Bayesian Networks for High-dimensional Data with Complex Mean Structure

Jessica E. Kasza

Thesis submitted for the degree of

Doctor of Philosophy

in

Statistics

at

The University of Adelaide

(Discipline of Statistics, School of Mathematical Sciences, Faculty of Engineering, Mathematical and Computer Sciences)

February 25, 2010

Contents

Abstract **xi**

Signed Statement **xiii**

Acknowledgements **xiv**

1 Introduction **1**

2 Graph Theory and Graphical Modelling **5**

2.1 Required Graph Theory 5

2.2 Graphical Models 8

2.2.1 Conditional Independence 9

2.2.2 Markov Properties 9

2.2.3 Independence Graphs 13

2.2.4 Gaussian Graphical Models 14

2.2.5 Directed Markov Properties 16

2.2.6 Equivalence of Directed Acyclic Graphs 18

2.2.7 Bayesian Networks 19

2.2.8 Linear Recursive Equations 20

2.3 Using Gaussian Graphical Models and Bayesian Networks to Model Genetic Regulatory Networks 23

3 Estimating Graphs for Gene Expression Data 26

3.1 The Bayesian Network Approach 27

3.1.1 Score-Based Methods 28

3.1.2 Constraint-Based Methods 33

3.2 The Gaussian Graphical Model Approach 33

3.2.1 Limited-Order Partial Correlation-Based Methods 34

3.2.2 Shrinkage-Based Methods 35

3.2.3 Other Methods 36

3.3 High-Dimensional Bayesian Covariance Selection 37

3.3.1 Construction of the High-dimensional Bayesian Covariance Selection Score Metric 38

3.3.2 Posterior Distributions 42

3.3.3 The High-dimensional Bayesian Covariance Selection Algorithm . . . 43

3.3.4 The High-dimensional Bayesian Covariance Selection Program 48

3.4 Extensions and Use of the Methods 49

4 Score Metrics for Data Sets with Complex Mean Structures 50

4.1 Motivation for the Inclusion of Complex Mean Structures 50

4.2 Derivation of the Score Metric 54

4.2.1 Assuming ϕ_i known: Derivation of S_1 57

4.2.2 Assuming $\boldsymbol{b}_i$ vary as $\boldsymbol{\gamma}_i$: Derivation of S_2 59

4.2.3 Assuming $\phi_i^{\frac{1}{2}} \sim$ Uniform $(0, \kappa)$: Derivation of S_3 62

4.2.4 Assuming $\phi_i \sim$ Inverse Gamma (α, β): Derivation of S_4 64

4.2.5 The score metrics when ϕ_i is small relative to ψ_i 65

4.3 Estimation of the Joint Covariance Matrix 68

4.4 Posterior Estimation of Parameters 72

4.4.1 Posteriors assuming ϕ_i known 74

4.4.2 Posteriors assuming $\boldsymbol{b}_i$ vary as $\boldsymbol{\gamma}_i$ 74

4.4.3 Posteriors assuming $\phi_i^{\frac{1}{2}} \sim$ Uniform $(0, \kappa)$ 75

4.4.4 Posteriors assuming $\phi_i \sim$ Inverse Gamma (α, β) 76

4.4.5 Gibbs sampling from the joint posterior distribution 76

4.5 Discussion 77

4.6 Implementation 80

5 Generalisation of the Distribution of the Random Effects 83

5.1 Exploring the Covariance Structure of the Random Effects 84

5.1.1 Assuming $\boldsymbol{b}_i|\phi_i \sim N_m(\mathbf{0}, \phi_i V)$, V known 84

5.1.2 A different variance parameter for each random effect 85

5.2 An Uninformative Random Effects Prior 87

6 Removal of Random Effects Through Analysis of Residuals 89

7 The Use of Score Metrics That Take Account of Complex Mean Structure 94

7.1 The Necessity of Taking Account of Complex Mean Structure 95

7.1.1 Analysis of the data sets 100

7.1.2 Using S_0 in the Estimation of Bayesian Networks 100

7.1.3 The Residual Approach to the Estimation of Bayesian Networks . . . 105

7.2 The Use of S_1 and S_2 in the Estimation of Bayesian Networks 107
7.2.1 The Use of S_1 108
7.2.2 The Use of S_2 112
7.3 Consequences of Misspecification of the Distribution of ϕ_i 115
7.3.1 The Effect of Model Misspecification on Posterior Estimation 119
7.4 Conclusions and Recommendations 134

8 Analysis of the Grape Gene Data **136**

8.1 The Grape Gene Data 138
8.2 Initial Analysis of the Grape Gene Data 142
8.3 Taking Account of Vineyard and Temperature Effects in the Analysis of the Grape Gene Data 145
8.3.1 Inclusion of the Effects of Vineyard and Temperature in the Model . 148
8.3.2 Using the Residual Approach to Estimate a Bayesian Network for the Grape Genes 151
8.3.3 Using the S_2 score metric to Estimate a Bayesian Network for the Genes 157
8.3.4 Using a Combination of S_2 and the Residual Approach in the Estimation of a Bayesian Network for the Grape Genes. 158
8.4 The Highest-Scoring Graphs Obtained 162
8.4.1 Biological Plausibility of the Graphs 164
8.5 Posterior Estimation of Vineyard Effects 165
8.6 Conclusions 172

