

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**Computationally tractable fitting of  
graphical models:  
the cost and benefits of decomposable Bayesian  
and penalized likelihood approaches.**

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Albany,

New Zealand.

Anne Marie Fitch

2012



## Abstract

Gaussian graphical models are a useful tool for eliciting information about relationships in data with a multivariate normal distribution. In the first part of this thesis we demonstrate that partial correlation graphs facilitate different and better insight into high-dimensional data than sample correlations. This raises the question of which method one should use to model and estimate the parameters. In the second, and major part, we take a more theoretical focus examining the costs and benefits of two popular approaches to model selection and parameter estimation (penalized likelihood and decomposable Bayesian) when the true graph is non-decomposable.

We first consider the effect a restriction to decomposable models has on the estimation of both the inverse covariance matrix and the covariance matrix. Using the variance as a measure of variability we compare non-decomposable and decomposable models. Here we find that, if the true model is non-decomposable, the variance of estimates is demonstrably larger when a decomposable model is used. Although the cost in terms of accuracy is fairly small when estimating the inverse covariance matrix, this is not the case when estimation of the covariance matrix is the goal. In this case using a decomposable model caused up to 200-fold increases in the variance of estimates.

Finally we compare the latest decomposable Bayesian method (the feature-inclusion stochastic search) with penalized likelihood methods (graphical lasso and adaptive graphical lasso) on measures of model selection and prediction performance. Here we find that graphical lasso is clearly outclassed on all measures by both adaptive graphical lasso and feature-inclusion stochastic search. The sample size and the ultimate goal of the estimation will determine whether adaptive graphical lasso or feature-inclusion stochastic search is better.



## Acknowledgements

First and foremost thanks to my supervisor Beatrix Jones . I have greatly benefited from her encouragement and help along the way. Without her I would never have started let alone finished! Thanks also to my co-supervisor James Curran, especially for his help with R-graphics and  $\text{\LaTeX}$ code.

Three authors made their code available to us. Thanks to Xianghong Zhou for the shortest path analysis code used by the authors of Zhou *et al.* (2002); James Scott for his C++ code for feature-inclusion stochastic search and Yang Feng for his R-code for adaptive graphical lasso.

Thanks to Robert McKibbin and Tony Norris (Heads of Institute) and Jeff Hunter, Howard Edwards and Marti Anderson (Statistics discipline leaders) who in these roles have both facilitated the opportunity for me to combine study and teaching and also gave much encouragement along the way. Thanks to fellow student Insha Ullah for sharing his R-code for finding a perfect elimination ordering. Thanks also to the rest of the Massey Statistics and Mathematics teams for your encouragement, especially over the write-up months.

Last, but certainly not least, a big thank you to my family. Thanks to Harmen for his help in getting Figure 3.3 into a printable form. Thanks to Nigel, Miriam and Harmen, Joanna, and Andrew, without their support and encouragement I could not have done this.

*Whaia te iti kahurangi ki te tuohu koe me he maunga teitei.*



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research objectives . . . . .	1
1.2	Gaussian graphical models . . . . .	2
1.3	Overview . . . . .	3
1.4	Publications . . . . .	6
<b>2</b>	<b>Background</b>	<b>7</b>
2.1	Gaussian graphical models . . . . .	7
2.1.1	Overview . . . . .	7
2.1.2	Penalized likelihood methods: graphical lasso and adaptive graphical lasso . . . . .	9
2.2	Graph theory . . . . .	12
2.2.1	Decomposability and decomposable graphs . . . . .	12
2.2.2	Gaussian graphical models . . . . .	16
2.2.3	Acyclic directed graphs . . . . .	18
2.3	Bayesian model selection . . . . .	19
2.4	Predictions . . . . .	22
<b>3</b>	<b>Shortest paths with partial correlations</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Data and methods . . . . .	25
3.2.1	Data . . . . .	25
3.2.2	Shortest path analysis . . . . .	26
3.2.3	Graphical lasso . . . . .	28



3.2.4	Using partial correlations in shortest path analysis . . . . .	29
3.3	Results . . . . .	29
3.4	Discussion . . . . .	35
<b>4</b>	<b>The cost of using a decomposable model</b>	<b>39</b>
4.1	Introduction . . . . .	39
4.2	Background . . . . .	42
4.2.1	General properties of graphs . . . . .	42
4.2.2	Parameter and variance estimation . . . . .	43
4.3	Theory for the four variable case . . . . .	44
4.4	Simulation study methods . . . . .	46
4.4.1	The four variable case . . . . .	46
4.4.2	20 and 50 variable cases . . . . .	48
4.5	Simulation study results . . . . .	49
4.5.1	Estimating $\Omega$ - the four variable case . . . . .	49
4.5.2	Estimating $\Omega$ - 20 and 50 variable cases . . . . .	52
4.5.3	Estimating the covariance matrix ( $\Sigma$ ) . . . . .	56
4.6	Case studies: Fisher's iris data and 12 node data . . . . .	58
4.6.1	Fisher's iris data . . . . .	58
4.6.2	12 node case . . . . .	60
4.7	Discussion . . . . .	62
<b>5</b>	<b>Decomposable covariance selection</b>	<b>65</b>
5.1	Introduction . . . . .	65
5.2	Background . . . . .	68
5.2.1	General properties of graphs . . . . .	68
5.2.2	Feature-inclusion stochastic search . . . . .	69
5.2.3	Kullback-Leibler divergence . . . . .	71
5.2.4	Graphical lasso and adaptive graphical lasso . . . . .	72
5.3	Data and methods . . . . .	74
5.3.1	Data . . . . .	74
5.3.2	Model selection . . . . .	75

5.3.3	Prediction . . . . .	76
5.4	Results . . . . .	77
5.4.1	Feature-inclusion stochastic search (FINCS) treatment of non-decomposable graphs . . . . .	77
5.4.2	Model selection comparison of FINCS with graphical lasso methods. . . . .	79
5.4.3	Comparison of predictions using FINCS and graphical lasso derived graphs . . . . .	83
5.5	Discussion . . . . .	87
<b>6</b>	<b>Concluding discussion</b>	<b>89</b>
6.1	Estimating the inverse covariance matrix . . . . .	89
6.2	Regularizing the estimate of the covariance matrix . . . . .	91
6.3	Concluding remarks . . . . .	92
 <b>Appendices</b>		
<b>A</b>	<b>Supplementary Tables and Figures for Chapter 4</b>	<b>95</b>
A.1	Tables and Figures for the four variable case . . . . .	95
A.2	Tables and Figures for 20 and 50 variable cases . . . . .	101
A.3	Tables and Figures for estimations of $\Sigma$ . . . . .	110
A.4	Tables and Figures for 12 node example . . . . .	113
<b>B</b>	<b>Supplementary Tables and Figures for Chapter 5</b>	<b>117</b>
B.1	$\Omega$ matrices . . . . .	117
B.2	Graphs and Tables for Section 5.4.1 . . . . .	118
B.3	Relative inclusion probabilities . . . . .	126
B.4	Maximum likelihood estimate (MLE) results . . . . .	131
<b>C</b>	<b>DRC forms</b>	<b>133</b>
<b>References</b>		<b>137</b>



# List of Figures

1.1	Example of a graphical model: Fisher’s <i>Iris virginica</i> dataset. . . . .	3
2.1	Graphs of five variables illustrating the idea of decomposition and decomposable versus non-decomposable. . . . .	13
3.1	Proportion of categorised paths less than upper bounds. . . . .	30
3.2	Percentage of transitive genes with annotations matching the annotations of the terminal genes on the same path. . . . .	31
3.3	Subgraphs of cytoplasm categorised path graphs. . . . .	33
4.1	A 4 variable non-decomposable graph(a) and a 4 variable decomposable graph (b). . . . .	45
4.2	Two different decomposable models when $p = 6$ . . . . .	48
4.3	OFI variances for $\hat{\omega}_{1,2}$ in $\Omega_{same}$ with sample sizes 10, 100 and 1000. . . . .	50
4.4	Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted. . . . .	54
4.5	Empirical variances for elements of $\Sigma$ when a decomposable model (Type A) vs when the true (non-decomposable) model is fitted. . . . .	57
4.6	The 12 variable models. . . . .	60
5.1	The 12 variable model. . . . .	74
5.2	Model score vs Kullback-Leibler divergence. . . . .	78
5.3	Model comparison measures: $p = 4, n = 50$ and $n = 1000$ . . . . .	80
5.4	Model comparison measures: $p = 20$ . . . . .	81
5.5	Model comparison measures: $p=12$ . . . . .	82
5.6	Mutual-funds data:edges vs total sum of squared errors. . . . .	83

5.7	Total sum of squared errors for 5 simulated samples of each $n$ and $p$ as specified. . . . .	85
5.8	Total sum of squared errors for 5 simulated samples: $n = 51$ and $p = 50$ . . . . .	86
A.1	OFI variances for $\widehat{\Omega}_{same}$ with sample sizes 10, 100 and 1000. . . . .	98
A.2	OFI variances for $\widehat{\Omega}_{big}$ with sample sizes 10, 100 and 1000. . . . .	99
A.3	OFI variances for $\widehat{\Omega}_{small}$ with sample sizes 10, 100 and 1000. . . . .	100
A.4	Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted and relative standard deviation (RSD) when fitting true (cycle) and decomposable models. . . . .	105
A.5	Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted and relative standard deviation (RSD) when fitting true and decomposable models. . . . .	106
A.6	Range of OFI variances when a cycle and two different decomposable models are fitted for $p=20, n=21$ . . . . .	107
A.7	Range of OFI variances when a cycle and two different decomposable models are fitted for $p=50, n=51$ . . . . .	107
A.8	Percentage of true non-zero elements declared zero when a cycle and two different decomposable models are fitted for $p=20, n=21$ . . . . .	108
A.9	Percentage of elements corresponding to ‘extra edges’ which are declared non-zero for $p=20, n=21$ . . . . .	108
A.10	Percentage of true non-zero elements declared zero when a cycle and two different decomposable models are fitted for $p=50, n=51$ . . . . .	109
A.11	Percentage of elements corresponding to ‘extra edges’ which are declared non-zero for $p=50, n=51$ . . . . .	109
A.12	Empirical variances for elements of $\Sigma$ when a decomposable model (type B) vs when the true (non-decomposable) model is fitted $p=20$ and $n=21$ . . . . .	111
A.13	Empirical variances for elements of $\Sigma$ when a decomposable model (type A) vs when the true (non-decomposable) model is fitted for $p=50$ and $n=51$ . . . . .	111

A.14	Empirical variances for elements of $\Sigma$ when a decomposable model (type B) vs when the true (non-decomposable) model is fitted for $p=50$ and $n=51$ . . . . .	112
A.15	Elements of $\Sigma$ when a decomposable model vs when the true (non-decomposable) model is fitted. . . . .	115
B.1	Top 10 graphs found by FINCS for $p=4$ , samples of $n=1000$ for different $ \tilde{\rho}_{ij} $ . . . . .	118
B.2	Top 50 graphs found by FINCS for $p=20$ , samples of $n=1000$ for different $ \tilde{\rho}_{ij} $ . . . . .	119
B.3	Top 10 graphs found by FINCS for $p=4$ , samples of $n=1000$ for $\Omega_{small}$ and $\Omega_{big}$ . . . . .	120
B.4	Top 50 graphs found by FINCS for samples of $n=1000$ for for $\Omega_{small}$ and $\Omega_{big}$ (both $p=20$ ) and for $\Omega_{twelve}$ ( $p=12$ ). . . . .	122



# List of Tables

4.1	$\Omega$ matrices used for simulating data. . . . .	47
4.2	$n = 1000$ empirical variances for cycle and percentage increase to decomposable. . . . .	49
4.3	$n = 1000$ Relative Standard Deviations for non-decomposable model and increase for decomposable model. . . . .	51
4.4	Percentage of times an element is declared zero ( $ \text{estimate}  < 2 \times$ standard error) for 1000 simulations ( $n = 10$ ). . . . .	53
4.5	Estimated $\Omega$ and $\Sigma$ matrices for <i>Iris virginica</i> dataset. . . . .	59
4.6	Estimates and standard error ( $\sqrt{\text{OFI variance}}$ ) for <i>Iris virginica</i> dataset. . . . .	59
4.7	Estimates and standard error ( $\sqrt{\text{OFI variance}}$ ) for $\hat{\omega}_{1,2}$ , $\hat{\omega}_{10,11}$ and $\hat{\omega}_{1,8}$ . . . . .	61
A.1	$n = 1000$ Relative Standard Deviations for non-decomposable model and increase for decomposable model. . . . .	96
A.2	Four variable empirical and EFI variances for elements of $\hat{\Omega}_{\text{same}}$ for cycle and decomposable, when $n=10$ . . . . .	96
A.3	Four variable empirical and EFI variances for elements of $\hat{\Omega}_{\text{big}}$ for cycle and decomposable, when $n=10$ . . . . .	96
A.4	Four variable empirical and EFI variances for elements of $\hat{\Omega}_{\text{small}}$ for cycle and decomposable, when $n=10$ . . . . .	97
A.5	Percentage of times an element is declared zero ( $ \text{estimate}  < 2 \times$ standard error) for 1000 simulations ( $\Omega_{\text{small}}$ ). . . . .	97
A.6	Non-zero elements of $\Omega$ matrices for $p=20$ . . . . .	102
A.7	Non-zero elements of $\Omega$ matrices for $p=50$ . . . . .	103



