncentered correlation is maximized.

The uncenteredness of the correlation arises because in the case of a perfect fit, y_i and z_i are directly proportional ($y_i = \alpha z_i$), instead of having an arbitrary linear relationship ($y_i = \alpha z_i + \beta$). Geometrically, this means that the red lines in Figure 2.17 pass through the origin. We note that by the Cauchy-Schwartz inequality [28], uncentered correlation takes values between $+1$ and -1, just as ordinary correlation does. Thus, $nFit^2$ takes values between 0 and 1, so its range of values is independent of the vertical scale.

Our method for computing $nFit$ is summarized in Algorithm 2.4.1 below. Using this algorithm as a subroutine, we estimate the optimal correlation threshold, τ, as follows:

1. For each value of τ in a range of values (say, 0.3 to 0.9 in increments of 0.1),

   (a) Construct a coexpression network, $\mathcal{G}$, with cutoff τ.

   (b) For each value of k in a range of values (say, 0.5 to 1.5 in increments of 0.1), let $nFit = \text{COMPUTE\_NFIT}(k, \mathcal{G})$.

2. Return the combination of τ and k that gives the smallest value of $nFit$.

Once optimal values of τ and k are returned, this procedure can be run again in a small neighbourhood around these values to obtain more accurate values of τ and k. In this way, we compute the optimal correlation threshold, τ, and the power law, d^{-k}, that best fits the resulting network.

Algorithm 2.4.1: COMPUTE_NFIT$(k, \mathcal{G})$

input: k, a positive integer

$\mathcal{G}$, a graph of nodes and edges

output: $nFit$, a measure of how well graph $\mathcal{G}$ fits the power law d^{-k}

Let $\{d_1, ..., d_n\}$ be the set of node degrees in graph $\mathcal{G}$

for each $i \in \{1, 2, ..., n\}$

do $\begin{cases} \text{Let } y_i \text{ be the number of nodes with degree at most } d_i \\ \text{Let } z_i = \sum_{d=1}^{d=d_i} d^{-k} \end{cases}$

Let $nFit = \sqrt{1 - (\sum y_i z_i)^2 / (\sum y_i^2 \sum z_i^2)}$

Return $nFit$

2.4.3 Estimating the false discovery rate

As described in Section 2.3.6, gene pairs whose expression profiles have an estimated Pearson correlation greater than a given threshold, τ, are called discoveries. Genes pairs whose expression levels are independent but whose estimated correlation is above τ are called false discoveries. This section describes our method for estimating an upper bound on the false discovery rate by using repeated permutation tests. Details are given in Algorithm 2.4.2 below. The algorithm has two inputs, a threshold, τ, and a set of gene pairs, GP. In practice, we set GP to be the set of all possible gene pairs, in which case the algorithm returns an estimated upper bound on the FDR at a threshold of τ. However, to understand the algorithm, it is helpful to consider its effect on other possible sets. In addition to its arguments, the algorithm assumes the existence of a data set for a set of genes, G.

Algorithm 2.4.2: ESTIMATE_FDR(GP, τ)

input: GP, a set of (unordered) gene pairs

τ, a threshold on correlation coefficient

output: an upper bound on the false discovery rate at threshold τ

1. Let G be the set of all genes in the data set

% Compute the number of discoveries, M.

2. Let GP_0 be the set of all (unordered) gene pairs (g_1, g_2) with g_1 and g_2 in G
3. For each pair of genes (g_1, g_2) in GP_0,

   let $\rho(g_1, g_2)$ be the correlation coefficient of their expression profiles
4. Let M be the number of gene pairs (g_1, g_2) in GP_0 for which $\rho(g_1, g_2) \geq \tau$

% Generate 1000 randomly permuted data sets and estimate a FDR for each.

5. For k from 1 to 1000, do
   a. For each gene in G, randomly permute its expression profile.

   % Count the number of false discoveries, m, in the permuted data set.

   b. For each pair of genes (g_1, g_2) in GP,

      let $\rho_r(g_1, g_2)$ be the correlation coefficient of their randomly permuted profiles

   c. Let m be the number of gene pairs (g_1, g_2) in GP for which $\rho_r(g_1, g_2) \geq \tau$.

   d. Let $fdr(k) = m/M$ % false discovery rate for the permuted data set

% Estimate the expected rate of false discoveries.

6. Let FDR be the average value of $fdr(k)$ over all k.
7. Return FDR

Algorithm 2.4.2 has two important properties. First, it is monotonic in GP. That is, suppose we run the algorithm twice, using the same threshold τ, but two different sets of gene pairs, GP_1 and GP_2, to give two different FDR estimates, FDR_1 and FDR_2. If $\mathrm{GP}_1 \subseteq \mathrm{GP}_2$, then $\mathrm{FDR}_1 \leq \mathrm{FDR}_2$. Second, if GP_1 is the (unknown) set of independent gene pairs, then FDR_1 will be an actual estimate of the FDR (not just an upper bound). Consequently, if GP_2 is the set of all possible gene pairs, then $\mathrm{GP}_1 \subseteq \mathrm{GP}_2$, and so $\mathrm{FDR}_1 \leq \mathrm{FDR}_2$. That is, FDR_2 is an upper bound on the FDR estimate.

To see the first property, that Algorithm 2.4.2 is monotonic in GP, it is enough to note that

in line 5.c. of the algorithm, m can only get larger as GP gets larger.

To see the second property, suppose that GP is the set of independent gene pairs. The algorithm uses permutation tests to estimate the number of pairs in GP that will be counted as discoveries. These are the false discoveries. Line 5.c. of the algorithm counts the number of such false discoveries, m, that are made when the expression profile of each gene is permuted randomly. This is done 1000 times for 1000 different random permutations, giving 1000 different counts, which are then averaged to give the expected number of false discoveries. This is converted to a false discovery rate by dividing by the number of discoveries, M.

2.5 Conclusions and biological interpretation

This chapter, published in part in [1], describes the construction and analysis of SeedNet, a coexpression network based, not on a diverse set of data, but on data gathered exclusively from imbibed mature seeds of Arabidopsis thaliana. In SeedNet, the nodes are genes and an edge between two genes means that their expression profiles are highly correlated, i.e., the correlation coefficient exceeds some threshold, which is chosen so that the network fits a scale-free graph as closely as possible [13]. The FDR of the edges at this threshold was estimated and shown to be extremely small (i.e., effectively zero).

We analyzed the correlation and covariance properties of the network and found, firstly, that the network consists of two main clusters, one corresponding to germination and one to non-germination. We showed, moreover, that the correlation between two genes is not due to preferential expression, but due to correlation during seed germination and during non-germination. Correlation, and the clusters based on it, is therefore not a proxy for preferential expression, but reflects other factors, specifically biological processes that operate during germination and during non-germination.

The analysis also revealed other intriguing properties in the correlation and covariance structure of the two clusters. For example, genes that are preferentially expressed during germination tend to be equally correlated during germination and non-germination. Thus, genes that are "turned up" during germination seem to work together in the same way both during germination and during non-germination. Likewise for genes that are preferentially

expressed during non-germination.

The chapter provides a detailed development of the methods used to construct and analyze SeedNet. The development includes mathematical proofs of results on covariance decomposition and preferential expression, and a robust algorithm for determining the optimum correlation threshold for approximating a scale-free graph. These methods and analysis techniques are not limited to seed germination in *Arabidopsis*, but are general and can be applied to other organisms and other condition-dependent data. SeedNet itself is available online as a community resource (http://vseed.nottingham.ac.uk).

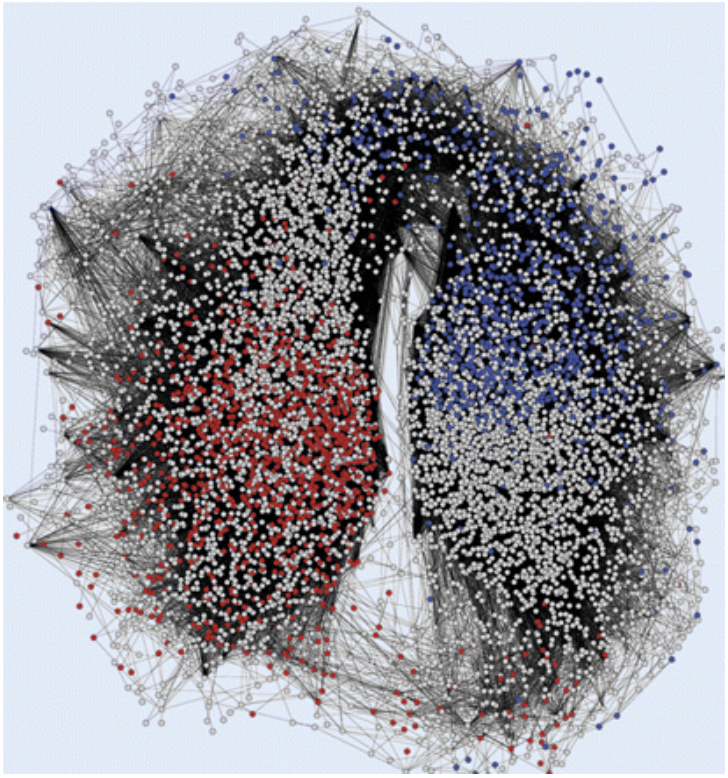


Figure 2.23: A visualization of SeedNet (image courtesy of George W. Bassel), reproduced from [1]. The left part of the network represents cluster D, and the right part represents cluster G. The red and blue dots represent genes that are significantly associated with non-germination and germination, respectively [1].

Figure 2.23 shows a visualization of SeedNet, in which the two large clusters are clearly present. [1] provides additional biological evidence that cluster D is associated with dormancy, and cluster G is associated with germination. For instance, it examines the distribution of

ABA[5] up/down-regulated genes and GA[6] up/down-regulated genes throughout the network. It also provides evidence suggesting seed dormancy evolved by incorporating older, existing pathways that regulate abiotic stress, since cluster D is enriched with vegetative abiotic stress response genes. Analysis of the network also suggests that staying in dormancy requires more transcriptional regulation than completing germination. Finally, [1] investigates substructure within the two large clusters. For instance, Figure 2.2 suggests the existence of a sub-cluster in the upper left corner of cluster D. [1] shows that this region corresponds to the top of the network in Figure 2.23. More importantly, it is shown to correspond to the biological transition between dormancy and germination.

In terms of SeedNet's predictive power, first, the condition-dependent approach used for SeedNet enabled us (in [1]) to capture more known interactions unique to seed germination than condition-independent approaches (such as AraNet [11]). SeedNet is therefore a richer source of new hypotheses about functional interactions between regulators of seed germination. In SeedNet, positive regulators and negative regulators of seed germination interact, and their net combined strength determines whether a seed remains dormant or completes germination [1]. Second, uncharacterized hub genes, i.e., the genes with the highest degree of transcriptional coordination in the network, are good candidates for predicted regulators of seed germination. In [1], 8 such hub genes are experimentally characterized, with a 50% success rate, much higher than the 22% success rate achieved by merely looking at genes with high preferential expression [1].

[5]Abscisic acid, which inhibits seed germination.
[6]Gibberellic acid, which promotes seed germination.

Chapter 3

Three-way interactions and bootstrap estimation of FDR

3.1 Introduction

In Chapter 2, we looked at gene coexpression in *Arabidopsis* under specific biological conditions, namely seed germination and non-germination. The idea was that although two genes may not be coexpressed in general, they might be under these specific conditions. That is, their coexpression may be condition dependent. In this chapter, we look at how coexpression depends on specific transcriptional conditions. In particular, we ask how the coexpression of a pair of genes depends on the expression level of a third gene. However, the third gene is not specified in advance. Instead, we search for triples of genes in which the expression level of one gene affects the coexpression of the other two. Such genes have a 3-way interaction. Discovering such interactions is more difficult, computationally and statistically, than discovering 2-way interactions, as we did in Chapter 2.

3.1.1 Three-way interactions

In general, a gene's expression level quantifies its activity level in a cell. Different gene expression results in different cell sizes, shapes and functions [29]. Activated or inhibited, silenced or induced, strong or weak, gene expression is regulated in an invisible manner to create visible

varieties in living organisms. Moreover, genes that are coexpressed may be involved in the same biological process, and thus, coexpression networks are often used to investigate gene function and regulation [1, 2, 4, 7, 8]. Methods for detecting, analyzing and clustering pairs of coexpressed genes are now well-developed. However, *pairwise* coexpression is clearly too simplistic to describe the complex relationships between gene expression levels, since these relationships can involve multiple genes and can vary depending on the biological context. In general, pairwise coexpression does not capture higher-order statistical dependencies or the complex biological relationships they reflect [19].

One way of viewing such higher-order dependencies is in terms of change in correlation. In general, the presence or absence of correlation between two genes depends on intrinsic cellular states. Two plant genes that are correlated during flowering may be uncorrelated during germination. Two human genes that are uncorrelated in normal tissues may be correlated in tumor tissues. With increasing amounts of transcriptome data, it becomes even more important to be able to compute correlations in a condition-relevant fashion. For example, Chapter 2 computed correlations under conditions of seed germination and non-germination.

But in many situations the relevant conditions are unknown or such conditions are not easily interpretable, so it can be difficult to automatically detect correlation from two gene profiles alone. For example, two genes may be correlated or anti-correlated, depending on the hormone level (determined by another gene) in the cell. The net correlation between these two genes may be close to zero if the hormone level varies randomly. More generally, the correlation between two genes can be affected by a third gene, usually called a modulator or controller gene [24]. This phenomenon is an example of three-way interaction, and can happen in post-translational regulation through post-translational modifications [20], where the activity level of a transcription factor can be modified by modulator proteins. For example, the MYB transcription factor PHR1, involved in plant response to phosphate starvation and an activator of IPS1, is controlled by a third gene SUMO E3 ligase SIZ1 through sumoylation, a post-translational modification process carried out by Small Ubiquitin-like Modifier (SUMO) proteins [30]. As a possible consequence, when SIZ1 is highly expressed, the transcription factor PHR1 and its downstream target gene IPS1 are highly correlated, but when the SIZ1 is lowly expressed, this correlation may disappear.

Three-way interactions can also happen in combinatorial regulation of target gene expression [31]. To illustrate, let g_1 and g_2 be two transcription factors of target gene g_3. Some combinatorial regulation can be modeled by a logical AND gate: $g_3 = g_1 \wedge g_2$ [32]. If g_2 is 1, then g_3 and g_1 have the same value. But if g_2 is 0, then g_3 is always 0, regardless of what the value g_1 takes. Likewise, some combinatorial regulation can be modeled by a logical OR gate: $g_3 = g_1 \vee g_2$. If g_2 is 0, then g_3 and g_1 assume the same value, but if g_2 is 1, then g_3 is always 1, regardless of what the value g_1 takes. So the output (g_3's expression) depends on a non-linear combination of the inputs (the expression levels of g_1 and g_2). In fact, combinatorial control of gene expression regulation through multiple transcriptional activators acting on several binding sites on the promoter of a target gene is prevalent in eukaryotic transcription, and this is a major level at which gene expression is controlled [29]. As a concrete example, in *Arabidopsis thaliana*, the synergistic interaction between protein rd22BP1 and protein AtMYB2 is required, along with other environmental signals, for the transcription of rd22 gene [33]. The binding sites for the two proteins are about 40 nucleotides apart in the rd22 promoter. This can be thought of as a typical three-way interaction and is one motivation for the present work.

For computational tractability, we primarily consider three-way interactions, although higher-order interactions that involve more than three genes exist in eukaryotic cells. We also do not address other gene expression regulation mechanisms such as RNA splicing, translation, degradation of mRNA or chromatin unwinding [29], since we focus on gene expression data, which is abundant and easily generated. Finally, our method generalizes coexpression analysis from pairs of genes to triples of genes. As such, it detects statistical dependencies in gene expression data, not physical interactions. (We shall often refer to these as indirect three-way interactions.) Sometimes, however, direct interaction is strongly suggested. For instance, if genes g_1 and g_2 are known transcription factors for gene g_3, then a three-way statistical dependence between the expression levels of g_1, g_2 and g_3 is likely to be the result of a direct three-way interaction between the genes (i.e., combinatorial regulation).

3.1.2 False Discovery Rate

One of the main challenges in detecting three-way interactions is the vast number of potential interactions, especially when carried out on a genome-wide scale. Given N genes, there are

$O(N^2)$ potential two-way interactions, but $O(N^3)$ potential three-way interactions, a much larger number when N is large. Any method for finding three-way interactions must deal with the computational and statistical problems posed by such large numbers of hypotheses.

Because of the vast number of possible three-way interactions, false positives can easily arise. It is therefore crucial to estimate the false discovery rate (FDR) [22, 34]. The FDR is the fraction of predicted interactions ("discoveries") that are not real interactions, but simply occur by chance, because of noisy or insufficient data. To estimate FDR, the usual first step is to estimate a p-value for each possible discovery [22]. One could do this analytically if the data were, say, Gaussian, but this is not generally the case for interactions in gene expression data, whose distributions are complex and unknown. Permutation tests are often used to overcome these problems [1, 20, 35]. This effectively tests a set of predictions under the null hypothesis that there are no dependencies between the genes. This is fine for two-way interactions, but for three-way interactions, this null hypothesis is too strong. For instance, to estimate the FDR for three-way dependencies, the correct null hypothesis is that there are no three-way dependencies, though pairwise dependencies may still exist. Permutation tests do not capture this more-complex null hypothesis, since they eliminate *all* dependencies. As we shall see, using permutation tests can seriously underestimate the FDR of three-way interactions, sometimes by several orders of magnitude.

To overcome these problems, we develop methods for estimating FDR based on the bootstrap [23]. The main advantage of the bootstrap is that it does not depend on the distribution of the data, which in the biological context is often complex and unknown. The bootstrap is most straightforward in estimating variance. Bootstrap techniques for estimating p-values and confidence intervals are more subtle, but can be applied in many situations [23]. We adapt these techniques to estimating the FDR of three-way interactions.

Because this is an initial study in using the bootstrap to estimate FDR for three-way interactions in gene expression data, we use a relatively simple model of three-way interaction based on regression, for which FDR estimates are relatively straightforward by a variety of methods, thus facilitating comparison. In addition, the vast number of possible three-way interactions greatly increases the computational complexity of the problem. To carry out a large number of computational experiments in reasonable time requires a model of relatively

low complexity. We show, however, that our model is robust in that it detects three-way interactions and accurately estimates FDR even when the data comes from more complex models including real biological systems.

3.1.3 Validation

We validate the model in two ways. First we show that it detects three-way interactions in real data, specifically in expression data from yeast and *Arabidopsis thaliana*. We provide both direct and circumstantial validation. For circumstantial validation, we use the putative three-way interactions detected by our model to infer related biological properties, such as gene function, protein-protein interactions, transcription factor-target pairs and potential combinatorial regulation. These inferences are readily testable using existing biological databases. For direct validation, we show that the putative three-way interactions detected by our model are significantly enriched with known three-way interactions from a curated dataset.

Second, we validate the FDR estimates. This is much harder, since we need to know *all* the three-way interactions in a data set. Since this is impossible for real biological data, we test the method on simulated data, for which all interactions are known. In this way, we can compare the FDR estimates of the bootstrap approach to those of other approaches under a wide range of statistical conditions. We show, for example, that all approaches produce accurate FDR estimates under ideal conditions. As the data becomes more complex and realistic (e.g., non-Gaussian, dependent noise samples, correlated predictors, non-linear dependencies, multimodal, etc), the bootstrap approach continues to give reasonable FDR estimates, while the other approaches rapidly break down.

Finally, we show that in addition to detecting three-way interactions, our method also detects new two-way interactions that cannot be detected by conventional means. The usual way to detect two-way interactions in gene expression data is to look for pairs of genes that are highly correlated. However, many of the three-way interactions detected by our method involve genes that are not correlated. Any two such genes interact in a way that cannot be detected by correlation-based methods. To illustrate the potential utility of the method, we use it to add over 64,000 edges to SeedNet (Section 2.3.4).

3.2 Background and related work

Section 3.2.1 briefly describes previous work on detecting three-way interactions in gene expression data. Section 3.2.2 gives an overview of the bootstrap [23].

3.2.1 Previous works and their limitations

A few works have looked at the problem of detecting three-way interactions in gene expression data [20, 21, 24], and several approaches can be discerned. The discretization approach partitions the samples into three groups based on the expression level of a controller gene [20, 21]. The High Group contains those samples in which the expression level of the controller gene is highest (e.g., in the top 35% of its values). The Low Group contains those samples in which the expression level of the controller gene is lowest (e.g., bottom 35%). The remaining samples (e.g., 30%) are ignored. The coexpression of two genes using the samples in High Group is then compared with the coexpression of the two genes using the samples in Low Group. If a great discrepancy in coexpression is seen between the two groups (as measured by correlation or mutual information, for example), a three-way interaction is declared. A drawback to this approach is that choosing the partition size can be quite arbitrary. There is also a loss of information due to discretization. Firstly, the continuous gene expression levels are totally lost for the controller gene. Secondly, for the other two genes, samples not in High Group and Low Group are ignored. The discretization approach also puts constraints on data, limiting its general applicability. For example, the controller gene must respect a range constraint and the other two genes must respect an independence constraint [20]. The gene expression of the controller gene must have a bimodal distribution and only genes with highest variance of expression levels are used [21]. The Liquid-Association approach [24] quantifies the correlation change of two genes conditioned on the increase or decrease of the gene expression level of a third gene, without explicitly partitioning the samples. However, its conversion of gene expression levels to ranks does lead to some information loss.

Another important question is how to assess the statistical significance of discoveries, a question not adequately addressed in most studies of three-way interaction. In [20], p-values obtained by permutation tests are Bonferroni corrected to measure the significance of discoveries

(putative three-way interactions), which is overly conservative. The permutation is done by choosing the partitions (High Group and Low Group) at random. This amounts to permuting the gene expression levels of the controller gene. FDR is briefly mentioned in [20], but variance of FDR estimates is not considered. The Liquid-Association approach adopts similar permutation tests to assign discoveries with p-values. Because there are so many triples, the expression profile of one gene within a triple is permuted as many as 10^6 times to obtain a distinguishable p-value. The permutation approach is not only computationally very expensive, but also tends to underestimate the false discovery rate (as we shall see). To circumvent this, some authors set aside a test dataset to validate their discoveries [21]. That is, some samples are first used to select triples of genes that are likely to interact. The remaining samples are then used to estimate a p-value for these putative interactions. Unfortunately, this approach is statistically inefficient since it uses only half the available samples to estimate p-values. Moreover, it makes two unrealistic assumptions about the null distribution: it assumes that all gene triples have the same null distribution, and it assumes that this null distribution is the same as the observed distribution, which is actually a mixture of null and non-null distributions. It is not hard to construct examples in which the first assumption is false. Moreover, the second assumption leads to overestimates of p-values and false discovery rates, sometimes by many orders of magnitude, especially when the number of non-null hypotheses (i.e., three-way interactions) is large.[1]

3.2.2 The bootstrap

The bootstrap is a computationally intensive statistical inference method that has found many applications since its birth in 1979 [23]. Statistical inference is used to draw conclusions about a population based on a sample from the population. There are two modes of bootstrap: parametric and nonparametric. In the parametric mode, we need to make restrictive assumptions about the underlying population [36]. The resulting inference can be erroneous if these assumptions do not meet reality. In fact, we observed that the parametric bootstrap method produces severely underestimated false discovery rates when the data disagree with the assumed model [data not shown]. In the nonparametric mode, however, we do not have to make such prior

[1] This is certainly the case for two-way interactions, in which the observed distribution of correlation coefficients is much wider than the null distribution of correlation coefficients derived by permutation tests.

assumptions. Since we have little knowledge about the underlying distributions in our study, we focus on the nonparametric bootstrap.

We denote a dataset as $\mathbf{x} = (x_1, x_2, ..., x_n)$, containing n data points, and denote a statistic on the dataset as $s(\mathbf{x})$. Note that because $\mathbf{x}$ is random, so is $s(\mathbf{x})$. We wish to perform inference on $s(\mathbf{x})$, such as estimating its variance, confidence intervals, etc. To do this, the bootstrap mimics sampling from a population by sampling from a dataset. Each bootstrap sample consists of n data points that are drawn with replacement n times from $\mathbf{x} = (x_1, x_2, ..., x_n)$. Roughly speaking, the bootstrap method for inference works as follows. First, B bootstrap samples $\mathbf{x}^{*b}$, $b = 1, 2, ..., B$, are generated from the original dataset, $\mathbf{x}$. Second, for each b, a bootstrap replicate of the statistic, $s(\mathbf{x}^{*b})$, is computed using the bth bootstrap sample. Finally, the B bootstrap replicates of the statistic, $s(\mathbf{x}^{*\mathbf{1}}), s(\mathbf{x}^{*\mathbf{2}}), ..., s(\mathbf{x}^{*\mathbf{B}})$, are used to do inference on the statistic, e.g., compute standard errors, confidence intervals and so on. For example, to estimate the standard error for a statistic $\hat{\theta} = s(\mathbf{x})$, let $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$ be the estimate of $\hat{\theta}$ on bootstrap sample $\mathbf{x}^{*b}$, and let $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$ be the average of these estimates. Then the bootstrap estimate of standard error for $\hat{\theta}$ is $\hat{se}(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}}$.

Confidence intervals. $\hat{\theta}$ is a point estimate for the true parameter of a population, θ. $\hat{\theta}$'s bootstrap standard error $\hat{se}$ is a way to say how accurate this estimate is. Another way is to construct a $1 - 2\alpha$ confidence interval for θ, meaning that $100 \cdot (1 - 2\alpha)\%$ of the time θ would be in the interval. As the number of bootstrap samples increases, the variance of bootstrap estimates decreases (Section 19.3 in [23]). As the data size tends to infinity, the bootstrap interval converges to the standard interval that is based on the normal distribution assumption for $\hat{\theta}$ (Section 12.1 in [23]). However, in most small-sample problems the bootstrap interval is more accurate than the standard interval [23].

The standard interval is constructed as follows. Suppose $\hat{\theta}$ is normally distributed with mean θ and variance se^2, then $Z = \frac{\hat{\theta} - \theta}{se} \sim \mathcal{N}(0, 1)$. Let $z^{(\alpha)}$ be the $100 \cdot \alpha$th percentile of $\mathcal{N}(0, 1)$, i.e., $z^{(\alpha)}$ is a value below which $100 \cdot \alpha$ percent of data points in the $\mathcal{N}(0, 1)$ distribution fall. Likewise, let $z^{(1-\alpha)}$ be the $100 \cdot (1 - \alpha)$th percentile of $\mathcal{N}(0, 1)$. The probability that Z would fall between $z^{(\alpha)}$ and $z^{(1-\alpha)}$ is $P(z^{(\alpha)} \leq \frac{\hat{\theta} - \theta}{se} \leq z^{(1-\alpha)}) = 1 - 2\alpha$, which is equivalent

to $P(\hat{\theta} - z^{(1-\alpha)} \cdot se \leq \theta \leq \hat{\theta} - z^{(\alpha)} \cdot se) = 1 - 2\alpha$. The interval $(\hat{\theta} - z^{(1-\alpha)} \cdot se, \hat{\theta} - z^{(\alpha)} \cdot se)$ is called a $1 - 2\alpha$ standard interval. In reality, we do not know se, but we can use the bootstrap estimate of it, $\hat{se}$. In addition, $\hat{\theta}$ usually does not strictly follow a normal distribution, so the standard interval is an approximation.

The bootstrap-t interval is $(\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \hat{se}, \hat{\theta} - \hat{t}^{(\alpha)} \cdot \hat{se})$. In contrast to the standard interval, $z^{(1-\alpha)}$ is replaced by $\hat{t}^{(1-\alpha)}$, and $z^{(\alpha)}$ is replaced by $\hat{t}^{(\alpha)}$. Rather than assuming that $Z = \dfrac{\hat{\theta} - \theta}{\hat{se}}$ follows a standard normal distribution (or a Student's t distribution), the distribution of Z is estimated empirically from the data at hand. $\hat{t}^{(1-\alpha)}$ and $\hat{t}^{(\alpha)}$ are the empirical $1 - \alpha$ and α quantiles of B bootstrap Z values, respectively. The bth bootstrap Z value is

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{se}^*(b)},$$

where $\hat{se}^*(b) = \hat{se}(\hat{\theta}^*(b))$ is the estimated standard error of $\hat{\theta}^*(b)$, which can be computed using bootstrap standard error estimate on $\mathbf{x}^{*b}$ (note: not on $\mathbf{x}$). The bootstrap-t method is not very reliable in practice for more general problems such as computing an interval for a correlation coefficient (see the last paragraph in Section 12.5 in [23]). In addition, it requires expensive nested bootstrapping (for estimating $\hat{se}^*(b)$) and sophisticated variance stabilization (see Algorithm 12.1 in [23]).

The percentile method, on the other hand, is more straightforward, more robust and easier for computing. The $1 - 2\alpha$ percentile interval is

$$(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}),$$

where $\hat{\theta}^{*(\alpha)}$ is the $100 \cdot \alpha$th percentile of $\hat{\theta}^*$'s distribution, and $\hat{\theta}^{*(1-\alpha)}$ is the $100 \cdot (1 - \alpha)$th percentile. The percentile interval is range-preserving: it is within the allowable range of any statistic. For example, the percentile interval for a correlation coefficient is always within $[-1, 1]$, whereas bootstrap-t intervals (or standard intervals) could contain values outside the $[-1, 1]$ range. Improved versions of the percentile method, namely, the BC_a and ABC method (Chapter 14 in [23]), have been developed, but we do not discuss them here.

Larger B (i.e., more bootstrap samples) produces more accurate bootstrap estimates, but

also takes more time to compute (which can be a particular problem when the statistic itself is expensive to compute). So there is a trade-off between accuracy and time. In general, $B = 200$ is usually good enough for estimating standard errors, but larger B is required for constructing confidence intervals [23].

The bootstrap is automatic, requires few distributional assumptions and can handle both easy inference problems (e.g., estimating standard errors) and complicated inference problems (e.g., estimating standard errors of regression coefficients, estimating confidence intervals, and hypothesis testing). With the bootstrap we can easily generate data-driven sampling distributions of a statistic (such as the sample correlation coefficient), rather than rely on complicated mathematical derivations or unrealistic assumptions on the data distribution.

Regression coefficients. Our interest will be mainly on the bootstrap applied to linear regression models. We wish to do hypothesis testing on regression coefficients, compute their standard errors, and perhaps also construct their confidence intervals. For ease of discussion, we first introduce some notation. The regression model is

$$y = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_p c_p + \varepsilon,$$

where y is a response variable, c_1 to c_p are predictor variables, β_1 to β_p are regression coefficients, and ε is noise. The regression coefficients are estimated from a sample data set. Let y_i be the ith observation for the response variable y, and let $\mathbf{c}_i = (c_{i1}, c_{i2}, ..., c_{ip})$ be the ith observation for the predictor variables $c_1, c_2, ..., c_p$, respectively. Let $\mathbf{x}_i = (y_i, \mathbf{c}_i)$, and let our sample data set $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, as shown in the following table:

	y	c_1	c_2	$\cdots$	c_p
$\mathbf{x}_1$	y_1	c_{11}	c_{12}	$\cdots$	c_{1p}
$\mathbf{x}_2$	y_2	c_{21}	c_{22}	$\cdots$	c_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_n$	y_n	c_{n1}	c_{n2}	$\cdots$	c_{np}

In regression, the bootstrap can be done in two ways. We can either bootstrap the data or bootstrap the residuals. To bootstrap the data, a bootstrap sample $\mathbf{x}^{*b}$ is formed by drawing n elements from $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ with replacement. This is also called bootstrapping pairs ((y_i, $\mathbf{c}_i$) is called the ith pair). To bootstrap the residuals, we fit the regression model on $\mathbf{x}$, estimate regression coefficients $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$, and compute residuals $\varepsilon_i = y_i - \mathbf{c}_i\hat{\beta} = y_i - \hat{y}_i$. A bootstrap sample $\mathbf{x}^{*b}$ is formed by bootstrapping n residuals and adding each of them to $\hat{y}_i$. Bootstrapping residuals makes stronger assumptions about the model and distribution of errors than bootstrapping pairs does, i.e., it assumes that the regression model is correct and that the errors are identically distributed and in particular that all the errors have the same variance [37]. With either method, we can obtain bootstrap estimates of regression coefficients, $\hat{\beta}^*(b)$, by fitting the regression model on $\mathbf{x}^{*b}$. Standard errors for the regression coefficient(s) can be computed in exactly the same way as outlined above in this section.

Hypothesis testing. Another important use of the bootstrap, especially in this chapter, is hypothesis testing, which is closely related to confidence intervals [23, 38]. In general, bootstrap hypothesis testing is carried out as follows. Suppose our null hypothesis is $H_0 : \theta = \theta_0$, and the alternative hypothesis is $H_1 : \theta \neq \theta_0$. Let the test statistic be

$$t = \frac{|\hat{\theta} - \theta_0|}{\hat{se}(\hat{\theta})},$$

where, as above, $\hat{\theta}$ is an estimate of θ, and $\hat{se}(\hat{\theta})$ is an estimate of the standard error of $\hat{\theta}$. This statistic is analogous to a t-statistic and to the statistic used in constructing studentized bootstrap confidence intervals [23]. Let B bootstrap replicates of this test statistic (under the null hypothesis) be

$$t_b^* = \frac{|\hat{\theta}^*(b) - \hat{\theta}|}{\hat{se}(\hat{\theta}^*(b))}, \quad b = 1, 2, ..., B.$$

These replicates approximate a sampling null distribution for t [38]. The p-value for a particular value of t is the proportion of t_b^* that are no less than t, i.e.,

$$\frac{\#\{t_b^* \geq t\}}{B}.$$

In the regression setting, testing the null hypothesis that a regression coefficient β_p is zero ($H_0 : \beta_p = 0$; $H_1 : \beta_p \neq 0$) simply amounts to replacing θ with β_p and replacing θ_0 with 0 in the above procedure. The standard error in the denominator of the formula for computing t or t_b^* also has an explicit mathematical formula and is therefore easy to compute [27]. The formula is an approximation assuming *i.i.d.* noise.[2] One contribution of this and next chapter is to show that this approximation is more than adequate for detecting three-way interactions in gene expression data and estimating their false discovery rates.

This chapter is organized as follows. Section 3.3 describes our regression-based detector for three-way interaction and our approach to estimating FDR. Section 3.4 demonstrates that the detected three-way interactions are biologically meaningful and useful. It describes FDR estimates of our bootstrap approach as well as other methods. The next chapter, Chapter 4, evaluates our bootstrap approach to estimating FDR over a wide range of statistical conditions.

3.3 Methods

3.3.1 Detecting three-way interactions in expression data using regression

Consider a triple of interacting genes, (g_1, g_2, g_3). For example, g_1 could be a transcription factor and g_2 a co-transcription factor in combinatorial regulation for the target gene g_3, or g_1 could be a transcription factor and g_2 could be g_1's modulator in post-translational regulation for the target gene g_3. We don't know how the predictor genes g_1 and g_2 are related to the target gene g_3. Nevertheless, we want to model their relationship.

A natural way to do this is with a regression model. The simplest is a first-order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_1, X_2 and Y are the expression levels of genes g_1, g_2 and g_3, respectively, and ε is independent random noise. This model captures linear, pairwise interactions between g_1, g_2 and g_3, but does not capture more-complex interactions, such as three-way interactions. To capture these, we can use a partial second-order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$. This model includes a quadratic interaction term, $X_1 X_2$, in addition to the linear non-interaction terms, X_1 and X_2. We found that FDR estimates based on this model are accurate when X_1 and X_2 are uncorrelated, but they rapidly become overly

[2] Independent and identically distributed noise.

optimistic as correlation increases [data not shown].[3] We have found that this problem can be substantially alleviated by introducing quadratic non-interaction terms, giving a full second-order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$. In this model, the X_1^2 term captures any quadratic dependence of Y on X_1. Without this term, any such dependence will show up in the interaction term, $X_1 X_2$, if X_1 and X_2 are correlated (since $X_1 X_2$ will then be correlated with X_1^2), leading to false positives. (Likewise for X_2^2.) Using a full second-order model prevents this.[4]

The second-order model is an approximation to some unknown relationship between the target gene and predictor genes since it can be considered as a Taylor series expansion of an unknown function $Y(X_1, X_2)$ [39]. The interaction term tells us how different levels of X_2 modify the effect of X_1 on Y (or symmetrically, how different levels of X_1 modify the effect of X_2 on Y). That is, the second-order term $\beta_5 X_1 X_2$ can be viewed as a first-order term $\beta_6 X_2$ where $\beta_6 = \beta_5 X_1$, i.e., β_6 is a coefficient whose value depends on X_1. In this way, X_1 modifies the interaction between X_2 and Y.

As we shall see, this second-order model works well in detecting three-way interactions and estimating FDR even when the data is generated by more-complex models. Also, as a special case, the second-order model includes the following model:

$$Y = a + b(X_1 - c)(X_2 - d) + \varepsilon,$$

where $\beta_0 = a + bcd, \beta_1 = -bd, \beta_2 = -bc$, $\beta_3 = \beta_4 = 0$ and $\beta_5 = b$. This model implies that the correlation between Y and X_2 is proportional to the value of $X_1 - c$.[5] In particular, Y and X_2 are correlated when $X_1 > c$, and anti-correlated when $X_1 < c$ (assuming b is positive).

In the second-order model, the highest power of the predictor variables is two. Although the response surface is curvilinear (because Y is non-linear in X_1 and X_2), it is still a linear regression problem (because Y is linear in the coefficients, β_i). To estimate the coefficients, we

[3] The problem does not arise when the data is generated by the partial second-order model itself, but when it is generated by more complex models.

[4] Of course, the X_1^2 term also provides a more complex model of the pairwise dependency between X_1 and Y, which in turn improves the FDR estimate, even when X_1 and X_2 are uncorrelated. (Likewise for X_2^2.) However, this improvement turns out to be quite minor, especially compared to the large improvement due to removing the effect of correlations between X_1 and X_2. [Data not shown]

[5] This is because the correlation between Y and X_2 is equal to the correlation between Y and $X_2 - d$, which is proportional to $b(X_1 - c)$, according to the regression model above.

fit the model to gene expression data $(y_i, x_{i1}, x_{i2}), i = 1, ..., N$. Let

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \varepsilon_i, \quad i = 1, ..., N \tag{3.1}$$

where y_i, x_{i1} and x_{i2} are (in our case) the ith gene expression levels for the target gene g_3, predictor gene g_1 and predictor gene g_2, respectively, and ε_i represents noise. Equation 3.1 can be conveniently expressed in a matrix form as follows:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where $\mathbf{y} = (y_1, y_2, ..., y_N)^T$ is a column vector, $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_N)^T$ is also a column vector, and $\mathbf{X}$ is an $N \times 6$ matrix with the ith row equal to $(1, x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2, x_{i1}x_{i2})$. It is well known that if we fit Equation 3.1 to the data, then $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_5)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the optimal estimate of the coefficients by the least squares method, which minimizes the residual sum of squares, $||\mathbf{y} - \mathbf{X}\beta||^2$ [27]. The fitted values $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_N)^T = \mathbf{X}\hat{\beta}$.

If the noise, ε, is *i.i.d.* and Gaussian, then the significance of the coefficient for the interaction term, β_5, can be tested using the following test statistic,

$$z = \frac{\hat{\beta}_5}{s.e.\{\hat{\beta}_5\}} = \frac{\hat{\beta}_5}{\hat{\sigma}\sqrt{\nu_5}},$$

where $\hat{\sigma} = \sqrt{\dfrac{1}{N-6}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$ and ν_5 is the last diagonal element of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ [27]. z has a t-distribution assuming the null hypothesis $H_0 : \beta_5 = 0$ is true [27]. A large observed value of $|z|$ will lead to rejection of the null hypothesis. Intuitively, a large value of $|z|$ indicates a strong three-way interaction. We shall show empirically that this statistic can be used to detect three-way interactions in gene expression data, even under non-Gaussian and non-*i.i.d.* assumptions, though care is needed to accurately estimate false discovery rates.

Like correlation coefficients, z can be positive or negative. Its sign also has a simple interpretation. Suppose (g_1, g_2, g_3) is a gene triple. If its associated z is non-zero, then the correlation between g_2 and g_3 depends on the expression level of g_1. In particular, if z is positive, then the correlation increases as the expression level of g_1 increases. Likewise, if z is negative, then the

correlation increases as the expression level of g_1 *decreases*. (These statements remain true if we swap g_1 and g_2, since values of z are symmetric in g_1 and g_2.)

In most of this chapter, we will not care about the sign of z, and we will detect three-way interactions by looking for triples with large values of $|z|$ (corresponding to a two-sided hypothesis test). This is not essential, however, and one might look for triples with a large, positive (or negative) value of z (Section 3.4.3.4).

3.3.2 Estimating FDR

This section describes different approaches to estimating FDR. FDR is a convenient measure of the statistical significance of the large number of discoveries made in genome-wide analyses. It is the proportion of discoveries expected to be false. For example, an FDR of 0.2 for ten discoveries says that we expect only two false discoveries among the ten. It might therefore be worthwhile to carry out (relatively expensive) targeted experiments to confirm these discoveries. So an accurate FDR estimate is of practical importance. Either underestimated or overestimated FDR can mislead us in our decision-making process. Recall that FDR is the (usually unknown) number of false discoveries divided by the (known) number of discoveries. In our application, to estimate the unknown number of false discoveries, we need to construct a null distribution of the test statistic, z (Section 3.3.1). Different ways of doing this amount to different ways of estimating FDR. The null distribution is the distribution of the test statistic assuming that there is no interaction effect, i.e., assuming that the null hypothesis $H_0 : \beta_5 = 0$ is true, where β_5 is the coefficient of the interaction term, X_1X_2, in our regression model (Equation 3.1). Estimating the number of false discoveries also requires an estimate of the number of true negatives. This is often approximated by the total number of possible discoveries, which leads to an overestimate of FDR [22]. We will use this approximation for the time being, and will correct for it later. To simplify further description, we let FD denote the expected number of false discoveries.