9 Conclusions and Future Work **175**

A Gaussian Quadrature **178**

A.1 R code for Gaussian Quadrature 183

B Random Effects Code **185**

B.1 Code for S_1 185

B.2 Code for S_4 192

B.3 Posterior Sampling Code 200

B.3.1 Posterior sampling when ϕ fixed 200

B.3.2 Posterior sampling when $\phi_i = \upsilon^{-1}\psi_i$ 201

B.3.3 Posterior sampling when $\phi_i^{\frac{1}{2}} \sim \text{Uniform}\,(0, \kappa)$ 202

C Grape Gene Data **204**

C.1 Boxplots of Gene Expression Levels 204

C.2 Differences Between Vineyards 206

C.3 Regressing the Gene Expressions on Temperature 211

Bibliography **236**

List of Figures

2.1 An undirected graph, discussed in Example 2.1 7

2.2 The graphs discussed in Example 2.2. 8

7.1 The connected component of the directed acyclic graph of Example 7.3 . . . 97

7.2 The connected components of the Bayesian network for Example 7.6, taking the covariates as vertices in the network. 105

7.3 Histograms of the samples from the marginal posterior distribution of b_{11}, ψ_1, b_{71}, ψ_7. 123

7.4 Medians and 90% posterior intervals for $b_{11}|\boldsymbol{x}_1$, $\psi_1|\boldsymbol{x}_1$, and ϕ_1 when $\boldsymbol{x}_1$ is generated under M_1. 124

7.5 Medians and 90% posterior intervals for $b_{71}|\boldsymbol{x}_7$, $\psi_7|\boldsymbol{x}_7$, and ϕ_7 when $\boldsymbol{x}_7$ is generated under M_1. 125

7.6 Histograms of the samples from the marginal posterior distribution of b_{51}, ψ_5, $b_{11,1}$, ψ_{11}. 127

7.7 Medians and 90% posterior intervals for $b_{51}|\boldsymbol{x}_5$, $\psi_5|\boldsymbol{x}_5$, and ϕ_5 when $\boldsymbol{x}_5$ is generated under M_2. 128

7.8 Medians and 90% posterior intervals for $b_{11,1}|\boldsymbol{x}_{11}$, $\psi_{11}|\boldsymbol{x}_{11}$, and ϕ_{11} when $\boldsymbol{x}_{11}$ is generated under M_2. 129

7.9 Histograms of the samples from the marginal posterior distribution of b_{61}, ψ_6, ϕ_6, $b_{17,1}$, ψ_{17}, $\phi_{17,1}$. 130

7.10 Medians and 90% posterior intervals for $b_{61}|\boldsymbol{x}_6$, $\psi_6|\boldsymbol{x}_6$, and ϕ_6 when $\boldsymbol{x}_6$ is generated under M_3. 132

7.11 Medians and 90% posterior intervals for $b_{17,1}|\boldsymbol{x}_{17}$, $\psi_{17}|\boldsymbol{x}_{17}$, and ϕ_{17} when $\boldsymbol{x}_{17}$ is generated under M_3. 133

8.1 A schematic representation of the development of grape berries. 141

8.2 The moral version of the highest-scoring graph obtained for the grape genes, when vineyard and temperature are not accounted for. 143

8.3 The temperatures at each vineyard at the times leading up to the picking of the grapes. 146

8.4 Histogram of the adjusted r^2s. 146

8.5 Histograms of the marginal standard deviations of the grape gene expression levels and the residual standard errors after regressing the expression levels on temperature and vineyard. 148

8.6 Scatterplots of the residuals after fitting the above model, with vineyard, main temperature and two-way temperature interaction effects for some pairs of genes. 149

8.7 The moral graphs of the highest-scoring Bayesian networks found for the grape genes, when the residual approach is taken. 154

8.8 The moral graphs of the highest-scoring Bayesian networks found for the grape genes, when the residual approach is taken. 155

8.9 The moral graphs of the highest-scoring Bayesian networks found for the grape genes when S_2 is used, for different values of v. 159

8.10 The moral graphs of the highest-scoring Bayesian networks found for the grape genes, when a combination of the residual approach and S_2 is used, for different values of v. 161

8.11 Connected components of Figure 8.8(b), with gene names included. 164

8.12 Scatterplots of the expression levels of some of the probes coding for the same genes. 166

8.13 90% posterior intervals for ψ_i, $i = 1, 2, \ldots, 26$, generated given the Bayesian networks found assuming $\upsilon = 0.5, 1, 10$. 169

8.14 90% posterior intervals for b^V_{i1}, $i = 1, 2, \ldots, 26$, the effect of the Clare vineyard on the expression level of gene i. 170

8.15 90% posterior intervals for b^V_{i2}, $i = 1, 2, \ldots, 26$, the effect of the Wingara vineyard on the expression level of gene i. 171