A.8	Four variable empirical variances for elements of $\widehat{\Sigma}$ for cycle and percentage increase to decomposable, when $n=10$ . . . . .	110
A.9	Estimates and OFI standard deviations for elements of $\widehat{\Omega}$ for cycle and three decomposable models , when $p=12$ and $n=250$ . . . . .	113
B.1	$\Omega_{twelve}$ . . . . .	117
B.2	True edges missing in top 10 graphs found by FINCS for $\Omega_{small}$ when $p=4$ . . . . .	121
B.3	True edges missing in top 10 graphs found by FINCS for $\Omega_{big}$ when $p=4$ . . . . .	121
B.4	True edges missing in top 10 graphs found by FINCS for $\Omega_{small}$ when $p=20$ . . . . .	123
B.5	True edges missing in top 10 graphs found by FINCS for $\Omega_{big}$ when $p=20$ . . . . .	124
B.6	True edges missing in top 10 graphs found by FINCS for $\Omega_{twelve}$ . . . . .	125
B.7	Relative inclusion probability matrices for $n=1000$ and $\tilde{\rho}_{ij}=-0.45$ . . . . .	126
B.8	Relative inclusion probability matrix for 12 node case when $n=1000$ and $\tilde{\rho}_{ij}=-0.4$ and for 20-node cycle when $n=1000$ and $\tilde{\rho}_{ij}=-0.45$ . . . . .	127
B.9	Relative inclusion probability matrices when $n=1000$ and top graphs are supersets. . . . .	128
B.10	Relative inclusion probability matrices when $n=50$ . . . . .	129
B.11	Relative inclusion probability matrices when $n=1000$ and partial correlations are small. . . . .	130
B.12	Model comparison measures for MLE results. . . . .	131
B.13	Sum of squared errors using the MLE. . . . .	131

# Chapter 1

## Introduction

Graphical models are a useful tool for providing understanding of joint distributions. A graph in this context is a set of vertices and a set of edges which join some of the vertices. In a graphical model the vertices represent random variables. We confine our interest to undirected graphs where the absence of an edge implies conditional independence and thus the edges represent the conditional independence structure. The focus of this research is Gaussian graphical models (GGMs). Here we assume the data is drawn from a multivariate normal distribution. In the Gaussian setting, edges between vertices are equivalent to non-zero elements in the inverse covariance matrix (Dempster, 1972). Thus partial correlations have the same zero pattern as the inverse covariance matrix.

### 1.1 Research objectives

The research objectives for this thesis fall into two parts. The first part consists of an initial motivational application. In the second, and major part, we take a more theoretical focus examining the costs and benefits of two popular approaches to model selection and parameter estimation (penalized likelihood and decomposable Bayesian).

The main objective of the first part (Chapter 3) was **to demonstrate that par-**

**tial correlation graphs facilitated different and better insight into high-dimensional data than sample correlations.** While the impetus for this was primarily motivational, it also gave a chance to begin comparing the performance of different approaches to parameter estimation.

In the second part (Chapters 4 and 5) we focused on situations where the true graph is non-decomposable. Here our first objective was **to understand what effect a restriction to decomposable models has on the estimation of both the inverse covariance matrix and the covariance matrix.** We consider the variability of estimates, as measured by the variance, comparing non-decomposable and decomposable models.

Finally we **compare the latest decomposable Bayesian and penalized likelihood methods on measures of model selection and prediction performance.**

## 1.2 Gaussian graphical models

We illustrate the relationship between the inverse covariance matrix and the graphical model with an low-dimensional example from Chapter 4 based on the Fisher's *Iris virginica* dataset. It consists of measurements of the sepal length, sepal width, petal width and petal length of 50 *Iris virginica* flowers. Figure 1.1 shows the graphical model and associated inverse covariance matrix ( $\Omega$ ) for this data set. In this simple example we can clearly see that the zero elements in the  $\Omega$  matrix correspond with the absence of edges in the graph. The graph shows that the direct relationships are between elements pertaining to the same part of the flower (sepal or petal) and between those pertaining to the same measurement (length or width). The properties of graphical models are explained more fully in Chapter 2.

There are two main attractions to using graphical models with high dimensional data (Lauritzen, 1996). Firstly graphical models are inherently modular in nature, thus in high-dimensions it is possible to work with simpler elements. Secondly graphical models provide natural data structures for digital computation which makes it

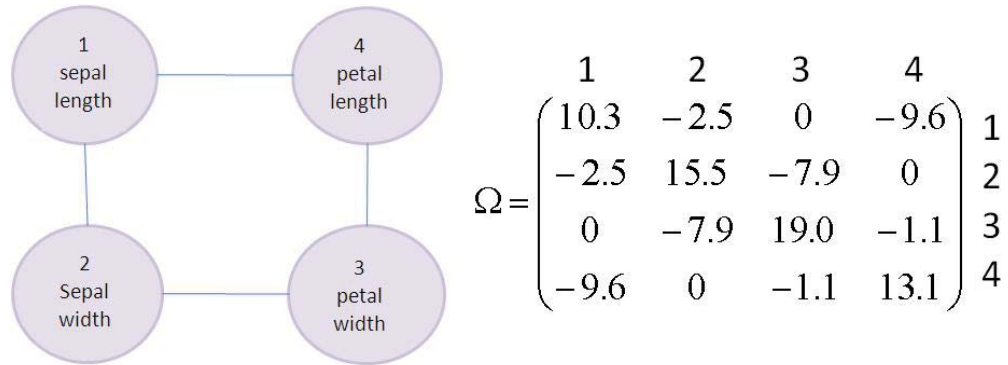


Figure 1.1: Example of a graphical model: Fisher’s *Iris virginica* dataset.

feasible to work with high-dimensional data. With lower dimensions, or very small subsets of high-dimensional data, we could also add that they give a simple visual representation of the independence structure.

Sparse graphs have a relatively small number of edges which, particularly in a high-dimensional setting, increases interpretability. They are commonly used in, for example, genomics and proteomics where they give some idea of cell pathways (Hastie *et al.*, 2009). This provides our entry point to consideration of GGMs. We begin with an application which motivates the use of GGMs: their use in representing gene association structures.

## 1.3 Overview

In Chapter 2 we give an overview of model selection and parameter estimation in the GGM framework, detailing the algorithms for the two penalized likelihood approaches used, namely graphical lasso and adaptive graphical lasso. We then highlight relevant graph theory, including more formally defining a Gaussian graphical model and defining a decomposable graph. Chapter 2 concludes with a section detailing feature-inclusion stochastic search, the decomposable Bayesian approach used in Chapter 5 and the formula used for obtaining predictions in the same chapter.

Chapters 3, 4 and 5 form the main body of work in this thesis. They are written as

three separate papers and, as such, each may be read independently of the others. Each chapter includes a brief review of the graph theory, model selection and/or parameter estimation methods pertinent to that chapter.

Chapter 3 details our motivational example: shortest path analysis using partial correlations for classifying gene functions from gene expression data. Graphical lasso is used to fit the GGMs and obtain partial correlations. We propose using the estimated partial correlations from these models to attach lengths to the edges of the GGM, where the length of an edge is inversely related to the partial correlation between the gene pair. The shortest paths between pairs of genes are found. Intermediate genes on the path are classified as having the same biological function as the terminal genes, if both the terminal genes have the same function. We validate the method using genes of known function from the Rosetta Compendium of yeast (*Saccharomyces cerevisiae*) gene expression profiles. We also compare our results with those obtained using a graph constructed using correlations. Using a partial correlation graph we are able to classify approximately twice as many genes to the same level of accuracy as when using a correlation graph. More importantly when both methods are tuned to classify a similar number of genes, the partial correlation approach can increase the accuracy of the classifications.

We move from motivating the use of sparse GGMs to consideration of computationally tractable methods for fitting them. Penalized likelihood approaches such as the graphical lasso (Friedman *et al.*, 2008b) used in Chapter 3 do not put any restriction on the configuration of edges in the graph. However, they are frequently criticized for including many small elements in the inverse covariance matrix. Bayesian approaches on the other hand generally yield a sparser model, but the computation time is longer and they often place restrictions on the models. In Chapter 3 we used graphical lasso to estimate the covariance matrix. As a comparison similar results were obtained using high-dimensional Bayesian covariance selection, which restricts model choice to those which can be represented as a directed acyclic graph.

Models that can be represented as a decomposable (triangulated) graph are more computationally tractable. In fact, in the high-dimensional Bayesian setting, it is

common to restrict model selection procedures to decomposable models (defined in Section 2.2). This restriction is made for computational convenience and research into the implications of this restriction is lacking. In Chapter 4 we focus on the cost, in terms of variability in parameter estimation, of using decomposable GGMs for computational convenience by examining the effect of adding extra edges to triangulate a non-decomposable graph.

We consider estimation of both the covariance matrix and the inverse covariance matrix, where the true model forms a cycle, but estimation is performed supposing that the pattern of zeros is a decomposable graphical model. We use a decomposable model where those elements not restricted to zero are a superset of those not restricted to zero in the true matrix. The variance of the maximum likelihood estimator based on the decomposable model is demonstrably larger than for the true non-decomposable model, and which decomposable model is selected affects the variance of particular elements of the matrix. When estimating the inverse covariance matrix the cost in terms of accuracy for using the decomposable model is fairly small, even when the difference in sparsity is large and the sample size is fairly small (for example the true model is a cycle of size 50, and the sample size is 51). However, when estimating the covariance matrix, the estimators for most elements had a dramatic increase in variance (200-fold in some cases) when a decomposable model was substituted. These increases become more pronounced as the difference in sparsity between models increases.

Chapter 5 completes our consideration of the costs and benefits of decomposable Bayesian and penalized likelihood approaches by comparing model selection and accuracy in prediction under both approaches when the true model is non-decomposable. Penalized likelihood approaches perform model selection in the context of parameter estimation. Bayesian approaches separate model selection from parameter estimation but frequently restrict consideration to decomposable models. When considering the variability of estimates, in Chapter 4 we assumed that the model selected was a superset of the true model. Here we begin by quantifying the scenarios under which this actually occurs when a decomposable Bayesian approach is used for model selection. We use the feature-inclusion stochastic search (Scott and Carvalho, 2008),

a Bayesian method, which Scott and Carvalho (2008) showed to have superior performance to both other Bayesian methods and early penalized regression methods. We find that for large samples, as expected, feature-inclusion stochastic search converges to supersets of the true model. For smaller sample sizes, and when there are elements with small partial correlations, at least one true edge is often missing from the graphs with the highest posterior probability. We then compare feature-inclusion stochastic search to two more recent penalized likelihood approaches, graphical lasso (Friedman *et al.*, 2008b) (also used in Chapter 3) and adaptive graphical lasso (Fan *et al.*, 2009). We make comparisons both in terms of model selection and prediction. Here we find that the graphical lasso is clearly outclassed by both feature-inclusion stochastic search and adaptive graphical lasso. The differences between feature-inclusion stochastic search and adaptive graphical lasso are not so clear cut.

We conclude in Chapter 6 by summarizing our findings, highlighting the original contribution this work makes to the knowledge of Gaussian graphical models. In our discussion here we draw together both parts by considering the implications of the properties observed in Chapters 4 and 5 in practical application such as the gene associations of Chapter 3. Included in this are suggestions for further research.

## 1.4 Publications

A version of Chapter 3: *Shortest path analysis using partial correlations* has been published in Bioinformatics as *Shortest path analysis using partial correlations for classifying gene functions from gene expression data* (Fitch and Jones, 2009). A version of Chapter 4: *The cost of using a decomposable model* is has been accepted for publication in The Journal of Computational Statistics and Data Analysis as *The cost of using decomposable Gaussian graphical models for computational convenience* (Fitch and Jones, 2012). We expect a further publication from the content of Chapter 5.

# Chapter 2

## Background

In this chapter we give material on key aspects of this thesis. In Section 2.1 we give an overview of model selection in the Gaussian graphical model (GGM) framework before focusing on penalized likelihood approaches to estimating the inverse covariance matrix. In Section 2.2 we give an overview of graph theory with an emphasis on decomposable graphs and the connections between graphical models and the inverse covariance matrix in the Gaussian setting (that is GGMs). We conclude this chapter by giving details of the the Bayesian model selection method used in Chapter 5 and by explaining how we go on to use our estimates to make predictions.

### **2.1 Gaussian graphical models: model selection and parameter estimation**

#### **2.1.1 Overview**

Dempster (1972), in his seminal paper *Covariance Selection*, introduced the concept of setting elements of the inverse covariance matrix to zero in order to reduce the number of parameters. He proposed using forward, or backward, selection and a likelihood ratio test to determine which elements should be set to zero. Since that time a large body of work has evolved, and various methods have been proposed



for model selection and parameter estimation. Traditional methods require the sample size ( $n$ ) to be greater than the number of parameters ( $p$ ). We focus here on methods that have been specifically developed for use with high-dimensional data, particularly where  $n < p$ .