3.3.2.1 The analytical approach and permutation approaches

To estimate FD, we could use an analytical approach or an approach based on permutation tests. Both approaches are widely used in the literature, and both require estimating p-values.

The analytical approach assumes the null distribution has a particular mathematical form, from which p-values can be computed. For example, in the context of our regression model, the analytical approach assumes that the null distribution of z values follows Student's t distribution. So the analytical FD estimate for a particular z value is equal to its two-tailed p-value times the number of true negatives. The two-tailed p-value is equal to the area under the t-distribution curve to the right of $|z|$ and to the left of $-|z|$.

The permutation approach differs in using permutation tests to estimate p-values. This has the virtue of not making assumptions about the form of the null distribution. Because we are dealing with three-way interactions, two types of permutation test can be distinguished, which we call total and partial permutation. The total permutation approach constructs a sampling null distribution of z values by independently permuting all gene expression profiles. This tests the null hypothesis that there are no interactions between genes g_1, g_2 and g_3. It works as follows. (1) For each gene triple (g_1, g_2, g_3), randomly permute the gene expression profiles of genes g_1, g_2 and g_3; (2) Fit the second-order model (Equation 3.1) to the permuted data to obtain a z value for that triple, denoted by z_{null}; (3) Repeat steps (1) and (2) P times so that for each triple we get P values of z_{null}. The P values of z_{null} form a sampling null distribution of z_{null}; (4) For each gene triple, the estimated p-value at threshold τ is the proportion of z_{null} values whose magnitudes are greater than τ, i.e., $\frac{\#\{|z_{null}| > \tau\}}{P}$. Finally, the estimated FD is the sum of these p-values over all possible gene triples. Note that this approach, unlike the analytical approach, can generate a different null distribution for each gene triple (i.e., different gene triples are not assumed to have the same null distribution.)

The partial permutation approach is similar to the total permutation approach, but it only permutes the gene expression profile of g_1, effectively removing all interactions involving g_1. This removes any three-way interaction, but preserves any two-way interaction between g_2 and g_3. Likewise if we only permute the expression profile of g_2. The partial permutation approach differs from the total permutation approach only in step (1). As an added note, permuting g_3 removes all interactions involving the target gene. This would test the null hypothesis that the target gene has no interaction with g_1 and g_2. Since this is similar to the total permutation approach (and has many of the same problems explained below), we do not permute g_3.[6]

[6]One could also permute the expression profiles of g_1 and g_2, but this removes all interactions between g_1, g_2

Both the analytical and the permutation approaches have serious weaknesses. As mentioned above, the analytical approach assumes that the true null distribution of z values follows Student's t distribution. More specifically, $z = \frac{\hat{\beta}_5}{s.e.\{\hat{\beta}_5\}}$ follows a Student's t distribution with $N-p-1$ degrees of freedom, where N is number of observations and p is the number of predictor variables in the regression model [27]. Thus, the FD at threshold τ is $n \cdot p_{value}(\tau)$, where n is total number of true negatives, and $p_{value}(\tau)$ is the two-tailed p-value for τ under the Student's t distribution. This assumption is quite strong, for it implies that the second-order model is exactly correct for the data and that the noise ε_i is independent and normally distributed with mean 0 and constant variance. In reality, we don't know how g_1, g_2 and g_3 are related, so the second-order model is at best an approximation to their true relationship. Moreover, the noise ε_i is not Gaussian, nor has mean 0, nor has constant variance. The accuracy of analytical FD estimates depends heavily on how well the real data satisfy these rather strong assumptions.

The permutation approaches do not assume *i.i.d.* unbiased Gaussian noise. However, the total permutation approach destroys all interactions, both two-way and three-way. The sampling null distribution of z values generated in this manner thus assumes no interaction effect and no main effect. That is, the null hypothesis implies that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ in Equation 3.1, which is too strong. Consequently, this approach underestimates the false discovery rate, sometimes by several orders of magnitude, as we shall see. The partial permutation approach destroys all interactions involving g_1. Thus, the null hypothesis implies that $\beta_1 = \beta_3 = \beta_5 = 0$. This is a weaker null hypothesis, but still too strong. In reality, we only want to remove the interaction effect but keep all the main effects. That is, our desired null hypothesis is $\beta_5 = 0$.

3.3.2.2 The bootstrap approach

The bootstrap approach makes fewer assumptions. It does not assume Gaussian noise, and does not destroy two-way interactions (main effects). It works as follows.

1. For each gene triple (g_1, g_2, g_3), do the following:

and g_3 and is equivalent to permuting all three profiles. Likewise for any other pair of genes.

(a) Construct the following table,

Y	X_1	X_2
y_1	x_{11}	x_{12}
y_2	x_{21}	x_{22}
y_3	x_{31}	x_{32}
$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}

where each entry (y_i or x_{ij}) is a gene expression level. Each row corresponds to one experimental condition, and each column corresponds to one gene. Thus, each column is a gene expression profile. As before, X_i is the expression level of gene g_i ($i = 1, 2$), and Y is the expression level of gene g_3. The above table is our original sample.

(b) Fit the second-order model (Equation 3.1) to this original sample to obtain $\hat{\beta}_5$, $\hat{\sigma}$ and ν_5, as described in Section 3.3.1. Compute the following test statistic:

$$z = \frac{\hat{\beta}_5}{\hat{\sigma}\sqrt{\nu_5}}.$$

(c) Draw n rows at random with replacement from the above table and make a new table. This amounts to taking a bootstrap sample.

(d) Fit the second-order model to the bootstrap sample to obtain $\hat{\beta}_5^*$, $\hat{\sigma^*}$ and ν_5^*, which are the bootstrap replicates of $\hat{\beta}_5$, $\hat{\sigma}$ and ν_5, respectively. The null z value from the current bootstrap sample, z_{null}^*, is defined as [37]

$$z_{null}^* = \frac{\hat{\beta}_5^* - \hat{\beta}_5}{\hat{\sigma^*}\sqrt{\nu_5^*}}.$$

(e) Repeat steps (c) and (d) B times to obtain B values of z_{null}^*, which are used to form the sampling null distribution of z values.

2. For each gene triple, the estimated p-value at threshold τ is the proportion of z_{null}^* values

whose magnitudes exceed τ, i.e., $\frac{\#\{|z^*_{null}| > \tau\}}{B}$. The estimated FD is the sum of these p-values over all possible gene triples.

Note that like the permutation approaches, the bootstrap approach will in general form a different null distribution for each gene triple. Also, since the number of triples is vast, the total number of z^*_{null} values used to estimate an FDR can be very large, even when the number bootstrap samples, B, is small. It is therefore possible to obtain accurate FD estimates with a relatively small number of bootstrap samples. This is reflected in Figure 3.12, which shows the variance of the FD estimates.

3.3.2.3 Estimating the number of true negatives

As mentioned above, the number of possible discoveries is often used as an approximation for the number of true negatives, which leads to an overestimate of FD and FDR. This overestimate is small when the total number of true positives is relatively small. However, it can be arbitrarily large when the number of true positives is large, as is often the case in the analysis of high-throughput genome data [34].

We therefore need to adjust the above FD estimates by the proportion of all true negatives, which we do not know but again can estimate. Let p_i be the p-value of triple i, as estimated in step 2 of the procedure in Section 3.3.2.2. The estimated number of true negatives is given by

$$\frac{\#\{p_i > \lambda\}}{1 - \lambda},$$

where λ is an adjustable parameter between 0 and 1 [34]. This formula is based on the observation that the distribution of p-values comprises two parts: p-values clustered around zero, which are from true positives, and p-values uniformly distributed on the interval $[0, 1]$, which are from true negatives. To illustrate, Figure 3.1 shows a histogram of p-values for 10,000 randomly chosen triples from a sample of seed germination/dormancy data of Chapter 2. True positives are above the red line, and true negatives are below the red line. λ is a threshold for the p-values that are considered to be predominantly from true negatives, so we usually choose a value of λ such that beyond this value the distribution of p_i's is almost uniform. Let A be

the total area under the red line, and let B be the total area under the red line and right of the yellow bar. Note that $A = \#\text{true negatives}$, $B = \#\{p_i > \lambda\}$ and $B = (1-\lambda)A$. Therefore, $A = \dfrac{B}{1-\lambda} = \dfrac{\#\{p_i > \lambda\}}{1-\lambda}$. To adjust the FD and FDR estimates, we simply multiply them by the proportion of true negatives, $\dfrac{A}{m}$, where m is the total number of possible discoveries.

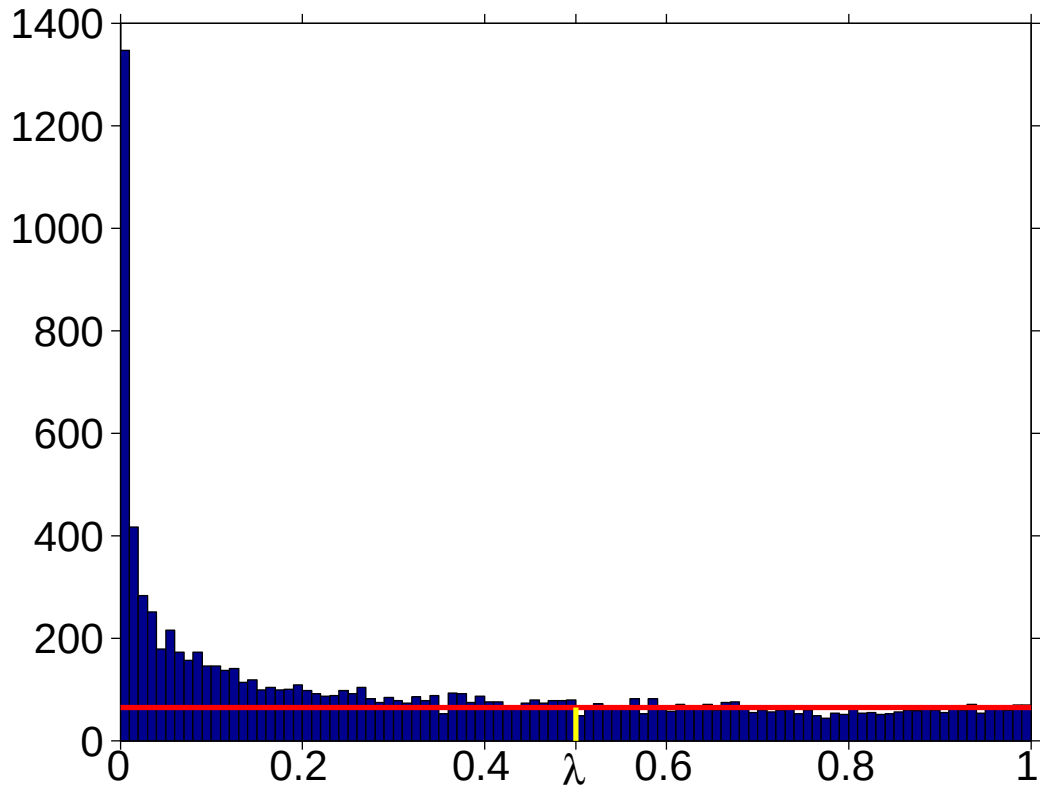


Figure 3.1: The distribution of 10,000 p-values for 10,000 randomly picked triples from the seed germination/dormancy data. This histogram shows a peaky uniform distribution, peaked near 0. The vertical yellow bar represents the value of λ.

3.4 Results and discussion

3.4.1 Correlation change

To illustrate our method, we construct a set of gene triples that are potentially involved in combinatorial regulation of seed germination and dormancy. The gene expression data used in this example is from germinating and dormant seeds of *Arabidopsis thaliana* as described in

Section 2.3.

A triple (g_1, g_2, g_3) is formed by choosing two predictor genes, g_1 and g_2, from a set of 882 transcription factors[7] of *Arabidopsis* and choosing a target gene, g_3, from a set of 74 target genes that are known to promote germination or dormancy. We have $\binom{882}{2}$ combinations for the two predictor genes, and $\binom{74}{1}$ combinations for the target gene, for a total of 28,750,554 triples. We use X_1, X_2 and Y to represent the expression levels of g_1, g_2 and g_3, respectively. For each triple, the second-order model in Equation 3.1 is fit to the germination/dormancy gene expression data and the value of z for the interaction term $\beta_5 X_1 X_2$ is estimated. (Recall that z is a test statistic for testing the null hypothesis $H_0 : \beta_5 = 0$, as described in Section 3.3.) The triples are sorted in descending order of $|z|$. So the triple estimated to have the strongest three-way interaction is on top.

As described in Section 3.3.1, three-way interactions can be viewed as condition-dependent correlations. That is, the correlation between Y and X_1 (X_2) depends on X_2 (X_1). We illustrate this using the *Arabidopsis* triples with strong $|z|$ values.

Specifically, let the High/Low Group of samples for X_2 be the 50 samples on which X_2 has the highest/lowest expression levels.[8] Let C_a be the Pearson correlation coefficient of Y and X_1 on the High Group, and let C_b be their correlation on the Low Group. Table 3.1 shows triples with the highest $|z|$ values, along with the values of C_a, C_b and $|C_a - C_b|$. This table provides a biologically meaningful interpretation for the $|z|$ values. The triples with strong three-way interactions (i.e., large $|z|$ values) also exhibit large correlation change (i.e., large $|C_a - C_b|$ value), from High Group to Low Group. In other words, the correlation between X_1 and Y when X_2 is low is significantly different from the correlation when X_2 is high. Moreover, the correlation coefficients in most cases change sign between column C_a and column C_b, going from correlation to anti-correlation, or vice versa, showing that the direction of correlation between X_1 and Y is controlled by X_2, the expression level of g_2.

For example, consider the third row in the table. Here, g_1 = AT2G23740 is strongly correlated with g_3 = AT2G32460 ($r = 0.87$) when g_2 = AT4G18390 is highly expressed, but becomes strongly anti-correlated ($r = -0.81$) when g_2 is lowly expressed. The absolute cor-

[7] We downloaded 1,515 TFs from AtTFDB [40], 882 of which are included in our seed data.
[8] The division of samples into High Group and Low Group is different for each gene triple.

relation change is 1.68. With a different controller gene, g_2 = AT5G63750, the direction of correlation change is reversed (see the second row), i.e., the correlation between the same g_1 and same g_3 is changed from negative ($r = -0.79$) to positive ($r = 0.89$). Such an interesting relationship between g_1 and g_3 is hardly detectable by pairwise correlation analysis. To see this, let C_1 be the Pearson correlation coefficient of X_1 and Y, and let C_2 be the Pearson correlation coefficient of X_2 and Y, on all samples. Notice that C_1 is only 0.21 in the second and third row, usually deemed insignificant, thus the relationship between g_1 and g_3 is undetectable by pairwise correlation analysis.

g_1	g_2	g_3	C_a	C_b	$\lvert C_a - C_b \rvert$	$\lvert z \rvert$	C_1	C_2
AT5G01960	AT2G23740	AT2G32460	-0.43	0.57	1.00	17.20	0.21	0.21
AT2G23740	AT5G63750	AT2G32460	-0.79	0.89	1.69	15.09	0.21	0.27
AT2G23740	AT4G18390	AT2G32460	0.87	-0.81	1.68	14.01	0.21	-0.36
AT2G45190	AT2G23740	AT2G32460	0.65	-0.44	1.10	13.73	-0.06	0.21
AT5G18830	AT5G01960	AT2G32460	-0.73	0.85	1.58	13.57	0.19	0.21
AT5G18830	AT5G63750	AT2G32460	-0.69	0.77	1.45	13.49	0.19	0.27
AT5G01960	AT4G32730	AT2G32460	-0.58	0.48	1.06	13.33	0.21	0.29
AT5G18830	AT2G28510	AT2G32460	-0.59	0.86	1.45	13.27	0.19	0.62
AT2G01650	AT4G18390	AT2G32460	-0.83	0.77	1.61	13.26	-0.05	-0.36
AT5G12840	AT2G28510	AT2G32460	-0.70	0.81	1.51	13.25	-0.00	0.62
AT4G14410	AT5G63750	AT2G32460	0.74	-0.69	1.42	13.24	-0.18	0.27
AT5G18830	AT4G18390	AT2G32460	0.82	-0.76	1.58	13.17	0.19	-0.36
AT4G18390	AT1G14685	AT2G32460	0.45	-0.79	1.24	13.13	-0.36	0.22
AT4G21750	AT4G18390	AT2G32460	0.90	-0.66	1.56	13.09	0.38	-0.36
AT2G20400	AT5G07680	AT2G32460	-0.71	0.89	1.60	13.07	0.02	0.66
AT5G63750	AT2G46530	AT2G32460	-0.63	0.71	1.34	12.94	0.27	0.24
AT2G02470	AT5G52510	AT2G32460	-0.75	0.79	1.53	12.89	0.03	0.39
AT1G08320	AT4G30080	AT1G31970	0.82	-0.14	0.96	12.82	0.27	0.07
AT1G14580	AT5G63750	AT2G32460	-0.78	0.76	1.54	12.80	0.09	0.27
AT4G21750	AT1G62300	AT2G32460	-0.59	0.86	1.45	12.80	0.38	0.30
AT1G14510	AT1G63900	AT3G47620	0.86	0.02	0.84	12.79	0.62	-0.18
AT2G20400	AT4G28890	AT2G32460	-0.68	0.88	1.56	12.75	0.02	0.66
AT5G01960	AT1G14685	AT2G32460	-0.47	0.73	1.21	12.74	0.21	0.22
AT4G04890	AT5G01960	AT2G32460	-0.65	0.84	1.49	12.72	0.24	0.21
AT2G28510	AT2G20400	AT2G32460	0.04	0.83	0.79	12.70	0.62	0.02
AT3G60030	AT2G23740	AT2G32460	-0.45	0.46	0.91	12.68	0.09	0.21
AT3G02790	AT2G23740	AT2G32460	0.50	-0.42	0.91	12.65	-0.06	0.21
AT2G23740	AT5G65510	AT2G32460	-0.60	0.85	1.45	12.63	0.21	0.42
AT5G51230	AT5G63750	AT2G32460	-0.72	0.79	1.51	12.60	0.13	0.27
AT5G52510	AT2G20400	AT2G32460	0.08	0.65	0.57	12.56	0.39	0.02
AT5G01960	AT4G36780	AT2G32460	-0.55	0.66	1.21	12.52	0.21	0.36
AT3G11200	AT5G01960	AT2G32460	-0.47	0.89	1.36	12.50	0.35	0.21
AT5G01960	AT3G16857	AT2G32460	-0.57	0.62	1.19	12.44	0.21	0.22
AT1G22985	AT3G04070	AT3G47620	-0.71	-0.22	0.49	12.43	-0.46	-0.21
AT1G14510	AT1G62300	AT2G32460	-0.74	0.79	1.53	12.42	0.25	0.30

AT5G52510	AT2G30470	AT2G32460	0.22	0.70	0.48	12.36	0.39	0.23
AT1G63840	AT4G21750	AT2G32460	-0.20	0.83	1.03	12.35	0.61	0.38
AT1G43850	AT5G52510	AT2G32460	-0.75	0.83	1.59	12.35	0.21	0.39
AT3G47600	AT5G63750	AT2G32460	-0.74	0.70	1.44	12.33	0.03	0.27
AT4G21750	AT5G01960	AT2G32460	-0.55	0.92	1.47	12.29	0.38	0.21
AT4G04890	AT1G62300	AT2G32460	-0.63	0.87	1.50	12.27	0.24	0.30
AT5G20910	AT3G04070	AT3G47620	-0.83	0.06	0.88	12.27	-0.45	-0.21
AT3G16770	AT3G62090	AT1G69260	0.70	-0.64	1.34	12.20	0.35	0.60
AT4G21750	AT5G63750	AT2G32460	-0.59	0.91	1.50	12.18	0.38	0.27
AT2G28510	AT5G55970	AT2G32460	0.11	0.84	0.73	12.18	0.62	0.16
AT2G23780	AT5G63750	AT2G32460	-0.67	0.73	1.39	12.15	0.07	0.27
AT5G65510	AT1G14580	AT2G32460	0.10	0.65	0.54	12.13	0.42	0.09
AT3G18290	AT2G28510	AT2G32460	-0.48	0.90	1.38	12.12	0.16	0.62
AT3G16770	AT3G60530	AT3G50500	0.67	-0.80	1.47	12.11	-0.24	0.17
AT3G18290	AT5G63750	AT2G32460	-0.75	0.78	1.54	12.10	0.16	0.27

Table 3.1: Correlation properties of the top 50 discoveries on the seed germination/dormancy data from *Arabidopsis thaliana* (i.e., the 50 three-way interactions with the highest values of $|z|$). The FDR for these discoveries is 0.006 (as estimated by the bootstrap method). g_1, g_2 and g_3 form a gene triple, where g_3 is the target gene, and g_1 and g_2 are predictor genes. C_a and C_b are the Pearson correlation between the profiles of g_1 and g_3 on the 50 High Group samples and the 50 Low Group samples, respectively. C_1 and C_2 are the correlation on all samples between the profiles of g_1 and g_3 and between the profiles of g_2 and g_3, respectively.

Table 3.2 shows full gene names and their current annotations for the gene IDs in Table 3.1 [41].

Figure 3.2 is a graphical representation of the relationships listed in Table 3.1. Target genes g_3 are shown in dark grey, and predictor genes g_1 and g_2 are shown in light grey. Each triple has three nodes and the three nodes are linked by two edges, $g_1 - g_3$ and $g_2 - g_3$. The triples are clustered with various degrees. The target gene AT2G32460 has the highest degree (i.e., is contained in many triples), suggesting that it participates in many distinct three-way interactions. Since it is affected by so many predictor genes, AT2G32460 might play a central

role in seed germination. A simple three-way interaction is depicted as an isolated triple, such as the one whose target gene is AT1G31970. Gene AT3G16770 affects two different target genes, in combination with two other predictor genes, AT3G60530 and AT3G62090 (see the chain of 5 nodes in the lower left corner). This is also the case for gene AT1G14510 (see the node bridging the two circular clusters).

Locus name	full gene name	gene annotations
AT1G08320	bZIP21	calcium-mediated signaling, response to mechanical stimulus [42]
AT1G14510	AL7/Alfin-like 7	methylated histone binding [43]
AT1G14580	C2H2-like zinc finger protein	sequence-specific DNA binding transcription factor activity [44]; zinc ion binding, nucleic acid binding (Communication:501714663)
AT1G14685	BPC2/Basic Pentacysteine 2	regulation of developmental process, response to ethylene [45]
AT1G22985	CRF7/Cytokinin response factor 7	protein binding [46]; DNA binding (Communication:501714663)
AT1G31970	STRS1/Stress response suppressor 1	RNA methylation [42]; ATP-dependent helicase activity (Communication:501714663)
AT1G43850	SEU/SEUSS	hydrogen peroxide catabolic process [42]; transcription cofactor activity, multicellular organismal development, ovule development [47]; protein binding, protein heterodimerization activity [48]; regulation of flower development [49]; gynoecium development [50]

AT1G62300	ATWRKY6	cellular response to boron-containing substance deprivation [51]; intracellular signal transduction, amino acid import, respiratory burst involved in defense response, toxin catabolic process [42]; response to chitin [52]; cellular response to phosphate starvation, protein binding [53]
AT1G63840	RING/U-box superfamily protein	abscisic acid-activated signaling pathway, response to ethylene, hyperosmotic salinity response, signal transduction, response to auxin, response to jasmonic acid, response to water deprivation [42]; response to abscisic acid [54]
AT1G63900	DAL1/DIAP1-like protein 1	protein import into chloroplast stroma, ubiquitin-protein transferase activity, chloroplast organization [55]
AT1G69260	AFP1/ABI five binding protein	response to water deprivation, negative regulation of programmed cell death, salicylic acid mediated signaling pathway, response to ethylene, jasmonic acid mediated signaling pathway, hyperosmotic salinity response, signal transduction, response to auxin [42]; abscisic acid-activated signaling pathway [56]
AT2G01650	PUX2/Plant UBX domain-containing protein 2	N-terminal protein myristoylation [42]; zinc ion binding, nucleic acid binding (Communication:501714663)
AT2G02470	AL6/Alfin-like 6	cellular response to phosphate starvation, root hair elongation, metal ion homeostasis [57]; methylated histone binding [43]
AT2G20400	MYB-like HTH transcriptional regulator family protein	pollen development [58]

AT2G23740	SUVR5/SU(VAR)3-9-related protein 5	chromatin silencing, regulation of histone H3-K9 dimethylation [59]; zinc ion binding (Communication:501714663)
AT2G23780	RING/U-box superfamily protein	zinc ion binding (Communication:501714663); phosphatidylinositol biosynthetic process [42]
AT2G28510	Dof-type zinc finger DNA-binding family protein	sequence-specific DNA binding transcription factor activity [44]
AT2G32460	MYB101/MYB domain protein 101	gibberellin biosynthetic process, gibberellic acid mediated signaling pathway [42]; positive regulation of programmed cell death [60]; positive regulation of abscisic acid-activated signaling pathway [61]; pollen development [58]
AT2G45190	AFO/Abnormal Floral Organs, FIL/FILAMENTOUS FLOWER, YAB1	embryo development ending in seed dormancy, stomatal complex morphogenesis, seed germination, chromatin assembly or disassembly, vegetative to reproductive phase transition of meristem, ovule development, seed dormancy process, response to abscisic acid, iron-sulfur cluster assembly [42]; meristem structural organization, cell fate commitment [62]; protein binding [63]; inflorescence meristem growth [64]
AT2G46530	ARF11/Auxin response factor 11	sequence-specific DNA binding transcription factor activity (Communication:501714663)
AT3G02790	MBS1/Methylene blue sensitivity 1	cellular response to singlet oxygen [65]

AT3G04070	NAC047/NAC domain containing protein 47	multicellular organismal development (Communication:501714663); organ senescence, amino acid transport [42]
AT3G11200	AL2/Alfin-like 2	mitotic nuclear division [42]; methylated histone binding [43]
AT3G16770	ERF72/Ethylene response factor 72	ethylene-activated signaling pathway, response to ethylene, protein binding [66]; response to cytokinin [67]; response to jasmonic acid [68]
AT3G16857	ARR1/Response regulator 1	protein N-linked glycosylation, cytokinin-activated signaling pathway, fatty acid beta-oxidation, regulation of seed germination, protein import into peroxisome matrix, regulation of shoot system development [42]; regulation of seed growth, regulation of root meristem growth [69]; phosphorelay response regulator activity [70]; regulation of anthocyanin metabolic process, regulation of chlorophyll biosynthetic process, primary root development [71]
AT3G18290	BTS/BRUTUS, EMB2454/Embryo defective 2454	cellular response to iron ion starvation [72]; embryo development ending in seed dormancy (Communication:501718471); zinc ion binding (Communication:501714663)
AT3G47600	ATMYB94/MYB domain protein 94	response to jasmonic acid, response to ethylene, response to salt stress, response to salicylic acid, response to cadmium ion, response to abscisic acid, response to auxin [73]; response to karrikin [74]

AT3G47620	TCP14/TEOSINTE branched, cycloidea and PCF (TCP) 14	inflorescence development, cell proliferation [75]; response to abscisic acid, regulation of seed germination, response to gibberellin [76]; regulation of defense response [77]; response to cytokinin [78]; protein binding [79]
AT3G50500	SNRK2-2/SNF1-related protein kinase 2-2	Golgi organization, glycolytic process, hyperosmotic response, water transport [42]; regulation of seed germination, response to gibberellin [80]; protein kinase activity, positive regulation of abscisic acid-activated signaling pathway [81]; protein binding [82]; response to abscisic acid [83]; protein phosphorylation [84]; protein binding [85]
AT3G60030	SPL12/Squamosa promoter-binding protein-like 12	xylan biosynthetic process, glucuronoxylan metabolic process [42]
AT3G60530	GATA4/GATA transcription factor 4	response to light stimulus [86]
AT3G62090	PIL2/Phytochrome interacting factor 3-like 2	xylem development, cell wall macromolecule metabolic process [42]; red or far-red light signaling pathway [87]; protein binding [88]
AT4G04890	PDF2/Protodermal factor 2	embryo development ending in seed dormancy, protein acetylation, vegetative to reproductive phase transition of meristem, ovule development, iron-sulfur cluster assembly, thylakoid membrane organization [42]; cotyledon development [89]; epidermal cell differentiation [90]; maintenance of floral organ identity [91]
AT4G14410	bHLH104/basic Helix-Loop-Helix 104	sequence-specific DNA binding transcription factor activity [44]

AT4G18390	TCP2/Teosinte branched 1	regulation of translation, leaf morphogenesis, plastid organization, cell differentiation, positive regulation of heterochronic development [92]
AT4G21750	ATML1/Meristem layer 1	vegetative to reproductive phase transition of meristem, thylakoid membrane organization, iron-sulfur cluster assembly, embryo development ending in seed dormancy, ovule development, regulation of meristem growth, protein acetylation [42]; cotyledon development [89]; epidermal cell differentiation [90]
AT4G28890	RING/U-box superfamily protein	developmental growth, root hair elongation, lateral root development, root hair cell differentiation [42]; protein ubiquitination, ubiquitin-protein transferase activity [93]
AT4G30080	ARF16/Auxin response factor 16	cell division, response to auxin, root cap development [94]; miRNA binding [95]
AT4G32730	MYB3R1/C-MYB-Like transcription factor 3R-1	cytokinesis by cell plate formation [42]; transcription coactivator activity [96]
AT4G36780	BEH2 BES1/BZR1 homolog 2	regulation of transcription [97]
AT5G01960	RING/U-box superfamily protein	zinc ion binding (Communication:501714663)[9]
AT5G07680	NAC4	multicellular organismal development (Communication:501714663)

[9]Annotated by TIGR Arabidopsis annotation team when no external reference is available.

AT5G12840	EMB2220/Embryo defective 2220, NFYA1/Nuclear factor Y, subunit A1	embryo development ending in seed dormancy (Communication:501718471) [41]; microgametogenesis, somatic embryogenesis, seed development [98]; CCAAT-binding factor complex [99]; regulation of timing of transition from vegetative to reproductive phase [100]
AT5G18830	SPL7/Squamosa Promoter binding protein-like 7	actin nucleation, Golgi vesicle transport, tissue development, protein desumoylation, organ morphogenesis, trichome morphogenesis, root hair cell differentiation, hydrogen peroxide biosynthetic process, vegetative to reproductive phase transition of meristem, glucuronoxylan metabolic process, xylan biosynthetic process, cell wall organization, cell growth, positive regulation of organelle organization [42]
AT5G20910	AIP2/ABI3-interacting protein 2	amino acid transport [42]; ubiquitin-protein transferase activity, protein ubiquitination [93]; negative regulation of abscisic acid-activated signaling pathway, protein binding [101]
AT5G52510	SCL8/Scarecrow-like 8	sequence-specific DNA binding transcription factor activity [44]
AT5G55970	RING/U-box superfamily protein	zinc ion binding (Communication:501714663)

AT5G63750	ARI13/Arabidopsis Ariadne 13	protein ubiquitination (AnalysisReference:501757242)[10]; ligase activity (AnalysisReference:501756968)[11]; zinc ion binding (AnalysisReference:501756966)[12]; expressed during flowering stage, petal differentiation and expansion stage [102]
AT5G65510	AIL7/Aintegumenta-like 7	organ morphogenesis [103]; auxin mediated signaling pathway involved in phyllotactic patterning [104]; maintenance of shoot apical meristem identity [105]
At5G51230	EMF2/Embryonic flower 2	histone methylation, vernalization response, regulation of gene expression by genetic imprinting [42]; protein binding [106]; negative regulation of flower development [107]

Table 3.2: Gene annotations for the genes in Table 3.1.

3.4.2 Extending SeedNet

In this section we look at three-way interactions in *Arabidopsis thaliana* using the seed germination/dormancy data of Chapter 2. The data, as described in Section 2.3, contain gene expression levels for 14,088 *Arabidopsis* genes (after filtering) on 138 seed samples, of which 73 are non-germinating seeds and 65 are germinating seeds, maintained in diverse physiological and environmental conditions and representing a wide range of developmental stages. We wish to extend the coexpression network, SeedNet, constructed in Chapter 2 using exactly the same data but taking account of three-way interactions. As we will see, many new edges, ignored by pairwise correlation analysis, emerge as a result of three-way interactions.

SeedNet has about 500,000 edges. We applied our second-order model to about 29 million triples made from all of the 882 *Arabidopsis* transcription factors in the seed dataset (used as predictor genes) and all of the 74 target genes used in [1] (known to promote germination or

[10]Transitive UniPathway annotation
[11]Transitive UniProtKB annotation
[12]Transitive UniProtKB annotation

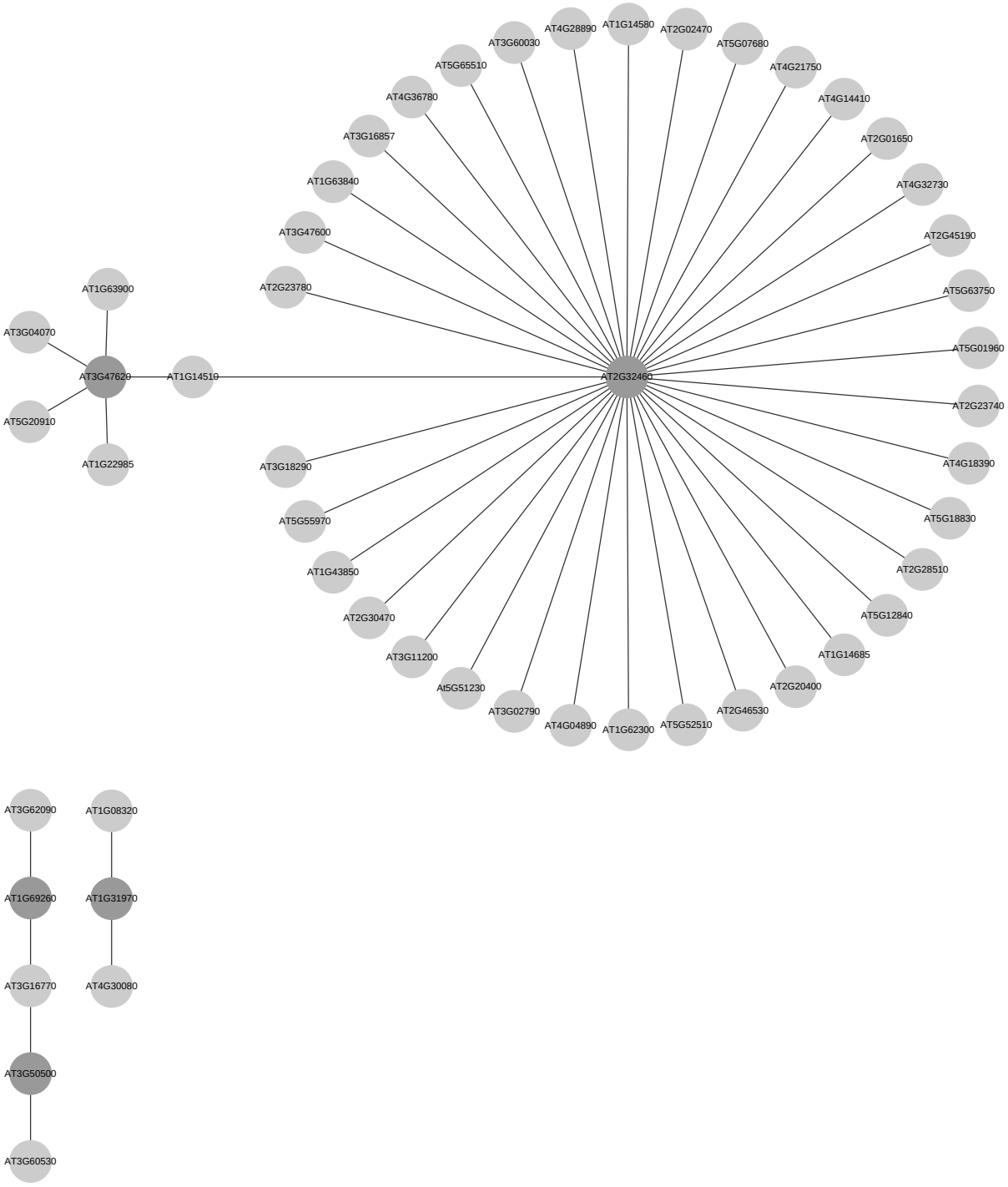


Figure 3.2: The graphical representation of three-way interactions from Table 3.1. The dark grey nodes represent target genes (g_3). The light grey nodes represent predictor genes (g_1 or g_2).

dormancy). We detected 1,912,656 significant triples with an estimated false discovery rate less than 0.05 (estimated by our bootstrap method), from which 64,698 distinct edges were extracted.

When we infer that triple (g_1, g_2, g_3) is significant, it implies that transcription factors g_1 and g_2 are involved in a (direct or indirect) three-way interaction with target gene g_3. We then create edges $g_1 - g_3$ and $g_2 - g_3$, and add them to SeedNet.[13] An edge $g_1 - g_3$ in SeedNet now means that genes g_1 and g_3 interact, possibly in combination with others genes. Likewise for edge $g_2 - g_3$. Interestingly, 99% of these distinct edges are new and not previously included in SeedNet. They are undetectable by pairwise correlation analysis alone, and thus serve as a potentially valuable augmentation to SeedNet.

3.4.3 Enrichment tests

We described our regression-based detector for three-way interactions (Section 3.3.1) and saw interesting patterns of correlation change among the top-ranked triples (Section 3.4.1). However, the question remains as to whether these putative three-way interactions are statistically significant, that is, whether they reflect biological reality within cells or are mere statistical anomalies. This question is particularly important in the present study because the vast number of possible three-way interactions means that a large number of anomalies can be expected.

To answer this question, we carried out a preliminary exploration by performing enrichment tests on two organisms, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* (yeast). For *Arabidopsis thaliana*, we do not have many curated three-way interactions, so we use inferred ones from other resources (e.g., protein-protein interactions, transcription factor-target pairs). For yeast, we have a limited set of 519 curated three-way interactions, and we use them. We show that our detected three-way interactions are enriched in terms of gene function, known protein-protein interactions, known transcription factor-target gene pairs, potential combinatorial regulation, and known three-way interactions.

In the remainder of this section, the first three subsections are about enrichment tests on *Arabidopsis thaliana*, and the last subsection is about yeast. To simplify further discussion,

[13] If triples (g_1, g_2, g_3) and (g_1', g_2, g_3) are both significant, then they both imply that edge $g_2 - g_3$ is significant. However, the edge is added to SeedNet only once.

we define a triple as three distinct genes (g_1, g_2, g_3). Let S_A be the set of all triples under consideration, and let S_I be a set of interesting triples (e.g., triples known to have three-way interactions). Obviously, $S_I \subset S_A$. We will look at a number of different criteria for interestingness, but in all cases, a positive result means that the detected triples are enriched with interesting triples.

More precisely, recall that each triple has a z score, measuring the confidence that the genes in the triple are involved in a three-way interaction (Section 3.3.1). We define the top N detections to be the N triples in S_A with the highest values of $|z|$. For each value of N, we define the number of interesting detections, X, to be the number of interesting triples amongst the top N detections. Enrichment means that X is significantly larger than expected by random chance. To demonstrate enrichment, we will compare plots of X vs. N for our detector against those for a random detector (e.g., Figure 3.3). We will also compute p-values of X for various values of N.

The gene expression data used in Sections 3.4.3.1, 3.4.3.2 and 3.4.3.3 are from *Arabidopsis* seeds as described in Section 2.3.

3.4.3.1 Gene function

In this section we ask the following question: are the detected three-way interactions enriched with genes promoting seed germination/dormancy? In other words, are the triples with higher $|z|$ scores more likely to contain genes promoting seed germination/dormancy? If we are detecting biological three-way interactions, then we would expect the detected triples to exhibit functional enrichment, just like the enrichment seen in significantly coexpressed gene pairs [1, 7]. This section shows that this is indeed the case.

Formally, let S_{TF} be the set of 882 transcription factors for *Arabidopsis thaliana* in the seed dataset. Let $S_{func} = S_{germ} \cup S_{dorm}$, where S_{germ} is a set of 39 genes known to promote seed germination, and S_{dorm} is a set of 35 genes known to promote seed dormancy.[14] Let the set of

[14]Data courtesy of George W. Bassel. These genes are collated from the literature by him (see the references in [1] for details).

all triples be:

$$S_A \;=\; \{(g_1, g_2, g_3) : g_1 \in S_{TF}, g_2 \in S_{TF}, g_3 \in S_{func}, g_1 \neq g_2, g_2 \neq g_3, g_1 \neq g_3\}.$$

That is, S_A consists of triples of distinct genes where the predictor genes (g_1 and g_2) are transcription factors and the target gene promotes germination/dormancy. The total number of triples in S_A is 28,734,696. Furthermore, we define two sets of interesting triples, S_I^1 and S_I^2.

$$S_I^1 \;=\; \{(g_1, g_2, g_3) : (g_1, g_2, g_3) \in S_A, g_1 \in S_{func} \text{ or } g_2 \in S_{func}\},$$

i.e., those triples in which at least one predictor gene promotes germination or dormancy. The total number of triples in S_I^1 is 1,146,312. Similarly,

$$S_I^2 \;=\; \{(g_1, g_2, g_3) : (g_1, g_2, g_3) \in S_A, g_1 \in S_{func} \text{ and } g_2 \in S_{func}\},$$

i.e., those triples in which both predictor genes promote germination or dormancy. The total number of triples in S_I^2 is 11,016.

Figure 3.3 shows the result of the enrichment test based on S_A and S_I^1. This figure plots X vs. N, where X is the number of interesting triples amongst the top N detections. The blue curve traces the values of X for our three-way interaction detector, whereas the red line traces the expected values of X for a random detector.[15] Suppose the total number of triples is M, K of which are interesting (i.e., size(S_A) = M and size(S_I) = K). The slope of the red line is K/M. Figure 3.4 shows the result of the enrichment test based on S_A and S_I^2. In both figures, the blue curve lies above the red curve, demonstrating that the top 100,000 triples are more likely to contain germination/dormancy promoting genes than by random chance. It also means that triples with a high value of $|z|$ have a higher density of such genes than triples with a low value of $|z|$. Thus, genes in triples with a high $|z|$ score exhibit functional enrichment, exactly as we would expect if we are detecting real three-way interactions.