8.16 90% posterior intervals for b^V_{i3}, $i = 1, 2, \ldots, 26$, the effect of the Willunga vineyard on the expression level of gene i. 173

C.1.1 Boxplots of the expression levels of genes 1 to 9 for grapes sampled at each of the vineyards. 204

C.1.2 Boxplots of the expression levels of genes 10 to 18 for grapes sampled at each of the vineyards. 205

C.1.3 Boxplots of the expression levels of genes 19 to 26 for grapes sampled at each of the vineyards. 205

List of Tables

7.1 Summary of the results obtained when S_0 is applied to data sets simulated according to Examples 7.1–7.6. 101

7.2 Mean and standard deviation of the number of edges in the highest-scoring network when data sets from Example 7.5 are analysed in halves. 102

7.3 Mean and standard deviation of the number of edges in the highest-scoring networks when covariates are included as vertices in the analysis of Example 7.6. 103

7.4 Mean and standard deviation of the number of edges in the highest-scoring networks when covariates are included as vertices in the analysis of Example 7.6, $\beta = 0.9$. 104

7.5 Summary of the results obtained when the residual approach is applied to data sets simulated according to Examples 7.4–7.6. 106

7.6 Summary of the results obtained when the residual approach is applied to data sets simulated according to Examples 7.1–7.3. 106

7.7 Summary of the results obtained when S_1 is applied to data sets using the true value of ϕ and quadrature of size 50. 109

7.8 Mean and standard deviation of the number of edges found in the analysis of data sets from Example 7.4 using S_1. 109

7.9 Mean and standard deviation of the number of edges found in the analysis of data sets from Example 7.6 using S_1. 110

7.10 Mean and standard deviation of the number of edges in the highest scoring networks found when the data sets generated by taking $\phi_i = \psi_i$ are analysed using S_0. 112

7.11 Mean and standard deviation of the number of edges in the highest scoring networks found when the data sets generated by taking $\phi_i = \psi_i$ are analysed using S_2. 113

7.12 Mean and standard deviation of the number of edges in the highest-scoring graphs found through the application of S_2 for varying values of υ^*. 114

7.13 Mean and standard deviation of the number of edges in the highest-scoring Bayesian networks for data sets generated from Example 7.7. 118

8.1 Grape heat shock genes. 139

8.2 Summaries of the highest-scoring graphs found for the grape genes 162

8.3 Parents of each gene in the highest-scoring Bayesian networks found using a combination of S_2 and the residual approach. 167

Abstract

In a microarray experiment, it is expected that there will be correlations between the expression levels of different genes under study. These correlation structures are of great interest from both biological and statistical points of view. From a biological perspective, the identification of correlation structures can lead to an understanding of genetic pathways involving several genes, while the statistical interest, and the emphasis of this thesis, lies in the development of statistical methods to identify such structures. However, the data arising from microarray studies is typically very high-dimensional, with an order of magnitude more genes being considered than there are samples of each gene. This leads to difficulties in the estimation of the dependence structure of all genes under study. Graphical models and Bayesian networks are often used in these situations, providing flexible frameworks in which dependence structures for high-dimensional data sets can be considered.

The current methods for the estimation of dependence structures for high-dimensional data sets typically assume the presence of independent and identically distributed samples of gene expression values. However, often the data available will have a complex mean structure and additional components of variance. Given such data, the application of methods that assume independent and identically distributed samples may result in incorrect biological conclusions being drawn. In this thesis, methods for the estimation of Bayesian networks for gene expression data sets that contain additional complexities are developed and implemented. The focus is on the development of score metrics that take account of these complexities for use in conjunction with score-based methods for the estimation of Bayesian networks, in particular the High-dimensional Bayesian Covariance Selection algorithm.

The necessary theory relating to Gaussian graphical models and Bayesian networks is reviewed, as are the methods currently available for the estimation of dependence structures

for high-dimensional data sets consisting of independent and identically distributed samples. Score metrics for the estimation of Bayesian networks when data sets are not independent and identically distributed are then developed and explored, and the utility and necessity of these metrics is demonstrated. Finally, the developed metrics are applied to a data set consisting of samples of grape genes taken from several different vineyards.

Signed Statement

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Jessica Kasza and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....

Acknowledgements

First and foremost, thanks are due to my supervisors, Gary Glonek and Patty Solomon. None of this would have been possible were it not for their wisdom, guidance and support. Gary, your patience is seemingly infinite, and your encouragement allowed me to do more than I thought possible. Patty, your faith in me gave me faith in myself. Thank you both for all of the effort you have put into supervising me over the years.

I would also like to acknowledge the help of Dr. Christopher Davies of the C.S.I.R.O., both for the grape gene data analysed in Chapter 8, and for his help in understanding the results of the analyses. Thank you, Chris!

The biggest thanks of all got to my family, for their support and love. Josh, your gifts of delicious chocolate were appreciated more than you know. Mum, Dad: thank you for everything.

This achievement would be empty without good friends to share it with: thank you all for being who you are. Thanks are especially due to Rhys Bowden, for being who he is.