One group of methods, scalable to high dimensional data, is based on the relationship between regression coefficients and elements of the inverse covariance matrix ( $\Omega$ ). If  $V$  is the set of all variables then:

$$\beta_{ij|V\setminus\{j\}} = \frac{-\omega_{ij}}{\omega_{ii}} \quad (2.1)$$

where  $\beta_{ij|V\setminus\{j\}}$  is the partial regression coefficient (Lauritzen, 1996). These methods all depend upon obtaining a (sparse) regression model for each variable in terms of the others. Key in this area, in terms of penalized likelihood approaches, is the work of Meinhausen and Bühlmann (2006) who use the lasso (Tibshirani, 1996) to obtain regression equations. Two different methods of obtaining the zero elements of  $\hat{\Omega}$  are proposed. Either  $\hat{\omega}_{ij}$  is deemed to be zero if one or both of  $\beta_{ij|V\setminus\{j\}} = 0$  and  $\beta_{ji|V\setminus\{i\}} = 0$  (an OR graph) or, alternatively, only if both of them are zero (an AND graph). They show that asymptotically the two models are the same. Equation (2.1) can be used to calculate estimated partial correlation coefficients from the partial regression coefficients giving:

$$\tilde{r}_{ij} = \frac{-\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}}$$

Other regression based methods use adaptations of the lasso such as the weighted lasso (Shimamura *et al.*, 2007), or other functions, for example a penalized loss function (Peng *et al.*, 2009) to determine the zero elements of  $\Omega$ . High-dimensional Bayesian covariance selection (Dobra *et al.*, 2004) is a Bayesian regression-based method. Unlike most Bayesian methods, high-dimensional Bayesian covariance selection does not restrict model selection to decomposable models. However this means that computation time is much greater than for other methods (see Chapter 3.4). Furthermore because model selection is done via an acyclic directed graph, this does result in some implied restrictions on the model space (albeit less than a restriction to decomposable models).

Another body of work focuses directly on the inverse covariance matrix. Schäfer and Strimmer (2005) apply a shrinkage algorithm to the correlation matrix, and use a false discovery rate estimate to determine when to set an element of the inverse covariance matrix to 0. A variety of different penalized likelihood methods have been proposed. In Section 2.1.2 we detail the graphical lasso (Friedman *et al.*, 2008b) and the adaptive graphical lasso (Fan *et al.*, 2009). These along with other adaptations such as the Smoothly Clipped Absolute Deviation (Fan *et al.*, 2009) apply an  $L_1$  penalty directly to the inverse covariance matrix. Others such as the fused lasso (Tibshirani *et al.*, 2005), the group lasso (Yuan and Lin, 2006) and the elastic net (Zou and Hastie, 2005) use a combination of  $L_1$  and  $L_2$  penalties. Many of these algorithms have been developed to suit data with particular structures.

A final group of Bayesian methods separate model selection from parameter estimation. Jones *et al.* (2005) used a Metropolis-based algorithm and Hans *et al.* (2007) used the parallel Shotgun Stochastic Search. We use the serial procedure feature-inclusion stochastic search (Scott and Carvalho, 2008) (see Section 2.3).

### 2.1.2 Penalized likelihood methods: graphical lasso and adaptive graphical lasso

We use two penalized likelihood methods to estimate  $\Omega$  in this thesis. These are the graphical lasso (in Chapters 3 and 5) and the adaptive graphical lasso (in Chapter 5).

The graphical lasso (Friedman *et al.*, 2008b) applies an  $L_1$  penalty directly to the inverse covariance matrix. Thus the objective function is

$$\log \det \Omega - \text{tr}(\Omega S) - \lambda \sum_{i=1}^p \sum_{j=1}^p |\omega_{ij}| \quad (2.2)$$

where  $\Omega$  is a  $p$ -dimensional positive definite matrix,  $S$  is the  $p$ -dimensional sample covariance matrix and  $\lambda > 0$  is the penalty.

The graphical lasso algorithm (Friedman *et al.*, 2008b), which estimates  $\Sigma$  rather than  $\Omega$  in the first instance, works as follows:

1. Start with  $D = S + \lambda I$ .
2. At each iteration, for  $j=1,2, \dots, p$ :
  - permute the rows and columns of  $D$  and  $S$  so that the target column ( $j$ ) is the last.
  - partition  $D$  and  $S$  as:

$$D = \begin{pmatrix} D_{11} & d_{12} \\ d_{12}^T & d_{22} \end{pmatrix} \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix} \quad (2.3)$$

- Solve the lasso problem

$$\min_{\beta} \left\{ \frac{1}{2} \|D_{11}^{\frac{1}{2}} \beta - b\|^2 + \lambda \|\beta\|_1 \right\} \quad (2.4)$$

where  $b = D_{11}^{-\frac{1}{2}} s_{12}$ .

- use the  $(p - 1)$ -dimensional solution vector  $\hat{\beta}$  to update  $D$  using  $d_{12} = D_{11}^{\frac{1}{2}} \hat{\beta}$ .

3. Continue until convergence.

The lasso problem (equation (2.4)) is solved using coordinate descent (Friedman *et al.*, 2007; Banjeee *et al.*, 2008). At each step the update has the form

$$\hat{\beta} \leftarrow \frac{St \left( u_j - \sum_{k \neq j} V_{kj} \hat{\beta}_k, \lambda \right)}{V_{jj}} \quad (2.5)$$

where  $V = D_{11}$ ,  $u = s_{12}$  and  $St$  is the soft threshold operator:

$$St(x, t) = \begin{cases} \text{sign}(x)(|x| - t) & \text{if } (|x| - t) \geq 0 \\ 0 & \text{if } (|x| - t) < 0 \end{cases} \quad (2.6)$$

$\hat{\Sigma}$  is estimated as  $D$  at convergence, with  $\hat{\Omega} = D^{-1}$  being calculated at convergence using the  $p$  stored  $\hat{\beta}$ , and the expressions

$$\hat{\omega}_{22} = 1/(d_{22} - d_{12}^T \hat{\beta}) \quad (2.7)$$

$$\hat{\omega}_{12} = -\hat{\beta} \hat{\omega}_{22} \quad (2.8)$$

Convergence is defined as occurring when the average absolute change in  $D$  is less than  $t$  times the average of the absolute value of the off-diagonal elements of the empirical covariance matrix  $S$ .  $t$  is a fixed threshold, set by default at 0.0001 in the R-package `glasso` (Friedman *et al.*, 2008a) which we use.

We note here that if the lasso problem (equation (2.4)) is solved using  $S_{11}$  rather than  $D_{11}$ , then the graphical lasso algorithm yields a solution to the Meinhausen and Bühlmann (2006) (OR-graph) approach. Thus their approach can be regarded as an approximation to graphical lasso.

If we set the penalty term  $\lambda$  to 0, then one iteration of the algorithm calculates the maximum likelihood estimate of  $\Omega$  ( $= S^{-1}$ ). If the graph structure is known, then the zero elements can be specified. In this case we replace  $D_{11}$  and  $\hat{\beta}$  with  $D_{11}^*$  and  $\hat{\beta}^*$  when solving the lasso problem at step 2.  $D_{11}^*$  is obtained from  $D_{11}$  by removing elements constrained to be zero.  $\hat{\beta}^*$  is then obtained by padding the solution  $\hat{\beta}^*$  with zeros in the appropriate positions. We use these two properties, and the R-package `glasso`, to obtain maximum likelihood estimates of graphs with a known structure in Chapter 4.

The algorithm can also be used with differing penalties for each element of the inverse covariance matrix by simply replacing  $\lambda$  in equation (2.5) with  $\lambda_{jk}$ . This fact enables us to also use the R-package `glasso` in our implementation of adaptive graphical lasso.

One criticism of the graphical lasso method that when a higher penalty is used to increase sparsity this results in all non-zero elements also shrinking further towards zero. Adaptive graphical lasso uses a weighted penalty in an attempt to overcome this bias. We follow the work of Fan *et al.* (2009) and define the weights to be

$$\zeta_{i,j} = 1/|\tilde{\omega}_{i,j}|^\gamma$$

where  $\tilde{\Omega} = (\tilde{\omega}_{i,j})_{1 \leq i,j \leq p}$  is any consistent estimate of  $\Omega$  and  $\gamma > 0$ . Thus for adaptive graphical lasso the objective function becomes

$$\log \det \Omega - \text{tr}(\Omega S) - \lambda \sum_{i=1}^p \sum_{j=1}^p \zeta_{i,j} |\omega_{ij}| \quad (2.9)$$

The advantage of using this weighted penalty is that at each iteration the weighting is applied in a manner which ensures that less shrinkage is applied to elements with current large magnitude estimates, thus reducing bias.

## 2.2 Graph theory

A graph  $\mathcal{G}$  consists of a set of vertices  $V$  and a set of edges  $E \subseteq (V \times V)$ . Our interest is in *undirected graphs* where  $e_{i,j} \in E \Leftrightarrow e_{j,i} \in E, \forall i, j \in V$ . In what follows we use the term graph to mean an undirected graph. Further we only consider graphs where the vertices represent continuous variables. Thus edges in our graphs are represented by lines and following the convention of Lauritzen (1996) vertices in our graphs are represented by circles. We stated in the introduction that an undirected graphical model may be used to represent the conditional independence structure whereby the absence of an edge implies that the two variables are conditionally independent. We now formally state this relationship as:

If  $\mathcal{G}=(V,E)$  is an undirected conditional independence graph with  $V=\{1,2,\dots,p\}$ , then  $(i,j) \notin E \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}, \forall i \neq j, i, j \in V$ .

A *complete* graph contains all possible edges. A *triangulated* (or *chordal*) graph (is an undirected graph that) contains no chordless cycles of four or more vertices, where a chord is an edge that joins two non-consecutive vertices.

$\mathcal{G}_A=(A,E_A)$  is a *subgraph* of  $\mathcal{G}$  if  $A \subseteq V$  and  $E_A=E \cap (A \times A)$ . A *clique* is a complete subgraph that is maximal with respect to the subset operator ( $\subseteq$ ).

### 2.2.1 Decomposability and decomposable graphs

The terms decomposable graph and triangulated graph are equivalent for undirected graphs(Lauritzen, 1996). Suppose  $A, B$  and  $C$  are disjoint subsets of  $V$ , each consisting of at least one vertex, and  $A \cup B \cup C = V$ .  $C$  *separates*  $A$  from  $B$  if all paths from  $A$  to  $B$  must go through  $C$ . If the subgraph  $\mathcal{G}_C$  is complete then  $A, B, C$  is a *decomposition* of  $\mathcal{G}$  into two components,  $A \cup C$  and  $B \cup C$  (see Figure

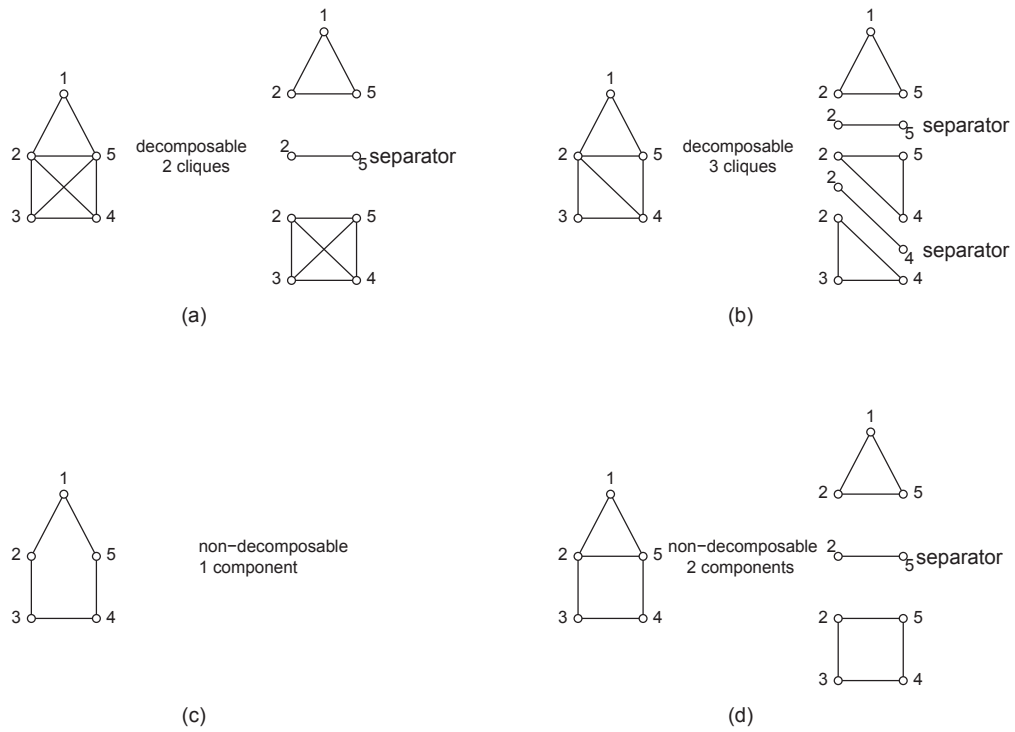


Figure 2.1: Graphs of five variables illustrating the idea of decomposition and decomposable versus non-decomposable.