In addition, p-values for gene function enrichment are extremely significant (as shown in

[15] In a random detector, we flip a coin in which the probability of heads is K/M (so that the expected proportion of predicted discoveries is equal to the proportion of true discoveries in the data).

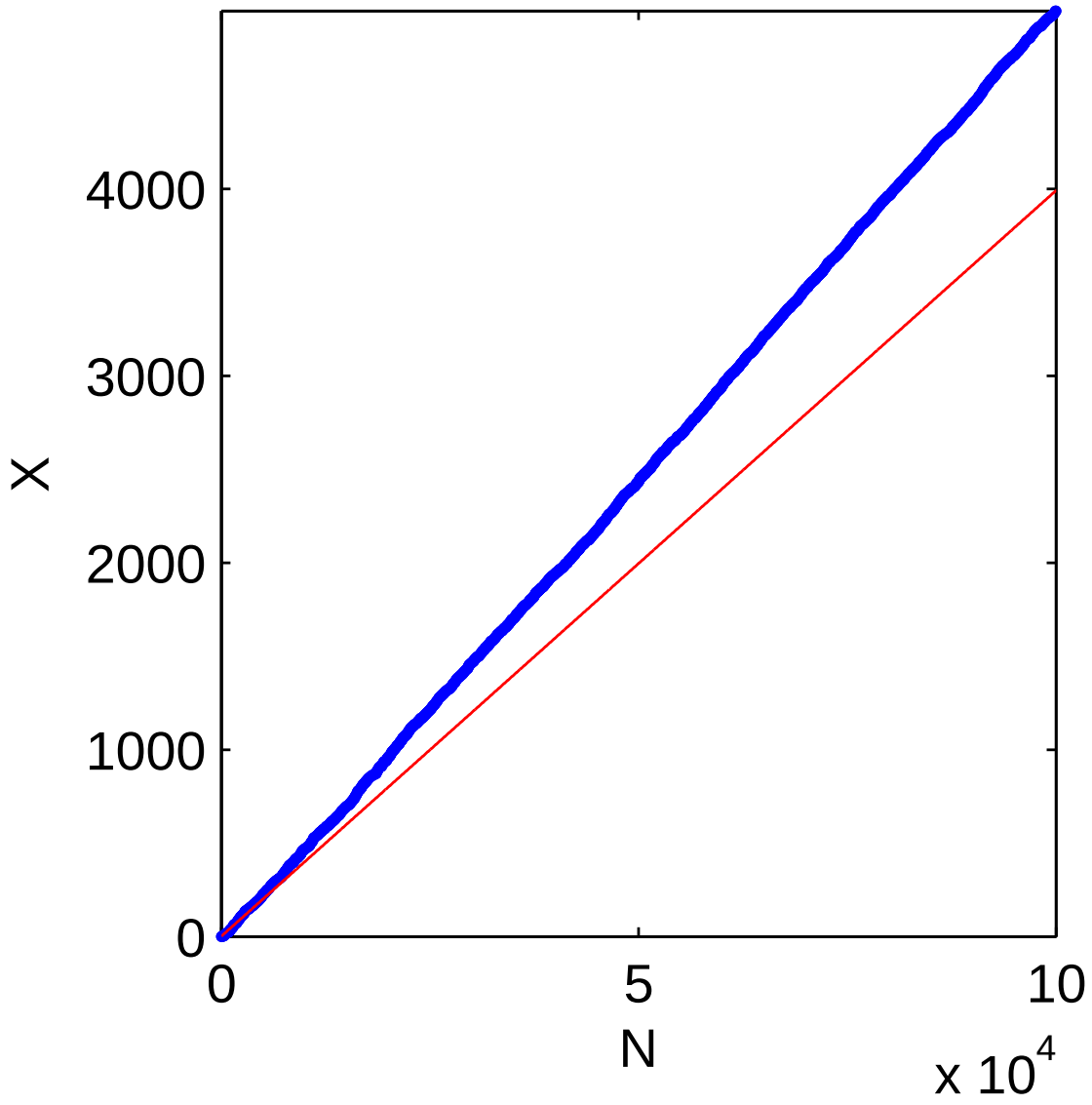


Figure 3.3: Gene-function enrichment curve (X vs. N). X is the number of interesting triples among the top N detections. A triple is interesting if it is in S_I^1 (i.e., at least one of its predictor genes promotes germination/dormancy). The blue line is the plot for our detector. The red line is the plot for a random detector of three-way interactions. P-values at 0.1%, 1%, 10% and 50% of all triples are 1.097E-13, 0, 0 and 0, respectively.

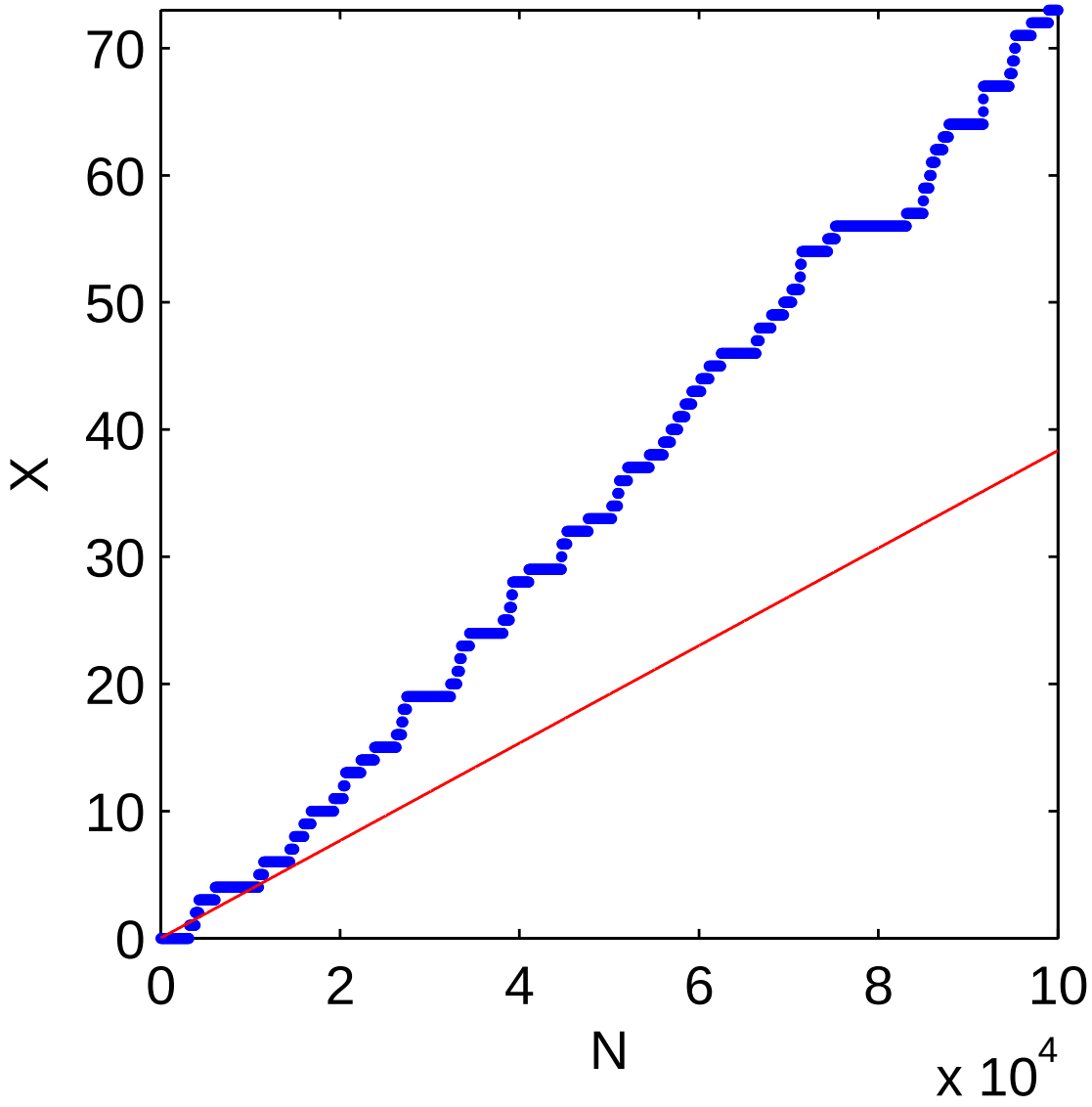


Figure 3.4: Gene-function enrichment curve. A triple is interesting if it is in S_I^2 (i.e., both predictor genes in the triple promote germination/dormancy). P-values at 0.1%, 1%, 10% and 50% of all triples are 0.017, 2.506E-13, 0 and 0, respectively.

the caption of both figures). To compute p-values, we used the hypergeometric test, which can be used to test for over-representation of interesting triples in the top N detections [108]. The hypergeometric p-value is

$$p = 1 - \sum_{i=0}^{X-1} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}. \tag{3.2}$$

Each figure gives p-values for several values of N. For example, Figure 3.3 says that when N is 0.1% of the total number of triples in S_A (i.e., N = 28,735), the p-value is 1.097×10^{-13}, which is extremely significant. For higher values of N, the p-values are essentially 0, so enrichment is virtually guaranteed.

3.4.3.2 Protein-protein interactions

This section shows that the detected triples are enriched with known protein-protein interactions. (That is, the gene triples are enriched with gene pairs whose protein products interact.) These results are understandable if we are detecting real three-way interactions, but not if we are detecting random triples. This is because three-way gene interactions may be associated with protein-protein interactions. For instance, the interaction between two proteins may affect the expression of a third gene, as in post-translational modification of transcription factors [20]. Similarly, two proteins that interact may have coexpressed genes, and this coexpression may be controlled by a third gene. If we are detecting real three-way interactions between genes, then such protein-protein interactions would appear in our detected triples.

To show that this does indeed happen, we construct interesting triples using protein-protein interaction (PPI) data [109, 110].[16] This PPI database contains 35,939 confirmed protein-protein interactions in *Arabidopsis*.

S_A is defined in the same way as in Section 3.4.3.1. S_I, the set of interesting triples, now consists of triples from S_A in which at least one edge is a known PPI interaction. Here one edge means any pair of genes within a triple (g_1, g_2, g_3). Figure 3.5 shows the result of the enrichment test based on S_I. The top one million triples are clearly enriched with the confirmed protein-protein interactions.

[16] Data courtesy of Nicholas Provart.

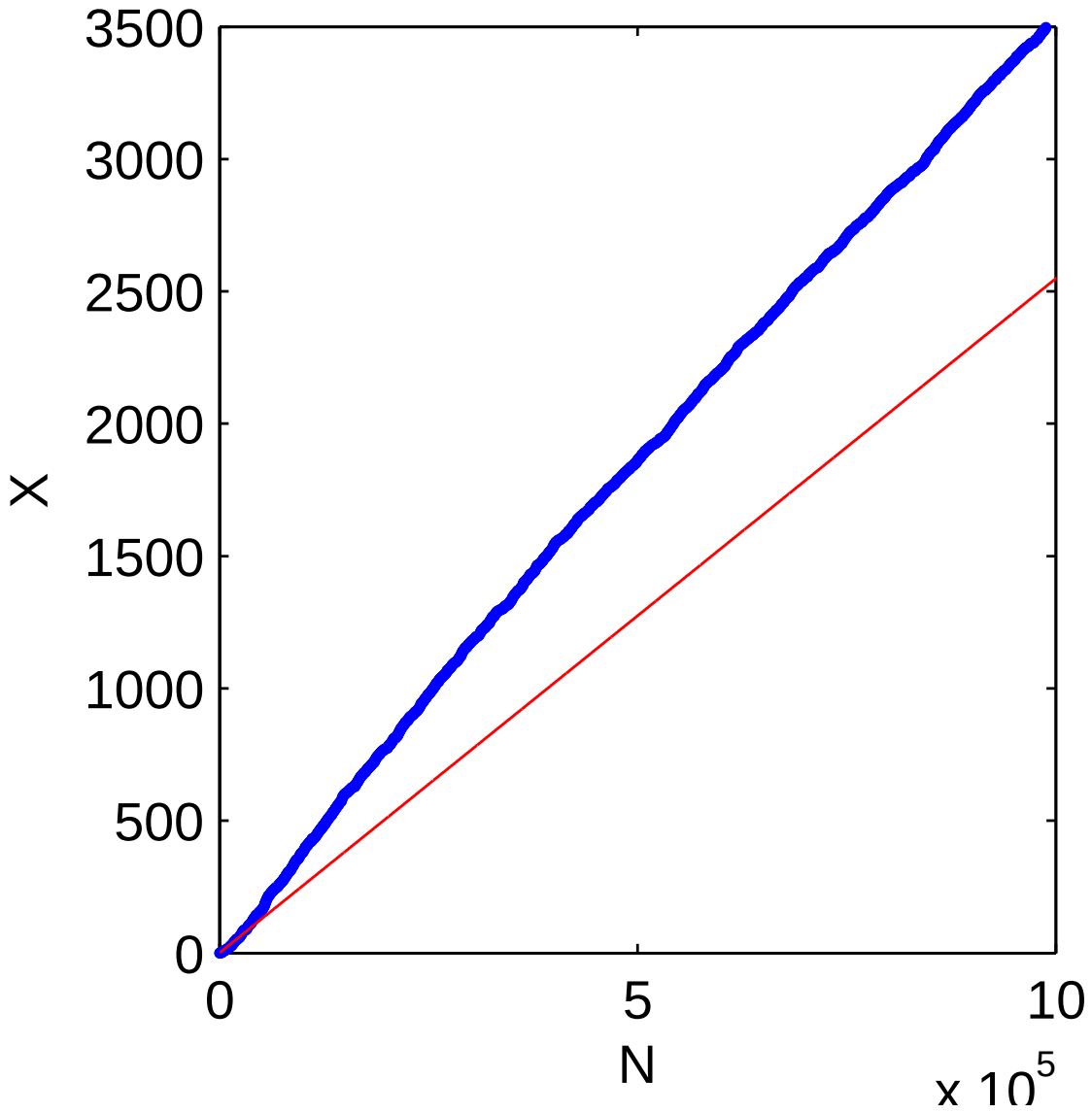


Figure 3.5: PPI enrichment curve. A triple is interesting if at least one pair of genes from the triple is a known protein-protein interaction. P-values at 1%, 10% and 50% of all triples are all 0.

3.4.3.3 Combinatorial regulation

This section shows that our discoveries are enriched with triples in which two transcription factors (TFs) regulate a common target gene, that is, triples of the form (g_1, g_2, g_3) in which g_1 and g_2 are known transcription factors for g_3. Such triples represent potential combinatorial regulation, which is a form of three-way interaction.

We define interesting triples using the TF-target database from the Arabidopsis Gene Regulatory Information Server (AGRIS) [111], which contains 11,483 direct interactions between TFs and target genes. For convenience of exposition, we call the database `agris`. S_A is the same as described in Section 3.4.3.1. S_I now consists of those triples (g_1, g_2, g_3) in S_A in which both (g_1, g_3) and (g_2, g_3) are in `agris`. Figure 3.6 shows that the top one million discoveries are enriched with interesting triples, as the blue curve is well above the red line. Moreover, because g_1 and g_2 directly interact with g_3,[17] it is likely that the detected three-way interactions are direct interactions, not indirect interactions, like those of earlier sections.

3.4.3.4 Transcription-factor modulation

Sections 3.4.3.1 to 3.4.3.3 provide circumstantial validation for our detector. In this section, we provide direct validation. The enrichment test is based on PTM-Switchboard [112], a curated dataset of known three-way interactions in yeast. Each of the 519 entries in PTM-Switchboard is a confident modulator-TF-target triple, representing a type of three-way interaction called post-translational modulation [20, 30, 112], in which the effect of a transcription factor on a target gene is modulated by a third gene. The detector is run on the yeast gene expression data, `cogrim`, which contain the gene expression levels for 6,026 yeast genes across 314 experiments [113], and have been used by the authors of PTM-Switchboard to test their modulator-detecting method on yeast [16].

We consider triples (g_1, g_2, g_3) in which g_2 is a transcription factor, g_3 is a potential target of g_2, and g_1 is a potential modulator of g_2. All possible targets and modulators are considered. That is, the set of all triples is

$$S_A \;=\; \{(g_1, g_2, g_3) : g_1 \in S_{mod}, g_2 \in S_{TF}, g_3 \in S_T, g_1 \neq g_2, g_2 \neq g_3, g_1 \neq g_3\},$$

[17] That is, the transcription factors produced by g_1 and g_2 physically interact with the promoter region of g_3.

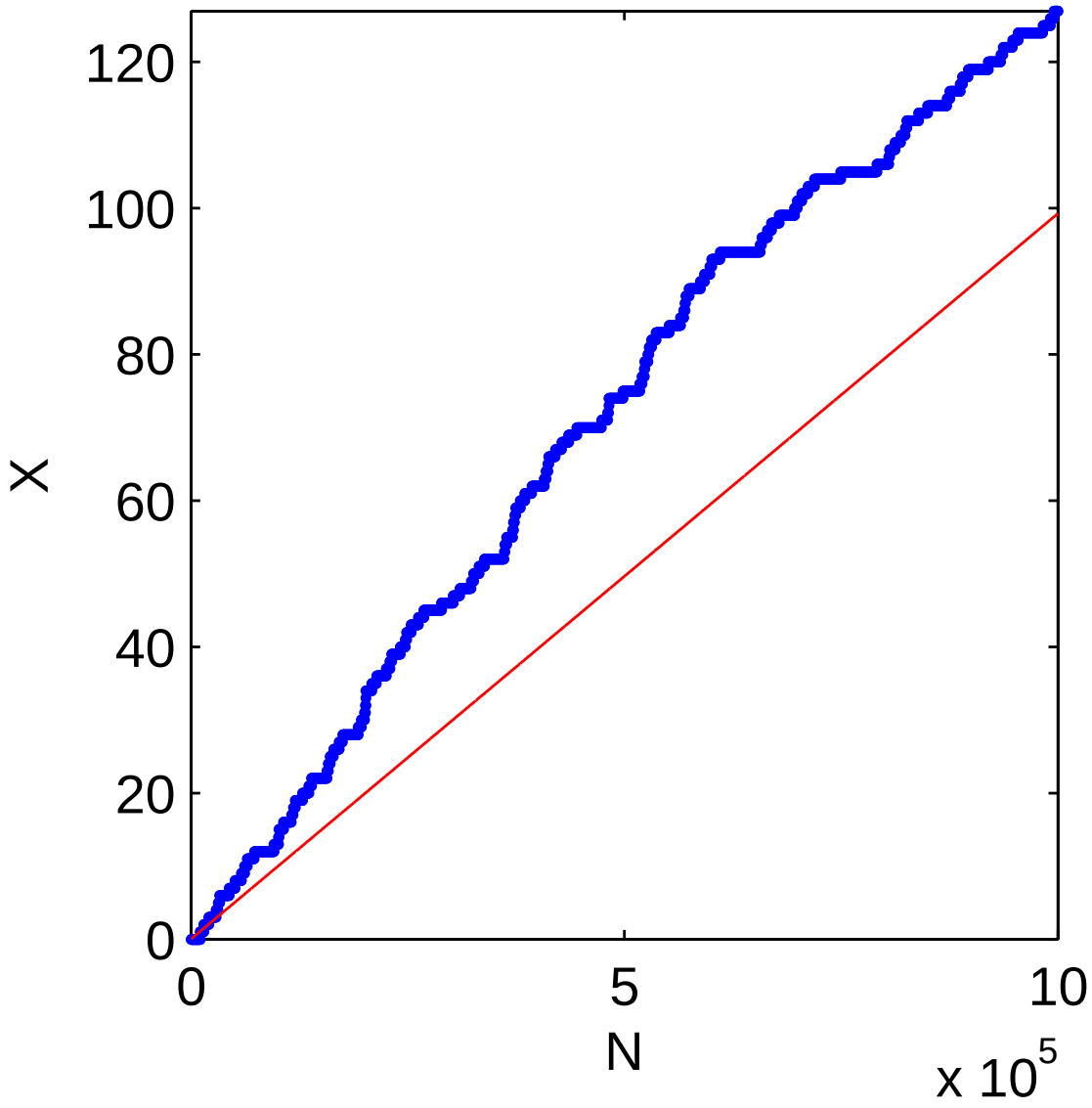


Figure 3.6: TF-target enrichment curve. A triple (g_1, g_2, g_3) is interesting if both (g_1, g_3) and (g_2, g_3) are in `agris`. P-values at 1%, 10% and 50% of all triples are 0.0025, 0.000196 and 0.0002, respectively.

where $S_{mod} = S_T$ is the set of all yeast genes, and S_{TF} is the set of Yeast transcription factors. S_A contains 11 billion triples. The set of interesting triples, S_I, consists of the 519 triples from PTM-Switchboard.

Figure 3.7 shows that the detected triples are enriched with the three-way interactions. Because the signal is weak (only 519 positives out of 11 billion triples), we have plotted the curve for all possible values of N. (The blue curve must therefore meet the red line at the top right of the figure, since at this point, all possible triples are deemed to be "detections".)

Recall that our three-way interaction detector computes a z value for each triple, reflecting the detector's confidence in the interaction. As described in Section 3.3.1, most of this chapter is focussed on triples with large values of $|z|$, indicating a strong three-way interaction. For the PTM-Switchboard triples, however, the z values have a significant positive bias [see Figure 3.8].[18] This may be a result of biologists focussing on testing for positive three-way interactions. In such cases, it can be more appropriate to look for triples with large, positive values of z, instead of $|z|$ (which also improves the signal-to-noise ratio). The blue curve in Figure 3.7 is based on a detector that looks for large values of z.

It is worth noting that the p-values will only decrease if new 3-way interactions are experimentally discovered and added to our dataset (i.e., if K increases in formula 3.2). In other words, the enrichment will become even more significant, and the blue curve will rise even further above the red line in Figure 3.7. A similar statement can be made for all of the enrichment tests in Section 3.4.3.

3.4.4 False Discovery Rate

The results of Section 3.4.3 show that the triples we detect are enriched with real three-way interactions. However, it is impossible for results of this kind to say exactly how many three-way interactions have been detected. For example, Section 3.4.3.4 tells us only how many PTM-Switchboard triples have been detected. However, these are just a tiny fraction of all the three-way interactions (direct and indirect) in yeast. How many of these unknown interactions have we detected? What is needed is a good estimate of the false discovery rate of our detector.

[18] In contrast, for all other data sources in this chapter, the z values have no such bias and their histograms are symmetric about 0 [data not shown].

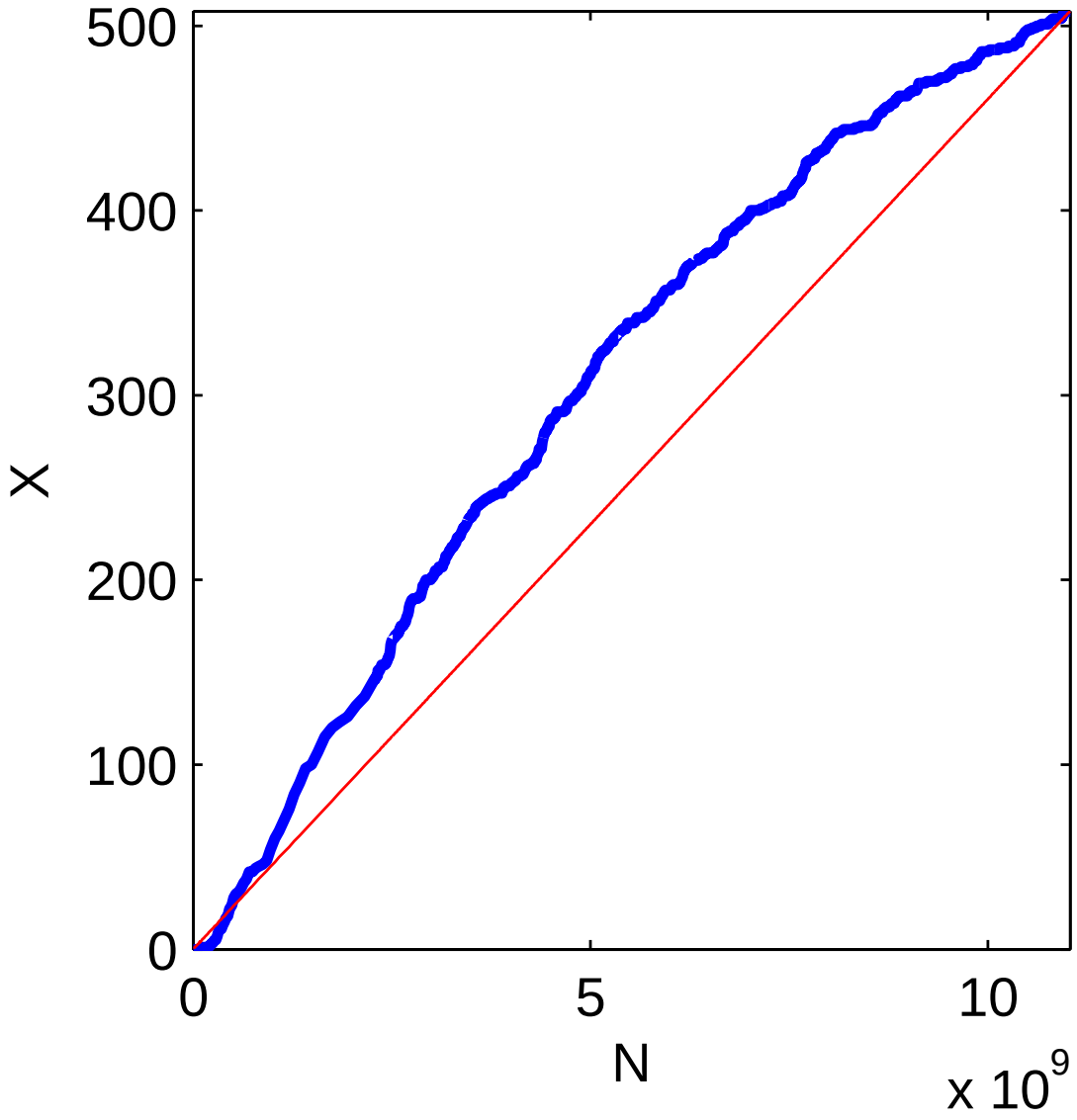


Figure 3.7: 3-way-interaction enrichment curve. A triple (g_1, g_2, g_3) is interesting if it is in PTM-Switchboard. P-values at 10%, 20% and 50% of all triples are 0.0086, 1.06E-05 and 1.89E-14, respectively.

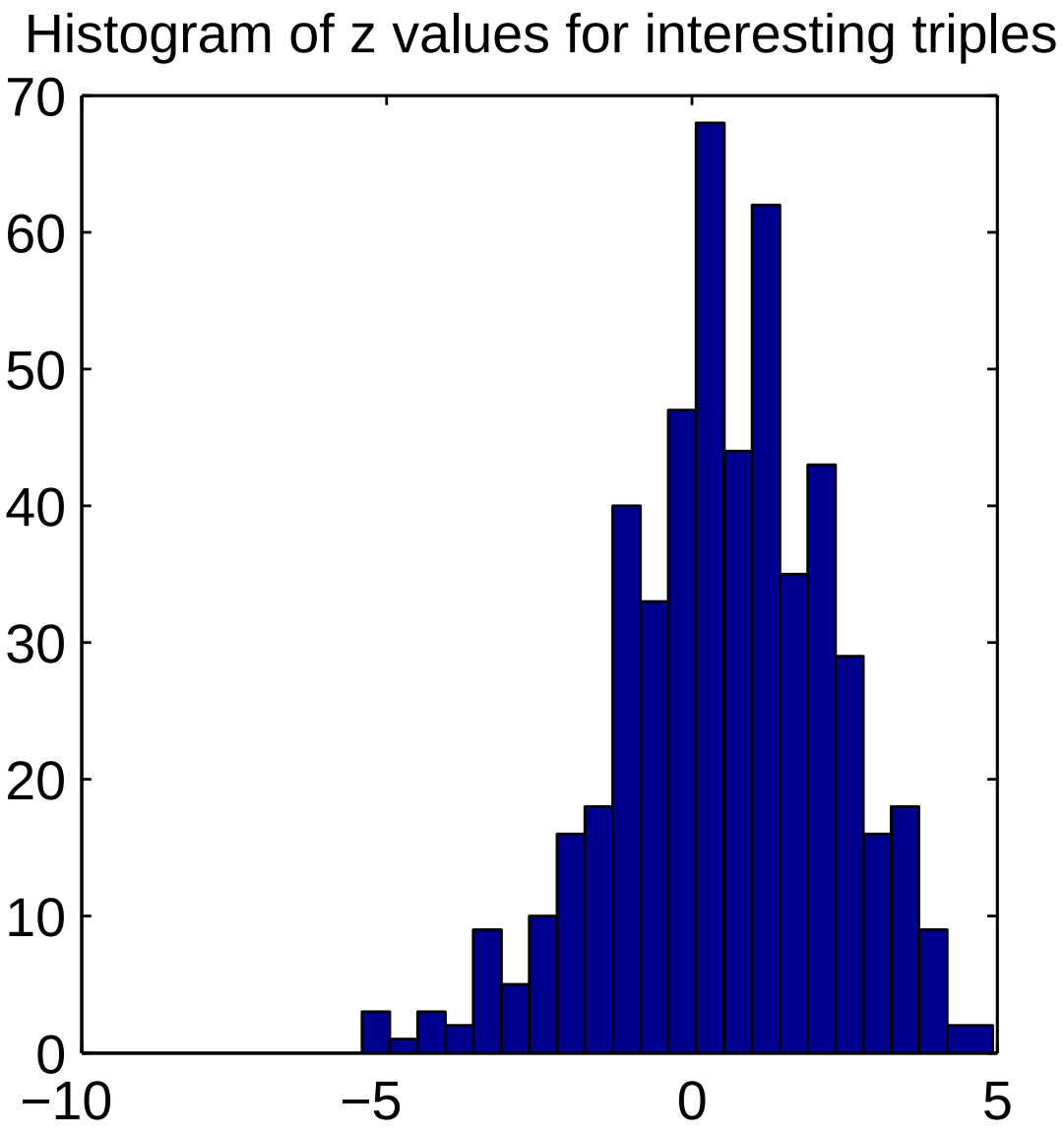


Figure 3.8: Histogram of z for the interesting triples, i.e., the triples in PTM-Switchboard.

This is the subject of the rest of this chapter.

The FDR is defined as "the expected proportion of errors among the rejected hypotheses" [22]. In our case, this is the expected proportion of discoveries that are false, where a "discovery" is a triple of genes predicted to have a three-way interaction. Recall that each gene triple is assigned a z value, reflecting the detector's confidence that there is a three-way interaction between the genes (Section 3.3.1). Formally, given a threshold, τ, we say that a triple is a discovery if and only if $|z| >= \tau$. If the triple does indeed represent a three-way interaction, then we say the discovery is *true*; otherwise, we say it is *false*. If we have R discoveries, V of which are false, then the FDR is formally and simply defined as $E(V/R)$, the expected value of V/R, where the value of V in most cases is unknown and needs to be estimated.

3.4.4.1 FDD curves

As the threshold, τ, varies, the number of discoveries (R) and the estimated number of false discoveries (V) vary, tracing out a curve. We call this curve an FDD curve. To generate such a curve, we simply need to choose a large set of thresholds and plot a point (x, y) for each one, where x is the number of discoveries at threshold τ, and y is the estimated number of false discoveries. It is convenient to use the values of $|z|$ for all the triples as the set of thresholds. Details are given in the following procedure, genFDD (Figure 3.9).

We plot FDD curves instead FDR curves for ease of comparison and interpretation. In contrast to FDR curves, FDD curves are smooth and monotonic. Figure 3.10, for example, shows four FDD curves corresponding to four approaches to estimating false discoveries, the bootstrap (green), partial permutation (pink), total permutation (blue) and analytical t (black).

3.4.4.2 Which estimate to use?

Figures 3.10 and 3.11 show a number of such curves for our *Arabidopsis* data and the yeast data (`cogrim`), respectively. Each curve corresponds to a different method of estimating the number of false discoveries. We have discussed these methods in Section 3.3.2. The main thing to note however is that they give widely varying estimates of the number of false discoveries. First, we see a clear separation of the green curve from the other three curves. In Figure 3.10, the

PROCEDURE genFDD($\mathbf{z}, \mathbf{z}_{null}$)

1. The input vector $\mathbf{z} = (z_1, z_2, ..., z_n)$ is a list of values of z for gene triples, where z_i is the value of z for the ith triple.
2. The input $\mathbf{z}_{null} = (\hat{z}_1, \hat{z}_2, ..., \hat{z}_n)$ is a list of null values of z for gene triples, where $\hat{z}_i$ is estimated by an FD estimation method (Section 3.3.2).
3. Estimate the number of true negatives (Section 3.3.2.3).
4. Let $\mathbf{z} \leftarrow |\mathbf{z}| = (|z_1|, |z_2|, ..., |z_n|)$, and let $\mathbf{z}_{null} \leftarrow |\mathbf{z}_{null}| = (|\hat{z}_1|, |\hat{z}_2|, ..., |\hat{z}_n|)$.
5. Let $\mathbf{z}_{sorted}$ be the set $\mathbf{z}$ sorted in descending order.
6. For each element, t, in $\mathbf{z}_{sorted}$,
   (a) Let x be its position in the list. (For example, if t is the first element in $\mathbf{z}_{sorted}$, then $x = 1$.) This is the number of discoveries at threshold t.
   (b) Let $y = \#\{\mathbf{z}_{null} > t\}$, the number of elements in $\mathbf{z}_{null}$ greater than t. This is an initial estimate of the number of false discoveries at threshold t.
   (c) Reduce y to account for the number of true negatives (Section 3.3.2.3). This is the final estimate of the number of false discoveries.
   (d) Plot a dot at position (x, y).
7. Connecting all dots above traces an FDD curve.

Figure 3.9: Generate an FDD curve.

total permutation and analytical t approaches give similar and the most optimistic estimates of false discoveries. In contrast, the bootstrap approach gives much higher estimates: about 3 times as many as that given by the partial permutation approach, and 9 times as many as that given by the total permutation and analytical t approaches. The partial permutation approach gives an estimate in between: the pink curve is clearly above the blue and black ones, but is well below the green curve. A similar pattern of separation can be seen in Figure 3.11. The practical question is: which curve (if any) is accurate? In Chapter 4, we provide evidence strongly suggesting that the bootstrap estimate is the most accurate. We also show that the permutation and analytical t approaches can grossly underestimate the true number of false discoveries, sometimes by several orders of magnitude.

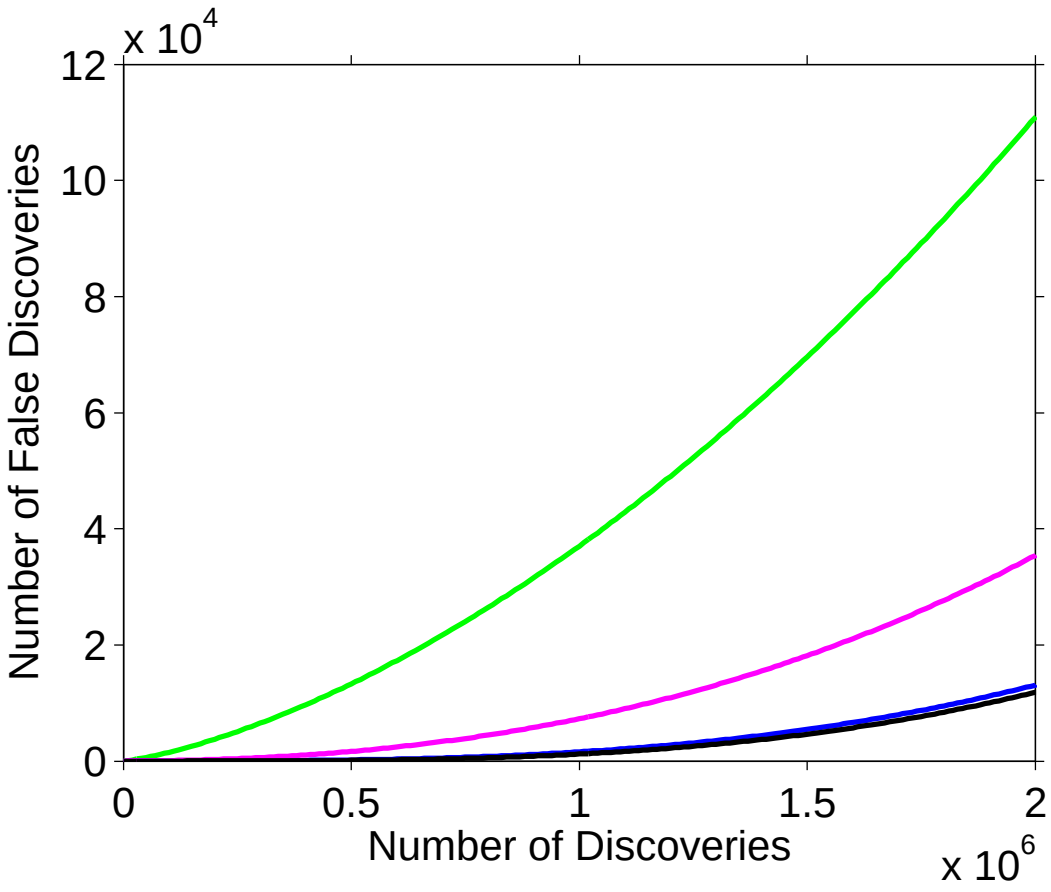


Figure 3.10: Number of False Discoveries versus number of Discoveries for the FD estimation methods on the seed germination/dormancy data from *Arabidopsis thaliana*. The green curve is estimated by the bootstrap, the blue by total permutation, the pink by partial permutation, and the black by analytical t.

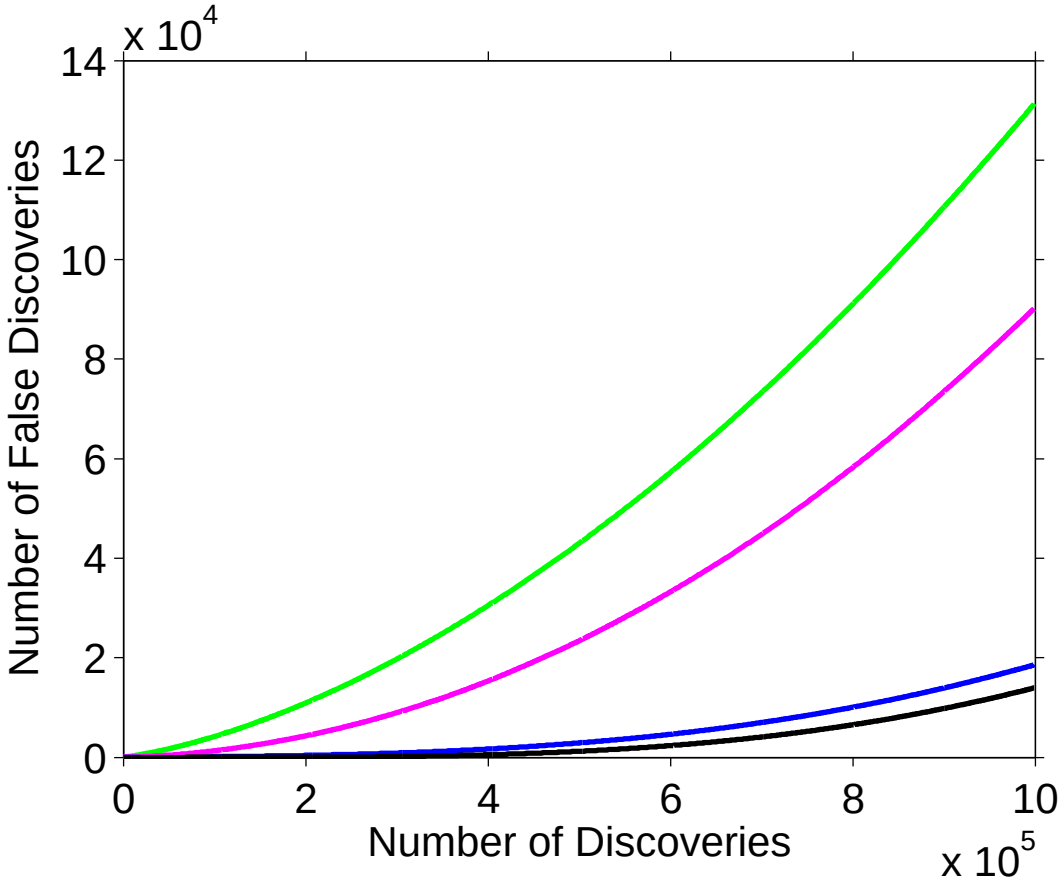


Figure 3.11: Number of False Discoveries versus number of Discoveries for the FD estimation methods on the yeast cogrim data. The green curve is estimated by the bootstrap, the blue by total permutation, the pink by partial permutation, and the black by analytical t.

3.4.4.3 Variance of FDD curves

In this section we look at the variance of the bootstrap FDD curve, and note that it is reasonably small. To do this, we repeat the bootstrap method and curve generation procedure ten times. The variance is illustrated in Figure 3.12. Each green curve represents one bootstrap estimate of false discoveries with $B = 1000$. Visually, the bootstrap FDD curves tend to cluster together, indicating a small variance. This is because the curves are based on a large number of discoveries. In general, the *coefficient of variation* [114] for the estimate of false discoveries decreases as number of discoveries increases. For instance, the coefficients of variation for the top 100, 1,000, 1,000,000 and 2,000,000 discoveries are 12.4, 8.5, 2.8 and 2.4, respectively. The coefficient of variation, defined as 100 times the ratio of standard deviation σ to mean μ (i.e., $100 \cdot \frac{\sigma}{\mu}$), measures the dispersion of the distribution of a random variable. Here, our random variable is the bootstrap estimate of false discoveries. Fewer bootstrap samples (smaller B) are needed to give a stable estimate of false discoveries when the number of discoveries is large, and more bootstrap samples (larger B) are needed when the number of discoveries is small (i.e., when we only consider the most confident discoveries). The total computation time increases linearly with B but this increase results in decreased variance of the estimate of false discoveries, a worthwhile trade-off between time and accuracy. Also, computation on each bootstrap sample is independent, allowing us to parallelize it.