2.1). A *prime component* is a subgraph that cannot be further decomposed. If we choose  $C$  so that it is the smallest complete subgraph that separates  $A$  and  $B$  then we can recursively decompose a graph into its prime components. Thus a complete prime component is a clique and a *decomposable graph* can be recursively decomposed into cliques (only). In Figure 2.1 graphs (a) and (b) are decomposable as all the subgraphs are complete. Figure 2.1 graph (c) is the 5-cycle, which has only one component and is non-decomposable. Many of the non-decomposable graphs we consider are cycles. Although we can find a decomposition for Figure 2.1 graph (d), as shown, the graph is non-decomposable because one of the components (the lower one) is neither complete nor able to be further decomposed.

Restricting consideration to decomposable graphs has advantages both in terms of tractability and computing time. Using these advantages usually requires not only identifying the cliques and separators, but obtaining a perfect ordering of them. If we have cliques  $(C_i)$  and separators  $(S_i)$  then in a perfect ordering  $(C_1, S_2, C_2, S_3, C_3, \dots, S_k, C_k)$  the separator  $S_i$  is the intersection of  $C_i$  with (the union of) all lower numbered cliques (see for example Jones *et al.* (2005)). There are many efficient algorithms for producing a perfect ordering, including those which order incomplete prime components as well as cliques (Dobra and Fienberg, 2000). It is necessary to order the vertices to obtain the ordering of the components; this is done via a perfect elimination ordering. We present here the maximum cardinality search algorithm of Berry *et al.* (2004). A maximal cardinality search algorithm is used to generate a perfect elimination ordering. If  $G$  is not decomposable then the maximum cardinality search adds edges so that a triangulated graph and its perfect elimination ordering are obtained. We note that it is possible to obtain more than one perfect elimination ordering for an undirected graph.

**Algorithm** MCS-M (Berry *et al.*, 2004) to obtain a perfect elimination ordering of the graph  $\mathcal{G}=(V,E)$ .

- Begin with all  $p$  vertices unnumbered and assigned a zero weight.
- Choose any vertex  $v \in V$  and number it  $p$ .
  1. Find all unnumbered vertices  $u \in V$ , where either  $uv \in E$  or there is a path from  $v$  to  $u$  and all the other vertices on the path have a weight less than the weight of  $u$ . Call this set of vertices  $U$ .
  2. Increase by 1 the weight of all vertices in  $U$ .
  3. For all the vertices  $u \in U$  if  $uv \notin E$  then triangulate by adding the edge  $uv$ .
- Select the (an) unnumbered vertex with the highest weight, number it  $p - 1$ , and repeat the numbered steps 1-3.
- Continue in the same manner decreasing the vertex number by 1 each time until all vertices have been numbered.

Once the vertices are numbered according to a perfect elimination ordering, then a perfect ordering of cliques proceeds as follows:

- Assign all vertices a weight of zero initially.
- Choose the highest numbered vertex ( $p$ ) as  $v$ :
  1. Let  $X$  be the set of all vertices  $x \in V$  where  $vx \in E$  and  $x$  has a lower numbering than  $v$ .
  2. If  $X$  is empty because it is not connected to any other vertices, then  $v$  forms its own clique.
  3. If  $X$  is empty because it is only connected to higher numbered vertices, then all cliques have been found.
  4. Otherwise let  $u$  be the highest numbered vertex in  $X$  and assign it a new weight equal to the maximum of its current weight and the number of other vertices in  $X$ .
  5. If the weight of  $v$  is less than the number of vertices in  $X$  then clique 1 is  $v \cup X$ .
- Repeat steps 1 to 5 for vertices  $p - 1$  to 1, increasing the clique number by 1 at step 5 each time a clique is added.

A perfect elimination ordering of the vertices and the corresponding perfect ordering of cliques enables the formation of a junction tree. The mathematical properties of a junction tree are set out in Lauritzen (1996, Chapter 2). It is sufficient here to say that a junction tree is an undirected tree, with vertices corresponding to cliques and edges to the existence of a separator between the cliques. The feature-inclusion stochastic search algorithm (see Section 2.3) uses the junction tree to improve efficiency. We also use a perfect elimination ordering of decomposable graphs in the identification of cliques and separators when estimating parameters using Bayesian model selection.



### 2.2.2 Gaussian graphical models

We restrict our interest to data drawn from a multivariate normal distribution with covariance matrix  $\Sigma$ . Without loss of generality, we assume all variables have been centred and thus have a mean of 0. If we scale the inverse covariance matrix so that the diagonals are all 1, then the off diagonal elements are the negative partial correlations. Thus the the pattern of zeros in the inverse covariance matrix ( $\Omega=\Sigma^{-1}$ ) represents the conditional independencies in the distribution and can also be seen in the absence of edges in the graph.

If  $X_p \sim \text{MVN}(0, \Omega^{-1})$ , then  $\mathcal{G}=(V, E)$  is a *Gaussian graphical model* means that

$$e_{i,j} \notin E \Leftrightarrow \omega_{i,j} = 0, \forall i \neq j \text{ and } i, j \leq p.$$

In the Gaussian setting decomposability means that the density can be written as a function of the densities of the prime components and separators. Thus

$$f(x) = \frac{\prod_{j=1}^k f(x_{P_j})}{\prod_{j=2}^k f(x_{S_j})} \quad (2.10)$$

where  $P_1, \dots, P_k$  is the sequence of  $k$  prime components and  $S_2, \dots, S_k$  the corresponding sequence of  $k - 1$  separators.

If  $S_j = P_i \cap P_j$  for some  $i < j$  means that the elements of  $\Sigma_{S_j}$  are common in  $\Sigma_{P_i}$  and  $\Sigma_{P_j}$ , and that  $(\Sigma|b, D, G) \sim \text{HIW}_G(b, D)$ , then the density of  $\Sigma$  can be written in a similar manner (Scott and Carvalho, 2008; Jones *et al.*, 2005) as:

$$p(\Sigma|b, D, G) = \frac{\prod_{P \in \mathcal{P}} p(\Sigma_P|b, D_P)}{\prod_{S \in \mathcal{S}} p(\Sigma_S|b, D_S)} \quad (2.11)$$

In this case each prime component  $P$  has an inverse Wishart distribution,  $\Sigma_P \sim \text{IW}(b, D_P)$ , with density,

$$p(\Sigma_P|b, D_P) \propto |\Sigma_P|^{(-b/2+|P|)} \exp\left\{-\frac{1}{2} \text{tr}(\Sigma_P^{-1} D_P)\right\}. \quad (2.12)$$

This density only has closed form for decomposable graphs (that is for cliques not incomplete prime components). If our prior is  $\Sigma \sim \text{HIW}_G(b, D)$ , then the conjugate posterior is  $(\Sigma|X) \sim \text{HIW}_G(b + n, D + X^T X)$  where  $X^T X$  is the sum of squares matrix for the data matrix  $X$ .

We make particular use of this fact in Chapter 5, when calculating the marginal likelihood of a graph using the fractional Bayes approach of Carvalho and Scott (2009). Fractional Bayes as introduced by O’Hagen (1995) is regarded as a useful model selection technique when prior information is weak. The basic idea of fractional Bayes is to train a non-informative prior for each model using a small fractional power ( $b$ ) of the likelihood function. This is done simultaneously for all models, converting all the non-informative priors into proper priors for selecting the model using the remainder of the likelihood. O’Hagen (1995) show that if  $b \rightarrow 0$  as  $n \rightarrow \infty$  then fractional Bayes is a consistent procedure.

In the GGM context the conventional choice of prior is  $\Sigma \sim HIW_G(\delta, \tau I)$ . In the fractional Bayes approach the hyper-inverse Wishart  $g$ -prior,  $(\Sigma|G) \sim HIW_G(gn, gX^T X)$ , is used, where  $g$  is the fractional power ( $b$ ). The  $g$ -prior is preferred because it gives a “sharper characterization of the model uncertainty” (Scott and Carvalho, 2008, page 794). In contrast the conventional prior induces a set of ridge-regression priors on each univariate conditional regression which leads to an artificial flattening of the modes in model space (Scott and Carvalho, 2008; Zellner and Siow, 1980).

When using the hyper-inverse Wishart  $g$ -prior the data is taken to be  $(1 - g)X^T X$  and represent  $(1 - g)n$  observations resulting in the marginal likelihood:

$$p(X|\mathcal{G}) = (2\pi)^{-np/2} \frac{h(\mathcal{G}, gn, gX^T X)}{h(\mathcal{G}, n, X^T X)} \quad (2.13)$$

where we set  $g$  as  $1/n$  and the function  $h$  is the normalizing constant of a hyper-inverse Wishart distribution. If  $\mathcal{P}$  represents the set of all prime components and  $\mathcal{S}$  the set of all separators in  $\mathcal{G}$  then

$$h(\mathcal{G}, b, D) = \frac{\prod_{P \in \mathcal{P}} \left| \frac{1}{2} D_P \right|^{\frac{(b+|P|-1)}{2}} \Gamma_{|P|} \left( \frac{(b+|P|-1)}{2} \right)^{-1}}{\prod_{S \in \mathcal{S}} \left| \frac{1}{2} D_S \right|^{\frac{(b+|S|-1)}{2}} \Gamma_{|S|} \left( \frac{(b+|S|-1)}{2} \right)^{-1}} \quad (2.14)$$

and  $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(x+(1-j)/2)$  is the multivariate gamma function.

Setting  $g = 1/n$  ensures that a minimum training sample of effective size 1 is used. The modified likelihood equation prevents prevents ‘double use’ of the data, with the implicit sample size being  $n - 1$ .

In the maximum likelihood framework, when the graph structure (and hence position of zero elements) is known, direct (analytic) estimates of the inverse covariance matrix are only possible if the graphical model is decomposable and the sample size ( $n$ ) is greater than the number of nodes in the largest clique (Whittaker, 2008; Wermuth, 1980; Lauritzen, 1996). In this case the estimate can be calculated using equation (2.15) where  $(K)^0$  is the extension of matrix  $K$  with zeros,  $\mathcal{C}$  is the set of all cliques and  $\mathcal{S}$  the set of all separators.

$$\hat{\Omega} = \sum_{C \in \mathcal{C}} (\hat{\Sigma}_C^{-1})^0 - \sum_{S \in \mathcal{S}} (\hat{\Sigma}_S^{-1})^0 \quad (2.15)$$

While equation (2.15) also holds in the situation where not all prime components are cliques, numerical solutions are required for finding  $\hat{\Sigma}_P^{-1}$ , when  $P$  is a prime component that is not a clique.

### 2.2.3 Acyclic directed graphs

In practical applications one is often interested in identifying causal relationships. In theory, causal relationships could be depicted using an acyclic directed graph (ADG), with the directed edges corresponding to causality. Although on their own Gaussian graphical models do not suggest causal relationships, in many cases it is possible to order the vertices and assign direction to edges. Such representations are not unique and are only possible if the graph is decomposable. ADGs can be grouped together into equivalence classes corresponding to the same inverse covariance structure. If the structure of the ADG corresponds to a decomposable structure then the corresponding undirected graph will have the same decomposable structure. If the structure of an ADG corresponds to a non-decomposable structure then the ‘moralizing’ algorithm used to convert ADGs to undirected graphs adds ‘moralized’ edges to the graph. The elements in the inverse covariance matrix corresponding to these ‘moralized’ edges are non-zero but not free to vary.

A non-decomposable undirected graph does not correspond directly with any ADG. Cox and Wermuth (2000), however, consider the case of the chordless four-cycle,

the simplest non-decomposable graph. They show that a chordless four-cycle may represent the situation where there are two unobserved variables. In this case there is a related ADG which assumes the existence of two latent variables and contains two additional vertices representing them.

The class of ADGs thus while containing more graphs than the class of decomposable graphs still represents a subset of all possible graphs. Furthermore, even in the situation where the data clearly implies a direction to the edge, care must be taken in assigning causality. As with any situation with observational data prior knowledge and scientific information should be used to confirm what the observed data indicates. These many instances when there is not a direct a correspondence between causality, ADGs and the inverse covariance matrix suggest that even when our ultimate goal is to understand a biological system it is worthwhile looking at unrestricted graphical models.

## 2.3 Decomposable Bayesian model selection using feature-inclusion stochastic search

In Chapter 5 we use the feature-inclusion stochastic search (FINCS) algorithm for Bayesian model selection (Scott and Carvalho, 2008). FINCS restricts the search to the space of decomposable graphs for computational convenience. FINCS, unlike many Bayesian algorithms, is not a sampling algorithm. The FINCS algorithm is a search algorithm which looks for posterior modes and retains a set of ‘top models’. It is a serial procedure that combines three types of moves through the space of all possible (decomposable) graphs. As each new model is found the model score (log of the non-normalized posterior probability) is calculated using:

$$\text{model score} = \log P(X|\mathcal{G}_k) + \log mc(\mathcal{G}_k) \quad (2.16)$$

where  $P(X|\mathcal{G}_k)$  is calculated using equation (2.13) and  $mc(\mathcal{G}_k)$ , the multiplicity correction prior over the graphs, using equation (2.18) As with all Bayesian algorithms normalizing of model scores to obtain the posterior probabilities can only occur

once the search (sampling) is complete. Exact normalization is only possible when all models have been visited. In this situation because only top models are retained normalization can only be done relative to the retained models.

Most moves are local moves which exploit the computational advantages of adding or deleting only one edge at a time. The decision to add or delete is randomly chosen. If addition is selected, then the edge to add is selected in proportion to its relative inclusion probability (equation (2.17)) and correspondingly if deletion, then the edge to delete is selected in proportion to the inverse of its relative inclusion probability. In both cases edges are also chosen so that decomposability is maintained. This is done by considering the effect of the added (or deleted) edge on the cliques that the two endpoint vertices belong to. Adding an edge either adds a totally new clique (and separator), results in two (or more) cliques amalgamating to one larger or, if allowed to occur, would create a non-decomposable component. Correspondingly deleting an edge either totally removes a clique (and separator), splits a clique into two or more smaller cliques or, if allowed to occur, would turn a clique into a non-decomposable component. The adding and subtracting are both done via the junction tree as only cliques directly connected to cliques containing the affected vertices may change. Computations are thus simplified by using the junction tree, which is updated after every local move. Updates to the model scores, therefore, only involve additions and subtractions relative to the changes in cliques and separators.

Global moves are used to move to another part of the graph space in order to avoid missing regions that are not easily found in stepwise moves. A global move is achieved by generating a randomized median triangulation pair. A randomized median triangulation pair is found by starting with an empty graph and adding edges in proportion to their current relative inclusion probability. The graph so formed ( $\mathcal{G}_N$ ) is usually not decomposable so a minimal decomposable supergraph ( $\mathcal{G}^+$ ) and a maximal decomposable subgraph ( $\mathcal{G}^-$ ) are found. Model scores are then calculated for both  $\mathcal{G}^+$  and  $\mathcal{G}^-$  and the one with the highest model score chosen as the new current graph.

Finally resampling moves revisit graphs in proportion to their model score and

thereby ensure that the global moves do not irretrievably direct the search away from ‘good’ graphs. Previously visited models are stored in a binary search tree indexed by a normalized model score. A beta distribution is used to represent the empirical distribution of these scores on the interval  $(0,1]$ . The parameters of the beta distribution are updated every time a substantial pocket of probability is found in the model space. The implementation of FINCS then uses an approximate resampling, by drawing a score from this beta distribution and resampling the model with the score closest to the one selected. This allows substantial gains in efficiency (see Scott and Carvalho (2008, Section 3.3)).

For each edge  $e_{i,j}$  the inclusion probability at step  $t$  is estimated by the relative inclusion probability

$$\hat{q}_{ij}(t) = \frac{\sum_{k=1}^{k=t} 1_{(i,j) \in \mathcal{G}_k} P(X|\mathcal{G}_k)\pi(\mathcal{G}_k)}{\sum_{k=1}^{k=t} P(X|\mathcal{G}_k)\pi(\mathcal{G}_k)} \quad (2.17)$$

$P(X|\mathcal{G}_k)$  is calculated using equation (2.13) and  $\pi(\mathcal{G}_k)$  is the prior probability of the graph. As the relative inclusion probabilities are only based on the graphs visited they do not converge to the true inclusion probabilities except in the trivial sense of all models eventually being enumerated.

Scott and Carvalho (2008) use a multiplicity-correction prior over graphs which is motivated by the standard binomial prior, where edge inclusions have a binomial distribution with success probability  $r$ . The multiplicity correction places a conjugate beta prior on  $r$ , with the default uniform prior (a beta distribution with parameters  $a = b = 1$ ) giving a marginal prior inclusion of 0.5 for all edges. This then means that

$$\pi(\mathcal{G}_k) \propto mc(\mathcal{G}_k) = \frac{(\kappa)!(m - \kappa)!}{(m + 1)(m!)} \quad (2.18)$$

where  $\mathcal{G}_k$  has  $\kappa$  edges out of  $m = p(p - 1)/2$  possible edges.

The FINCS algorithm thus generates a set of ‘top’ graphs and their model scores. The more usual Bayesian approach is a Markov chain Monte Carlo (MCMC) sampling algorithm where posterior probabilities can be estimated using a graph’s frequency in the sample of graphs produced. However, for problems with even a moderate number of variables and a restriction to decomposable models, the sampling

space is so large that a graph's frequency in the sample does not reflect its posterior probability (Jones *et al.*, 2005). In search algorithms such as FINCS, which only retain a set of top graphs, the exponentiated model scores can only be normalized with respect to the retained graphs. Only at very small dimensions is there any sense that the retained graphs represent most of the posterior probability and only in these situations can the normalized scores be thought of as real posterior probabilities. Nevertheless, the relative size of the model scores is true. We, therefore, refer to normalized exponentiated model scores as relative posterior probabilities. Model averaging using the relative posterior probabilities as weights can be useful to reflect some degree of the model uncertainty if the retained models are truly different models. We will see this is less effective when the retained models are very similar.

For any graph we use its adjacency matrix to identify cliques and separators. Parameter estimation is then done by finding the posterior mean for each clique ( $\bar{\Omega}_C$ ) and separator ( $\bar{\Omega}_S$ ) and combining them using equation (2.15) where for any clique  $C$  (or similarly separator  $S$ ):  $\bar{\Omega}_C = \frac{1}{n+p_C-1}(\text{ssd}_C)^{-1}$ , with  $\text{ssd}$  the sample sums of squares matrix and  $p_C$  the number of vertices in  $C$ .

## 2.4 Predictions

Using Gaussian data also enables simple predictions to be made. We first estimate  $\hat{\Sigma}$  as  $\hat{\Omega}^{-1}$  where  $\hat{\Omega}$  has been estimated from a data set with observation for all variables. We then partition the prediction data set into two groups  $X_1$  and  $X_2$  where  $X_1 = x_1$  are assumed known. In Chapter 5 we compute the conditional expectation of  $X_2$  based on the 'observed' values ( $X_1 = x_1$ ). This is done using the standard relationship:

$$(\mu_2|X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$

where because our data is centred  $\mu_1 = \mu_2 = 0$ .

# Chapter 3

## Shortest path analysis using partial correlations

A version of this material appeared in 2009 as *Shortest path analysis using partial correlations for classifying gene functions from gene expression data* by A. Marie Fitch and Beatrix Jones in *Bioinformatics* 25(1), pages 42-47.

### 3.1 Introduction

Gaussian graphical models (GGMs) and related methods are a popular tool for representing gene association structures in microarray experiments (Matusno *et al.*, 2006; Toh and Horimoto, 2002; Aburatani *et al.*, 2003; Wille *et al.*, 2004; de la Fuente *et al.*, 2004; Dobra *et al.*, 2004; Schäfer and Strimmer, 2005; Shimamura *et al.*, 2007; Banjeree *et al.*, 2008). In a GGM edges correspond to non-zero elements in the inverse covariance matrix (Dempster, 1972). The absence of an edge represents conditional independence between two genes. Biological validation of the use of GGMs with gene expression data has tended to be done on a small scale. Banjeree *et al.* (2008) focus on 2 sub-graphs, Shimamura *et al.* (2007) and Wille *et al.* (2004) use data sets with small numbers of genes and others such as Toh and Horimoto (2002) and Aburatani *et al.* (2003) validate the use of a network of clusters rather than individual genes.



We introduce a method for classifying gene functions using GGMs. The results of this procedure on expression data from genes with known function validate the use of GGMs with gene expression data on a large scale basis. The procedure would provide a method of ‘mining’ the graph for functional information when used with data where the the function of some (but not all) genes are unknown.

Our method is motivated by the shortest path analysis of Zhou *et al.* (2002). In their approach, a graph is created where edges occur between genes (nodes) with high expression correlations. The length of an edge is inversely related to the correlation between the gene pair. The shortest paths between pairs of genes are found. In the context of this study the existence of *transitive genes* implies that the shortest path between two genes will include at least one other gene. For example, the shortest path from gene 5 to gene 9 may be gene 5, gene 13, gene 192, gene 245, gene 9. In this example gene 5 and gene 9 are the terminal genes and gene 13, gene 192 and gene 245 are the *transitive genes*. If the shortest path between two genes includes at least one transitive gene, then it is postulated that the transitive genes on the path will be involved in the same biological process as the terminal genes. Gene function is described using categories from the Gene Ontology database (see section 3.2.1).

We propose using a GGM rather than a correlation based graph, with the edge lengths inversely related to the partial correlations derived from the inferred inverse covariance matrix. Because a GGM reflects the conditional independence structure, it is a more natural way in which to capture the situation when the relationship between genes is mediated by another gene. In the multivariate normal distribution setting, the partial correlations enable us to fully quantify the relationship of two variables given all the other variables. When dealing with gene expression data we typically have the situation where the number of variables ( $p$ ) is much greater than the sample size ( $n$ ) which can create problems in obtaining a consistent estimator of the inverse covariance matrix. Many researchers work with subgroups, or smaller networks (Matusno *et al.*, 2006; Toh and Horimoto, 2002; Aburatani *et al.*, 2003), or propose methods using the more easily calculated first (or second) order partial correlations to visualise conditional independencies rather than the inverse

covariance matrix (Wille *et al.*, 2004; de la Fuente *et al.*, 2004). We have identified several different methods for obtaining an estimator of the inverse covariance matrix. Dobra *et al.* (2004), Meinhausen and Bühlmann (2006), and Shimamura *et al.* (2007) use regression based methods to estimate the partial regression coefficients and thence inverse covariance matrix, Schäfer and Strimmer (2005) use shrinkage of the covariance matrix, and Friedman *et al.* (2008b) apply an  $L_1$  penalty to the inverse covariance matrix. We use this final method, graphical lasso, to obtain the partial correlations.

We find that using the partial correlation graph rather than the correlation graph enables more transitive genes to be identified. In addition, when the correlation and partial correlation graphs are tuned to attempt a similar number of gene classifications, the classifications derived from the partial correlation graph are more accurate for some cellular compartments.

The rest of this chapter is organised as follows. In Section 3.2 we detail the dataset used and overview the use of shortest path analysis, using graphical lasso to derive partial correlations and then using these derived partial correlations to obtain shortest paths between gene pairs. In Section 3.3 we present results using the derived partial correlations and compare them to those using correlations. Finally in Section 3.4 we discuss the advantages of using partial correlations.

## 3.2 Data and methods

### 3.2.1 Data

In this section we use the yeast (*Saccharomyces cerevisiae*) gene expression profiles from the Rosetta Compendium (Hughes *et al.*, 2000). The Rosetta Compendium data gives results ( $\log_{10}$  (ratios)) from 300 deletion and drug treatment experiments. The dataset that we use is restricted to those genes with known functions and, with minor exceptions detailed below, is the same dataset as used by Zhou *et al.* (2002) to validate their work.

The functional categories for each gene from the Gene Ontology (GO) database are used (<http://www.yeastgenome.org>). We used the 2002 version to enable straightforward comparisons with the work of Zhou *et al.* (2002). Although the database includes three ontologies based on molecular function, biological process, and cellular component, only the biological processes ontology is used. The ontology is organised so that parent-child relationships between the descriptive terms are defined. This enables the ontology to be represented as a tree with annotations becoming more detailed as one moves down the tree. Furthermore a gene may have more than one annotation within the ontology.

In order to facilitate comparisons with the results of Zhou *et al.* (2002) we organised our datasets using similar criteria. We excluded any genes with no GO annotations, those with the GO annotation ‘biological process unknown’ and those for which there were less than 100 experimental results. We increased the minimum number of experimental results required, from the 80 used by Zhou, to 100 as we found that at least 100 experimental results were required for graphical lasso regression, implemented using the R-package `glasso` (R Development Core Team, 2009; Friedman *et al.*, 2008a), to work. Three datasets were created - one for each of the major cellular compartments (mitochondria, cytoplasm and nucleus). The cytoplasm dataset was identical to that used by Zhou. However our nucleus data set contained two fewer genes due to the requirement for more experiments and our mitochondria dataset was slightly smaller (261 genes compared to 266). The datasets for each of the three cellular compartments are worked with separately. *Informative* annotations are found by identifying GO nodes that satisfy the properties that (i) the node contains more than  $\gamma = 30$  genes and (ii) each of the node’s children contains fewer than  $\gamma$  genes.