3.4.5 Correlated predictors

As before, let X_1, X_2 and Y be the gene expression profiles of genes g_1, g_2 and g_3, where g_1 and g_2 are predictor genes for target gene g_3. In general, the more correlated X_1 and X_2 become, the more difficult it is to distinguish the effects of X_1 and X_2 on Y. It therefore becomes more difficult to detect three-way interactions, and the false discovery rate rises. This is true for any method of detecting three-way interactions from gene expression profiles.

This section illustrates the difficulty by applying our second-order detector to simulated data (described in Section 4.1.2). Because the data is simulated, we can control the correlations. Moreover, we know which discoveries are true and which are false, so we can plot the true FDR.

Figure 3.13 shows true FDD curves for four different values of correlation, ρ, between X_1

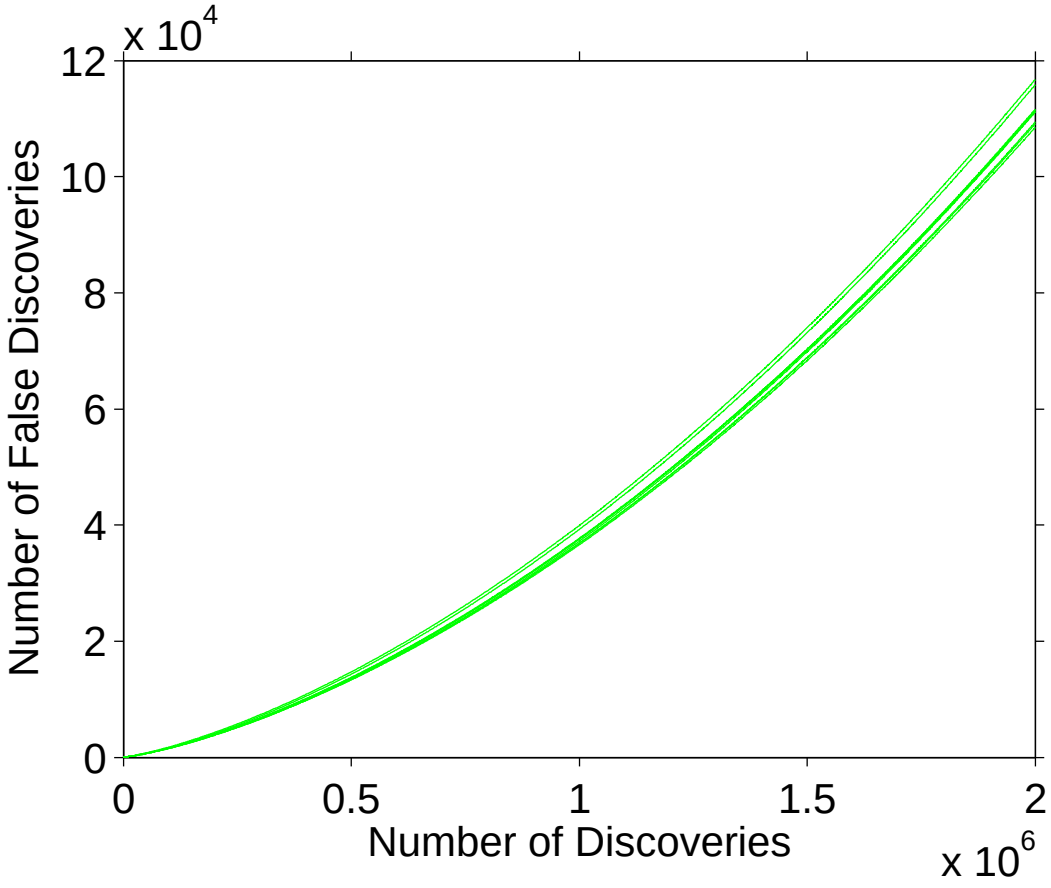


Figure 3.12: Variance of the bootstrap FDD curve on the germination/dormancy data. The ten green curves are generated by repeating the bootstrap method ten times, each with a different randomization. The FDD curves for top two million discoveries are shown.

and X_2. The high curves correspond to higher correlations, showing that FD increases with correlation. Note that these curves show true FD, not estimated FD, so they are not an anomaly of any particular FD estimation method.

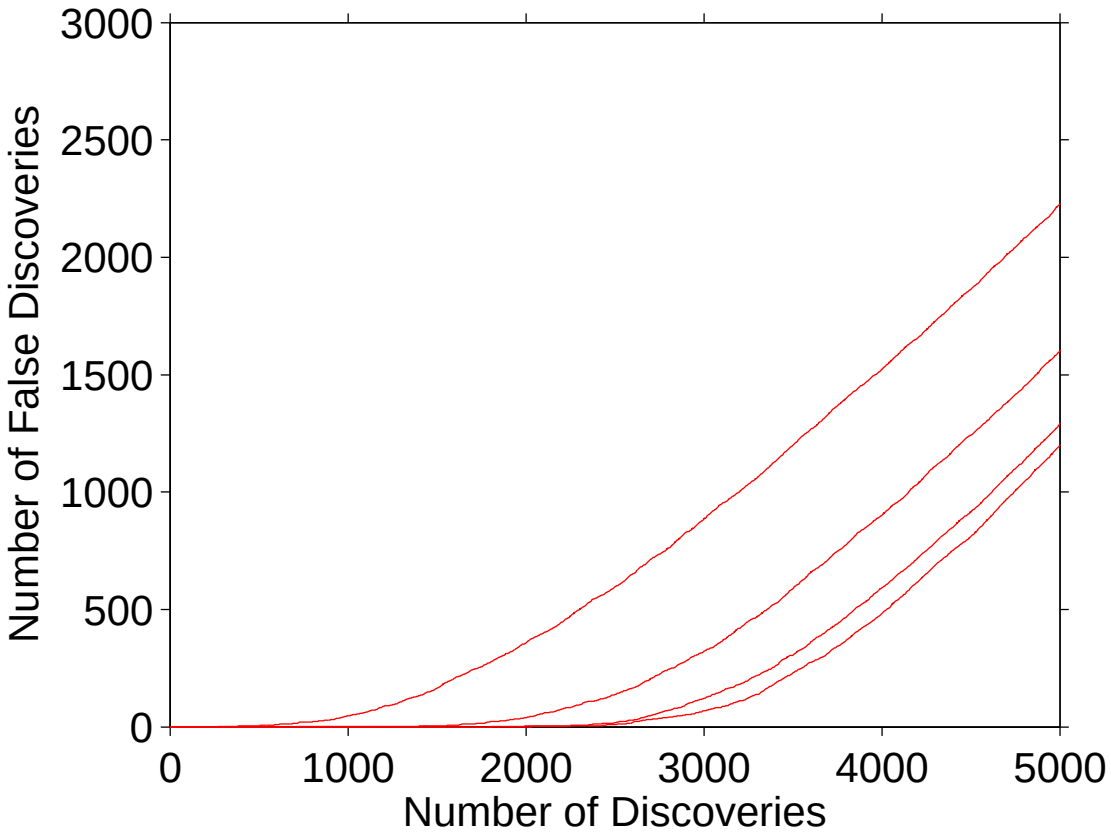


Figure 3.13: True FDD curves under different correlations, ρ, between two predictor genes. $\rho = 0, 0.3, 0.6$ and 0.9. Higher curves correspond to larger values of ρ.

To gain more insight into this phenomenon, Figures 3.14 to 3.16 show histograms of z for interacting and non-interacing triples for $\rho = 0, 0.5$ and $\rho = 0.99$, respectively. When the predictor genes are uncorrelated, the z values for interacting triples are distributed in a totally different manner from the z values for non-interacting triples (Figure 3.14). In particular, the majority of the z values for the interacting triples are larger (in magnitude) than the z values for the non-interacting triples. This results in low p-values and low FDR. In contrast, when the predictor genes are highly correlated, the predictor variables in the second-order model also become highly correlated, a phenomenon called multicollinearity [39]. In this situation, the

second-order model can barely distinguish the non-interacting triples from interacting triples. With multicollinearity, the estimated regression coefficient of the interaction term, $\hat{\beta}_5$, becomes highly inaccurate. This can be seen in Figure 3.16, which shows that the distributions of z for interacting and non-interacting triples are very similar. Recall that the distribution of z is the null distribution when testing for three-way interactions. Thus, Figure 3.14 (no correlation) results in low p-values and low FDR, while Figure 3.16 (high correlation) results in high p-values and high FDR.

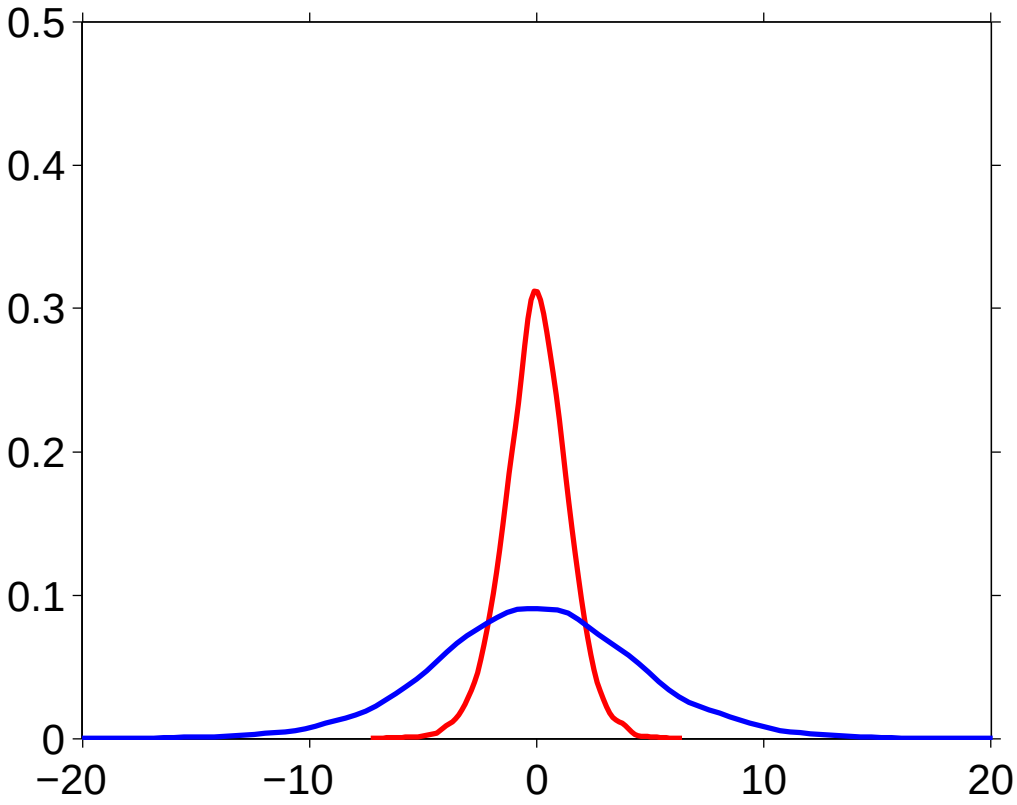


Figure 3.14: Two distributions of z values of the interaction term in the second-order model, when the predictor genes are uncorrelated ($\rho = 0$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples.

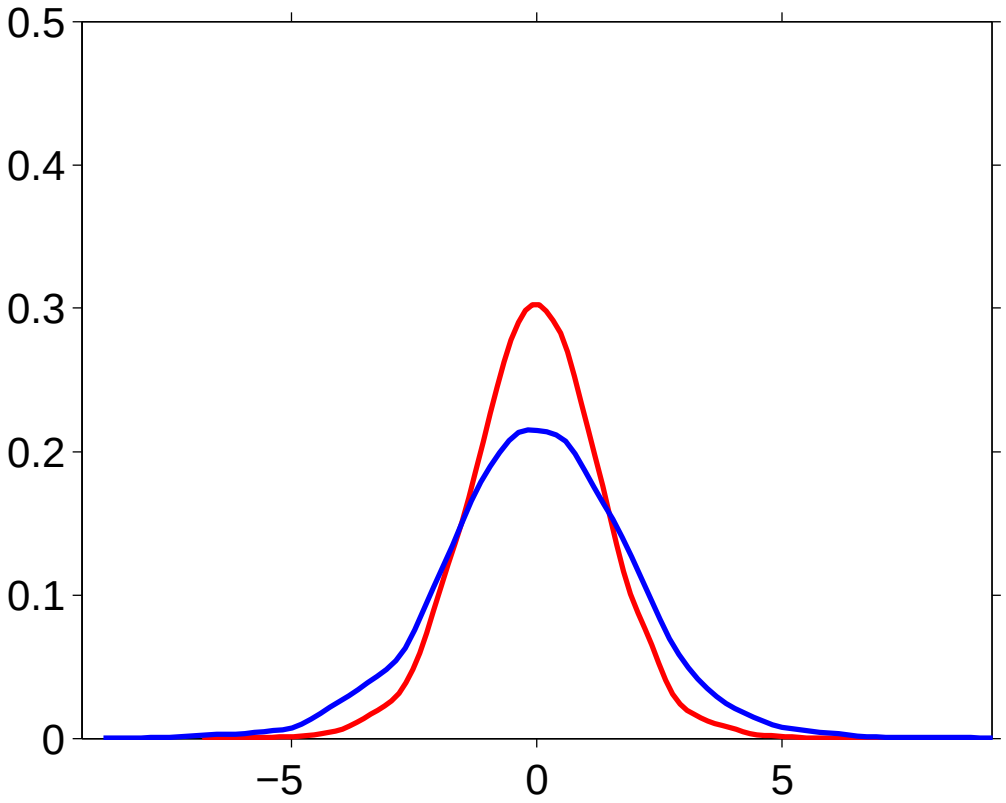


Figure 3.15: Two distributions of z values of the interaction term in the second-order model, when the predictor genes are moderately correlated ($\rho = 0.8$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples.

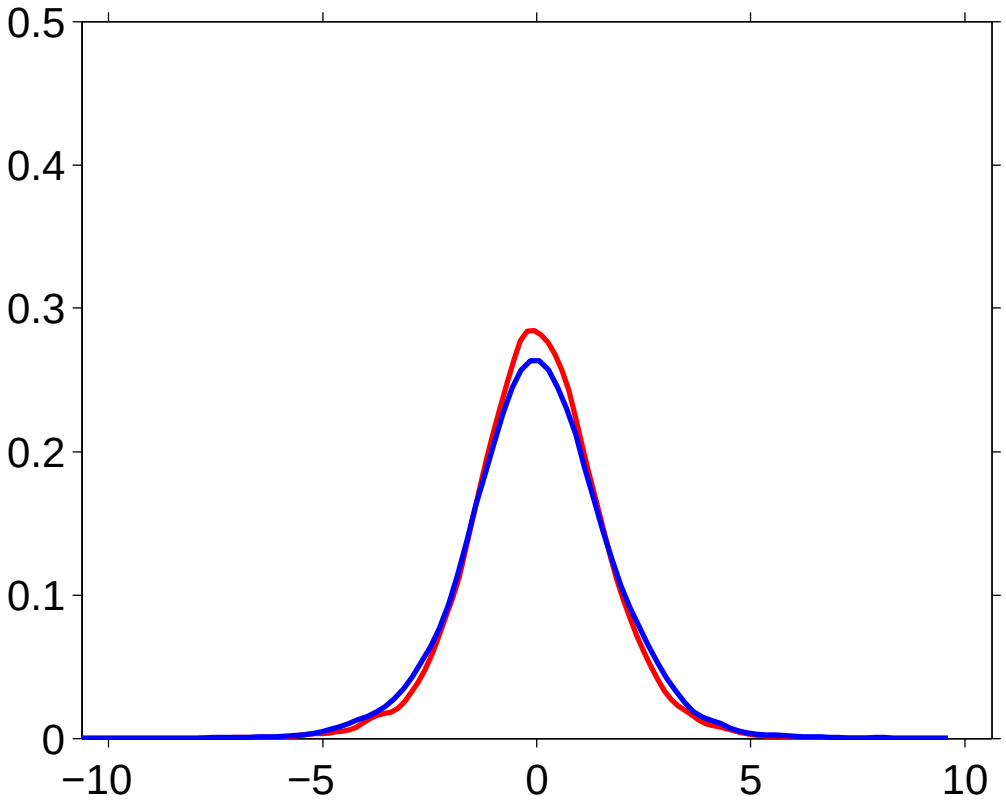


Figure 3.16: Two distributions of z values of the interaction term in the second-order model, when the predictor genes are highly correlated ($\rho = 0.99$). The red curve is the histogram of 10,000 z values obtained from regressing 10,000 non-interacting triples. The blue curve is the histogram of 10,000 z values obtained from regressing 10,000 interacting triples.

Chapter 4

Validating FDR estimates

To validate FDR estimates, we need to know *all* the three-way interactions in a data set. Unfortunately, large-scale, well-curated sets of interacting and non-interacting gene triples are unavailable. Instead, we test our method on simulated data, for which all interactions are known. In this way, we can compare the FDR estimates of the bootstrap approach to those of other approaches under a wide range of statistical conditions. We show, for example, that all approaches produce accurate FDR estimates under ideal conditions. As the data becomes more complex and realistic (e.g., non-Gaussian, dependent noise samples, correlated predictors, non-linear dependencies, multi-modal, etc), the bootstrap approach continues to give reasonable FDR estimates, while the other approaches rapidly break down. Typically, the other approaches give estimates that underestimate the true FDR by a considerable amount, often by several orders of magnitude. For example, an FDR estimate near zero may be given, when the true FDR is in fact quite large (say 50% or 90%), rendering the estimate useless. When the true FDR is near zero, the bootstrap method may sometimes give an overestimate, but it is still small, and therefore useful.

4.1 Results on simulated data

As in Section 3.3.1, let X_1 and X_2 be the expression levels of two predictor genes, g_1 and g_2, and let Y be the expression level of a target gene, g_3. Our general framework for data simulation is as follows. For a non-interacting triple, Y is modeled as an additive function of X_1, X_2 and

noise, that is, $Y = f(X_1) + g(X_2) + \varepsilon$, where $f(X_1)$ and $g(X_2)$ describe the main effects of X_1 and X_2 on Y, respectively, and ε is random noise. For an interacting triple, we add an interaction term, $h(X_1, X_2)$, to the sum, so that $Y = f(X_1) + g(X_2) + h(X_1, X_2) + \varepsilon$. By using different functions, f, g and h, we simulate a range of data models, from simple to complex. In section 4.1.6, we consider more-complex, non-additive models. The simulations are intended to be extensive rather than exhaustive. To get stable results, we use 500 bootstrap samples for the bootstrap method, and 100 permutations for the permutation methods. Note that, although random, the noise need not be Gaussian, need not be *i.i.d.*, need not be independent of X_1 and X_2, need not have constant variance or zero bias, etc. Our simulations include examples of all these cases. Likewise for the distributions of X_1 and X_2.

Our procedures for generating simulated data are given in Figures 4.1, 4.6, 4.10, 4.12, 4.14, 4.17 and 4.19. The procedures generate data that is more complex than the quadratic model used in our regression-based detector (Section 3.3.1), as is almost certainly the case for biological data. However, although some of the procedures are biologically inspired, they do not attempt to be biologically realistic. Instead, the goal is to test the various FDR estimation methods over a wide range of well-defined statistical conditions.

Each procedure creates a set of gene names, creates gene triples from these names, and then generates expression data for each gene. In Figure 4.12, the procedure assigns each gene to many different triples. Thus the expression data of different triples can be mutually dependent. In all other figures, the procedures assign each gene to exactly one triple, so the data from different triples is mutually independent.

Each simulation procedure generates M_1 non-interacting triples and M_2 interacting triples. For all triples, the parameter N specifies the length of each gene expression profile, and nl_1 controls the amount of additive noise, ε, in the model above. We call this "biological noise", since it represents biological influences on the target gene that we do not model. The correlation between the two predictor genes, X_1 and X_2, is usually determined by the parameter ρ, but sometimes by other parameters, depending on the procedure. At this point, the values of X_1, X_2 and Y represent expression levels in the cell. As a final step, we add (or multiply) noise to each of them, to represent measurement error. The parameter nl_2 controls the amount of this measurement noise. All simulated data in this section use $M_1 = 15,000$, $M_2 = 5,000$, and

$N = 100$. The values of other parameters are shown in individual figures. In general, parameter values were chosen with several criteria in mind: (i) to be roughly similar in all simulations (so as not to be arbitrary), (ii) to give true FDR values that are not almost 0 or almost 1, (iii) to give true FDR values that are similar in all plots, and (iv) to produce FD curves that are similar in form to those of the real data in Figures 3.10 and 3.11.

4.1.1 Quadratic data

First, we look at a simple case of interacting and non-interacting gene triples, where two predictor genes have a quadratic relationship with a target gene. We consider this case first because it is the model our detector uses, so we expect the performance of the detector to be good and the estimates of FDR to be accurate. For an interacting triple, the data is generated using the full second-order model (Section 3.3.1), i.e., $Y = a + bX_1 + cX_2 + dX_1^2 + eX_2^2 + fX_1X_2 + \varepsilon$. The coefficients a, b, ..., f are randomly chosen and are different for each gene triple. For a non-interacting triple, $f = 0$, thus removing the interaction term. The noise, ε, is *i.i.d.* Gaussian.[1] Simulation details are given in Figure 4.1. We call this simulated data `DataQuad`.

To avoid dominance by any one variable, when generating Y, each predictor (i.e., X_1, X_2, X_1^2, X_2^2 and X_1X_2) is normalized to have variance 1. Formally, $Y = a + \hat{b}X_1 + \hat{c}X_2 + \hat{d}X_1^2 + \hat{e}X_2^2 + \hat{f}X_1X_2 + \varepsilon$, where a hat above a coefficient means that it has been scaled to normalize the predictor. For example, $\hat{f} = f/sd(X_1X_2)$, where sd denotes standard deviation (see steps 1(d) and 2(b) in Figure 4.1). To model systematic measurement errors, measurement noise is incorporated in each simulated gene expression level (see step 3). ρ is the correlation between X_1 and X_2.

Ideal data. The ideal data is generated using PROCEDURE genDataQuad in Figure 4.1 with $nl_2 = 0$, i.e., no measurement noise is added. It is ideal because it is exactly the model used by our detector, so we expect all methods of estimating FD to work well. Figure 4.2 shows the FDD curves of the FD estimation methods described in Section 3.3.2. The red curve is the true FDD curve. Each of the other colors corresponds to one estimation method: green - bootstrap, blue - total permutation, pink - partial permutation, black - analytical t. Each

[1] In this and other simulations throughout this thesis, non-Gaussian noise was also tried (e.g., exponential and beta-distributed noise) but had very little effect on the results.

PROCEDURE genDataQuad($M_1, M_2, N, \rho, nl_1, nl_2$)

1. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Let $\mu = [\mu_1, \mu_2]$, where μ_1 and μ_2 are randomly chosen from a uniform distribution on [-0.5,0.5]. μ_1 and μ_2 represent the mean expression levels of the two predictor genes, g_1 and g_2, respectively.

   (b) Let r be a random number from a uniform distribution on [0,1]. r represents the variance in the expression levels of g_1 and g_2. Let $\Sigma = r \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

   (c) Generate N bivariate Gaussian data points $x_1, ..., x_N$, where $x_i = (x_{i1}, x_{i2}) \sim \mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times 2$ matrix. The first column $(x_{11}, x_{21}, ..., x_{N1})^T$ is the expression profile for gene g_1. The second column $(x_{12}, x_{22}, ..., x_{N2})^T$ is the expression profile for gene g_2.

   (d) Let a, b, c, d and e be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let

   $$y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \varepsilon_i, \quad i = 1, ..., N,$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates that it has been scaled to normalize its predictor (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting triples (true positives), as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Repeat steps 1(a) to 1(c).

   (b) Let a, b, c d, e and f be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let

   $$y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}x_{i2} + \varepsilon_i, \quad i = 1, ..., N,$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.1: Generate data in which two predictor genes have a quadratic relationship with a target gene. Note that the coefficients of the quadratic (a through f) can be arbitrarily close to 0, thus allowing for arbitrarily-weak interactions.

estimation method gives an estimated FDD curve that is close to the true FDD, consistent with our expectation. Figure 4.2 also provides a sanity check. If our implementations of the various FDR estimation methods are correct, then all the curves should agree on ideal data, which they do. In addition, if we were over- or under-estimating the true FDR, we would not expect the true FDR curve to agree with all the FDR estimates, especially the analytical estimate, which can be regarded as the best estimate in this case.

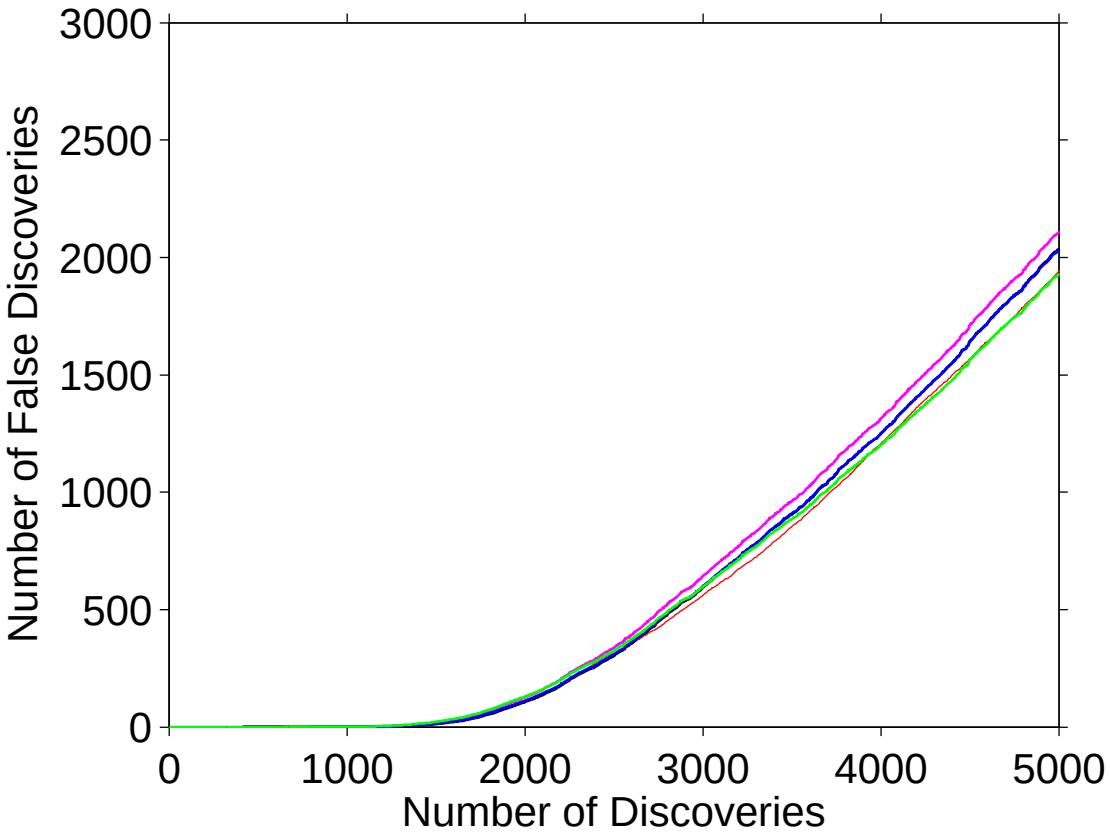


Figure 4.2: FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0$, $nl_1 = 2$, $nl_2 = 0$ (no measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

Additive measurement noise. On top of this ideal data, we add a small amount of Gaussian measurement noise (ε) to each simulated expression level (see step 3 in PROCEDURE genDataQuad). The FDD curves of the FD estimation methods begin to separate, with the bootstrap curve remaining closest to the true curve (Figure 4.3).

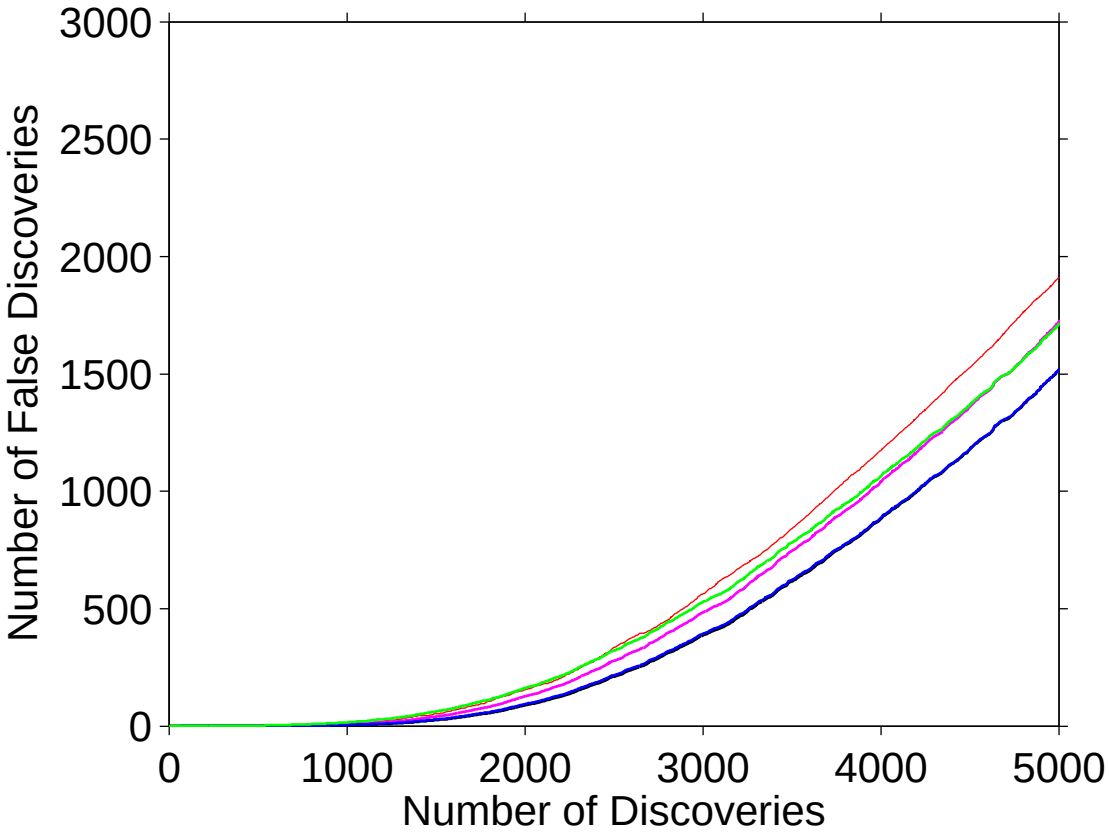


Figure 4.3: FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0$, $nl_1 = 0.5$, $nl_2 = 0.5$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

Multiplicative measurement noise. Another kind of measurement noise is multiplicative, which scales each simulated expression level by a random amount. Specifically, each expression level is multiplied by log-normal noise, that is, by noise whose logarithm has a normal distribution [115] (see step 3 in PROCEDURE genDataQuad). This multiplicative measurement noise causes much greater separation of the FDD curves than does the additive noise (Figures 4.4 and 4.5). The bootstrap method still gives good FD estimates, but the other methods underestimate the true FD to various extents, and sometimes by an order of magnitude (Figure 4.5).

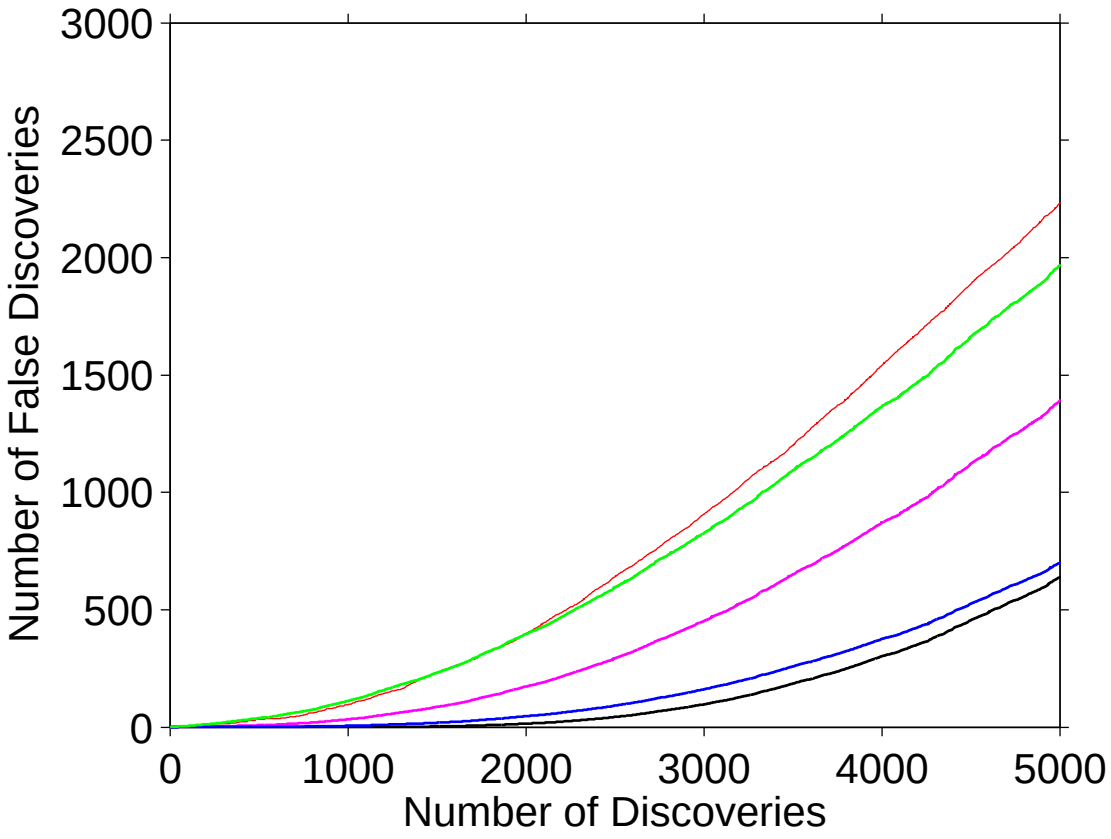


Figure 4.4: FDD curves of the FD estimation methods on `DataQuad`: $\rho = 0, nl_1 = 0.5, nl_2 = 0.5$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.1.2 Cubic data

The purpose of this section and succeeding sections is to study the performance of our detector and of the methods of estimating FDR when the data model is different from the model used in

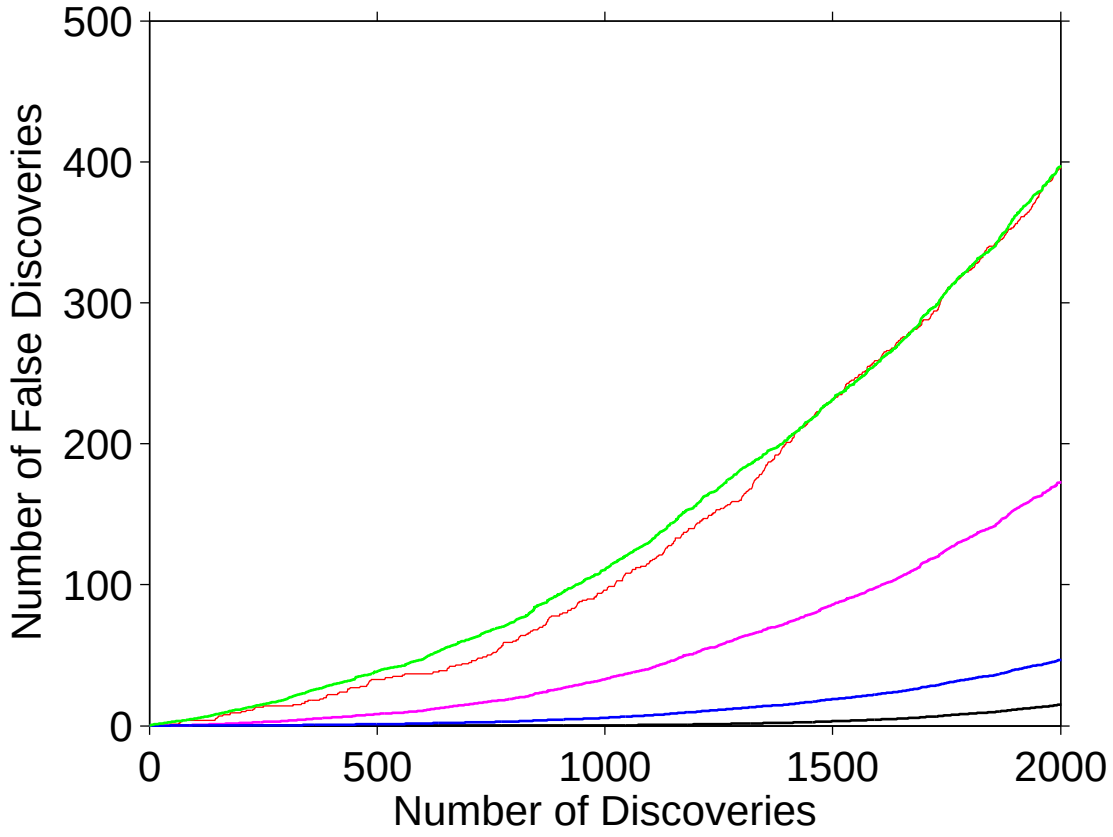


Figure 4.5: Same as Figure 4.4 but showing only the top 2,000 discoveries. Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

the detector (as is the case for real data). The previous section began the study by looking at multiplicative noise (the detector assumes additive noise). This section continues the study by using data based on a cubic data model. Although cubic data is a simple and natural extension of the quadratic model used in the detector, we shall see that the various FDR estimates already diverge substantially, even with additive measurement noise. For an interacting triple, $Y = a + bX_1 + cX_2 + dX_1^2 + eX_2^2 + fX_1^3 + gX_2^3 + hX_1X_2 + iX_1X_2^2 + jX_2X_1^2 + \varepsilon$, where hX_1X_2, $iX_1X_2^2$ and $jX_2X_1^2$ are interaction terms. As with quadratic data, the coefficients a, b, c, ..., j are randomly chosen and are different for each gene triple. Also, to avoid dominance by any one variable, each predictor is normalized to have variance 1. The model for non-interacting triples is the same except that $h = i = j = 0$. Simulation details are given in Figure 4.6, and we call the simulated data `DataCubic`.

The effect of correlation. Figures 4.7 to 4.9 show FDD curves for cubic data with various amounts of correlation between the predictor variables ($\rho = 0, 0.3, 0.6$). Most noticeably, the pink curve, corresponding to the partial permutation estimate used in [24], is highly sensitive to correlation, moving from an underestimate at low correlation, to an overestimate at high correlation. In contrast, the green curve, corresponding to our bootstrap estimate, remains close to the true FDD curve regardless of correlation. In other words, the partial permutation estimate is much less stable and accurate than the bootstrap estimate. The curves for the other two methods, total permutation and analytical t, are well below the true FDD curve for all values of ρ. More details on the effect of correlated predictors can be found in Section 3.4.5.

4.1.3 Many predictor genes

In our simulations so far, a target gene has had exactly two predictor genes. We now consider the more-realistic case in which a target gene can have many predictor genes. Of course our detector and all the FD estimation methods are based on two predictor genes, and the goal is to see how well they perform on data generated by this more-complex model. We note that with many predictor genes, gene interactions can be much more complex and can include not just 3-way, but 4-way and other higher-order interactions.

In general, when there are P predictor genes, we call the combination of target gene and

PROCEDURE genDataCubic($M_1, M_2, N, \rho, nl_1, nl_2$)

1. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Let $\mu = [\mu_1, \mu_2]$, where μ_1 and μ_2 are randomly chosen from a uniform distribution on [-0.5,0.5]. μ_1 and μ_2 represent the mean expression levels of the two predictor genes, g_1 and g_2, respectively.

   (b) Let r be a random number from a uniform distribution on [0,1]. r represents the variance in the expression levels of g_1 and g_2. Let $\Sigma = r \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

   (c) Generate N bivariate Gaussian data points $x_1, ..., x_N$, where $x_i = (x_{i1}, x_{i2}) \sim \mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times 2$ matrix. The first column $(x_{11}, x_{21}, ..., x_{N1})^T$ is the expression profile for gene g_1. The second column $(x_{12}, x_{22}, ..., x_{N2})^T$ is the expression profile for gene g_2.

   (d) Let a, b, c, d, e, f and g be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let

   $$y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 + \varepsilon_i, \quad i = 1, ..., N,$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting triples (true positives), as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Repeat steps 1(a) to 1(c).

   (b) Let a, b, c d, e, f, g, h, i and j be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let

   $$\begin{aligned} y_i = a &+ \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 \\ &+ \hat{h}x_{i1}x_{i2} + \hat{i}x_{i1}x_{i2}^2 + \hat{j}x_{i2}x_{i1}^2 + \varepsilon_i, \quad i = 1, ..., N, \end{aligned}$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.6: Generate data in which two predictor genes have a cubic relationship with a target gene.