### 3.2.2 Shortest path analysis

For each pair of genes Zhou *et al.* (2002) use the minimum of the absolute value of the leave-one-out Pearson’s correlation coefficients, as a robust measure of their correlation. A separate graph is formed for each of the three datasets. Zhou *et al.*

define three tuning parameters. The first of these is a threshold for including edges in the graph: an edge is included when the absolute value of their correlation ( $|\rho|$ ) is greater than 0.6. The ‘length’ of that edge is set to be  $(1-|\rho|)^6$ . The second tuning parameter, the exponent in this expression, is used to enhance the differences between low and high correlations, thus creating shortest paths that are likely to cover more transitive genes. The shortest path between pairs of vertices is calculated using Dijkstra’s algorithm (Dijkstra, 1959). A final tuning parameter discards paths with length above some threshold: to be used for classification a path needed to have an overall length less than 0.008. We apply Zhou’s method to the datasets described in section 3.2.1 and vary the restriction on the overall length from 0.001 to no restriction at all for comparison purposes. Consequently, our results show minor variations to those published in Zhou *et al.* (2002)

The shortest path is found between each gene pair with at least one common informative annotation. The common informative annotation is ‘attached’ to this shortest path when the path includes more than the two terminal genes (i.e. the path contains transitive genes). The path is recorded separately with each different annotation when the two terminal genes share more than one identical informative annotation. We call each recorded instance a categorised path. The number of transitive gene categorisations is determined by counting all transitive genes from all categorised paths.

Transitive genes are classified as having the same biological function as the terminal genes. To assess the validity of the shortest path method all transitive gene categorisations are then checked to see if they match exactly to the informative annotation of the terminal genes (a level 0 match) or if they share the same direct parent ontology node with the terminal genes (a level 1 match). We adopt the convention used by Zhou *et al.* (2002) whereby all level 0 matches are also included as level 1 matches.

### 3.2.3 Graphical lasso

We use estimated partial correlations, rather than correlations to create our graph. Partial correlations are a more natural way of capturing the situation where the relationship between two genes is mediated by another gene. We obtain a sparse inverse covariance matrix using graphical lasso as proposed by Friedman *et al.* (2008b). The lasso (“least absolute shrinkage and selection operator”) method was first proposed by Tibshirani (1996) in the context of regression and uses an  $L_1$  penalization. An attractive attribute of  $L_1$  penalization is that it shrinks some elements of the inverse covariance to exactly zero, which corresponds to an absence of an edge in our graph.

Friedman *et al.* (2008b) follow Banjeree *et al.* (2008) and apply the  $L_1$  penalty directly to the inverse covariance matrix. We denote the penalty parameter as  $\lambda$ . If  $\Omega = \Sigma^{-1}$  denotes the inverse covariance and  $S$  denotes the empirical covariance matrix, then the inverse covariance is estimated by maximising

$$\log \det \Omega - \text{trace}(S \Omega) - \lambda \|\Omega\|_1 \quad (3.1)$$

over positive semi-definite matrices  $\Omega$ .

Friedman *et al.* (2008b) recommend using a ‘likelihood approach’ 10-fold cross validation for estimation of the penalty parameter  $\lambda$ . Using the R-package `glasso`, we apply the graphical lasso to nine-tenths of the data for different values of  $\lambda$ , then evaluate the penalized log-likelihood using equation 3.1 over the validation set. For each value of  $\lambda$  the 10 values obtained in this way are averaged. The value of  $\lambda$  which maximizes the average validation penalized log-likelihood is selected as the penalty parameter. These optimal values of  $\lambda$  (cytoplasmic 0.08, mitochondrial 0.085 and nuclear 0.13) are used in graphical lasso to obtain an estimate of the partial correlations as described in the next section.

### 3.2.4 Using partial correlations in shortest path analysis

Like Zhou *et al.* (2002), we use a leave-one-out approach in order to obtain a model which is robust to outliers. Once the optimal penalty parameters have been determined as described above, the inverse covariance matrices for each cellular compartment are estimated by repeatedly applying glasso to the data leaving out one experiment at a time, using the same penalty parameter each time. The minimum leave-out-one partial covariances are found by finding the element-wise minimum of the absolute values of the leave out one inverse covariance matrices. Let  $\tilde{\omega}_{ij}$  be the elements of the minimum absolute value matrix, but with their signs restored. A graphical model is then obtained with edges where every  $\tilde{\omega}_{ij}$  is non-zero, and edge lengths set as  $1-|\tilde{r}_{ij}|$ ; where the robust partial correlation  $\tilde{r}_{ij}$  is computed as:

$$\tilde{r}_{ij} = \frac{-\tilde{\omega}_{ij}}{\sqrt{\tilde{\omega}_{ii}\tilde{\omega}_{jj}}}.$$

As for the correlation graphs, the shortest path between pairs of genes in this graphical model is calculated using Dijkstra’s algorithm. Following Zhou *et al.* (2002), we tune our classifier by using only paths below some maximum overall length for classification. The number of categorised paths, transitive gene categorisations, Level 0 and Level 1 matches are calculated as described in Section 3.2.2. We repeat these calculations for a variety of upper bounds on the path length. This was done separately for each cellular compartment.

## 3.3 Results

The number of categorised paths obtained varied considerably between cellular compartments. With no restriction on the total path length, mitochondria had 6 058 categorised paths, cytoplasm 31 712 and nuclear 27 805. The number of categorised paths was calculated for path length upper bounds between 1.7 and 2.1. Figure 3.1 displays these as a fraction of the number of categorised paths obtained with no upper bound. For all three cellular compartments the number of categorised paths

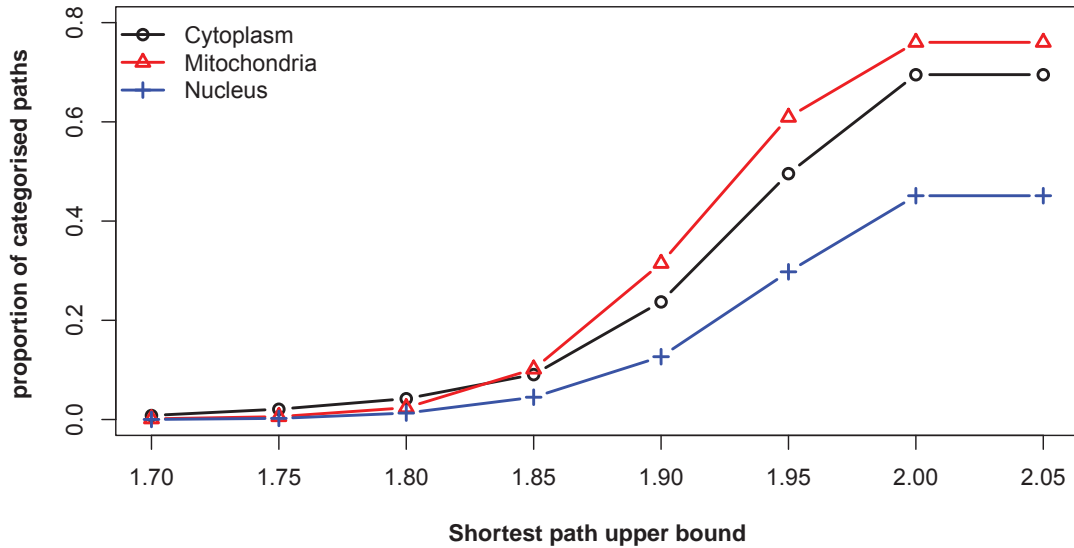


Figure 3.1: For each cellular compartment the number of categorised paths that are less than upper bounds between 1.7 and 2.1 is expressed as a proportion of the number of categorised paths obtained with no upper bound.

begins to rise steeply around an upper bound of 1.8 and flattens after 2. An upper bound of 2 includes all paths containing two edges. A similar pattern of steepening and flattening can be observed around higher integer values as all paths with a given number of edges are included.

The number of transitive gene categorisations identified is directly related to the number of categorised paths. If we remove the upper bound on the path length, then the maximum number of transitive gene categorisations identified was 6 to 100 times greater using partial correlations. In Figure 3.2 we show the percentage of matches (level 0 and level 1 matches as defined in Section 3.2.2) obtained as a function of the number of transitive gene categorisations identified using correlations and partial correlations. The number of categorisations was varied by changing the upper bound on path length, independently for each method. The percentage of matches decreases as the number of transitive gene categorisations increases. When the two methods are compared for a similar number of transitive gene categorisations (i.e. a vertical slice of a panel in Figure 3.2) we observed that over most of the range

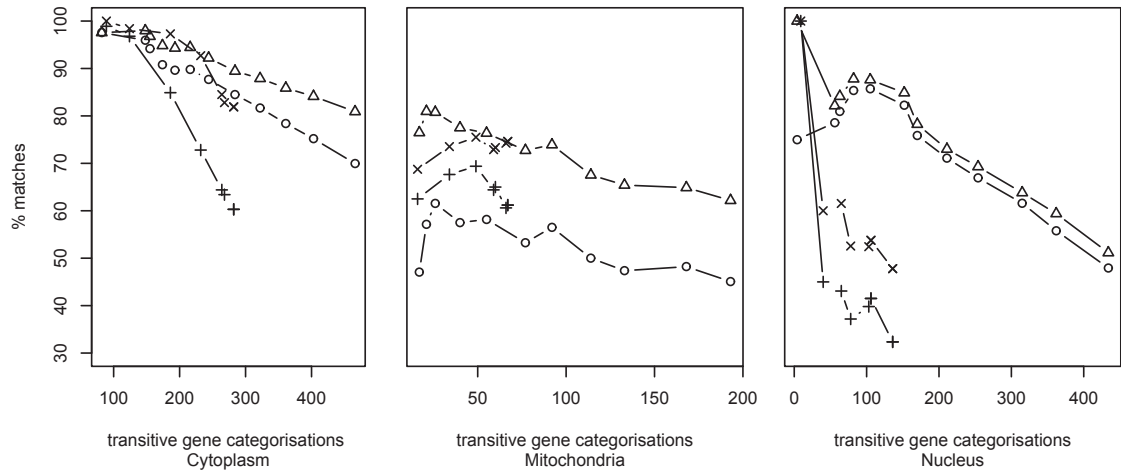


Figure 3.2: Percentage of transitive genes with annotations matching the annotations of the terminal genes on the same path.  $\circ$  = level 0 matches (partial correlations),  $\triangle$  = level 1 matches (partial correlations),  $+$  = level 0 matches (correlations)  $\times$  = level 1 matches (correlations)

using partial correlations gave a similar (mitochondria) or higher (cytoplasm and nucleus) percentage of matches compared with using correlations.

If no upper bound is used for the correlation graph, and an upper bound of 1.85 used for the partial correlation graph, then the two approaches produce a similar percentage of matches (correct or near correct categorisations), but the partial correlation approach identifies categories for approximately twice as many genes. (This is a comparison between the last point in each of the traces shown in each panel of figure 3.2.) Thus using partial correlations enables the possibility of identifying more transitive genes than using correlations. More importantly, using partial correlations can increase the percentage of matches when the same number of transitive genes are identified.

The graphs obtained using partial correlations behaved quite differently to those obtained using correlations. To illustrate and explore this we consider the cytoplasm genes RPS29B, RPL37B, RPS27A, RPS21A, RPL37A, RPS29A and RPL29. Zhou *et al.* (2002) identify the shortest path RPL37A-RPL37B-RPS29B-RPS29A-RPL29-RPS21A as an example where their method works well because it contains genes that are not tightly coregulated and yet all are involved in the GO process



protein synthesis. We consider only categorised path graphs, that is graphs where only edges that are part of a categorised path (see Section 3.2.2), and thus impact on the categorisations made, are included. For the correlation categorised path graph we use the same upper bound on shortest path length (0.008) as Zhou *et al.* (2002). With this cut off 64.4% of transitive genes are correctly categorised (Level 0 match) and a further 20.1% are almost correctly categorised (Level 1 match). For the partial correlation categorised path graph we use an upper bound on the shortest path length of 1.85, which gives similar percentages of correct categorisations; 70.0% of transitive genes are correctly categorised and a further 10.9% almost correctly categorised.

The nodes for the example genes are shown in the subgraphs in Figure 3.3 (A) and (B). Edges between genes on the graph are displayed, while the presence (but not exact number) of edges connecting to other genes is indicated. Colour is used to distinguish distinct categorised paths between the example genes. Both categorised paths in the partial correlation subgraph contain two edges, while two of the three in the correlation subgraph contain five edges. In the full correlation categorised path graph gene RPS27A is not transitive on any path and thus is not categorised. Gene RPL29 is the only gene from the subgraph that is transitive in more than two distinct categorised paths in the full correlation categorised path graph.

The subgraphs suggest that the partial correlation categorised path graph may be less sparse than the correlation categorised path graph. Consideration of the full categorised path graphs showed that the correlation categorised path graph contained 302 edges while the partial correlation categorised path graph contained 1211. Consideration of categorised path graphs with the same upper bounds on maximum path length for the other two cellular compartments revealed that the mitochondrial partial correlation categorised path graph also contained four times as many edges as the correlation categorised path graph, while the nuclear partial correlation categorised path graph contained five times as many edges as the corresponding correlation categorised path graph. For the cytoplasmic and mitochondrial compartments both correlation and partial correlation categorised path graphs included approximately 25% of the edges in the graph used for shortest path calculations. The nuclear graphs

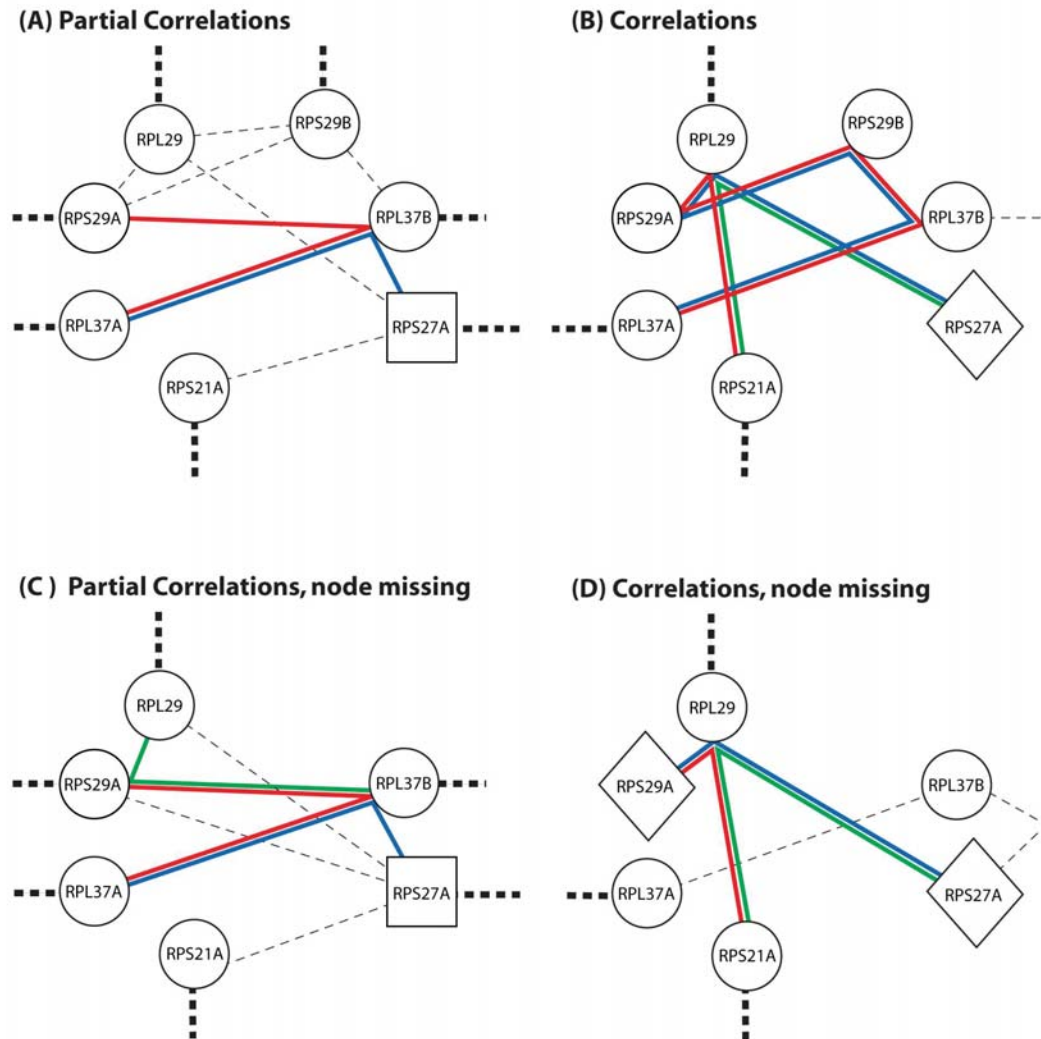


Figure 3.3: Subgraphs of cytoplasm categorised path graphs. Shortest paths between genes in the subgraph are indicated by coloured lines. Dashed lines = an edge that is part of a shortest path with other genes. Thick dashed lines leaving the graph represent more than 1 edge.  $\circ$  = genes which are transitive in at least one path and correctly categorised in all or almost all instances.  $\square$  = genes which are transitive and incorrectly categorised  $\diamond$  = genes which are not transitive in any path

behaved quite differently with 32% of edges included in the correlation categorised path graph and only 6% in the partial correlation categorised path graph.

The only change when a gene is deleted in a graph where an edge represents a correlation between two genes is the deletion of edges connected to that gene; the weights of other edges are unchanged. The same cannot be said for a graph where edges represent partial correlations. The weights of edges will change and potentially change the structure as well. We consider the effect of deleting a gene on the categorised path graphs by deleting gene RPS29B and constructing new categorised path graphs using both correlations and partial correlations. In both cases we use the same upper bounds on the path length (1.85 and 0.008) as in the previous example and show subgraphs of the same genes (Figure 3.3 (C) and (D)). The partial correlation categorised path graph now has 1 203 edges and the correlation categorised path graph 301. In the correlation categorised path graph RPS29B was only connected to genes shown on the subgraph and all changes when RPS29B is deleted are visible in the ensuing subgraph (Figure 3.3 (D)). The edges that included this gene are now missing from the categorised paths, meaning the paths include fewer edges and RPS29A is now not transitive on any path and thus unable to be categorised. One extra edge is included in the graph on a new categorised path from RPL37B to RPS27A. The situation with the partial correlation categorised path graph is more complex. RPS29B was connected to 11 genes in the original categorised path graph. In addition to the edges involving RPS29B the new categorised path graph has another 5 edges deleted and 8 new edges added. Most of these changes involve at least one gene that was previously connected to the deleted gene. Only one of these changes, an edge between RPS29A and RPS27A, is visible in the subgraph (Figure 3.3(C)). In the sub-graph all genes that were previously correctly categorised are still correctly categorised, and even in the full graph there are only three genes that are no longer transitive on any path. (One of these had been correctly categorised, one almost correctly and one incorrectly.)

## 3.4 Discussion

Both shortest path methods considered use tuning parameters. For the correlation graphs, the first tuning parameter is the minimum correlation at which an edge is placed in the graph. Our estimation procedure for the partial correlations naturally includes model selection (setting some partial covariances to zero, and eliminating the corresponding edges from the graph) governed by the penalty parameter  $\lambda$ . Zhou *et al.* (2002) used a powering factor on the edge lengths to increase the number of transitive genes. Using partial correlations, more transitive genes were identified without the need for a similar tuning parameter. Both methods use an upper bound on the length of paths used to make categorisations.

Zhou *et al.* (2002) give little guidance on their choice of tuning parameters. In contrast, the tuning parameter ( $\lambda$ ) for graphical lasso can be successfully chosen using 10-fold cross validation. Choice of the upper bound on path length is more delicate: as the bound increases, the number of transitive gene categorisations identified increases but the percentage of matches decreases. The optimal trade off between accuracy and number genes classified will depend on the application. However, for all three examples considered, the number of categorised paths (and thus transitive gene categorisations) identified rose sharply as the upper bound on path length was increased from 1.8 to 2, and using a upper bound just below 2 produced both a large number of categorisations and a reasonable degree of accuracy. Regardless of the stringency chosen, the partial correlation approach had similar accuracy, or better accuracy, than the correlation approach when the two methods were tuned to predict categories for the same number of genes.

The correlation graph is more sparse than our partial correlation graph when the graphs are tuned to categorise a similar number of genes. This sparsity carries over into the categorised path graphs. The path weights in the correlation graph are scaled so that shortest paths include a maximum number of transitive genes. No such tuning is used in the partial correlation graph. All paths in a partial correlation categorised path include only 1 transitive gene when the upper bound on shortest path length is less than 2. Most paths are longer and include at least two transitive

genes in the correlation categorised path graphs. The combined result of these differences is that in the partial correlation categorised path graphs most genes are transitive on many 3 gene (2 edge) paths, whereas in the correlation categorised path graphs most genes are transitive on 1 or 2 longer paths (and their sub-paths).

The categorised path graph is key to the categorisations made. It is only when the structure of the categorised path graph changes that categorisations alter. When a node is removed, the changes in both the correlation and partial correlation categorised path graphs are greater than just the deletion of edges incident at the node removed. Although these changes involve both the inclusion and deletion of edges in both cases the observed changes were relatively minor and local.

The partial correlations of cytoplasmic genes were also estimated using covariance shrinkage (Schäfer and Strimmer, 2005) and High-dimensional Bayesian Covariance Selection (HdBCS) (Dobra and West, 2004) for comparative purposes. We obtained fewer non-zero partial correlations using covariance shrinkage and thence fewer transitive genes than using graphical lasso. The percentage of gene categorisation matches was also lower for this approach. We obtained similar results using HdBCS to those using graphical lasso. However HdBCS is very computationally intensive. Obtaining partial correlations for the 398 cytoplasmic genes using HdBCS took a week using a 20 node parallel computer. By comparison using R with glasso obtaining the leave-one-out partial correlations for the same dataset took 150 minutes on a desktop computer.

GGMs assume that the data has a multivariate normal distribution. The datasets we use, as is typical for gene expression data, violate this assumption. Most genes in each dataset have long tailed noise. We have partially dealt with this by robustifying the data using leave-one-out partial correlations. Meinhausen and Bühlmann (2006) found, in simulations with long-tailed noise added, that the properties of their approximated lasso graph estimator did not appear to be critically affected by deviations from the multivariate normal assumption. We would expect the graphical lasso to exhibit similar properties.

A weakness of lasso methods is that a high number of extra weak edges tend to be

included (Schäfer and Strimmer, 2005). In fact, when the number of variables is thought of as growing with the sample size, realistic conditions can be found where the lasso is not consistent for edge selection (Meinhausen and Yu, 2006). We do not regard this as a serious problem here. Using shortest paths means that pairs of genes with a relatively low estimated partial correlation (corresponding to high edge length) are unlikely to lie on shortest paths. Furthermore, the threshold on shortest path lengths effectively rules out all pairs of genes with a very low estimated partial correlation. On the other hand in the situation where the partial correlation between genes is truly very small it is likely that graphical lasso, as with other estimation methods, may fail to distinguish the small partial correlation from zero. Robins *et al.* (2003) highlight the difficulty of obtaining uniformly consistent estimators for weak edges in the context of causal inference. Our results in Chapter 4 suggest that the presence of small partial correlations can increase the variability of estimates for other elements as well. We leave investigation of the impact of this phenomena on shortest path algorithms to future research.

The high percentages of correctly classified genes validate the notion that GGMs capture information relevant to the biological structure. The method presented also offers excellent prospects for predicting the functional category for genes with unknown function from a set of expression data where a mixture of well understood and unstudied genes are present.



# Chapter 4

## The cost of using decomposable Gaussian Graphical Models for computational convenience

A version of this material will appear in *The Journal of Computational Statistics and Data Analysis* as *The cost of using decomposable Gaussian graphical models for computational convenience* by A. Marie Fitch and Beatrix Jones. <http://dx.doi.org/10.1016/j.csda.2012.01.020>

### 4.1 Introduction

Estimation of a sparse inverse covariance matrix is a useful analysis in the multivariate setting, both as a means of obtaining a regularized estimate of the covariance matrix ( $\Sigma$ ), and also for the insights to be gained into the patterns of conditional independence. When the data is Gaussian this conditional independence structure can be represented by a Gaussian graphical model. The absence of an edge in a Gaussian graphical model corresponds to a zero element in the inverse covariance matrix ( $\Omega$ ) and thus represents conditional independence between two variables (Dempster, 1972). The sparse structure of the inverse covariance is typically obtained via a



model selection procedure. Our focus in this chapter is on estimation of the  $\Omega$  and  $\Sigma$  parameters after model selection has been performed, and assuming that model selection has been successful in the sense that all elements that are non-zero in the true model are also non-zero in the selected model. The increased computational burden of working with non-decomposable models leads us to restrict ourselves to decomposable models, as outlined briefly below.

A decomposable graph (defined in Section 4.2.1) has properties which make the estimation of the parameters of  $\Omega$  simpler. The maximum likelihood estimator only exists in closed form when the graph is decomposable (Wermuth, 1980; Lauritzen, 1996, Sec 5.3); it must be computed iteratively for non-decomposable graphs (Lauritzen, 1996; Dahl *et al.*, 2008). The maximum likelihood estimate is unique and can be calculated from the maximum likelihood estimates for individual prime components and separators when the number of variables ( $p$ ) is greater than the sample size ( $n$ ), if all prime components have  $p_i < n$ , where  $p_i$  is the number of variables in component  $i$ . A penalized likelihood approach such as the graphical lasso described in Friedman *et al.* (2008b), can be undertaken in other cases. Triangulating a graph (making a graph decomposable by adding edges to the graph, or equivalently removing the restriction that certain elements of  $\Omega$  must be zero) typically reduces the number of vertices in each prime component. This means that triangulation not only enables the existence of the maximum likelihood estimator in closed form, but also frequently clears the way for the application of ‘simple’ maximum likelihood.

While treatment of non-decomposable graphs in the likelihood framework is less straightforward than for decomposable graphs, the computational burden is not prohibitive. However, in the Bayesian setting, the marginal likelihood of a particular graph structure has closed form only for decomposable graphs. The additional computational burden of a high dimensional search in the space of non-decomposable graphs is sufficiently large that estimation has often been restricted to decomposable models (see for example Dawid and Lauritzen (1993); Giudici and Green (1999); Rajaratnam *et al.* (2008); Scott and Carvalho (2008)). When this restriction to decomposable models has not been made consideration has usually been restricted to

either very low dimensional cases (Roverato, 2002; Dellaportas *et al.*, 2003), or higher dimensional cases where the sample size was much larger (Wong *et al.*, 2003).