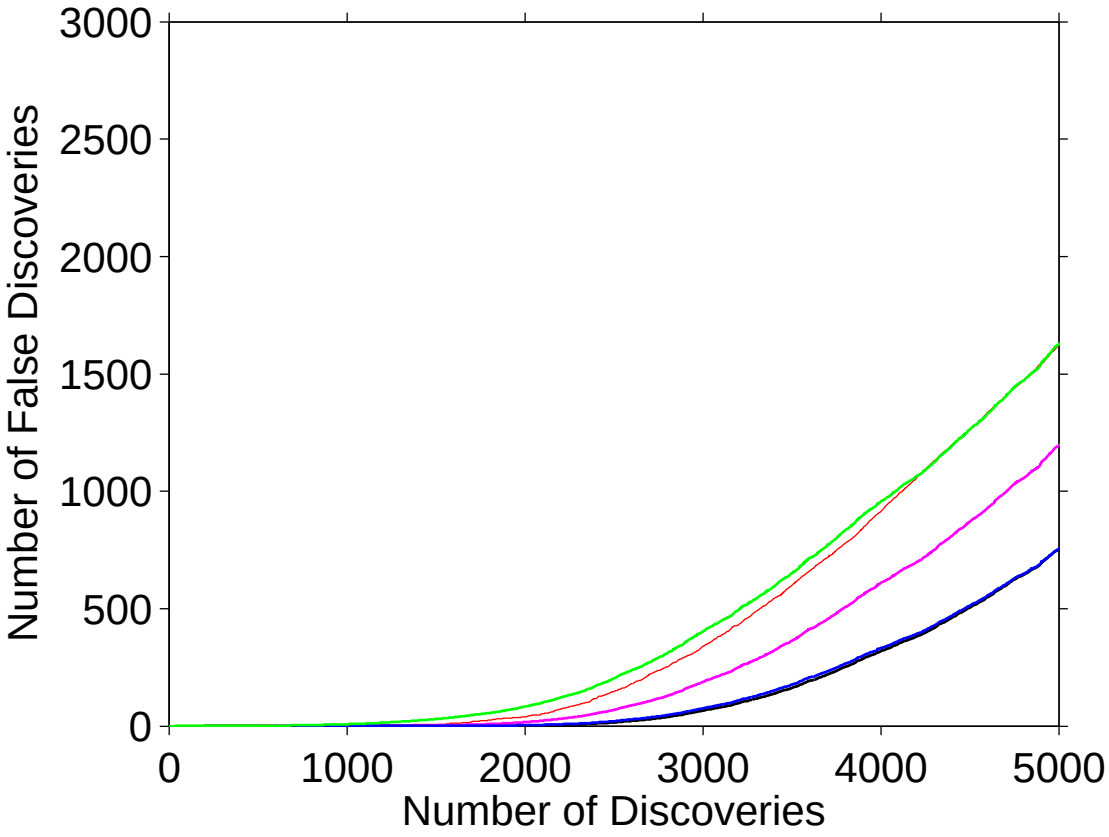


Figure 4.7: FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

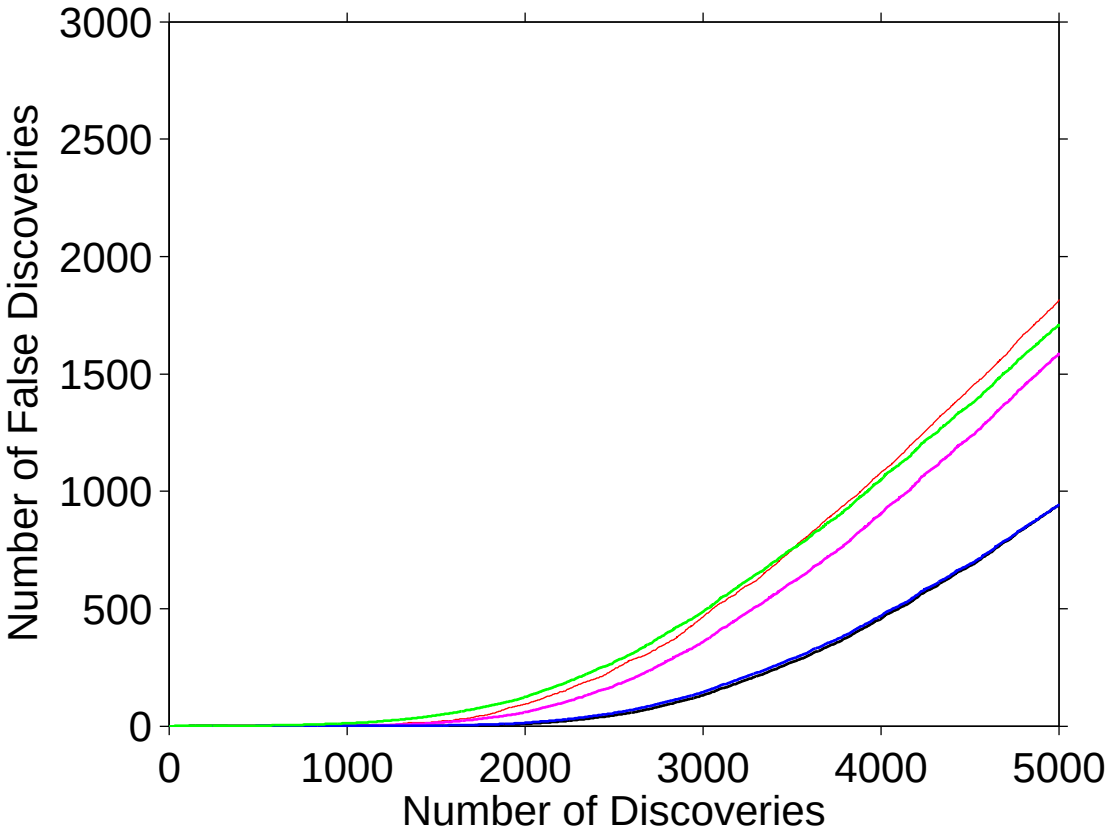


Figure 4.8: FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0.3$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

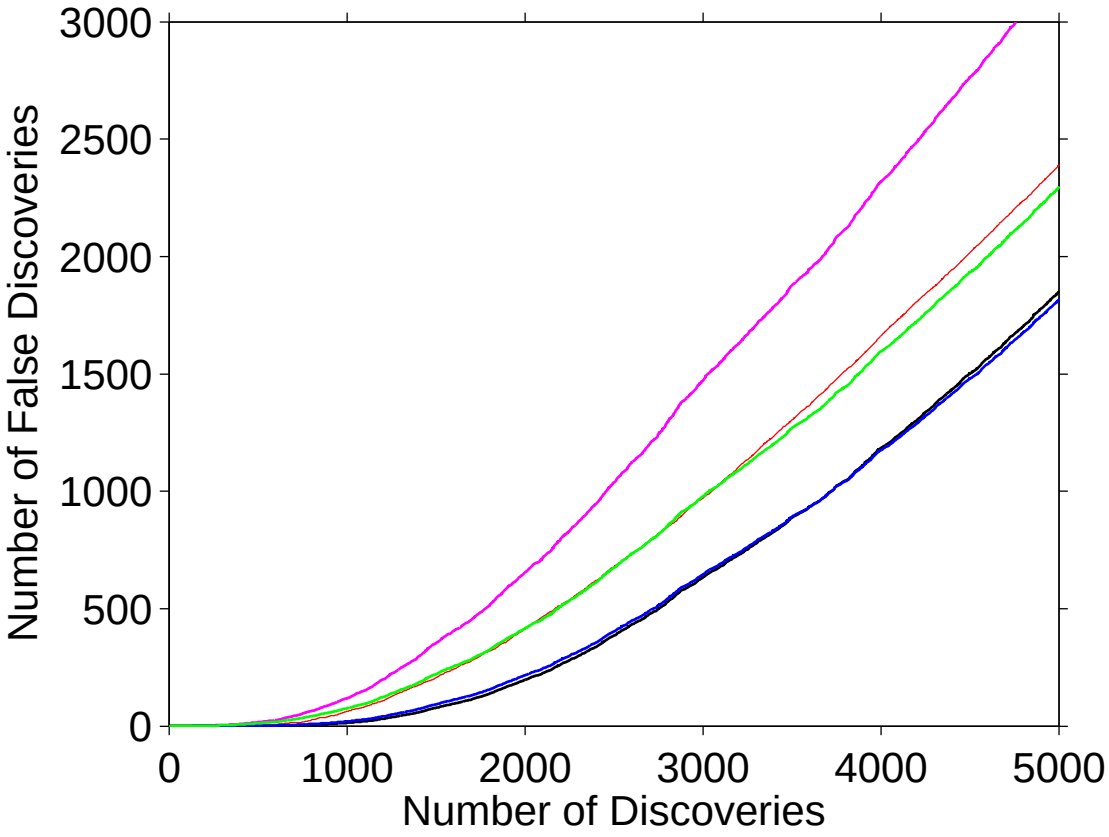


Figure 4.9: FDD curves of the FD estimation methods on `DataCubic`: $\rho = 0.6$, $nl_1 = 1$, $nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

predictor genes a P-way unit. Each P-way unit has an associated $P \times P$ matrix to describe the correlations between predictor genes. Note that P-way units are used during data simulation, but 2-way units (i.e., one target gene and two predictor genes) are still used during detection. In each P-way unit, the extra predictor gene X_3 (if $P = 3$) or extra predictor genes $X_3, ..., X_p$ (if $P > 3$) can be thought of as additive noise, since the model used for detecting three-way interactions includes only two predictor genes, g_1 and g_2. This is a form of non-Gaussian biological noise.

For the target gene, we use a cubic data model. Thus, for a non-interacting P-way unit, $Y = a + b_1X_1 + \cdots + b_pX_p + c_1X_1^2 + \cdots + c_pX_p^2 + d_1X_1^3 + \cdots + d_pX_p^3 + \varepsilon$. For an interacting P-way unit, $Y = a + b(b_1X_1 + \cdots + b_pX_p) + c(c_1X_1 + \cdots + c_pX_p)^2 + d(d_1X_1 + \cdots + d_pX_p)^3 + \varepsilon$.[2] The noise, ε, is *i.i.d.* Gaussian. Note that when the cubic term is expanded, it includes terms of the form $X_1X_2X_3$, which means the simulated data includes 4-way interactions. Simulation details are given in Figure 4.10, and we call the simulated data `DataMany`. When $P = 2$, i.e., when there are no extra predictor genes, `DataMany` is similar to `DataCubic`. As usual, to avoid dominance by any one predictor, each predictor is normalized to have variance 1. Formally, $Y = a + \hat{b}(b_1X_1 + \cdots + b_pX_p) + \hat{c}(c_1X_1 + \cdots + c_pX_p)^2 + \hat{d}(d_1X_1 + \cdots + d_pX_p)^3 + \varepsilon$, where a hat above a coefficient means that it has been scaled to normalize the corresponding predictor. For example, $\hat{c} = c/sd((c_1X_1 + \cdots + c_pX_p)^2)$, where sd denotes standard deviation.

To provide a simple way of specifying the amount of correlation between predictor genes, we allow arbitrary covariance matrices, Σ, but provide a single parameter for adjusting the intensity of the off-diagonal elements. Specifically, we let $\Sigma = \lambda B + (1 - \lambda)\text{diag}(B)$, where B is a symmetric, positive-definite matrix (randomly chosen) and λ is a real number between 0 and 1. (See step 1(b) in Figure 4.10.) Here, diag(B) is a diagonal matrix (representing uncorrelated predictors) whose diagonal is the diagonal of B. When $\lambda = 1$, B is the covariance matrix, and when $\lambda = 0$, diag(B) is the covariance matrix. When $0 < \lambda < 1$, the covariance matrix is a weighted sum of B and diag(B). Note that the diagonal of Σ is always equal to the diagonal of B. Thus, λ controls the amount of correlation between predictor genes while maintaining the variance in gene expression levels.

[2] For a p-way unit, the number of possible terms is $O(p^3)$, which requires specifying $O(p^3)$ coefficients. The approach taken here simplifies the specification by requiring only $O(p)$ coefficients.

PROCEDURE genDataMany ($M_1, M_2, N, P, \lambda, nl_1, nl_2$)

1. Generate data for M_1 non-interacting P-way units, as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Let $\mu = [\mu_1, ..., \mu_p]$, where $\mu_1, ..., \mu_p$ are randomly chosen from a uniform distribution on $[-0.5, 0, 5]$. $\mu_1, ..., \mu_p$ represent the mean expression levels of the P predictor genes. Here, $p = P$.

   (b) Let A be a $P \times P$ matrix whose elements are randomly chosen from a uniform distribution on $[-0.5, 0, 5]$. Let $B = A^T A$ (thus B is symmetric and positive definite). Let $\Sigma = \lambda B + (1 - \lambda)\text{diag}(B)$, where $\text{diag}(B)$ is a diagonal matrix whose main diagonal consists of the diagonal elements in B.

   (c) Generate N data points $x_1, ..., x_N$, where $x_i = (x_{i1}, ..., x_{ip}) \sim \mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times P$ matrix. The jth column $(x_{1j}, x_{2j}, ..., x_{Nj})^T$ is the expression profile for the jth predictor gene, where $j = 1, ..., P$. In particular, the first column is the expression profile for gene g_1, and the second column is the expression profile for gene g_2.

   (d) Let a, $b_1, ..., b_p$, $c_1, ..., c_p$ and $d_1, ..., d_p$ be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let $y_i = a + \hat{b}_1 x_{i1} + \cdots + \hat{b}_p x_{ip}$ $+ \hat{c}_1 x_{i1}^2 + \cdots + \hat{c}_p x_{ip}^2 + \hat{d}_1 x_{i1}^3 + \cdots + \hat{d}_p x_{ip}^3 + \varepsilon_i,\ i = 1, ..., N$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting P-way units, as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Repeat steps $1(a)$ to $1(c)$.

   (b) Let a, b, $b_1, ..., b_p$, c, $c_1, ..., c_p$, d and $d_1, ..., d_p$ be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let $y_i = a + \hat{b}(b_1 x_{i1} + \cdots + b_p x_{ip}) + \hat{c}(c_1 x_{i1} + \cdots + c_p x_{ip})^2 + \hat{d}(d_1 x_{i1} + \cdots + d_p x_{ip})^3 + \varepsilon_i,\ i = 1, ..., N$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization as described in Section 4.1.3. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.10: Generate data in which each target gene has more than two predictor genes. Here, $P > 2$ is the number of predictor genes for each target gene, and λ is in $[0, 1]$ and determines the average correlation between the P predictor genes. This generates multi-way interactions between $P + 1$ genes.

Figure 4.11 shows FDD curves for `DataMany` data with $P = 3$. The curve produced by the bootstrap method (in green) is closest to the true FDD curve (in red). The bootstrap method is therefore the most robust to the extra biological noise introduced by the extra predictor gene.

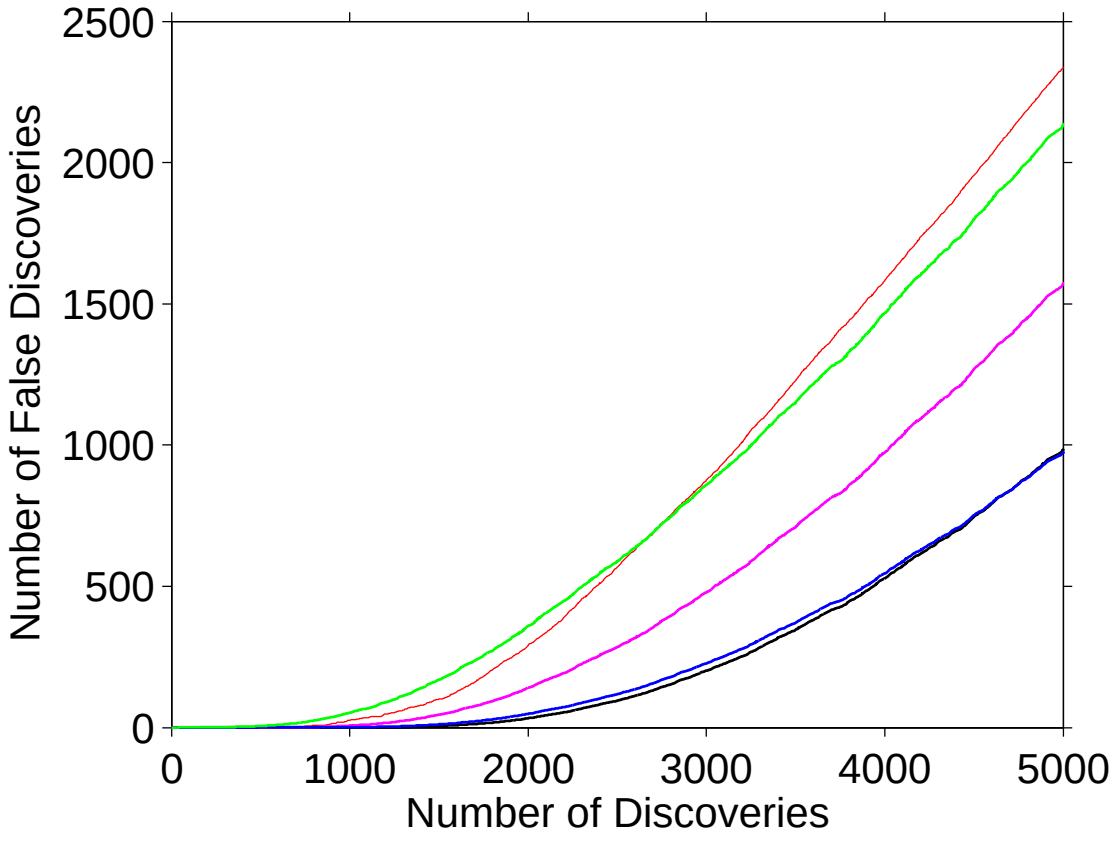


Figure 4.11: FDD curves of the FD estimation methods on `DataMany`: $P = 3$, $\lambda = 0.2$, $nl_1 = 0.1$, $nl_2 = 0.1$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.1.4 Dependent data

In the previous sections, we assumed that all gene triples are independent. This section introduces dependencies between triples, which is more typical of real expression data. That is, we generate data in which the expression levels in different triples are dependent. We do this in two ways: (i) by having triples share genes, and (ii) by introducing additional correlations between genes in different triples. Simulation details are given in Figure 4.12, and we call the simulated data `DataDependent`.

PROCEDURE genDataDependent $(M_1, M_2, N, L, \lambda, nl_1, nl_2)$

1. Generate data for a pool of L predictor genes, as follows. Let $\mu = [\mu_1, ..., \mu_L]$, where μ_i, $i = 1, ..., L$, is randomly chosen from a uniform distribution on $[-0.5, 0.5]$. These L values represent the mean expression levels of the L predictor genes. Let A be an $L \times L$ matrix whose elements are randomly chosen from a uniform distribution on $[-0.5, 0.5]$. As in Figure 4.10, let $\Sigma = \lambda A^T A + (1 - \lambda)\text{diag}(A^T A)$. Generate N multivariate Gaussian data points $x_1, ..., x_N$, where $x_i = (x_{i1}, ..., x_{iL}) \sim \mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times L$ matrix, each column representing an expression profile for one predictor gene. Generate L gene names and associate each one with one of these expression profiles.

2. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate M_1 target gene names. For each target gene, g_3, randomly choose two different genes, g_1 and g_2, from the pool of L predictor genes. Form the triple (g_1, g_2, g_3). For each such triple, generate data for g_3 as follows:

   Let a, b, c, d, e, f and g be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let

   $$y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 + \varepsilon_i, \quad i = 1, ..., N,$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. Generate data for M_2 interacting triples (true positives), as follows. First, generate M_2 target gene names. For each target gene, g_3, randomly choose two different genes, g_1 and g_2, from the pool of L predictor genes. Form the triple (g_1, g_2, g_3). For each such triple, generate data for g_3 as follows:

   Let $a, b, c, d, e, f, g, h, i$ and j be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let

   $$\begin{aligned} y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 \\ + \hat{h}x_{i1}x_{i2} + \hat{i}x_{i1}x_{i2}^2 + \hat{j}x_{i2}x_{i1}^2 + \varepsilon_i, \quad i = 1, ..., N, \end{aligned}$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

4. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.12: Generate dependent data where triples can share genes and genes can be correlated across triples. Here, L is the size of the pool of shared genes, and λ is in $[0, 1]$ and determines the average correlation between genes in the pool.

Step 1 in PROCEDURE genDataDependent produces expression profiles for a pool of L genes. The expression levels of these L genes follows a multivariate Gaussian distribution, $\mathcal{N}(\mu, \Sigma)$. Here, Σ is the covariance matrix of the L genes, and μ is the mean of the L gene expression profiles. As in section 4.1.3, we allow arbitrary covariance matrices, but provide a single parameter, λ, for adjusting the intensity of the off-diagonal elements. Gene sharing is achieved in steps 2 and 3, where predictor genes are chosen from this pool of L genes. Two predictor genes are chosen for each triple, and the target gene depends on them through a cubic data model. When the number of triples is greater than L, there is significant gene sharing among the triples.

Figure 4.13 shows the FDD curves of the FD estimation methods. The expression data used for this example was generated from a pool of 1,000 predictor genes shared by 20,000 triples, so each predictor gene belongs to 40 triples on average. Overall, the bootstrap curve (in green) is much closer to the correct FDD curve (in red) than the other curves are.

4.1.5 Multi-modal data

Until now we have only considered gene expression levels that are Gaussian, whereas real expression levels are often non-Gaussian and sometimes multi-modal. This section looks at data from Gaussian mixture models, which are both non-Gaussian and multi-modal and are a standard statistical model [27, 116]. (Section 4.3 looks at other forms of non-Gaussian data.) Simulation details are given in Figure 4.14, and we call the simulated data `DataMixture`. In PROCEDURE genDataMixture, the parameter K specifies the number of Gaussian components in the mixture, and μ_{size} determines the average distance between these components. In each gene triple, the expression levels of predictor genes have a Gaussian mixture model, and the target gene depends on them through a cubic model. Each gene triple has a different, randomly chosen mixture model and a different, randomly chosen cubic model.

Figure 4.16 gives an example of the data for two predictor genes, from which we can clearly see three distinct modes, or Gaussian components, both in the joint distribution and in the marginals. Figure 4.15 shows that, on the `DataMixture` data, although all the methods underestimate the true FD, the bootstrap method does so the least. Moreover, the bootstrap method gives very accurate FD estimates up to the top 2,000 discoveries, whereas the other methods

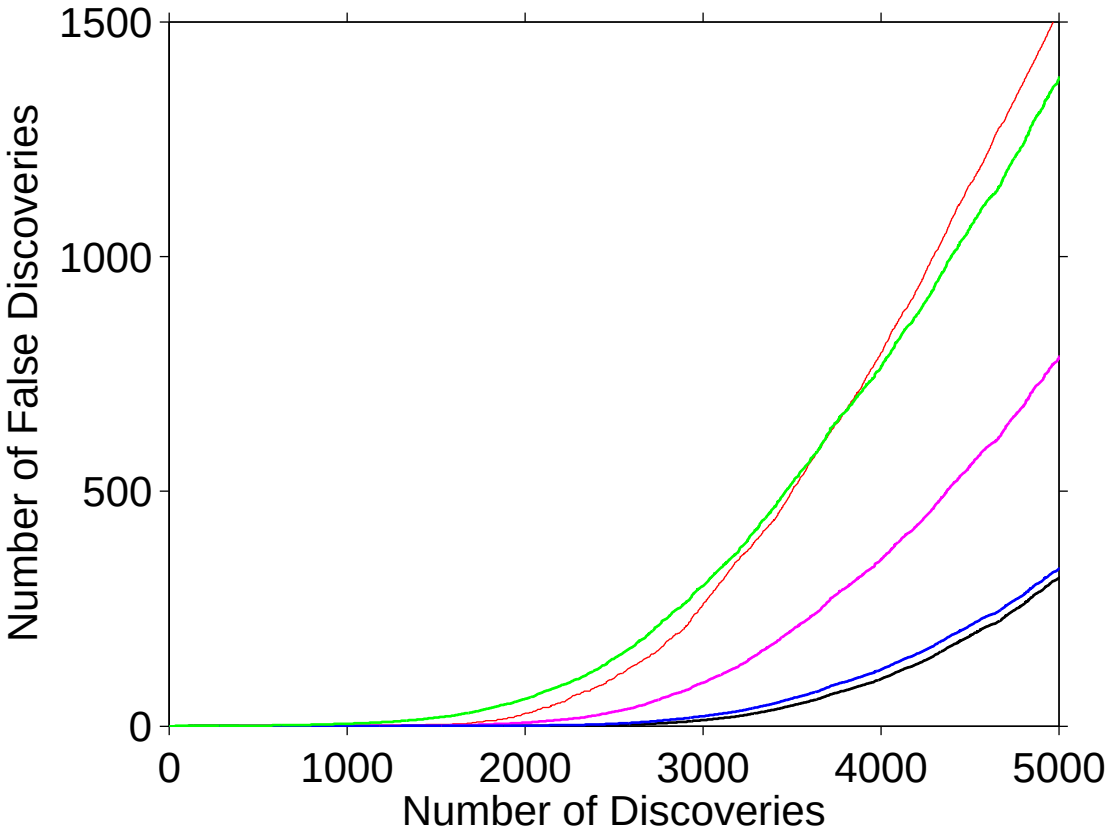


Figure 4.13: FDD curves of the FD estimation methods on `DataDependent`: $L = 1000$, $\lambda = 0.2$, $nl_1 = 0.3$, $nl_2 = 0.3$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

PROCEDURE genDataMixture ($M_1, M_2, N, nl_1, nl_2, K, \mu_{\text{size}}$)

1. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for the three genes as follows:

   (a) Let $\mu_k = (\mu_{k1}, \mu_{k2})$, where μ_{k1} and μ_{k2} are random numbers from a uniform distribution on $[-0.5 \cdot \mu_{\text{size}}, 0.5 \cdot \mu_{\text{size}}]$. $k = 1, 2, .., K$. μ_k is the mean of the kth mixture component.

   (b) Let $\Sigma_k = \left[\begin{smallmatrix} r_1 & 0 \\ 0 & r_2 \end{smallmatrix}\right] \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \cdot \left[\begin{smallmatrix} r_1 & 0 \\ 0 & r_2 \end{smallmatrix}\right]$, where r_1 and r_2 are random numbers from a uniform distribution on $[0, 1]$, and ρ is a random number from a uniform distribution on $[-1, 1]$. Σ_k is the covariance matrix for the kth mixture component. Within this component, ρ is the correlation of g_1 and g_2, and r_i is the standard deviation of g_i.

   (c) Let $\frac{p_k}{\sum_{i=1}^{K} p_i}$ be the probability of the kth mixture component, where $p_1, ..., p_K$ are random numbers from a uniform distribution on $[0, 1]$.

   (d) Generate N bivariate data points drawn from the mixture of K Gaussian components. The resulting data form an $N \times 2$ matrix. The first column $(x_{11}, x_{21}, ..., x_{N1})^T$ is the expression profile for gene g_1. The second column $(x_{12}, x_{22}, ..., x_{N2})^T$ is the expression profile for gene g_2.

   (e) Let a, b, c, d, e, f and g be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let $y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 + \varepsilon_i$, $i = 1, ..., N$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting triples (true positives), as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Repeat steps $1(a)$ to $1(d)$.

   (b) Let $a, b, c, d, e, f, g, h, i$ and j be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on $[0, 1]$. Let $y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 + \hat{h}x_{i1}x_{i2} + \hat{i}x_{i1}x_{i2}^2 + \hat{j}x_{i2}x_{i1}^2 + \varepsilon_i$, $i = 1, ..., N$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.14: Generate data in which the expression levels of predictor genes have a Gaussian mixture model. Here, K specifies the number of mixture components, and μ_{size} determines the average distance between the components. Since each target gene has two predictor genes, the mixture models are two-dimensional.

underestimate the true FD by at least an order of magnitude..

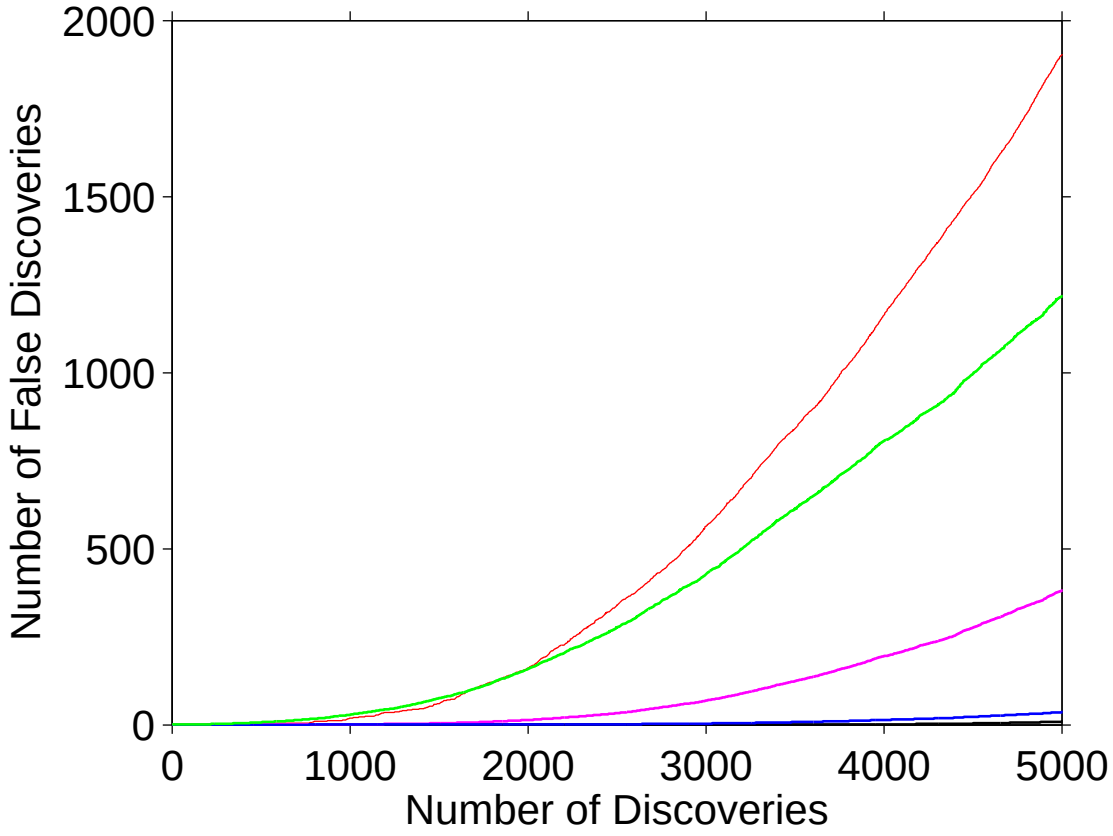


Figure 4.15: FDD curves of the FD estimation methods on `DataMixture`: $K = 3$, $\mu_{\text{size}} = 0.5$, $nl_1 = 0.1$, $nl_2 = 0.1$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.1.6 More-complex non-linearities

The above sections describe three-way gene interactions in terms of smooth, low-order polynomials, e.g., quadratic functions and cubic functions. The bootstrap method works well in these cases. In reality, the interaction is likely to be more complex and less smooth. In this section, we look at two forms of interaction that are more biologically inspired. In these data models, three-way interactions are described in terms of discontinuous and non-differentiable functions, in order to demonstrate the robustness of the bootstrap method in the face of non-smoothness. As usual, X_1, X_2 and Y represent the expression levels of two predictor genes and a target gene, respectively, and ε represents biological noise.

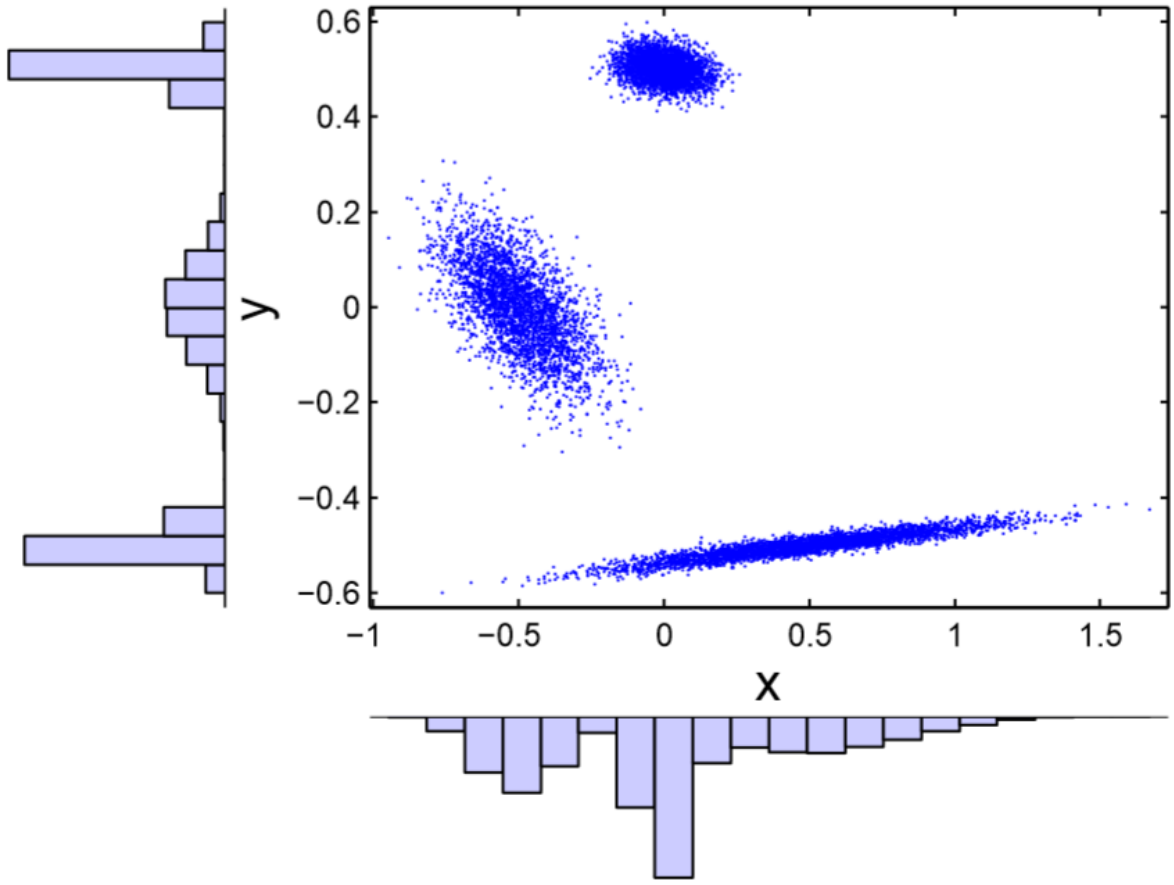


Figure 4.16: Representative multi-modal data for two predictor genes, x and y, generated using a Gaussian mixture model with three components. $N = 10,000, K = 3, \mu_{\text{size}} = 1$.

Max data. Here we consider a non-differentiable data model. Specifically, for an interacting triple, $Y = \max(f(X_1), g(X_2)) + \varepsilon$, for some functions f and g. In our simulations, we use quadratic polynomials for f and g, and *i.i.d.* Gaussian noise for ε. Note that the max function is an extension of the OR logic in combinatorial regulation, since OR logic can be rewritten as $\max(a, b)$, where a and b are binary numbers. For a non-interacting triple, $Y = a + bX_1 + cX_2 + dX_1^2 + eX_2^2 + \varepsilon$, as before, where ε is *i.i.d.* Gaussian random noise. Simulation details for max data are given in Figure 4.17, and we call the simulated data `DataMax`. Notice that this max model of three-way interactions goes beyond the additive model of earlier sections, in which $Y = f(X_1) + g(X_2) + h(X_1, X_2) + \varepsilon$, where $h(X_1, X_2)$ is a smooth interaction term.

Figure 4.18 shows the FDD curves of the FD estimation methods. Clearly, the bootstrap curve (in green) is much closer to the correct FDD curve (in red) than the other curves are. Moreover, the green curve almost coincides with the red curve for the top 1,000 discoveries, whereas the other curves underestimate the true FD by at least an order of magnitude.

Switch data. Here we consider a discontinuous data model, in which the interaction between genes g_1 and g_3 switches between two modes, depending on the level of gene g_2. Specifically, given a threshold, τ, if $X_2 > \tau$, then $Y = f(X_1) + \varepsilon$; otherwise, $Y = g(X_1) + \varepsilon$. In this way, g_2 controls the interaction between g_1 and g_3. In our simulations, f and g are distinct quadratic polynomials. This form of three-way interaction can be viewed as an idealization of TF modulation, in which the interaction between a transcription factor and a target gene is modulated by a third gene [20, 24, 30, 112]. Simulation details for switch data are given in Figure 4.19, and we call the simulated data `DataSwitch`.

Figure 4.20 shows the FDD curves of the FD estimation methods on `DataSwitch`. The bootstrap method gives the most accurate FDD curve, whereas the other methods give significantly underestimated FDD curves (underestimated by at least an order of magnitude for the top 2,000 discoveries).

PROCEDURE genDataMax($M_1, M_2, N, \rho, nl_1, nl_2$)

1. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:
   (a) Let $\mu = [\mu_1, \mu_2]$, where μ_1 and μ_2 are randomly chosen from a uniform distribution on [-0.5,0.5]. μ_1 and μ_2 represent the mean expression levels of the two predictor genes, g_1 and g_2, respectively.
   (b) Let r be a random number from a uniform distribution on [0,1]. r represents the variance in the expression levels of g_1 and g_2. Let $\Sigma = r \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.
   (c) Generate N bivariate Gaussian data points drawn from $\mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times 2$ matrix. The first column $(x_{11}, x_{21}, ..., x_{N1})^T$ is the expression profile for gene g_1. The second column $(x_{12}, x_{22}, ..., x_{N2})^T$ is the expression profile for gene g_2.
   (d) Let a, b, c, d and e be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let $y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \varepsilon_i$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates normalization (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.
2. Generate data for M_2 interacting triples (true positives), as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:
   (a) Repeat steps $1(a)$ to $1(c)$.
   (b) Let a_1, a_2, b_1, b_2, c_1 and c_2 be randomly chosen from a uniform distribution on [-0.5,0.5]. Let s be randomly chosen from a uniform distribution on [0,1]. Let $y_i = \max(a_1 + \hat{b}_1 x_{i1} + \hat{c}_1 x_{i1}^2,\ a_2 + \hat{b}_2 x_{i2} + \hat{c}_2 x_{i2}^2) + \varepsilon_i$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.
3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.17: Generate data for which, in an interacting triple, the target gene is controlled by the stronger of the two predictor genes.

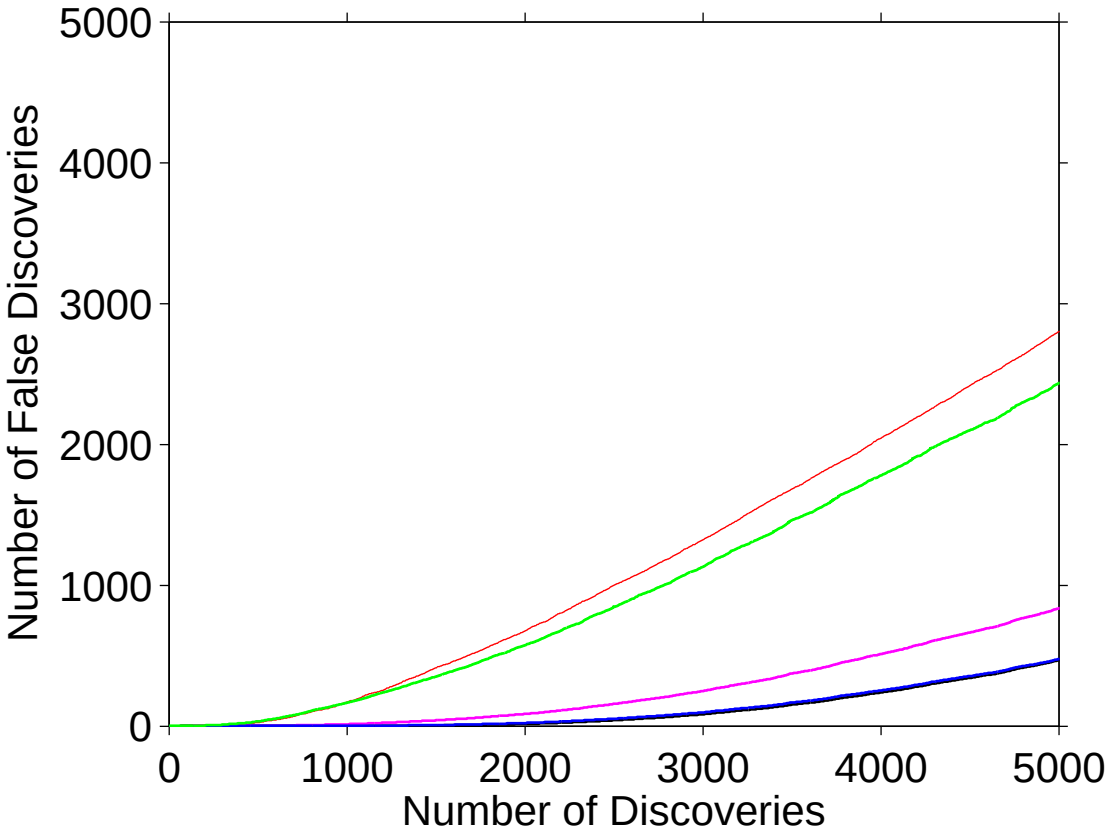


Figure 4.18: FDD curves of the FD estimation methods on `DataMax`: $\rho = 0$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

PROCEDURE genDataSwitch $(M_1, M_2, N, \rho, nl_1, nl_2)$

1. Generate data for M_1 non-interacting triples (true negatives), as follows. First, generate $3M_1$ gene names. From these names, create M_1 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Let $\mu = [\mu_1, \mu_2]$, where μ_1 and μ_2 are randomly chosen from a uniform distribution on [-0.5,0.5]. μ_1 and μ_2 represent the mean expression levels of the two predictor genes, g_1 and g_2, respectively.

   (b) Let r be a random number from a uniform distribution on [0,1]. r represents the variance in the expression levels of g_1 and g_2. Let $\Sigma = r \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

   (c) Generate N bivariate Gaussian data points drawn from $\mathcal{N}(\mu, \Sigma)$. The resulting data form an $N \times 2$ matrix. The first column $(x_{11}, x_{21}, ..., x_{N1})^T$ is the expression profile for gene g_1. The second column $(x_{12}, x_{22}, ..., x_{N2})^T$ is the expression profile for gene g_2.

   (d) Let a, b, c, d and e be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on [0,1]. Let $y_i = a + bx_{i1} + cx_{i2} + dx_{i1}^2 + ex_{i2}^2 + \varepsilon_i$, where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting triples (true positives), as follows. First, generate $3M_2$ gene names. From these names, create M_2 triples, with each name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for g_1, g_2 and g_3 as follows:

   (a) Repeat steps $1(a)$ to $1(c)$.

   (b) Let τ, a_1, b_1, c_1, a_2, b_2 and c_2 be randomly chosen from a uniform distribution on $[-0.5, 0.5]$. Let s be randomly chosen from a uniform distribution on [0,1].

   (c) If $x_{i2} \geq \tau$, let
   $$y_i = a_1 + b_1 x_{i1} + c_1 x_{i1}^2 + \varepsilon_i,$$
   otherwise, let
   $$y_i = a_2 + b_2 x_{i1} + c_2 x_{i1}^2 + \varepsilon_i,$$
   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each of x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.19: Generate data for which, in an interacting triple, the interaction between genes g_1 and g_3 switches between two modes, depending on the expression level of gene g_2.

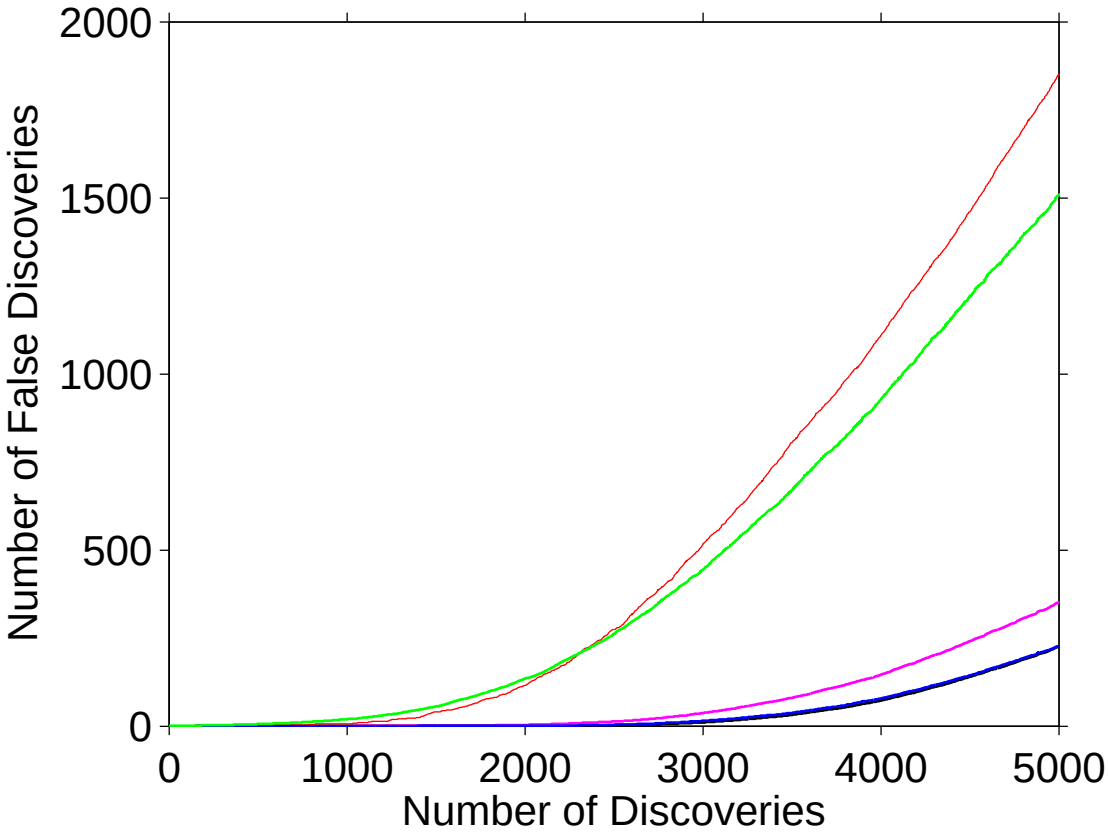


Figure 4.20: FDD curves of the FD estimation methods on `DataSwitch`: $\rho = 0$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.2 Results on real/simulated data

Using simulated gene expression based on relatively simple statistical models, Section 4.1 demonstrates the superiority of the bootstrap method for estimating the number of false discoveries under a variety of statistical conditions, and demonstrates that other estimation methods underestimate the number of false discoveries, often by more than an order of magnitude. In this section we use real data, i.e., the log-transformed germination/dormancy gene expression data of Chapter 2, as the data for predictor genes. In this way, we can test the FD estimation methods under the specific, and real conditions of this data set. Of course, although the expression levels of some genes are now real, others must be simulated in order to have known three-way interactions. To this end, data for the target genes are generated from the predictor genes as in Section 4.1. Since the expression data is now half real and half simulated (for predictor genes, the expression data is real, whereas for target genes, it is simulated), we call it real/simulated data.

In this section, we look at four types of real/simulated data and the estimates of false discoveries for them. All figures in this section use $M_1 = M_2 = 10,000$, and the length of each expression profile is $N = 138$, the same as that for the seed germination/dormancy data of Chapter 2. The procedures below all assume that the real data is stored in a matrix, X, whose ij^{th} entry is the expression level of gene j in experiment i. That is, the columns of X are gene expression profiles.

We shall show that the bootstrap method of estimating false discoveries outperforms the other methods on this data, and that the other methods often underestimate FD by more than an order of magnitude.

4.2.1 Cubic data

In this section, we look at how the FD estimation methods perform on the real/simulated version of cubic data. The data generation procedure, genRealCubic (Figure 4.21), is similar to genDataCubic in Section 4.1.2, except that now expression data for the predictor genes are sampled from real expression data. We call the simulated data `RealCubic`.

PROCEDURE genRealCubic(M_1, M_2, nl_1, nl_2)

1. Generate data for M_1 non-interacting triples (true negatives), as follows:

   First, generate $3M_1$ gene names. From these names, create M_1 triples, with each gene name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for genes g_1, g_2 and g_3 as follows:

   (a) Randomly choose two columns, $(x_{11}, x_{21}, ..., x_{N1})^T$ and $(x_{12}, x_{22}, ..., x_{N2})^T$, from the real data matrix, X. These are the expression profiles for predictor genes g_1 and g_2, respectively. N is the length of each profile.

   (b) Let a, b, c, d, e, f and g be random numbers from a uniform distribution on $[-0.5, 0.5]$. Let s be a random number from a uniform distribution on $[0, 1]$.

   (c) Let the ith gene expression level for target gene g_3 be

   $$y_i = a + \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 + \varepsilon_i, \quad i = 1, ..., N,$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. A hat above a coefficient indicates that it has been scaled to normalize its predictor (Section 4.1.1). The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

2. Generate data for M_2 interacting triples (true positives), as follows:

   First, generate $3M_2$ gene names. From these names, create M_2 triples, with each gene name assigned to one triple. For each triple, (g_1, g_2, g_3), generate expression profiles for genes g_1, g_2 and g_3 as follows:

   (a) Repeat step 1(a).

   (b) Let $a, b, c, d, e, f, g, h, i$ and j be random numbers from a uniform distribution on $[-0.5, 0.5]$. Let s be a random number from a uniform distribution on $[0, 1]$.

   (c) Let the ith expression level for target gene g_3 be

   $$\begin{aligned} y_i = a &+ \hat{b}x_{i1} + \hat{c}x_{i2} + \hat{d}x_{i1}^2 + \hat{e}x_{i2}^2 + \hat{f}x_{i1}^3 + \hat{g}x_{i2}^3 \\ &+ \hat{h}x_{i1}x_{i2} + \hat{i}x_{i1}x_{i2}^2 + \hat{j}x_{i2}x_{i1}^2 + \varepsilon_i, \quad i = 1, ..., N, \end{aligned}$$

   where ε_i is Gaussian random noise with mean 0 and standard deviation $s \cdot nl_1$. The column vector $(y_1, y_2, ..., y_N)^T$ is the expression profile for the target gene, g_3.

3. For each x_{i1}, x_{i2} and y_i, add a different ε (or multiply by a different e^{ε}), where ε is random Gaussian noise with mean 0 and standard deviation nl_2.

Figure 4.21: Generate cubic data where two predictor genes have a cubic relationship with a target gene. The expression profiles for the two predictor genes are from real data.

The FDD curves of the FD estimation methods are shown in Figure 4.22 (multiplicative measurement noise) and in Figure 4.23 (additive measurement noise). In both cases, the bootstrap curve is closest to the true FDD curve. The total permutation method and analytical t method give almost zero FD estimates for top 5,000 discoveries. While clearly above zero, the partial permutation method significantly underestimates the true FDD curve.

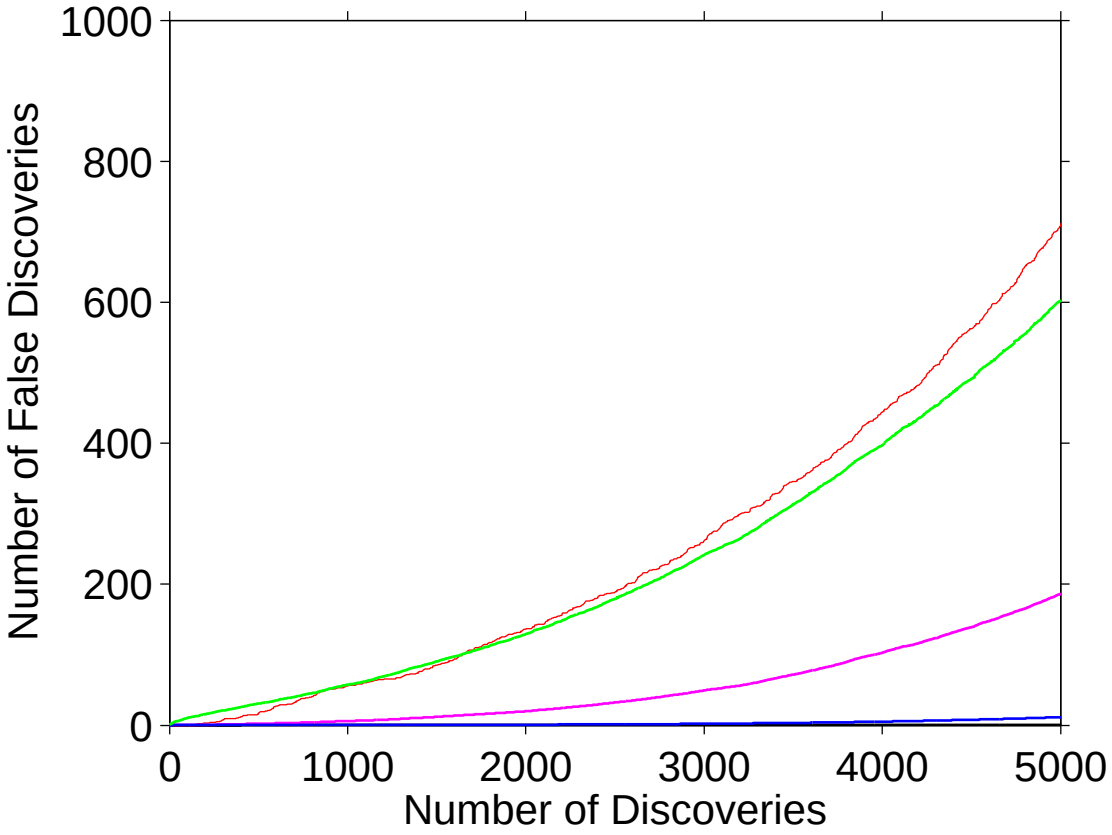


Figure 4.22: FDD curves of the FD estimation methods on `RealCubic`: $nl_1 = 0.4, nl_2 = 0.4$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.2.2 Many predictor genes

The procedure for simulating data when target genes have many predictor genes is similar to PROCEDURE genDataMany in Section 4.1.3. However, instead of being sampled from a multivariate Gaussian distribution, the gene expression profiles of each P-way unit are sampled from real expression data (as in step 1(a) of PROCEDURE genRealCubic in Section 4.2.1).

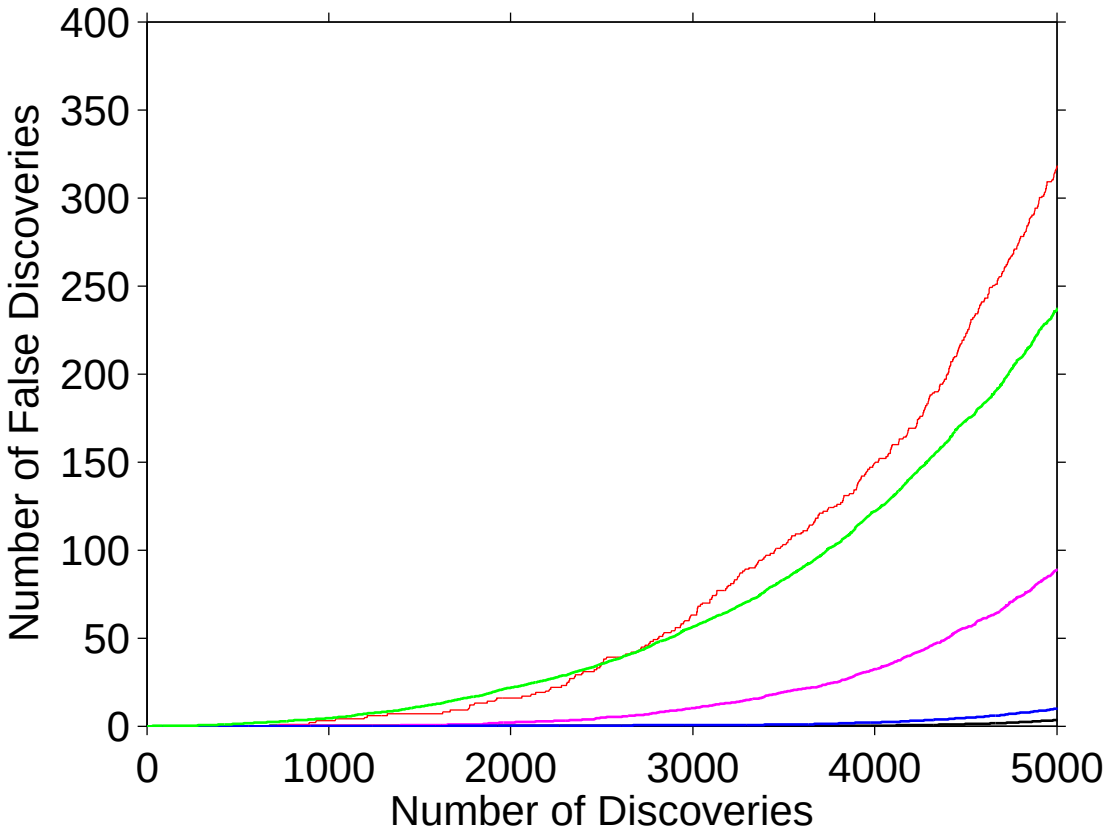


Figure 4.23: FDD curves of the FD estimation methods on `RealCubic`: $nl_1 = 1, nl_2 = 0.2$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

We call the simulated data `RealMany`. Figure 4.24 shows the FDD curves of the FD estimation methods on `RealMany`, from which we can see that the bootstrap method performs significantly better than the other methods.

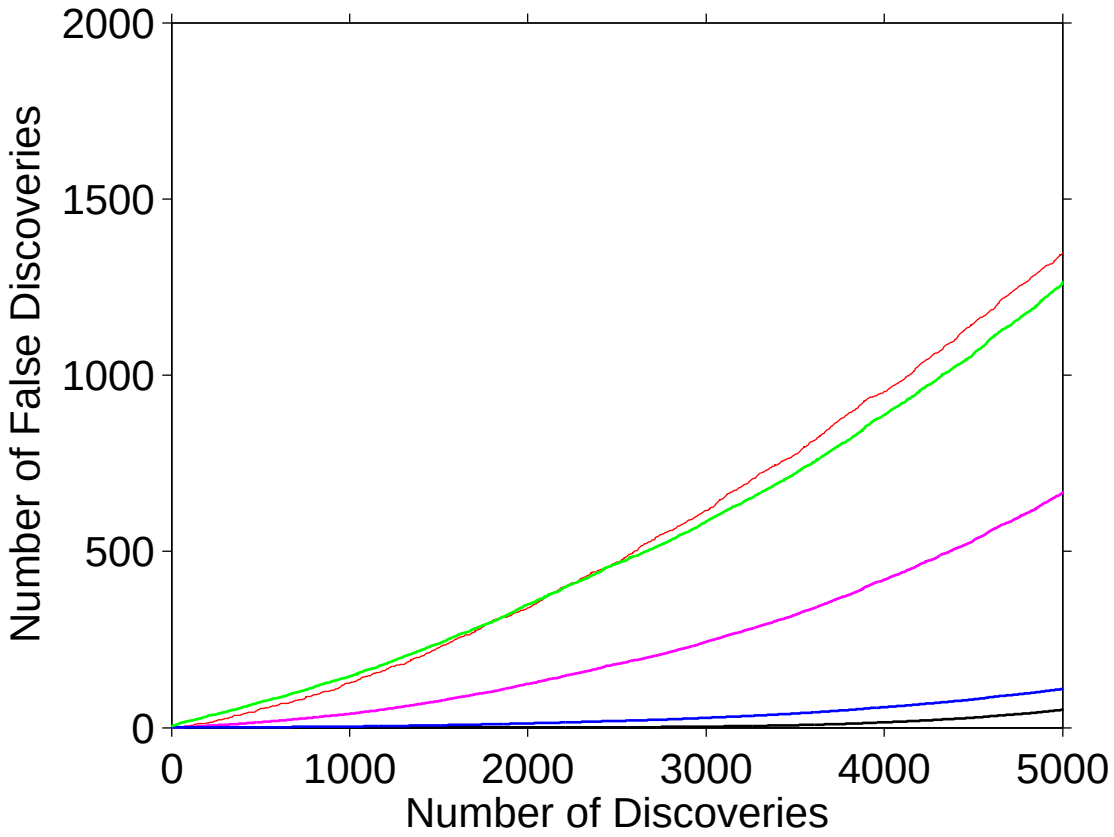


Figure 4.24: FDD curves of the FD estimation methods on `RealMany` data: $nl_1 = 0.5, nl_2 = 0.5$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.2.3 Max data

The simulation procedure for real max data is similar to PROCEDURE genDataMax in Section 4.1.6, except that now expression levels for predictor genes are sampled from real expression data (as in step 1(a) in PROCEDURE genRealCubic in Section 4.2.1). We call the simulated data `RealMax`. The FDD curves of the FD estimation methods on `RealMax` are shown in Figure 4.25. The bootstrap method gives a very accurate FDD curve, whereas the other methods significantly underestimate the true FDD curve.

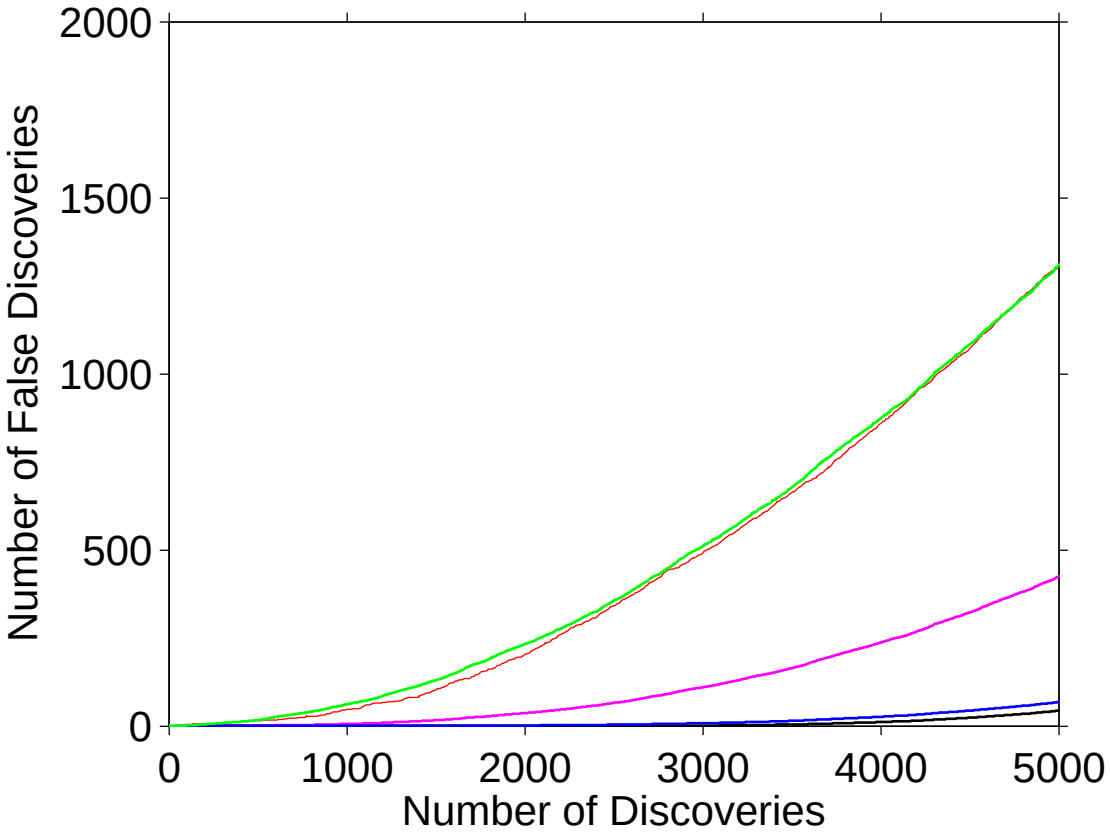


Figure 4.25: FDD curves of the FD estimation methods on `RealMax`: $nl_1 = 0.1, nl_1 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.2.4 Switch data

The simulation procedure for real switch data is similar to PROCEDURE genDataSwitch in Section 4.1.6, except that now expression levels for predictor genes are sampled from real expression data (as in step 1(a) in PROCEDURE genRealCubic in Section 4.2.1). We call the simulated data `RealSwitch`. Figures 4.26 and 4.27 show FDD curves of the FD estimation methods on `RealSwitch`. The bootstrap method gives the best FDD curve, while the other methods significantly underestimate the true curve.

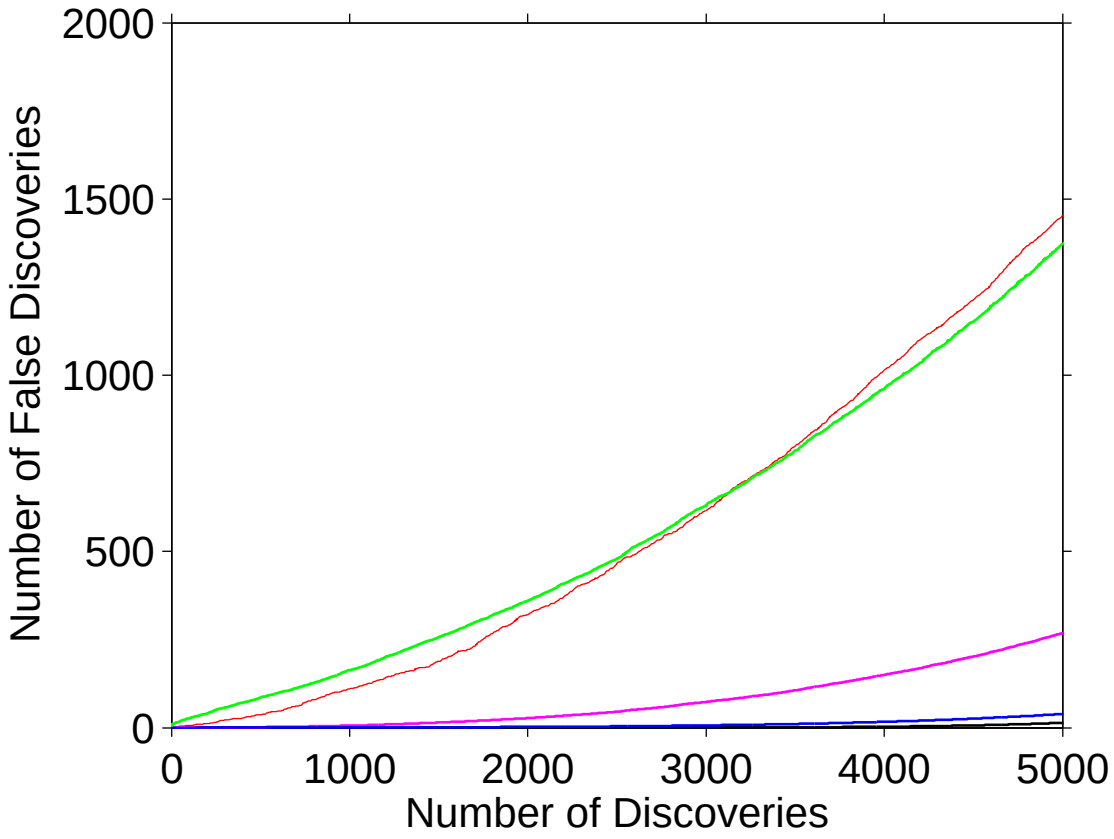


Figure 4.26: FDD curves of the FD estimation methods on `RealSwitch`: $nl_1 = 0.2, nl_2 = 0.2$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

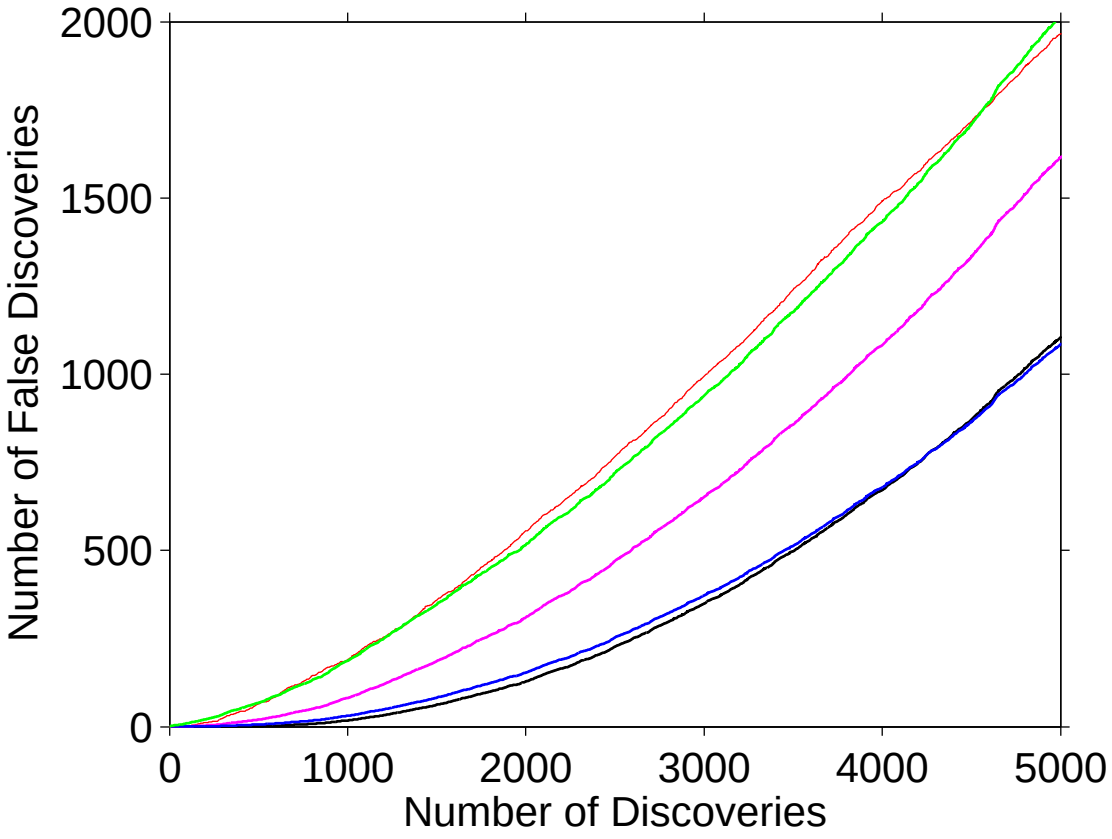


Figure 4.27: FDD curves of the FD estimation methods on `RealSwitch`: $nl_1 = 0.5, nl_2 = 0.5$ (additive measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

4.3 Results on other non-Gaussian data

Section 4.1.5 looked at a particular form of non-Gaussian gene expression data (Gaussian mixture models). This section considers other forms of non-Gaussian data. Specifically, we look at data from Laplacian distributions (which have longer tails) and beta distributions (which are skewed and have shorter tails). While it is easy to simulate data for correlated Gaussian variables, correlations between non-Gaussian variables are much harder to simulate. One way to do this is with copulas [117].

4.3.1 Copulas

A copula is a multivariate distribution for which the marginal probability of each variable is uniform [117]. There are a number of ways of constructing copulas. One popular method is to start with a multivariate Gaussian distribution and then transform each variable so that it has a uniform distribution. Specifically, if $(x_1, ..., x_n)$ are the variables of a multivariate Gaussian, and if we let $y_i = \phi_i(x_i)$, where ϕ_i is the marginal CDF of x_i, then each y_i has a uniform distribution on $[0, 1]$ [117]. Because each ϕ_i is strictly monotonic, Spearman correlation among the variables is preserved by this transformation (though Pearson correlation is not). The distribution of $(y_1, ..., y_n)$ is known as a Gaussian copula [117].

In a similar fashion, we can transform the y_i. For instance, let $z_i = F^{-1}(y_i)$, where F^{-1} is the inverse CDF of a univariate distribution, f. Then each z_i has distribution f. Again, because F^{-1} is strictly monotonic, the Spearman correlation among the z_i is the same as among the y_i and the x_i. In this way, we can generate multivariate distributions with a given (Spearman) correlation structure and with any marginal distributions we like.

For example, applying the inverse CDF of the beta distribution to a copula transforms the variables so they have a beta distribution with dependence preserved among variables. This process is illustrated in Figures 4.28, 4.29 and 4.30. First, bivariate Gaussian data is converted to copula data through the CDF of a Gaussian distribution, then the copula data is converted to bivariate beta-distributed data through the inverse CDF of the beta distribution.

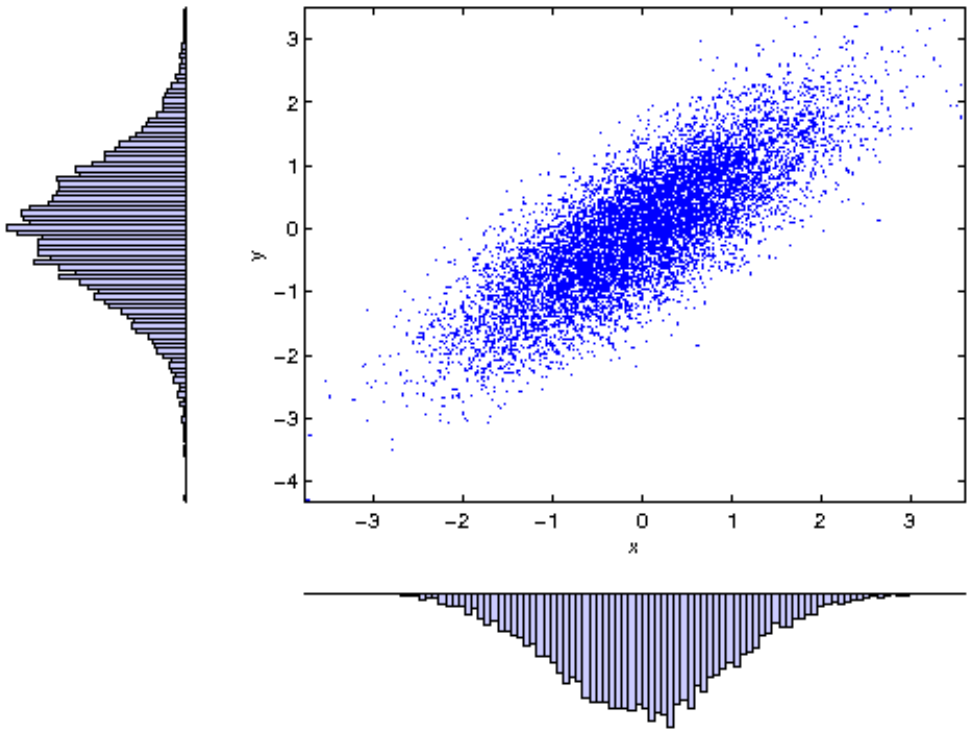


Figure 4.28: The scatter plot and marginal distribution of 10,000 2-dimensional normally distributed data points with mean $\mu = (0, 0)$ and covariance $\Sigma = [1.0,\ 0.8; 0.8,\ 1.0]$. Notice that the two variables, x and y, are strongly correlated, and each follows a Gaussian distribution.

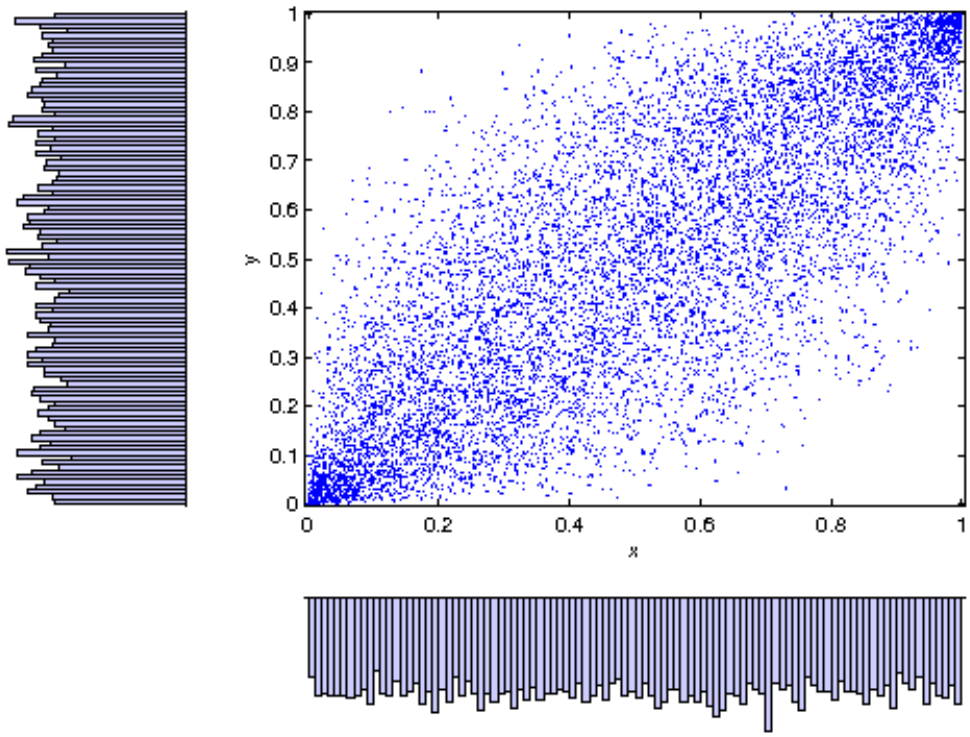


Figure 4.29: The scatter plot and marginal distribution of 10,000 2-dimensional uniform data points from a Gaussian copula (constructed using the data points in Figure 4.28). Notice that the two variables, x and y, are still correlated, but now each has a uniform distribution.

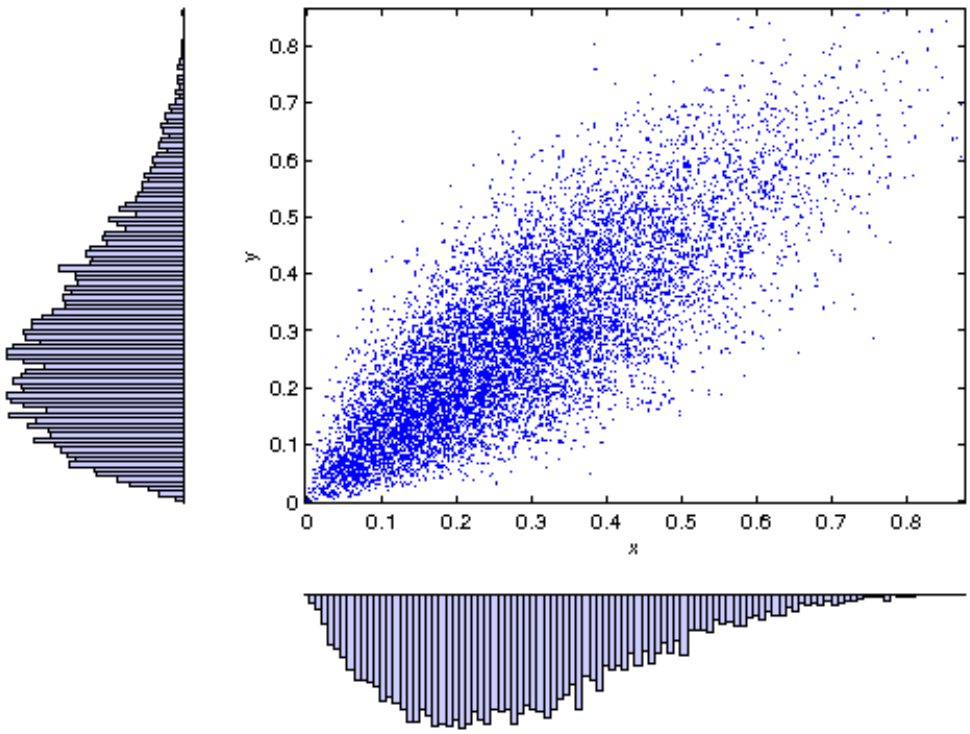


Figure 4.30: The scatter plot and marginal distribution of 10,000 2-dimensional data points following a beta distribution. The data points are generated by applying the inverse CDF of a beta distribution (with parameters $A = 2$ and $B = 5$) to the data in Figure 4.29. Notice that the two variables, x and y, are still correlated, and each follows a beta distribution.

4.3.2 Multivariate beta distributions

The expression levels for predictor genes in Section 4.1.2 are Gaussian, and therefore symmetric. In this section, we examine how well the FD estimation methods work on beta-distributed data, which is skewed and has shorter tails than Gaussians (as seen in the marginal distributions of Figure 4.30).

A random variable $x \in [0, 1]$ follows a beta distribution if its density function is $f(x; A, B) = \frac{1}{\mathcal{B}(A, B)} x^{A-1}(1-x)^{B-1}$, where $A > 0$ and $B > 0$ control the shape of the distribution. At the start of Section 4.3, we outlined how to generate bivariate beta-distributed data. The details are given in PROCEDURE genBetaData (Figure 4.31). The procedure returns an $N \times 2$ matrix, D, representing two gene expression profiles of length N. To generate this data, the procedure first draws N pairs of values independently from a bivariate Gaussian distribution with correlation ρ. These values are then transformed to values from a bivariate uniform distribution, and then to values from a bivariate beta distribution. Finally, the N pairs are scaled and shifted, so that each gene has a different range of expression levels. The scaling factor is chosen randomly from a uniform distribution on $[0, \sigma_{max}]$, and the amount of shift is chosen randomly from a uniform distribution on $[-\mu_{max}, \mu_{max}]$.

The method for generating the simulated data is the same as PROCEDURE genDataCubic (Section 4.1.2), except that the data for predictor genes, g_1 and g_2, is now bivariate beta distributed (with $A = 2$ and $B = 6$). Moreover, the correlation, ρ, is not fixed but is different for each gene pair, and is randomly chosen from a compressed Gaussian distribution, as described in Section 4.4 (with $\sigma_u = 0.3$). We call the simulated data `DataBetaRandomRho`. The FDD curves of the FD estimation methods are shown in Figure 4.32. It can be seen that the bootstrap estimate of FD is much larger than the others, does not underestimate the true FD at small numbers of discoveries, and is more accurate at higher numbers of discoveries.

4.3.3 Multivariate Laplacian distributions

In this section, we examine how well the FD estimation methods work on Laplacian data, which has longer tails than Gaussians and is not smooth at the origin.

PROCEDURE genBetaData $(N, \rho, A, B, \sigma_{max}, \mu_{max})$

1. Let D be an $N \times 2$ matrix in which each row is drawn randomly from a bivariate Gaussian copula with underlying covariance $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The entries of D are therefore uniformly distributed on $[0, 1]$, and the columns of D are correlated.
2. Replace each entry, d_{ij}, of D by $F^{-1}(d_{ij})$, where F^{-1} is the inverse CDF of a beta distribution with parameters A and B. The entries of D are now beta distributed, and the columns have the same Spearman correlation.
3. Randomly choose two values, a and b, from a uniform distribution on $[0, \sigma_{max}]$. Multiply the first column of D by a, and multiply the second column by b.
4. Randomly choose two values, c and d, from a uniform distribution on $[-\mu_{max}, \mu_{max}]$. Add c to the first column of D, and add d to the second column.
5. Return D.

Figure 4.31: Generate bivariate beta-distributed data.

A random variable x has a Laplacian distribution if its density function is $f(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$, where $b \geq 0$ is a scale parameter (larger b producing longer tails) and μ is a location parameter. Note that the Laplacian distribution is an exponential distribution in which negative values are possible. It has longer tails than the Gaussian since $\exp(-|x|)$ decreases much more slowly than $\exp(-x^2)$. The procedure for generating Laplacian data, genLaplacianData, is given in Figure 4.33. It is similar to genBetaData except that it uses a standard Laplacian distribution (with density function $f(y) = \frac{1}{2}\exp(-|y|)$) in step 2. In this case, the inverse CDF is $F^{-1}(y) = \text{sign}(y - 0.5)G^{-1}(2|y - 0.5|)$, where G^{-1} is the inverse CDF of an exponential distribution with unit variance (i.e., with density function $g(y) = \exp(-y)$). As in genBetaData, the data is finally scaled and shifted to give each gene a different range of expression levels. Figure 4.34 illustrates correlated bivariate Laplacian data, with the following parameters for genLaplacianData: $N = 10,000, \rho = 0.6, \sigma_{max} = 5, \mu_{max} = 5$, where σ_{max} and μ_{max} have the same meaning as in Section 4.3.2.

The method for generating the simulated data is the same as PROCEDURE genDataCubic, except that now the data for predictor genes, g_1 and g_2, is Laplacian. Moreover, the correlation, ρ, is not fixed but is different for each gene pair, and is randomly chosen from a compressed Gaussian distribution, as described in Section 4.4 (with $\sigma_u = 0.3$). We call the simulated

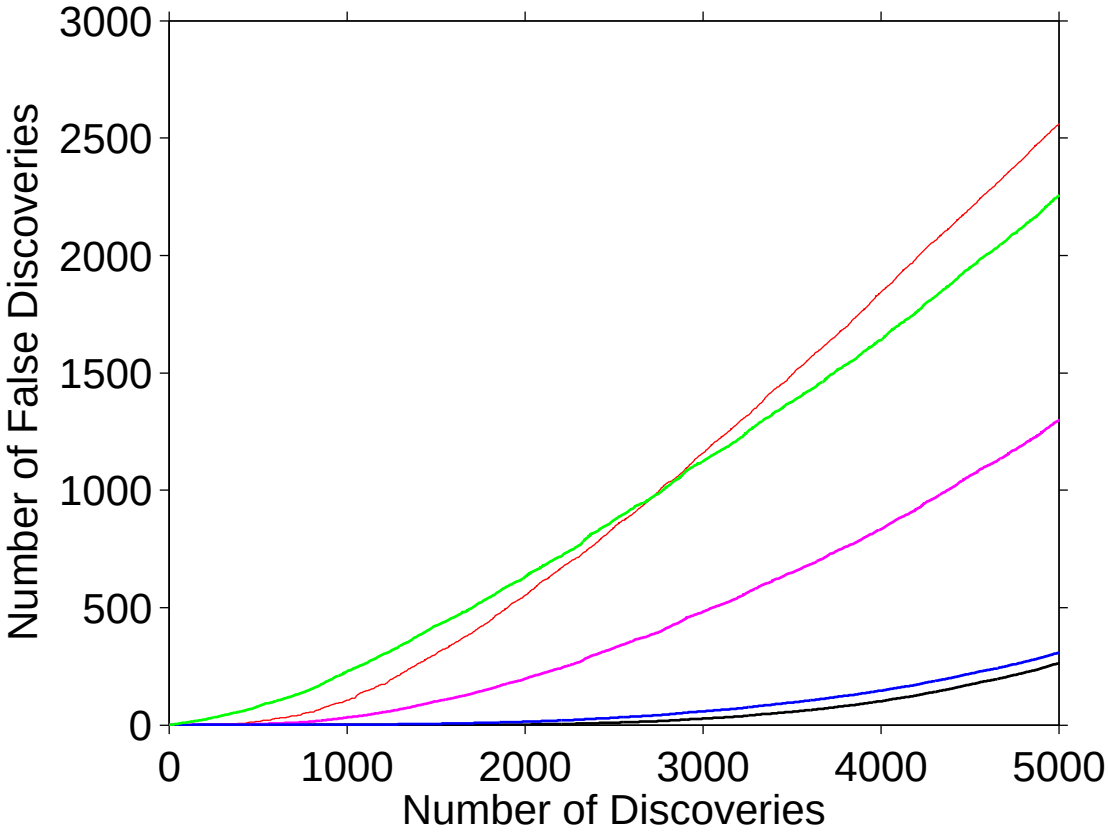


Figure 4.32: FDD curves of the FD estimation methods on `DataBetaRandomRho`: $M_1 = 15000, M_2 = 5000, N = 100, \sigma_u = 0.3, nl_1 = 0.1, nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

PROCEDURE genLaplacianData $(N, \rho, \sigma_{max}, \mu_{max})$

1. Let D be an $N \times 2$ matrix in which each row is drawn randomly from a bivariate Gaussian copula with underlying covariance $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The entries of D are therefore uniformly distributed on $[0, 1]$, and the columns of D are correlated.
2. Replace each element, d_{ij}, of D by $F^{-1}(d_{ij})$, where F^{-1} is the inverse CDF of a Laplacian distribution with parameters $b = 1$ and $\mu = 0$. The entries of D are now Laplacian distributed, and the Spearman correlation of the columns is unchanged.
3. Randomly choose two values, a and b, from a uniform distribution on $[0, \sigma_{max}]$. Multiply the first column of D by a, and multiply the second column by b.
4. Randomly choose two values, c and d, from a uniform distribution on $[-\mu_{max}, \mu_{max}]$. Add c to the first column of D, and add d to the second column.
5. Return D.

Figure 4.33: Generate bivariate Laplacian data.

data `DataLaplaceRandomRho`. The FDD curves of the FD estimation methods are shown in Figure 4.35. Again, the bootstrap method gives the most accurate FD estimate overall.

4.4 Results on mixtures of correlations

The above data generation procedures, e.g., genDataCubic, fix the correlation parameter, ρ, for all interacting and non-interacting triples. That is, for each gene triple, (g_1, g_2, g_3), the correlation between the predictor genes, g_1 and g_2, is the same. In this section, we generate data in which the predictor genes have many different correlations. Besides being more realistic, this allows us to test the performance of the FD estimation methods on data containing a mixture of many different correlations, not just a single correlation. For each gene triple, we choose the correlation of the predictor genes randomly from a probability distribution. We consider two distributions: (i) a compressed Gaussian distribution, and (ii) an empirical distribution of correlations in real expression data.

Compressed Gaussian distributions (e.g., Figure 4.36) are commonly seen in the correlation distributions of real data (e.g., Figure 4.38), where higher correlation arises with lower probability and lower correlation arises with higher probability. To simulate this, we let ρ be the

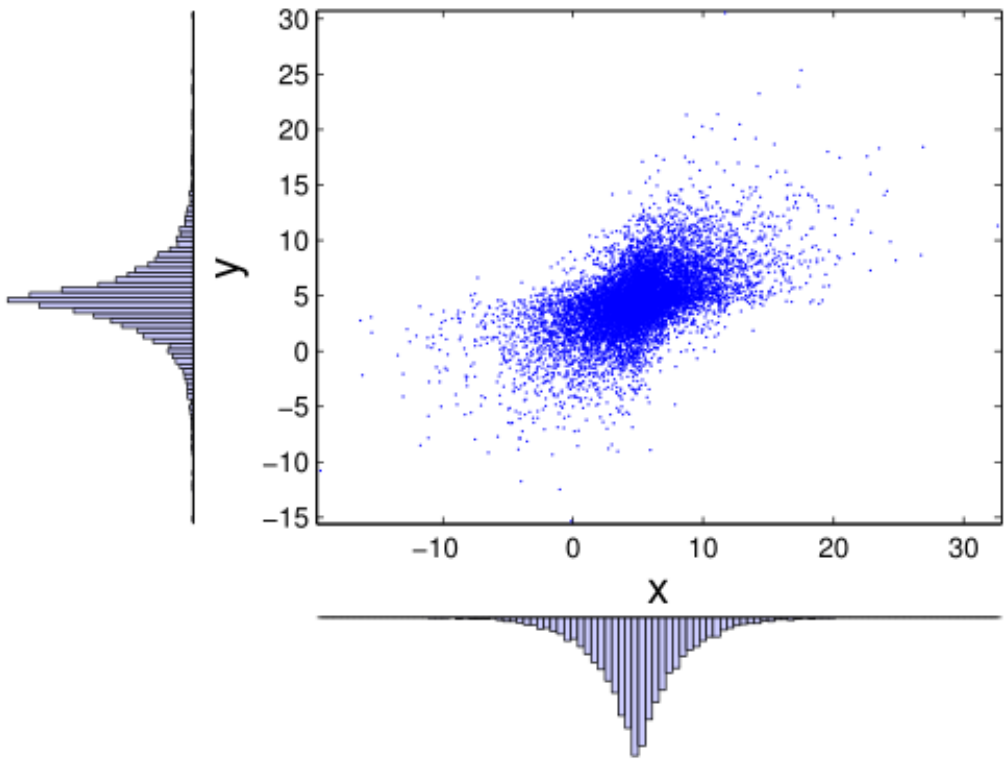


Figure 4.34: The scatter plot and marginal distribution of 10,000 2-dimensional data points following a Laplacian distribution. The data points are generated by applying the inverse CDF of the Laplacian distribution to the data from Figure 4.29. The two variables, x and y, are correlated ($r = 0.57$), and each follows a Laplacian distribution.

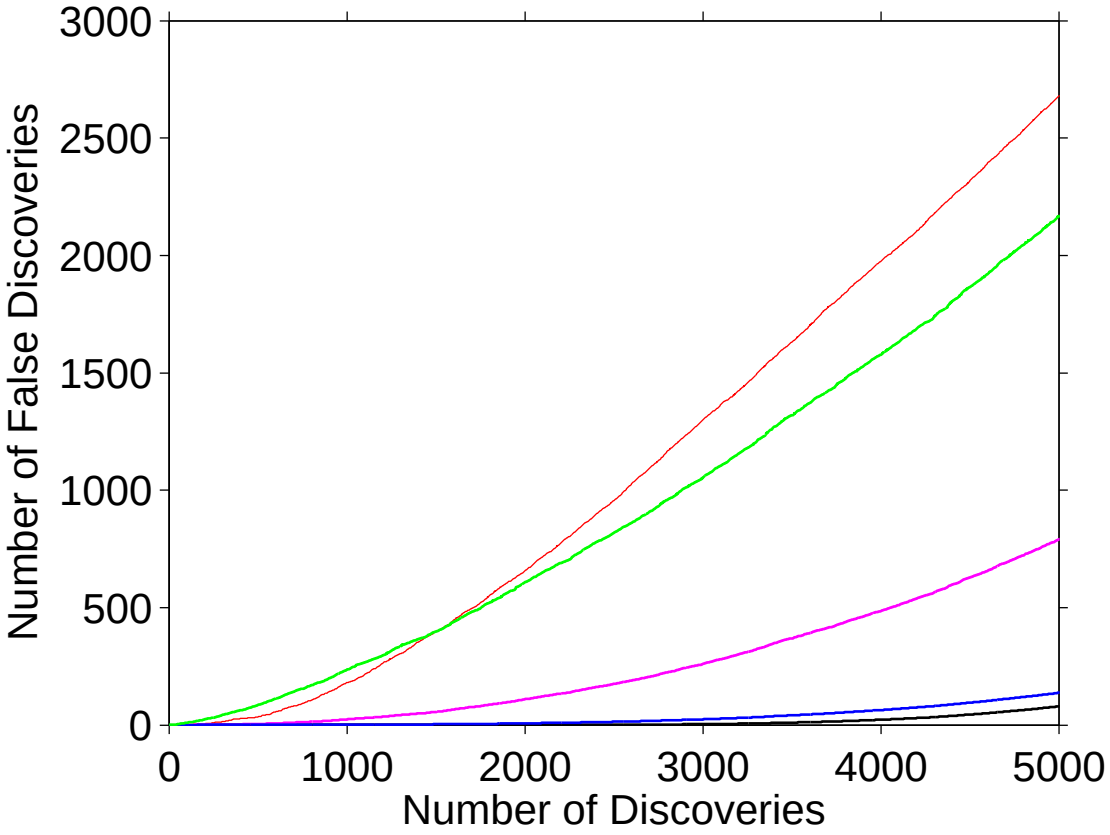


Figure 4.35: FDD curves of the FD estimation methods on `DataLaplaceRandomRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $\sigma_u = 0.3$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

inverse Fisher transformation of u:

$$\rho = \frac{e^{2u} - 1}{e^{2u} + 1},$$

where u is normally distributed with mean zero and standard deviation σ_u. In this way, ρ is bounded within [-1, 1], the valid range for correlation coefficients.

We generate simulated data, called `DataCubicRandomRho`, by incorporating this ρ into PROCEDURE genDataCubic. Specifically, in step 1(b) in Figure 4.6, ρ is chosen randomly as just described, so that it has a different value for each gene triple. Figure 4.37 shows the FDD curves of the FD estimation methods applied to this data. The bootstrap curve (green) is closest to the true FD curve (red), while the other curves deviate downward from the true curve considerably.

In our second approach, we choose ρ randomly from a population of real correlation coefficients. This population consists of all pairwise correlation coefficients of expression profiles (from the seed germination/dormancy data of Chapter 2) of all the 882 transcription factors in our seed dataset for *Arabidopsis*. This population of real ρ's has a similar distribution to the population of simulated ρ's, as shown in Figures 4.36 and 4.38. Figure 4.39 shows the FDD curves when real ρ's are used. Not surprisingly, the bootstrap curve is again the most accurate.

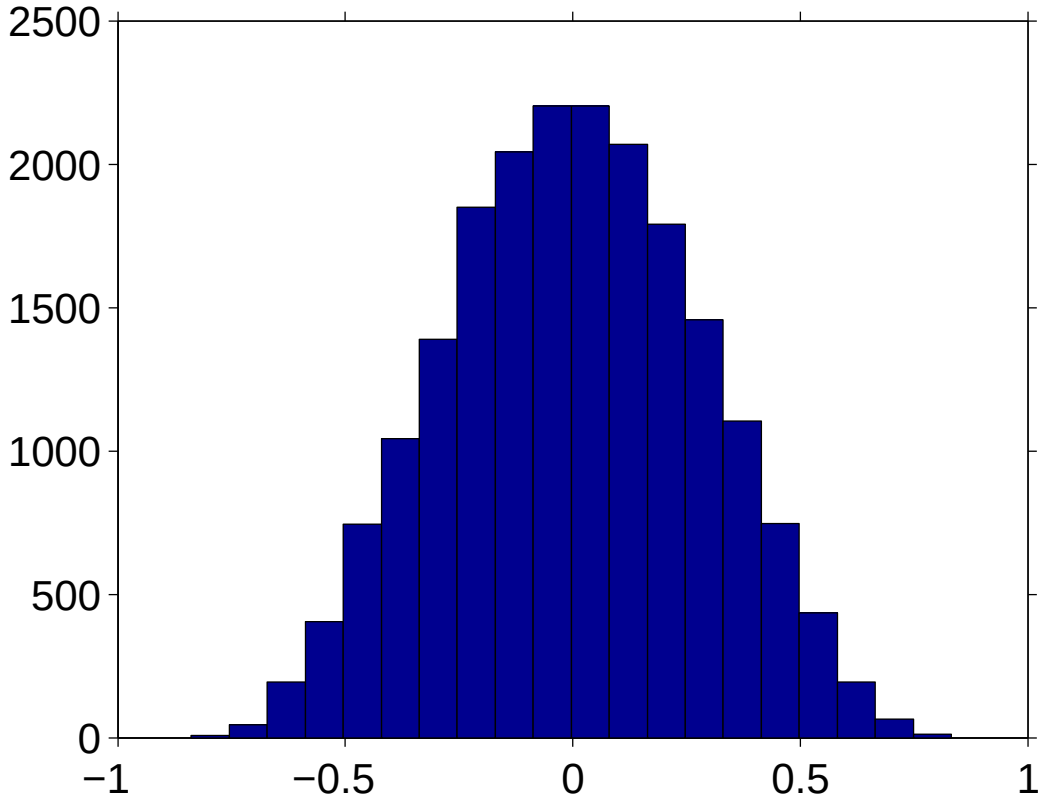


Figure 4.36: Distribution of 20,000 simulated correlation coefficients, ρ, where $\rho = (e^{2u} - 1)/(e^{2u} + 1)$, and $u \sim \mathcal{N}(0, \sigma_u)$ for $\sigma_u = 0.3$.

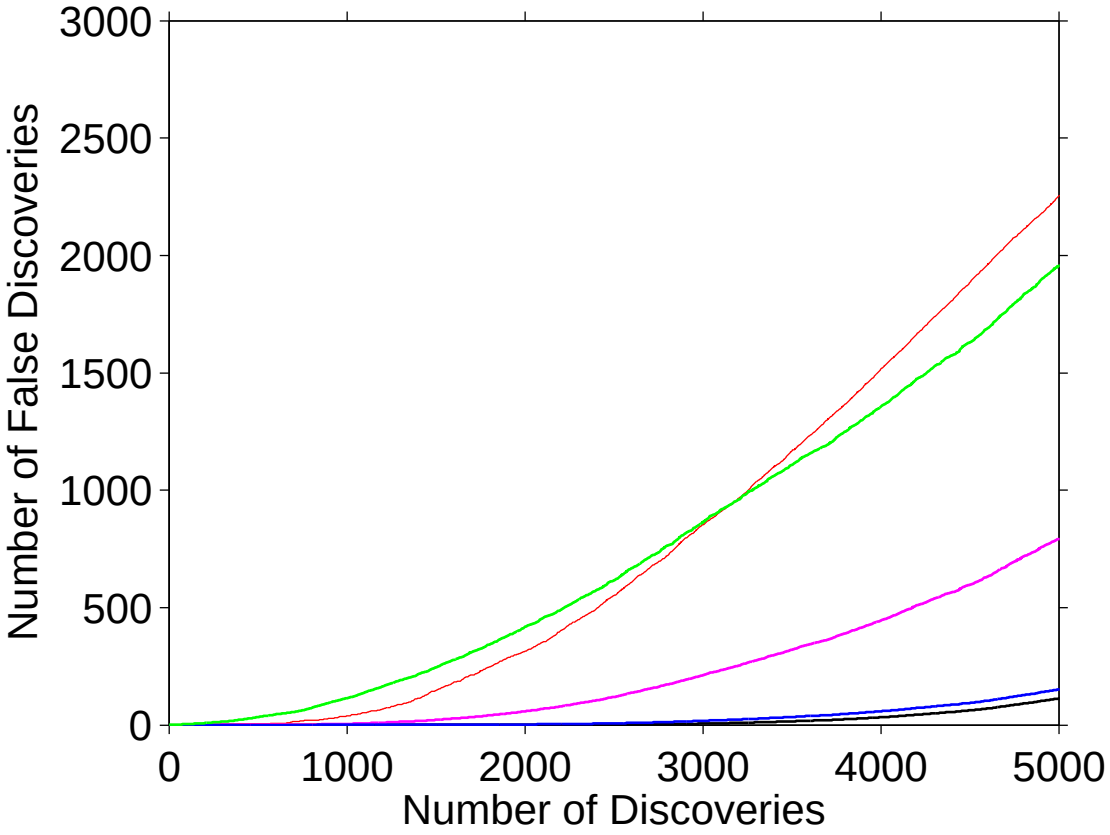


Figure 4.37: FDD curves of the FD estimation methods on `DataCubicRandomRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $\sigma_u = 0.3$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

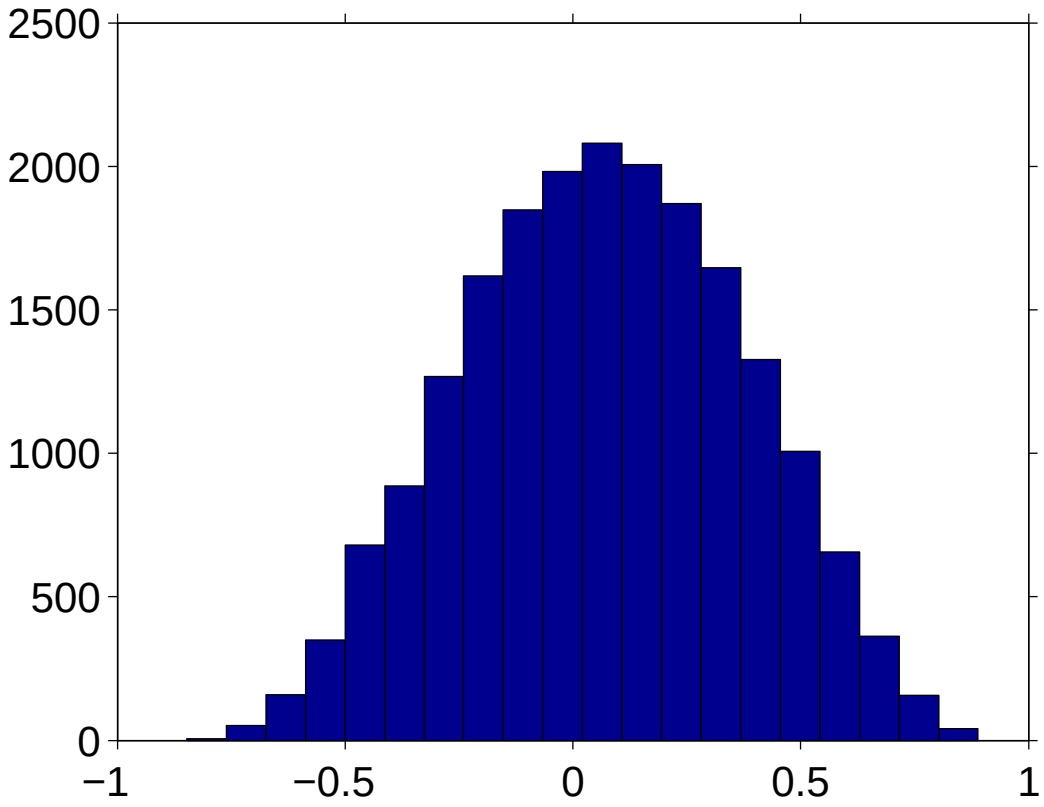


Figure 4.38: Distribution of 20,000 real correlation coefficients in the seed germination/dormancy data for *Arabidopsis*.

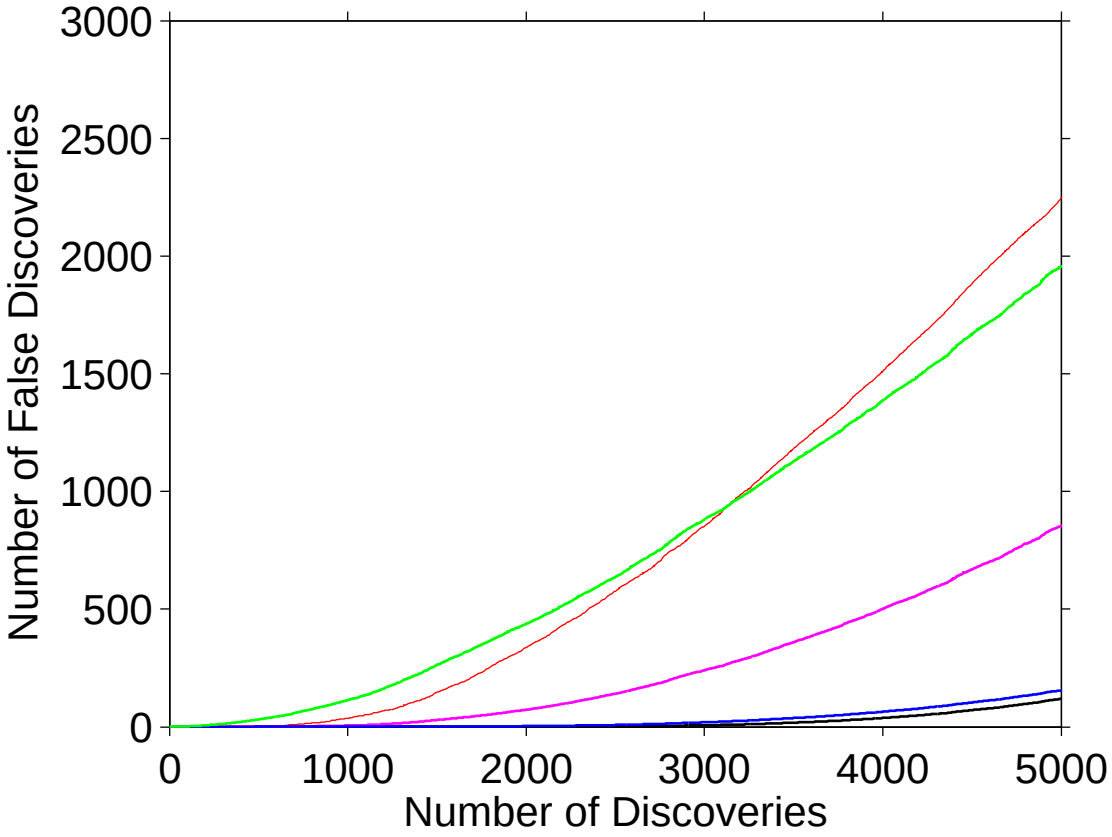


Figure 4.39: FDD curves of the FD estimation methods on `DataCubicRealRho`: $M_1 = 15000$, $M_2 = 5000$, $N = 100$, $nl_1 = 0.1$, $nl_2 = 0.1$ (multiplicative measurement noise). Bootstrap (green), partial permutation (pink), total permutation (blue), analytical t (black), true FD (red).

Chapter 5

Summary and future work

Conventional coexpression analysis looks at the expression levels of pairs of genes over a diverse set of conditions. Although such "condition independent" analyses provide useful generalized information, the coexpression of a pair of genes depends on the biological context, including environmental factors and the expression levels of other genes. Looking at gene coexpression in a "condition dependent" fashion should more precisely identify gene interactions relevant to answering more-specific biological questions [1]. This thesis addressed this issue in the specific context of seed germination in *Arabidopsis thaliana*, the model organism of plant biology.

We studied two aspects of the problem: (i) detecting two-way interactions, and (ii) detecting three-way interactions. Chapter 2 addresses the first problem by detecting two-way interactions using gene expression data, not from a diverse range of sources, but exclusively from imbibed mature seeds of *Arabidopsis* in a state of either germination or non-germination. The interactions thus detected are more accurate in answering specific biological questions about seed germination [1]. A coexpression network, SeedNet [1], now available online as a community resource, is constructed by including all two-way interactions whose strength of correlation exceeds some threshold. This threshold is chosen so that the network fits a scale-free graph as closely as possible. At this threshold, all edges in SeedNet have extremely high significance, as measured by FDR.

SeedNet consists of two main clusters, one associated with germination, and the other with non-germination. The thesis shows that the correlation between genes is not due to preferential expression, but due to correlation during seed germination and non-germination.

Correlation, and the clusters based on it, are therefore not a proxy for preferential expression, but reflect other factors, specifically biological processes that operate during germination and during non-germination. The thesis also examines the correlation and variance structure of the two clusters in more detail, revealing intriguing properties. We show, for example, that genes that are preferentially expressed during germination tend to be equally correlated during germination and non-germination. Thus, genes that are "turned up" during germination seem to work together in the same way during both germination and non-germination. Likewise for genes that are preferentially expressed during non-germination.

The thesis gives a detailed development of the methods used to construct and analyze SeedNet. The development includes mathematical proofs of results on covariance decomposition and preferential expression, and a robust algorithm for determining the optimum correlation threshold for approximating a scale-free graph. These methods form the computational contribution of [1]. Chapter 2 is also being submitted for publication as a separate paper.

In our second approach, detecting three-way interactions, the coexpression of a pair of genes can depend on the expression level of a third (unspecified) gene. That is, we search for triples of genes in which the expression level of one gene affects the coexpression of the other two. In this way, the thesis addresses another limitation of conventional coexpression analysis: it focusses on pairwise relationships between genes. Pairwise coexpression is clearly too simplistic to describe the complex relationships between gene expression levels, which can involve multiple genes and can vary depending on the biological context. In general, pairwise coexpression does not capture higher-order statistical dependencies or the complex biological relationships they reflect.

As a first step, Chapter 3 develops a quadratic regression model to detect three-way interactions in gene expression data, and applies it to data from *Arabidopsis* and yeast. The model is relatively straightforward and computationally inexpensive, enabling us to detect a vast number of possible three-way interactions in a reasonable amount of time and to compare FDR estimates by a variety of methods. The thesis provides both direct and circumstantial evidence demonstrating that the model detects real biological signals, validating its efficacy. We also show that discoveries made by the detector exhibit the expected correspondence between three-way interaction and transcriptionally-dependent coexpression; that is, when three genes

interact, the coexpression of two of them depends on the expression level of the third. Finally, we show that the detector works even when predictor genes are correlated (unlike [20]), though the FDR may increase.[1]

A particular challenge in discovering three-way interactions on a genome-wide scale is the large potential for false positives, since the number of possible three-way interactions grows as $O(N^3)$, as opposed to $O(N^2)$ for pairwise interactions, where N is the number of genes under study. Thus, a crucial step in the discovery process is accurately estimating FDR, since only if the FDR is low and stable can a discovery be confidently declared. However, as the thesis shows, extending the approaches used for two-way interactions can seriously underestimate the FDR for three-way interactions, sometimes by several orders of magnitude.

Estimating the FDR of three-way interactions in gene expression data faces two main challenges: (i) the underlying distribution of the data is unknown, and (ii) estimating the null distribution is considerably more subtle than for two-way interactions. The bootstrap is a well-known solution to the first problem, and this thesis explores its utility in addressing the second problem. In particular, Chapter 3 develops a method based on the bootstrap for estimating the FDR of our regression-based detector. Chapter 4 then tests the method and compares it to other methods used in the literature, including permutation tests and an analytical t-test. The tests show that these methods produce widely differing estimates of FDR on both real and simulated data, often differing by more than an order of magnitude. In particular, the bootstrap method consistently produces by far the largest estimates. This indicates that either the bootstrap method is overestimating or the other methods are underestimating the number of false discoveries.

It is impossible to determine which method is more accurate without knowing all the three-way interactions in a dataset. Since this is unknown for large biological datasets, we test the methods on simulated data, for which all interactions are known. The thesis develops numerous procedures for simulating data with a wide variety of statistical properties and provides a clear description of these procedures. While the simulations themselves are a highly simplified approximation of biological reality, they do provide strong evidence that the bootstrap method

[1] When two genes are correlated, it is difficult to unravel their effects on a third gene. In particular, regression is known to be difficult when predictor variables are correlated [27].

is the most accurate over a wide range of statistical conditions. In particular, while all the methods give good estimates on idealized data, the bootstrap method consistently gives the most reasonable estimates on more complex data, while the other methods rapidly break down, consistently underestimating the true number of false discoveries, often by more than an order of magnitude. In sum, our bootstrap method provides the best available estimate of FDR for three-way interactions over a wide range of conditions.

In addition, the simulation procedures themselves can be regarded as a community resource for generating benchmark data for testing methods that estimate the FDR of three-way interactions in gene expression data. Such data is crucial for comparing FDR estimation methods and estimating their accuracy.

Finally, for the same *Arabidopsis* data used in SeedNet, our detector of three-way interactions discovers a large number of high-confidence three-way interactions, from which over 64,000 new edges can be added to SeedNet, significantly enlarging the set of edges in SeedNet. These new edges usually do not represent high pairwise correlation, and consequently are out of reach by conventional means, but they fall well within the spectrum of our three-way interaction detector, which instead looks for transcriptionally-dependent correlations. This greatly extends the usability of gene expression data as it reveals a large number of relationships that cannot be detected by traditional coexpression analysis. The material in Chapters 3 and 4, on three-way interactions, is being submitted for publication.

Several possibilities can be explored in the future. The most tangible one is that we can continue the current study on the bootstrap method for estimating FDR by using a combination of our data simulation models. For example, for some gene triples, the two predictor genes could have a quadratic relationship with the target gene, while for other triples, the relationship could be cubic. More generally, when generating simulated data, the gene triples could have a mixture of the relationships described in Chapter 4. Also, from a practical point of view, it would more directly benefit the bioinformatics community to develop easy-to-use software packages for detecting interactions and estimating their FDR, either in the form of application program interfaces (APIs), a stand-alone program, or online web tools.

While the multi-way interactions explored in this thesis are mostly three-way interactions

having one target gene and two predictor genes, we can consider extending the second-order model to higher-order models to capture four-way interactions with three predictor genes, five-way interactions with four predictor genes, etc. Unfortunately, this would dramatically increase the number of potential hypothesis tests. To address this, we could first cluster genes, and then, for each small cluster select a representative gene (profile), then detect multi-way interactions for these representatives. This can be thought of as multi-way interactions between groups. Another way to build higher-order models is to join two existing three-way interactions by letting the target gene of one triple be a predictor gene of another. (In contrast, we cannot simply join two two-way interactions to form a three-way interaction.) Furthermore, can we devise a unified framework which enables us to harvest 2,3,4,5,...-way interactions simultaneously? Then how do we tailor the bootstrap method to estimate FDR?

The bootstrap method of estimating FDR developed in this thesis was applied to 3-way interactions. In principle, it could be adapted for 2-way interactions and compared to the permutation-based method used in Chapter 2. However, the 2-way detector of Chapter 2 is based on correlation coefficients, while the 3-way detector of Chapter 3 is based on regression. Applying the bootstrap method of estimating FDR to 2-way interactions would require reformulating the 2-way detection problem in terms of regression.

While the second-order model developed in Chapter 3 enjoys simplicity and speed, thus enabling us to test hundreds of millions of hypotheses within a reasonable amount of time, we certainly can consider more-complex but more computationally expensive models. For example, some undirected graphical models have found applications in learning the structure of dependency for a small number of discrete (response) variables [118, 119]. However, just to do so already requires using nonstandard optimization and using cross-validation to tune their regularization parameters [118], or imposes restrictions on selecting subsets (of variables) [119]. Furthermore, how can we adapt these models so that they can discover multi-way interactions in gene expression data? Do these models scale well to a large number of genes? More importantly, how do we estimate FDR? All these questions are worth exploring.

Some guidelines are already available. For example, detecting three-way interactions on a genome-wide scale requires simple models, because of the vast number of possible triples. However, the number of triples decreases rapidly as the number of genes decreases, so more-

complex models could be used if the number of genes of interest is small (e.g., hundreds instead of thousands). The bootstrap can then be used to estimate p-values, from which an FDR is easily estimated (Section 3.3.2.2). Moreover, if an analytical formula (like the t test in Section 3.3.2.1) is available for estimating p-values assuming the expression data is Gaussian, then a bootstrap version of the method may provide good p-value estimates for real (non-Gaussian) data. This approach was taken in Section 3.3.2.2.

For the two-way interactions of Chapter 2, there are a number of other correlation measures that could be tried, e.g., mutual information [35], Kendall rank correlation [120] and Spearman rank correlation [120]. For instance, although Spearman rank correlation throws away information, it might find non-linear monotonic relationships that are not detected well by Pearson correlation. The statistical significance (FDR) of these correlations could also be estimated by permutation tests. However, the biological significance of the results would have to be confirmed by additional laboratory experiments.

Another avenue for future research is to further develop the simulation procedures of Chapter 4, with the aim of capturing a wider range of statistical properties, especially those found in real gene expression data. For example, in simulating interacting gene triples, the procedures randomly choose values for the interaction coefficients (e.g., coefficient f in the interaction term $\hat{f}x_{i1}x_{i2}$ in Figure 4.1). In this thesis, these coefficients were chosen from a uniform distribution on [-0.5,0.5]. This means that the coefficients can in principle be 0. If all the interaction coefficients in an interacting triple are 0, then the triple is not interacting at all. Fortunately, this situation has probability 0. Nevertheless, there is a non-zero probability of choosing interaction coefficients that are all very close to 0. Such triples represent very weak three-way interactions, and one can ask whether they should count as real three-way interactions at all. This, of course, is a biological question: how weak does a three-way interaction have to be before it is no longer of biological interest. In some ways the question is mute, since in our simulations, the number of very-weak three way interactions is miniscule compared to the number of interacting and non-interacting gene triples, and their effect on the results is correspondingly tiny. Nevertheless, one could certainly consider modifying the simulation procedures so that for interacting triples, there is zero probability of choosing a value for an interaction coefficient in a small interval around 0. This raises the question of how small the interval should be, which is precisely the

biological question mentioned above. Since the question may not have a definitive answer, one could simulate many data sets, each with a different-sized interval around 0. For each such data set, one would estimate the FDR using a variety of methods, as in Chapter 4, and see if interval size has any effect on the results, and if so, whether the bootstrap method is still uniformly the most accurate.

Lastly, another avenue of future research is to investigate whether the 64,000 new edges added to SeedNet can be used to predict gene function in the traditional manner (guilt by association) or by looking for hub genes. Perhaps they can also be used to help find potential combinatorial regulation (as in Section 3.4.3.3). More generally, for any method that uses coexpression edges to make biological inferences, it may be possible to adapt the method to use new edges inferred from three-way interactions in gene expression data.

Appendix A

Appendix for Chapter 2

A.1 Covariance decomposition

Let x_{ij} and y_{ij} be real numbers, for $i = 1 \cdots m$ and $j = 1 \cdots n_i$. Intuitively, there are m groups (e.g., m types of seeds) and x_{ij} and y_{ij} are a pair of measurements on group i (e.g., x_{ij} is the expression level of a gene on seed j in group i. Likewise for y_{ij}.) We let $N = \sum_i n_i$ be the total number of measurements. This appendix shows that the total covariance of the x_{ij} and y_{ij} can be decomposed into a sum of covariances within groups and between groups.

Lemma 1. *For each i,*

$$\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}) \;=\; \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \;+\; n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

where $\bar{x}_{\bullet\bullet} = \sum_{ij} x_{ij}/N$ is the average of x_{ij}, and $\bar{x}_{i\bullet} = \sum_j x_{ij}/n_i$ is the average of x_{ij} on group i (and likewise for $\bar{y}_{\bullet\bullet}$ and $\bar{y}_{i\bullet}$).

Proof.

$$\begin{aligned}
&\sum_j (x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}) \\
&= \sum_j [(x_{ij} - \bar{x}_{i\bullet}) + (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})][(y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})] \\
&= \sum_j [(x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \;+\; (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \\
&\quad + \; (x_{ij} - \bar{x}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \;+\; (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{i\bullet})] \\
&= \sum_j (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \;+\; n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \\
&\quad + \; (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \sum_j (x_{ij} - \bar{x}_{i\bullet}) \;+\; (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) \sum_j (y_{ij} - \bar{y}_{i\bullet}) \\
&= \sum_j (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) \;+\; n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})
\end{aligned}$$

since $\sum_j (x_{ij} - \bar{x}_{i\bullet}) \;=\; n_i \bar{x}_{i\bullet} - n_i \bar{x}_{i\bullet} \;=\; 0$, and likewise for $\sum_j (y_{ij} - \bar{y}_{i\bullet})$. □

Corollary 1. $Cov_{total} \;=\; Cov_{within} + Cov_{between}$, *where*

$$\begin{aligned}
Cov_{total} &= \frac{1}{N} \sum_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}) && \textit{is the total covariace} \\
Cov_{within} &= \frac{1}{N} \sum_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet}) && \textit{is the covariance within groups} \\
Cov_{between} &= \frac{1}{N} \sum_i n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) && \textit{is the covariance between groups}
\end{aligned}$$

In the special case in which $x_{ij} = y_{ij}$, we get the following standard result of analysis of variance (ANOVA) about the variance of x_{ij} [114].

Corollary 2. $Var_{total} \;=\; Var_{within} + Var_{between}$, *where*

$$\begin{aligned}
Var_{total} &= \frac{1}{N} \sum_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\bullet\bullet})^2 && \textit{is the total variance} \\
Var_{within} &= \frac{1}{N} \sum_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})^2 && \textit{is the variance within groups} \\
Var_{between} &= \frac{1}{N} \sum_i n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 && \textit{is the variance between groups}
\end{aligned}$$

When there are only two groups, the expression $\bar{x}_{\bullet\bullet}$ can be eliminated, leading to specialized results, which are used in Section 2.3.2.1 and Appendix A.2.

Lemma 2. *If $m = 2$, then*

$$\begin{aligned}
\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet} &= n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N \\
\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet} &= n_1(\bar{x}_{2\bullet} - \bar{x}_{1\bullet})/N
\end{aligned}$$

Proof. Recalling that $N = n_1 + n_2$,

$$\begin{aligned}
\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet} &= \bar{x}_{1\bullet} - \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}/N \\
&= \bar{x}_{1\bullet} - \sum_{i=1}^{m} n_i\bar{x}_{i\bullet}/N \\
&= \bar{x}_{1\bullet} - (n_1\bar{x}_{1\bullet} + n_2\bar{x}_{2\bullet})/(n_1 + n_2) \\
&= n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/(n_1 + n_2) \\
&= n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N
\end{aligned}$$

The proof of the second equation is similar. □

Corollary 3. *If $m = 2$, then*

$$Var_{between} = \frac{n_1 n_2}{N^2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2$$

Proof.

$$\begin{aligned}
Var_{between} &= \frac{1}{N}\sum_{i=1}^{m} n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 \\
&= \frac{1}{N}[n_1(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + n_2(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2] \\
&= \frac{1}{N}[n_1 n_2^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2/N^2 + n_2 n_1^2(\bar{x}_{2\bullet} - \bar{x}_{1\bullet})^2/N^2] \\
&= \frac{1}{N^3} n_1 n_2(n_1 + n_2)(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2 \\
&= \frac{n_1 n_2}{N^2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2
\end{aligned}$$

since $N = n_1 + n_2$. □

A.2 Preferential expression

As described in Section 2.3.1.1, we use gene significance [7] (abbreviated as GS) to quantify the preferential expression of a gene. Formally, GS is the correlation coefficient between gene expression levels and binary sample traits. In our case, the samples are seeds, and the traits are "germinating" and "non-germinating." Intuitively, GS is high if the gene's expression levels tend to be high on germinating seeds and low on non-germinating seeds. This section defines GS formally and shows that it is proportional to the difference in the mean expression level on germinating seeds and the mean expression level on non-germinating seeds, a result that is needed in Section 2.3.2.2. It follows as a corollary that GS is equivalent to a t statistic for the difference in mean expression levels, a result that we include for completeness.

As in Section 2.3.2.1, we divide the seeds into two groups, with group 1 consisting of germinating seeds, and group 2 consisting of non-germinating seeds. The seeds in group 1 are labelled $1, 2, ..., n_1$, and the seeds in group 2 are labelled $1, 2, ..., n_2$. We let $N = n_1 + n_2$ be the total number of seeds. Given a gene, we let x_{ij} denote it's expression level in seed j of group i. We also let y_{ij} denote the trait of seed j in group i. That is, $y_{ij} = 1$ for germinating seeds, and -1 for non-germinating seeds. Consequently, $y_{1j} = 1$ and $y_{2j} = -1$, for all j. The averages $\bar{x}_{i\bullet}$, $\bar{y}_{i\bullet}$, $\bar{x}_{\bullet\bullet}$ and $\bar{y}_{\bullet\bullet}$ are defined as in Lemma 1 of Appendix A.1.

Gene significance is defined as the Pearson correlation coefficient of the x_{ij} and the y_{ij} [7].

Definition 1. *$GS = S_{xy}/\sqrt{S_x S_y}$ where*

$$\begin{aligned} S_{xy} &= \sum_{ij}(x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}) \\ S_x &= \sum_{ij}(x_{ij} - \bar{x}_{\bullet\bullet})^2 \\ S_y &= \sum_{ij}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \end{aligned}$$

Because y_{ij} has a special form, we can simplify the expressions for S_{xy} and S_y.

Lemma 3. $S_{xy} = 2n_1 n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N$

Proof. First recall that $\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet} = n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N$, by Lemma 2 in Appendix A.1. Thus, since $y_{1j} = 1$,

$$\begin{aligned}\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{\bullet\bullet})(y_{1j} - \bar{y}_{\bullet\bullet}) &= (1 - \bar{y}_{\bullet\bullet})\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{\bullet\bullet}) \\ &= (1 - \bar{y}_{\bullet\bullet})n_1(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet}) \\ &= (1 - \bar{y}_{\bullet\bullet})n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N\end{aligned}$$

Likewise, since $y_{2j} = -1$,

$$\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{\bullet\bullet})(y_{2j} - \bar{y}_{\bullet\bullet}) = (1 + \bar{y}_{\bullet\bullet})n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N$$

Adding these two expressions, we get

$$\begin{aligned}S_{xy} &= \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet}) \\ &= \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{\bullet\bullet})(y_{1j} - \bar{y}_{\bullet\bullet}) \;+\; \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{\bullet\bullet})(y_{2j} - \bar{y}_{\bullet\bullet}) \\ &= (1 - \bar{y}_{\bullet\bullet})n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N \;+\; (1 + \bar{y}_{\bullet\bullet})n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N \\ &= 2n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N\end{aligned}$$

□

Lemma 4. $S_y = 4n_1n_2/N$

Proof. First note that $\bar{y}_{\bullet\bullet} = (n_1 - n_2)/N$, since $y_{1j} = 1$ and $y_{2j} = -1$. Hence,

$$\begin{aligned}
S_y &= \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n_1}(1 - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2}(-1 - \bar{y}_{\bullet\bullet})^2 \qquad \text{since } y_{1j} = 1 \text{ and } y_{2j} = -1 \\
&= \sum_{j=1}^{n_1}(1 - 2\bar{y}_{\bullet\bullet} + \bar{y}_{\bullet\bullet}^2) \; + \; \sum_{j=1}^{n_2}(1 + 2\bar{y}_{\bullet\bullet} + \bar{y}_{\bullet\bullet}^2) \\
&= (n_1 + n_2) + 2(n_2 - n_1)\bar{y}_{\bullet\bullet} + (n_1 + n_2)\bar{y}_{\bullet\bullet}^2 \\
&= (n_1 + n_2) + 2(n_2 - n_1)(n_1 - n_2)/N + (n_1 + n_2)(n_1 - n_2)^2/N^2 \\
&= (n_1 + n_2)^2/N - 2(n_1 - n_2)^2/N + (n_1 - n_2)^2/N \qquad \text{since } N = n_1 + n_2 \\
&= (n_1 + n_2)^2/N - (n_1 - n_2)^2/N \\
&= 4n_1n_2/N
\end{aligned}$$

□

Having simplified S_{xy} and S_y, we can now simplify the expression for GS.

Corollary 4.

$$GS = \alpha\frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet}}{\sqrt{Var_{total}}}$$

where $Var_{total} = S_x/N$ is the variance in the expression level, and $\alpha = \sqrt{n_1n_2}/N$ is a positive constant.

Proof. By the above lemmas,

$$GS = \frac{S_{xy}}{\sqrt{S_xS_y}} = \frac{2n_1n_2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/N}{\sqrt{S_x\, 4n_1n_2/N}} = \frac{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})\sqrt{n_1n_2}/N}{\sqrt{S_x/N}}$$

□

Thus, the gene significance, GS, is proportional to the drop in mean expression level from

group 1 to group 2 ($\bar{x}_{1\bullet} - \bar{x}_{2\bullet}$) relative to the total variation in expression levels (Var_{total}). This is a measure of the significance of the drop. It is closely related to a more well-known measure of significance, the t test statistic, as shown in the next corollary. This corollary and its proof use the notation for variance defined in Corollary 2 of Appendix A.1.

Corollary 5. *$GS = \alpha t/\sqrt{1+(\alpha t)^2}$ where $t = (\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/\sqrt{Var_{within}}$ and $\alpha = \sqrt{n_1 n_2}/N$.*

Proof.

$$\begin{aligned}
GS^2 &= \alpha^2 \frac{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{Var_{total}} && \text{by Corollary 4} \\
&= \frac{\alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{Var_{within} + Var_{between}} && \text{by Corollary 2} \\
&= \frac{\alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{Var_{within} + (\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2 n_1 n_2/N^2} && \text{by Corollary 3} \\
&= \frac{\alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{Var_{within} + \alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2} \\
&= \frac{\alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2/Var_{within}}{1 + \alpha^2(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2/Var_{within}} \\
&= \frac{\alpha^2 t^2}{1+\alpha^2 t^2}
\end{aligned}$$

□

The important point here is that $(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})/\sqrt{V_{within}}$ is proportional to a t statistic for the difference of two population means [114]. In our case, it measures the significance of the difference in mean expression levels on germinating and non-germinating seeds. Equally important, the function $h(t) = \alpha t/\sqrt{1+(\alpha t)^2}$ is monotonically increasing. Thus, the statistic GS is *equivalent* to the t statistic. In particular, $t > t_0$ if and only if $GS > h(t_0)$, for any threshold t_0. Thus, any test on t can be converted to a test on GS, and vice versa.

A.3 Degree distributions in geometric random graphs

This appendix shows that geometric random graphs are a poor model for coexpression networks such as SeedNet. In a geometric random graph, $G(n, r)$, n nodes are independently and uniformly distributed in a d-dimensional unit cube, and any two nodes are joined by an edge if the Euclidean distance between them is less than r. For example, $G(1000, 0.1)$ in 2-dimensional

space contains 1000 nodes embedded in a unit square, and two nodes have an edge if their Euclidean distance is less than 0.1. We look at how well geometric random graphs model the node-degree distribution of coexpression networks derived from the SeedNet data using a variety of correlation cutoffs, τ. Figure A.1 shows histograms of node-degree distribution for $G(n, r)$ in 2-dimensional space for $n = 14,088$ and several values of r. Here, 14088 is the number of genes in the SeedNet data, and r is chosen so that the average node degree in $G(n, r)$ is roughly equal to the average node degree in one of the coexpression networks. Likewise, Figure A.2 and Figure A.3 show degree distributions for geometric random graphs in 3-dimensional space and 4-dimensional space, respectively. These histograms are vastly different from those in Figure 2.20 and the right-hand side of Figure 2.21, which show the true node-degree distribution of the coexpression networks. In particular, the true histograms are sharply peaked at the left, decrease monotonically from left to right, and are convex. In contrast, the histograms for geometric random graphs are peaked at the right, increase from left to right before suddenly plunging to zero, and three of them are highly non-convex, having two plateaux and an up-down peak, and the fourth one is concave.

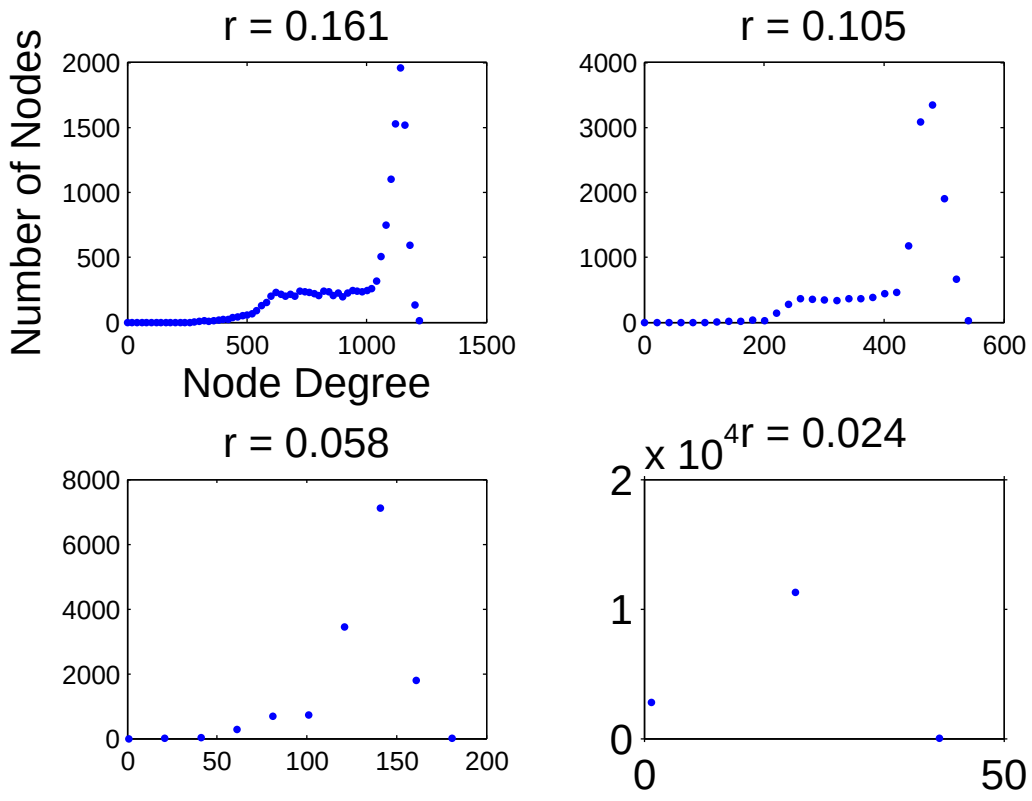


Figure A.1: Histograms of degree distribution in four geometric random graphs in 2-dimensional space, $G(14088, 0.161)$, $G(14088, 0.105)$, $G(14088, 0.058)$, and $G(14088, 0.024)$. The values of the second parameter in G, $r = 0.161, 0.105, 0.058$, and 0.024, are chosen such that the resulting graphs have roughly the same average node degree as the coexpression networks with correlation cutoffs $\tau = 0.5, 0.6, 0.7$, and 0.8, respectively.

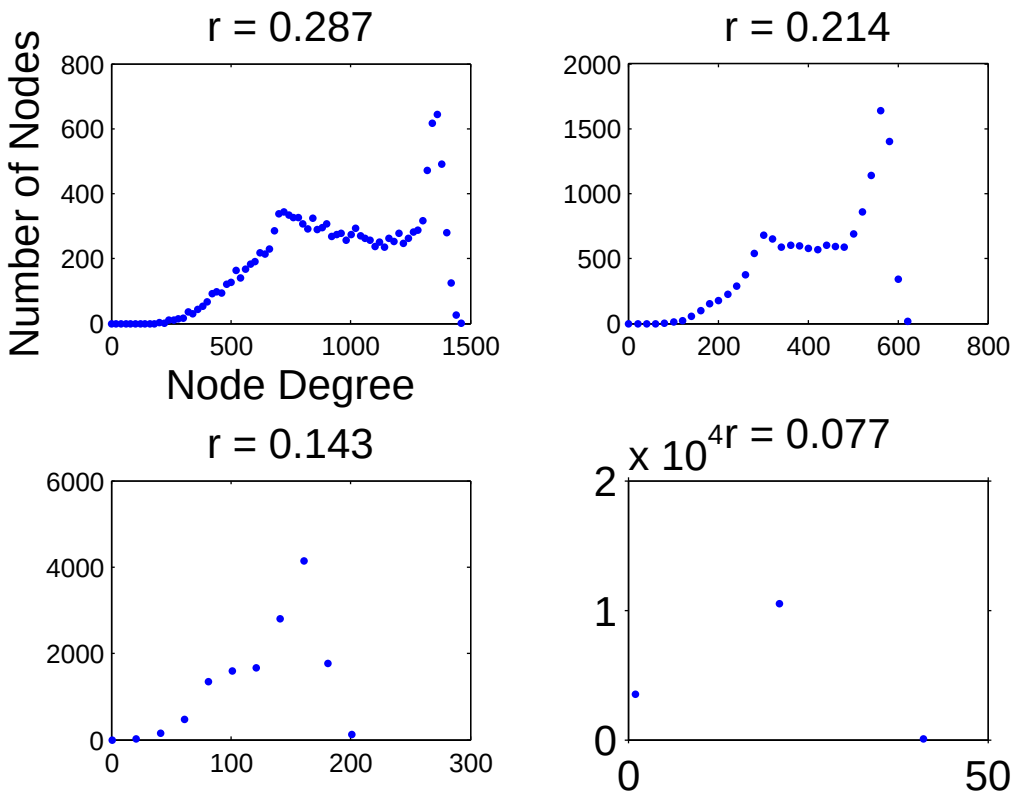


Figure A.2: Similar to Figure A.1, but the geometric random graphs are in 3-dimensional space.

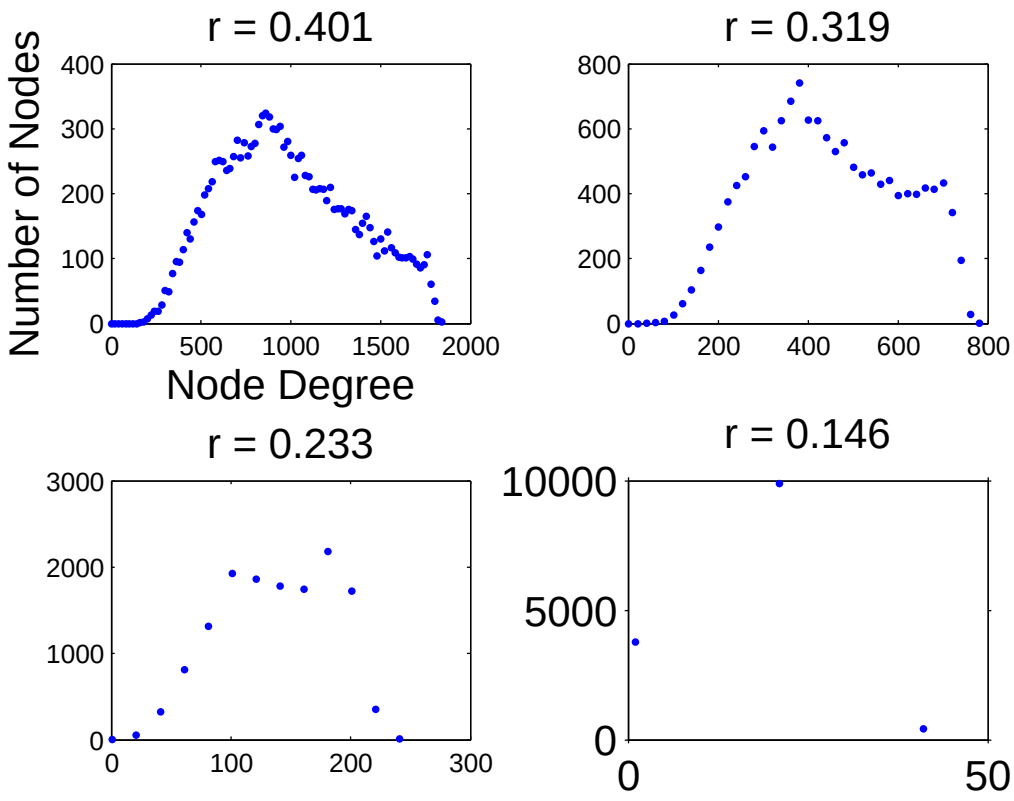


Figure A.3: Similar to Figure A.1, but the geometric random graphs are in 4-dimensional space.

Bibliography

[1] G. W. Bassel, H. Lan, E. Glaab, D. J. Gibbs, T. Gerjets, N. Krasnogor, A. J. Bonner, M. J. Holdsworth, and N. J. Provart, "Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 23, pp. 9709–9714, 2011.

[2] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of National Academy of Sciences of the United States of America*, vol. 95(25), pp. 14863–14868, 1998.

[3] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, *et al.*, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[4] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks," *BMC Bioinformatics*, vol. 6, no. 1, p. 227, 2005.

[5] K. Aoki, Y. Ogata, and D. Shibata, "Approaches for extracting practical information from gene co-expression networks in plant biology," *Plant and Cell Physiology*, vol. 48, no. 3, pp. 381–390, 2007.

[6] K. Saito, M. Y. Hirai, and K. Yonekura-Sakakibara, "Decoding genes with coexpression networks and metabolomics–majority report by precogs," *Trends in Plant Science*, vol. 13, no. 1, pp. 36–43, 2008.

[7] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[8] B. Usadel, T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. J. Provart, "Co-expression tools for plant biology: opportunities for hypothesis generation and caveats," *Plant, Cell & Environment*, vol. 32, no. 12, pp. 1633–1651, 2009.

[9] K. Mochida and K. Shinozaki, "Advances in omics and bioinformatics tools for systems analyses of plant functions," *Plant and Cell Physiology*, vol. 52, no. 12, pp. 2017–2038, 2011.

[10] M. T. Weirauch, "Gene coexpression networks for the analysis of DNA microarray data," *Applied Statistics for Network Biology: Methods in Systems Biology*, pp. 215–250, 2011.

[11] I. Lee, B. Ambaru, P. Thakkar, E. M. Marcotte, and S. Y. Rhee, "Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana," *Nature Biotechnology*, vol. 28, no. 2, pp. 149–156, 2010.

[12] J. D. Bewley, "Seed germination and dormancy.," *The Plant Cell*, vol. 9, no. 7, pp. 1055–1066, 1997.

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[14] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?," *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[15] A. Blais and B. D. Dynlacht, "Constructing transcriptional regulatory networks," *Genes & Development*, vol. 19, no. 13, pp. 1499–1511, 2005.

[16] M. Hansen, L. Everett, L. Singh, and S. Hannenhalli, "Mimosa: mixture model of co-expression to detect modulators of regulatory interaction.," *Algorithms for Molecular Biology*, vol. 5, no. 4, 2010.

[17] S. Istrail and E. H. Davidson, "Logic functions of the genomic cis-regulatory code," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4954–4959, 2005.

[18] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6286–6291, 2010.

[19] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[20] K. Wang, M. Saito, B. C. Bisikirska, M. J. Alvarez, W. K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. A. Margolin, *et al.*, "Genome-wide identification of post-translational modulators of transcription factor activity in human B-cells," *Nature Biotechnology*, vol. 27, no. 9, pp. 829–837, 2009.

[21] J. Zhang, Y. Ji, and L. Zhang, "Extracting three-way gene interactions from microarray data," *Bioinformatics*, vol. 23, no. 21, pp. 2903–2909, 2007.

[22] 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*, vol. 57, no. 1, pp. 289–300, 1995.

[23] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, vol. 57. Florida: CRC Press, 1994.

[24] K.-C. Li, "Genome-wide coexpression dynamics: theory and application," *Proceedings of the National Academy of Sciences*, vol. 99, no. 26, pp. 16875–16880, 2002.

[25] E. Nambara, R. Hayama, Y. Tsuchiya, M. Nishimura, H. Kawaide, Y. Kamiya, and S. Naito, "The role of abi3 and fus3 loci in arabidopsis thaliana on phase transition from late embryo development to germination," *Developmental Biology*, vol. 220, no. 2, pp. 412–423, 2000.

[26] S. Horvath and P. Langfelder, "Tutorial for the WGCNA package for R." 2011.

[27] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, vol. 2. New York: Springer, 2009.

[28] G. Strang, *Linear Algebra and Its Applications (4th Edition)*. Brooks Cole, 2005.

[29] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*. New York: Garland Science, 2013.

[30] K. Miura, A. Rus, A. Sharkhuu, S. Yokoi, A. S. Karthikeyan, K. G. Raghothama, D. Baek, Y. D. Koo, J. B. Jin, R. A. Bressan, *et al.*, "The Arabidopsis SUMO E3 ligase SIZ1 controls phosphate deficiency responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7760–7765, 2005.

[31] P. M. Bowers, S. J. Cokus, D. Eisenberg, and T. O. Yeates, "Use of logic relationships to decipher protein network organization," *Science*, vol. 306, no. 5705, pp. 2246–2249, 2004.

[32] M. A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, no. 2, pp. 185–198, 2004.

[33] J. L. Riechmann, "Transcriptional regulation: a genomic overview," *The Arabidopsis Book/American Society of Plant Biologists*, vol. 1, 2002.

[34] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.

[35] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5, pp. 418–429, 2000.

[36] P. Buzkova, T. Lumley, and K. Rice, "Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions," *Annals of Human Genetics*, vol. 75, no. 1, pp. 36–45, 2011.

[37] J. Fox, *Applied regression analysis and generalized linear models*. California: Sage Publications, 2008.

[38] P. Hall and S. R. Wilson, "Two guidelines for bootstrap hypothesis testing," *Biometrics*, vol. 47, no. 2, pp. 757–762, 1991.

[39] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, vol. 821. New Jersey: Wiley, 2012.

[40] R. V. Davuluri, H. Sun, S. K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, and E. Grotewold, "AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors," *BMC Bioinformatics*, vol. 4, no. 1, p. 25, 2003.

[41] E. Huala, A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, *et al.*, "The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant," *Nucleic Acids Research*, vol. 29, no. 1, pp. 102–105, 2001.

[42] K. S. Heyndrickx and K. Vandepoele, "Systematic identification of functional plant modules through the integration of complementary data sources," *Plant Physiology*, vol. 159, no. 3, pp. 884–901, 2012.

[43] W. Y. Lee, D. Lee, W.-I. Chung, and C. S. Kwon, "Arabidopsis ING and Alfin1-like protein families localize to the nucleus and bind to H3K4me3/2 via plant homeodomain fingers," *The Plant Journal*, vol. 58, no. 3, pp. 511–524, 2009.

[44] J. Riechmann, J. Heard, G. Martin, L. Reuber, C.-Z. Jiang, J. Keddie, L. Adam, O. Pineda, O. Ratcliffe, R. Samaha, *et al.*, "Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes," *Science*, vol. 290, no. 5499, pp. 2105–2110, 2000.

[45] M. M. Monfared, M. K. Simon, R. J. Meister, I. Roig-Villanova, M. Kooiker, L. Colombo, J. C. Fletcher, and C. S. Gasser, "Overlapping and antagonistic activities of BASIC PENTACYSTEINE genes affect a range of developmental processes in Arabidopsis," *The Plant Journal*, vol. 66, no. 6, pp. 1020–1031, 2011.

[46] J. W. Cutcliffe, E. Hellmann, A. Heyl, and A. M. Rashotte, "CRFs form protein–protein interactions with each other and with members of the cytokinin signalling pathway in Arabidopsis via the CRF domain," *Journal of Experimental Botany*, vol. 62, no. 14, pp. 4995–5002, 2011.

[47] R. G. Franks, C. Wang, J. Z. Levin, and Z. Liu, "SEUSS, a member of a novel family of plant regulatory proteins, represses floral homeotic gene expression with LEUNIG," *Development*, vol. 129, no. 1, pp. 253–263, 2002.

[48] V. Gregis, A. Sessa, L. Colombo, and M. M. Kater, "AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in Arabidopsis," *The Plant Cell Online*, vol. 18, no. 6, pp. 1373–1382, 2006.

[49] V. V. Sridhar, A. Surendrarao, and Z. Liu, "APETALA1 and SEPALLATA3 interact with SEUSS to mediate transcription repression during flower development," *Development*, vol. 133, no. 16, pp. 3159–3166, 2006.

[50] F. Bao, S. Azhakanandam, and R. G. Franks, "SEUSS and SEUSS-LIKE transcriptional adaptors regulate floral and embryonic development in Arabidopsis," *Plant Physiology*, vol. 152, no. 2, pp. 821–836, 2010.

[51] I. Kasajima, Y. Ide, M. Yokota Hirai, and T. Fujiwara, "WRKY6 is involved in the response to boron deficiency in Arabidopsis thaliana," *Physiologia Plantarum*, vol. 139, no. 1, pp. 80–92, 2010.

[52] M. Libault, J. Wan, T. Czechowski, M. Udvardi, and G. Stacey, "Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor," *Molecular plant-microbe interactions*, vol. 20, no. 8, pp. 900–911, 2007.

[53] I.-F. Chang, A. Curran, R. Woolsey, D. Quilici, J. C. Cushman, R. Mittler, A. Harmon, and J. F. Harper, "Proteomic profiling of tandem affinity purified 14-3-3 protein complexes in Arabidopsis thaliana," *Proteomics*, vol. 9, no. 11, pp. 2967–2985, 2009.

[54] Z. Xin, Y. Zhao, and Z.-L. Zheng, "Transcriptome analysis reveals specific modulation of abscisic acid signaling by rop10 small GTPase in Arabidopsis," *Plant Physiology*, vol. 139, no. 3, pp. 1350–1365, 2005.

[55] Q. Ling, W. Huang, A. Baldwin, and P. Jarvis, "Chloroplast biogenesis is regulated by direct action of the ubiquitin-proteasome system," *Science*, vol. 338, no. 6107, pp. 655–659, 2012.

[56] L. Lopez-Molina, S. Mongrand, N. Kinoshita, and N.-H. Chua, "AFP is a novel negative regulator of ABA signaling that promotes ABI5 protein degradation," *Genes & Development*, vol. 17, no. 3, pp. 410–418, 2003.

[57] N. N. P. Chandrika, K. Sundaravelpandian, S.-M. Yu, and W. Schmidt, "ALFIN-LIKE 6 is involved in root hair elongation during phosphate deficiency in Arabidopsis," *New Phytologist*, vol. 198, no. 3, pp. 709–720, 2013.

[58] D. Reňák, N. Dupláková, and D. Honys, "Wide-scale screening of t-DNA lines for transcription factor genes affecting male gametophyte development in Arabidopsis," *Sexual Plant Reproduction*, vol. 25, no. 1, pp. 39–60, 2012.

[59] E. Caro, H. Stroud, M. V. Greenberg, Y. V. Bernatavichute, S. Feng, M. Groth, A. A. Vashisht, J. Wohlschlegel, and S. E. Jacobsen, "The SET-domain protein SUVR5 mediates H3K9me2 deposition and silencing at stimulus response genes in a DNA methylation–independent manner," *PLoS Genetics*, vol. 8, no. 10, p. e1002995, 2012.

[60] M. M. Alonso-Peral, J. Li, Y. Li, R. S. Allen, W. Schnippenkoetter, S. Ohms, R. G. White, and A. A. Millar, "The microrna159-regulated GAMYB-like genes inhibit growth and promote programmed cell death in Arabidopsis," *Plant Physiology*, vol. 154, no. 2, pp. 757–771, 2010.

[61] J. L. Reyes and N.-H. Chua, "ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination," *The Plant Journal*, vol. 49, no. 4, pp. 592–606, 2007.

[62] S. Sawa, K. Watanabe, K. Goto, E. Kanaya, E. H. Morita, and K. Okada, "FILAMENTOUS FLOWER, a meristem and organ identity gene of Arabidopsis, encodes a protein with a zinc finger and HMG-related domains," *Genes & Development*, vol. 13, no. 9, pp. 1079–1088, 1999.

[63] P. Sieber, M. Petrascheck, A. Barberis, and K. Schneitz, "Organ polarity in Arabidopsis. NOZZLE physically interacts with members of the YABBY family," *Plant Physiology*, vol. 135, no. 4, pp. 2172–2185, 2004.

[64] N. Lugassi, N. Nakayama, R. Bochnik, and M. Zik, "A novel allele of FILAMENTOUS FLOWER reveals new insights on the link between inflorescence and floral meristem organization and flower morphogenesis," *BMC Plant Biology*, vol. 10, no. 1, p. 131, 2010.

[65] N. Shao, G. Y. Duan, and R. Bock, "A mediator of singlet oxygen responses in Chlamydomonas reinhardtii and arabidopsis identified by a luciferase-based genetic screen in algal cells," *The Plant Cell Online*, vol. 25, no. 10, pp. 4209–4226, 2013.

[66] M. Büttner and K. B. Singh, "Arabidopsis thaliana ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein," *Proceedings of the National Academy of Sciences*, vol. 94, no. 11, pp. 5961–5966, 1997.

[67] W. G. Brenner, G. A. Romanov, I. Köllmer, L. Bürkle, and T. Schmülling, "Immediate-early and delayed cytokinin response genes of arabidopsis thaliana identified by genome-wide expression profiling reveal novel cytokinin-sensitive processes and suggest cytokinin action through transcriptional cascades," *The Plant Journal*, vol. 44, no. 2, pp. 314–333, 2005.

[68] H.-Y. Li, S. Xiao, and M.-L. Chye, "Ethylene-and pathogen-inducible Arabidopsis acyl-CoA-binding protein 4 interacts with an ethylene-responsive element binding protein," *Journal of Experimental Botany*, vol. 59, no. 14, pp. 3997–4006, 2008.

[69] K. Hill, D. E. Mathews, H. J. Kim, I. H. Street, S. L. Wildes, Y.-H. Chiang, M. G. Mason, J. M. Alonso, J. R. Ecker, J. J. Kieber, *et al.*, "Functional characterization of type-B

response regulators in the arabidopsis cytokinin response," *Plant Physiology*, vol. 162, no. 1, pp. 212–224, 2013.

[70] I. Hwang, H.-C. Chen, and J. Sheen, "Two-component signal transduction pathways in Arabidopsis," *Plant Physiology*, vol. 129, no. 2, pp. 500–515, 2002.

[71] R. D. Argyros, D. E. Mathews, Y.-H. Chiang, C. M. Palmer, D. M. Thibault, N. Etheridge, D. A. Argyros, M. G. Mason, J. J. Kieber, and G. E. Schaller, "Type B response regulators of Arabidopsis play key roles in cytokinin signaling and plant development," *The Plant Cell Online*, vol. 20, no. 8, pp. 2102–2116, 2008.

[72] T. A. Long, H. Tsukagoshi, W. Busch, B. Lahner, D. E. Salt, and P. N. Benfey, "The bHLH transcription factor POPEYE regulates response to iron deficiency in arabidopsis roots," *The Plant Cell Online*, vol. 22, no. 7, pp. 2219–2236, 2010.

[73] C. Yanhui, Y. Xiaoyuan, H. Kun, L. Meihua, L. Jigang, G. Zhaofeng, L. Zhiqiang, Z. Yunfei, W. Xiaoxiao, Q. Xiaoming, *et al.*, "The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family," *Plant Molecular Biology*, vol. 60, no. 1, pp. 107–124, 2006.

[74] D. C. Nelson, G. R. Flematti, J.-A. Riseborough, E. L. Ghisalberti, K. W. Dixon, and S. M. Smith, "Karrikins enhance light responses during germination and seedling development in Arabidopsis thaliana," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 7095–7100, 2010.

[75] M. Kieffer, V. Master, R. Waites, and B. Davies, "TCP14 and TCP15 affect internode length and leaf shape in Arabidopsis," *The Plant Journal*, vol. 68, no. 1, pp. 147–158, 2011.

[76] K. Tatematsu, K. Nakabayashi, Y. Kamiya, and E. Nambara, "Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in Arabidopsis thaliana," *The Plant Journal*, vol. 53, no. 1, pp. 42–52, 2008.

[77] M. S. Mukhtar, A.-R. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M. T. Nishimura, *et al.*, "Independently evolved virulence effectors

converge onto hubs in a plant immune system network," *science*, vol. 333, no. 6042, pp. 596–601, 2011.

[78] E. Steiner, I. Efroni, M. Gopalraj, K. Saathoff, T.-S. Tseng, M. Kieffer, Y. Eshed, N. Olszewski, and D. Weiss, "The Arabidopsis O-linked N-acetylglucosamine transferase spindly interacts with class I TCPs to facilitate cytokinin responses in leaves and flowers," *The Plant Cell Online*, vol. 24, no. 1, pp. 96–108, 2012.

[79] M. Dreze, A.-R. Carvunis, B. Charloteaux, M. Galli, S. J. Pevzner, M. Tasan, Y.-Y. Ahn, P. Balumuri, A.-L. Barabási, V. Bautista, *et al.*, "Evidence for network evolution in an Arabidopsis interactome map," *Science*, vol. 333, no. 6042, pp. 601–607, 2011.

[80] U. Piskurewicz, Y. Jikumaru, N. Kinoshita, E. Nambara, Y. Kamiya, and L. Lopez-Molina, "The gibberellic acid signaling repressor RGL2 inhibits arabidopsis seed germination by stimulating abscisic acid synthesis and ABI5 activity," *The Plant Cell Online*, vol. 20, no. 10, pp. 2729–2745, 2008.

[81] H. Fujii, P. E. Verslues, and J.-K. Zhu, "Identification of two protein kinases required for abscisic acid regulation of seed germination, root growth, and gene expression in Arabidopsis," *The Plant Cell Online*, vol. 19, no. 2, pp. 485–494, 2007.

[82] S.-Y. Park, P. Fung, N. Nishimura, D. R. Jensen, H. Fujii, Y. Zhao, S. Lumba, J. Santiago, A. Rodrigues, F. C. Tsz-fung, *et al.*, "Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins," *Science*, vol. 324, no. 5930, pp. 1068–1071, 2009.

[83] M. Boudsocq, H. Barbier-Brygoo, and C. Laurière, "Identification of nine sucrose nonfermenting 1-related protein kinases 2 activated by hyperosmotic and saline stresses in Arabidopsis thaliana," *Journal of Biological Chemistry*, vol. 279, no. 40, pp. 41758–41766, 2004.

[84] H. J. Joshi, M. Hirsch-Hoffmann, K. Baerenfaller, W. Gruissem, S. Baginsky, R. Schmidt, W. X. Schulze, Q. Sun, K. J. Van Wijk, V. Egelhofer, *et al.*, "MASCP Gator: an ag-

gregation portal for the visualization of Arabidopsis proteomics data," *Plant Physiology*, vol. 155, no. 1, pp. 259–270, 2011.

[85] N. Nishimura, A. Sarkeshik, K. Nito, S.-Y. Park, A. Wang, P. C. Carvalho, S. Lee, D. F. Caddell, S. R. Cutler, J. Chory, *et al.*, "PYR/PYL/RCAR family members are major in-vivo ABI1 protein phosphatase 2C-interacting proteins in Arabidopsis," *The Plant Journal*, vol. 61, no. 2, pp. 290–299, 2010.

[86] I. W. Manfield, P. F. Devlin, C.-H. Jen, D. R. Westhead, and P. M. Gilmartin, "Conservation, convergence, and divergence of light-responsive, circadian-regulated, and tissue-specific expression patterns during evolution of the Arabidopsis GATA gene family," *Plant Physiology*, vol. 143, no. 2, pp. 941–958, 2007.

[87] K. A. Franklin and G. C. Whitelam, "Light-quality regulation of freezing tolerance in Arabidopsis thaliana," *Nature genetics*, vol. 39, no. 11, pp. 1410–1413, 2007.

[88] T. Yamashino, A. Matsushika, T. Fujimori, S. Sato, T. Kato, S. Tabata, and T. Mizuno, "A link between circadian-controlled bHLH factors and the APRR1/TOC1 quintet in Arabidopsis thaliana," *Plant and Cell Physiology*, vol. 44, no. 6, pp. 619–629, 2003.

[89] M. Nakamura, H. Katsumata, M. Abe, N. Yabe, Y. Komeda, K. T. Yamamoto, and T. Takahashi, "Characterization of the class IV homeodomain-leucine zipper gene family in Arabidopsis," *Plant Physiology*, vol. 141, no. 4, pp. 1363–1375, 2006.

[90] M. Abe, H. Katsumata, Y. Komeda, and T. Takahashi, "Regulation of shoot epidermal cell differentiation by a pair of homeodomain proteins in Arabidopsis," *Development*, vol. 130, no. 4, pp. 635–643, 2003.

[91] N. Kamata, H. Okada, Y. Komeda, and T. Takahashi, "Mutations in epidermis-specific HD-ZIP IV genes affect floral organ identity in Arabidopsis thaliana," *The Plant Journal*, vol. 75, no. 3, pp. 430–440, 2013.

[92] I. Efroni, E. Blum, A. Goldshmidt, and Y. Eshed, "A protracted and dynamic maturation schedule underlies Arabidopsis leaf development," *The Plant Cell Online*, vol. 20, no. 9, pp. 2293–2306, 2008.

[93] S. L. Stone, H. Hauksdóttir, A. Troy, J. Herschleb, E. Kraft, and J. Callis, "Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis," *Plant Physiology*, vol. 137, no. 1, pp. 13–30, 2005.

[94] J.-W. Wang, L.-J. Wang, Y.-B. Mao, W.-J. Cai, H.-W. Xue, and X.-Y. Chen, "Control of root cap formation by microRNA-targeted auxin response factors in Arabidopsis," *The Plant Cell Online*, vol. 17, no. 8, pp. 2204–2216, 2005.

[95] A. C. Mallory, D. P. Bartel, and B. Bartel, "MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes," *The Plant Cell Online*, vol. 17, no. 5, pp. 1360–1375, 2005.

[96] N. Haga, K. Kato, M. Murase, S. Araki, M. Kubo, T. Demura, K. Suzuki, I. Müller, U. Voß, G. Jürgens, *et al.*, "R1R2R3-Myb proteins positively regulate cytokinesis through activation of KNOLLE transcription in Arabidopsis thaliana," *Development*, vol. 134, no. 6, pp. 1101–1110, 2007.

[97] Z.-Y. Wang, T. Nakano, J. Gendron, J. He, M. Chen, D. Vafeados, Y. Yang, S. Fujioka, S. Yoshida, T. Asami, *et al.*, "Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis," *Developmental Cell*, vol. 2, no. 4, pp. 505–513, 2002.

[98] J. Mu, H. Tan, S. Hong, Y. Liang, and J. Zuo, "Arabidopsis transcription factor genes NF-YA1, 5, 6, and 9 play redundant roles in male gametogenesis, embryogenesis, and seed development," *Molecular Plant*, vol. 6, no. 1, pp. 188–201, 2013.

[99] D. Edwards, J. A. Murray, and A. G. Smith, "Multiple genes encoding the conserved CCAAT-box transcription factor complex are expressed in Arabidopsis," *Plant Physiology*, vol. 117, no. 3, pp. 1015–1022, 1998.

[100] S. Wenkel, F. Turck, K. Singer, L. Gissot, J. Le Gourrierec, A. Samach, and G. Coupland, "CONSTANS and the CCAAT box binding complex share a functionally important

domain and interact to regulate flowering of arabidopsis," *The Plant Cell Online*, vol. 18, no. 11, pp. 2971–2984, 2006.

[101] X. Zhang, V. Garreton, and N.-H. Chua, "The AIP2 E3 ligase acts as a novel negative regulator of ABA signaling by promoting ABI3 degradation," *Genes & Development*, vol. 19, no. 13, pp. 1532–1543, 2005.

[102] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, and J. U. Lohmann, "A gene expression map of Arabidopsis thaliana development," *Nature Genetics*, vol. 37, no. 5, pp. 501–506, 2005.

[103] R. C. Elliott, A. S. Betzner, E. Huttner, M. P. Oakes, W. Tucker, D. Gerentes, P. Perez, and D. R. Smyth, "AINTEGUMENTA, an APETALA2-like gene of Arabidopsis with pleiotropic roles in ovule development and floral organ growth.," *The Plant Cell Online*, vol. 8, no. 2, pp. 155–168, 1996.

[104] V. Pinon, K. Prasad, S. P. Grigg, G. F. Sanchez-Perez, and B. Scheres, "Local auxin biosynthesis regulation by PLETHORA transcription factors controls phyllotaxis in Arabidopsis," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 1107–1112, 2013.

[105] J. S. Mudunkothge and B. A. Krizek, "Three arabidopsis AIL/PLT genes act in combination to regulate shoot apical meristem function," *The Plant Journal*, vol. 71, no. 1, pp. 108–121, 2012.

[106] Y. Chanvivattana, A. Bishopp, D. Schubert, C. Stock, Y.-H. Moon, Z. R. Sung, and J. Goodrich, "Interaction of Polycomb-group proteins controlling flowering in arabidopsis," *Development*, vol. 131, no. 21, pp. 5263–5276, 2004.

[107] L. Chen, J.-C. Cheng, L. Castle, and Z. R. Sung, "EMF genes regulate arabidopsis inflorescence development.," *The Plant Cell Online*, vol. 9, no. 11, pp. 2011–2024, 1997.

[108] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "GO::TermFinder–open source software for accessing gene ontology information and find-

ing significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[109] J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar, and M. Geisler, "A predicted interactome for arabidopsis," *Plant Physiology*, vol. 145, no. 2, pp. 317–329, 2007.

[110] "The Arabidopsis Interactions Viewer." http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi. Accessed 15 Apr 2015.

[111] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold, "AGRIS: the Arabidopsis gene regulatory information server, an update," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1118–D1122, 2011.

[112] L. Everett, A. Vo, and S. Hannenhalli, "PTM-Switchboard–a database of posttranslational modifications of transcription factors, the mediating enzymes and target genes," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D66–D71, 2009.

[113] G. Chen, S. T. Jensen, and C. J. Stoeckert, "Clustering of genes into regulons using integrated modeling - COGRIM," *Genome Biology*, vol. 8, no. 1, p. R4, 2007.

[114] H. Frank and S. C. Althoen, *Statistics: concepts and applications*. Cambridge, England: Cambridge university press, 1994.

[115] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: keys and clues," *BioScience*, vol. 51, no. 5, pp. 341–352, 2001.

[116] G. McLachlan and D. Peel, *Finite mixture models*. New York: John Wiley & Sons, 2004.

[117] C. Genest, R. Silva, G. Elidan, Z. Ghahramani, and J. Lafferty, "NIPS 2011 workshop on copulas in machine learning." [http://pluto.huji.ac.il/∼galelidan/CopulaWorkshop/].

[118] M. W. Schmidt, K. P. Murphy, G. Fung, and R. Rosales, "Structure learning in random fields for heart motion abnormality detection.," in *CVPR*, vol. 1, p. 2, 2008.

[119] S. Ding, G. Wahba, and X. Zhu, "Learning higher-order graph structure with features by structure penalty," in *Advances in Neural Information Processing Systems*, pp. 253–261, 2011.

[120] G. K. Kanji, *100 statistical tests*. Sage, 2006.