This chapter considers the cost, in terms of accuracy of the inferred  $\Omega$  and  $\Sigma$  parameters, of fitting a decomposable model to data when the true underlying graph is non-decomposable. The decomposable model we fit is a ‘true’ decomposable model in that the edges are a superset of the non-decomposable edges. For some edges we have removed the restriction that the corresponding parameter must equal zero. For these edges, however it is still possible, and indeed almost certain, that for sufficiently large samples the estimates for the associated parameters will converge to zero. Our interest is in what type of differences may be observed at smaller sample sizes, particularly in the situation where the  $n \approx p = p_1$  (i.e the true model has a single component). High dimensional models, where  $n \ll p$ , typically consist of many such components.

In Section 4.2 we give background material on graphical models and maximum likelihood estimation. We use the variance of maximum likelihood estimates and the inverse curvature of the log likelihood surface as measures of the quality of estimation. We expect that, because of the central role of the likelihood in Bayesian inference, the pattern of variation would be similar for maximum *a posteriori* Bayes’ estimates.

We begin with consideration of the simplest non-decomposable graph, a 4 variable cycle. Both theory and simulations show that the variances using the decomposable model are greater than those for the true non-decomposable model. Both the actual difference and the percent difference in variance are influenced by underlying parameter values. We then consider two cases where  $n = p + 1$  ( $p = 20$  and  $50$ ) and the true model is a cycle. We only use the results of computational experiments for these larger cycles. We also consider, in both of these cases, whether the nature of the decomposable model has any effect on the quality of the estimation. Here the situation is more complex with the pattern of differences depending on the particular decomposable model used. We complete this section by comparing each estimate to twice its standard deviation as a simple post-processing assessment of

the model. Here we see that when the decomposable model is used there are strong pointers to the edges added to make the model decomposable being superfluous. It is a concern that we find that there are also occasions (particularly with small sample sizes) where a true edge also looks superfluous only in the decomposable model.

In section 4.5.3 we consider estimation of the covariance matrix. Although some elements showed negligible, if any, difference between models, for most elements the variance increased dramatically (more than 200 fold in some cases) when a decomposable model was used. We conclude, in Section 4.6, by considering two case studies (Fisher's Iris data and a 12 node data set) before summarizing our findings in Section 4.7.

## 4.2 Background

### 4.2.1 General properties of graphs

Let  $\mathcal{G} = (V, E)$  represent a graph, where  $V$  is the set of vertices representing continuous variables and  $E$  the set of edges. We are restricting our interest to undirected graphs where all edges are undirected, thus  $(u, v) \in E$  always implies  $(v, u) \in E$ .

All vertices are joined in a complete graph, thus a complete undirected graph contains all possible edges. In any incomplete graph  $\mathcal{G} = (V, E)$ , given three disjoint subsets  $(A, B, C)$  of  $\mathcal{G}$  such that  $A \cup B \cup C = \mathcal{G}$ ,  $C$  is a separator of  $A$  and  $B$  if  $C$  is complete and for every  $\alpha \in A$  and  $\beta \in B$ , all paths from  $\alpha$  to  $\beta$  intersect  $C$ .  $A \cup C$  is a prime component if the separator  $C$  is chosen so that it does not contain a subgraph that separates  $A$  and  $B$ , and if also  $A \cup C$  cannot be further decomposed. We can find the prime components  $(P_i)$  of  $\mathcal{G}$  by iterative decomposition (Dobra and Fienberg, 2000). Thus prime components are a collection of subgraphs which cannot be further decomposed. The prime components of a decomposable graph are all complete.

### 4.2.2 Parameter and variance estimation

A Gaussian graphical model is one where the variables have a multivariate normal distribution, with covariance matrix  $\Sigma$ . Without loss of generality in what follows we assume the data to be centred (that is all  $p$  variables have mean 0). The maximum likelihood estimator of the inverse covariance matrix  $\Omega = \Sigma^{-1}$  is obtained by maximizing the log likelihood equation.

$$L(\Omega) = \log \det \Omega - \text{tr}(\Omega S) \quad (4.1)$$

where  $\Omega$  is a positive definite matrix and  $S$  is the sample covariance matrix.

The likelihood equations are then obtained from the matrix of partial derivatives

$$\frac{\partial L(\Omega)}{\partial \Omega} = 2(\Omega^{-1} - S) - (\Omega^{-1} - S) \circ I \quad (4.2)$$

We use the R-package `glasso` (with `rho = 0`, and zero elements specified) (R Development Core Team, 2009; Friedman *et al.*, 2008a) to obtain the maximum likelihood estimates.

If  $\hat{\Omega}$  is the maximum likelihood estimate of  $\Omega$ , then we define the three estimators of  $\text{Cov}(\hat{\Omega})$  as follows:

- The expected Fisher information (EFI) covariance of  $\hat{\Omega}$  is the inverse of the expected Fisher information.

$$\text{EFI Cov}(\hat{\Omega}) = \left( E \left[ -\frac{\partial^2 L(\Omega)}{\partial \Omega^2} \right] \right)^{-1}$$

This quantity depends on knowledge of (or algebraically representing) the true  $\Omega$  matrix.

- The observed Fisher information (OFI) covariance of  $\hat{\Omega}$  is the inverse of the observed Fisher information.

$$\text{OFI Cov}(\hat{\Omega}) = \left( - \left[ \frac{\partial^2 L(\Omega)}{\partial \Omega^2} \right]_{\Omega=\hat{\Omega}(x)} \right)^{-1}$$

- The empirical covariance (empirical  $\text{Cov}(\widehat{\Omega})$ ), only available in a simulated setting, is found by computing  $\widehat{\Omega}$  for a large number of different samples simulated with true inverse covariance  $\Omega$ , and calculating the sample covariance matrix for this sample of estimates.

If  $L(\Omega)$  is defined as in equation (4.1) then

$$\frac{\partial^2 L(\Omega)}{\partial \Omega^2} = -\frac{1}{2} Q' [\Omega^{-1} \otimes \Omega^{-1}] Q \quad (4.3)$$

where  $Q$  is a matrix with entries  $\{0, 1\}$  satisfying  $\text{vec}(\Omega) = Q(\omega)$  and  $\omega$  is the vector of elements  $\omega_{ij}$  of  $\Omega$  such that  $i \geq j$  and  $\omega_{ij} \neq 0$  (Drton and Eichler, 2006)

The empirical estimate of the variance, based upon a large number of repeated samples, should give a good idea of the sampling distribution of  $\widehat{\Omega}$  regardless of the sample size ( $n$ ) used for each individual estimate; the OFI and EFI estimates of the variance are expected to be similar when the  $n$  contributing to each individual estimate is large. Efron and Hinkley (1978) give a frequentist justification for preferring the inverse of the observed information over the inverse of the expected total information for one-parameter estimation problems. We use the EFI variance and the OFI variance in different ways. The EFI variance allows us to observe the effect of changes in the true  $\Omega$  without introducing variability from simulation. OFI is the most readily available variance estimate when analyzing data, where the true  $\Omega$  is unknown, so we examine its performance especially when  $n$  is not large.

### 4.3 Theory for the four variable case

The simplest non-decomposable graph is a 4-cycle shown in Figure 4.1(a). Suppose we have four variables  $X_1, X_2, X_3, X_4$  with inverse covariance matrix  $\Omega_{nd}$  given by equation (4.4). If, for the purposes of estimation, we remove the restriction that  $\omega_{2,4} = 0$ , then we will have the decomposable graph shown in Figure 4.1 (b) with the related inverse covariance matrix elements to be estimated symbolically represented by  $\Omega_d$ . We use  $t$  to denote element  $\omega_{2,4}$ , the element associated with the additional

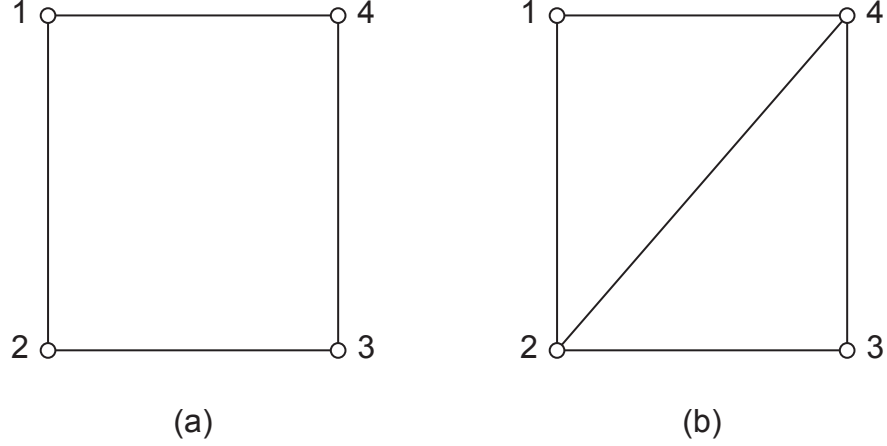


Figure 4.1: A 4 variable non-decomposable graph(a) and a 4 variable decomposable graph (b).

edge. Thus:

$$\Omega_{nd} = \begin{pmatrix} a & m & 0 & s \\ m & b & q & 0 \\ 0 & q & c & r \\ s & 0 & r & d \end{pmatrix} \quad \Omega_d = \begin{pmatrix} a & m & 0 & s \\ m & b & q & t \\ 0 & q & c & r \\ s & t & r & d \end{pmatrix} \quad (4.4)$$

Using equation (4.3), with assistance from the symbolic algebra functions of MATLAB(The MathWorks), we now compute the difference between the EFI variance of the the non-zero off diagonal elements common to both matrices. (In computing this difference, we do not use the fact that the true value of  $t$  is zero.) Formulae are given for a sample of size  $n$ .

$$\text{EFI Var}(m \text{ in } \Omega_d) - \text{EFI Var}(m \text{ in } \Omega_{nd}) = \frac{a(q^2s - bcs + mqr)^2}{c\delta} \quad (4.5)$$

$$\text{EFI Var}(s \text{ in } \Omega_d) - \text{EFI Var}(s \text{ in } \Omega_{nd}) = \frac{a(r^2m - cdm + qrs)^2}{c\delta} \quad (4.6)$$

$$\text{EFI Var}(q \text{ in } \Omega_d) - \text{EFI Var}(q \text{ in } \Omega_{nd}) = \frac{c(m^2r - abr + mqs)^2}{a\delta} \quad (4.7)$$

$$\text{EFI Var}(r \text{ in } \Omega_d) - \text{EFI Var}(r \text{ in } \Omega_{nd}) = \frac{c(s^2q - adq + mrs)^2}{a\delta} \quad (4.8)$$

where  $\delta = n(abcd - 2mqr s - m^2r^2 - q^2s^2)$

$\Omega$  is an inverse covariance matrix, therefore we know by definition that it is positive definite and that  $a$  and  $c$  are positive. This property ensures that the expected Fisher information calculated using equation (4.3) will be positive definite. Since the determinant of a positive definite matrix is positive, the determinants of both  $\Omega_{nd}$  and the expected Fisher information of  $\Omega_{nd}$  are positive.

$$\delta = \det(\text{expected Fisher information of } \Omega_{nd}) \times 16(\det(\Omega_{nd}))^5 \quad (4.9)$$

Since the right hand side of equation (4.9) must be positive,  $\delta$  is positive. Thus equations (4.5) to (4.8) are positive and the variance of each term in the decomposable model is larger than that for the non-decomposable model of the same sample size. We also note that, because the variance of each of these elements in  $\Omega_d$  is independent of  $t$ , each of the difference terms is independent of  $t$ , and so remains positive no matter what value  $t$  takes. It is thus the fact that we are estimating a value for  $t$  that induces the added variance not the actual value of  $t$ . Unsurprisingly, having to estimate additional parameters increases the variance of the estimates. The only difference in the calculation of the OFI rather than EFI variance is that the maximum likelihood estimate of  $\Omega$  is used rather than the true  $\Omega$  matrix.

## 4.4 Simulation study methods

### 4.4.1 The four variable case

Data were simulated in order to compare the OFI, EFI, and empirical variances. Data were simulated from three inverse covariance matrices ( $\Omega$ ) with different characteristics (see Table 4.1). These matrices were:

1. A matrix with all partial correlations the same ( $\Omega_{same}$ );
2. A matrix with large average absolute partial correlations ( $\Omega_{big}$ );
3. A matrix with small average absolute partial correlations ( $\Omega_{small}$ ).

We ensured that valid (positive definite)  $\Omega$  matrices were obtained by using the algorithm of Atay-Kayis and Massam (2005) to generate  $\Omega_{big}$  and  $\Omega_{small}$ .

$$\begin{array}{ccc}
 \Omega_{same} & \Omega_{big} & \Omega_{small} \\
 \begin{pmatrix} 20 & -9 & 0 & -9 \\ -9 & 20 & -9 & 0 \\ 0 & -9 & 20 & -9 \\ -9 & 0 & -9 & 20 \end{pmatrix} & \begin{pmatrix} 67 & -18 & 0 & -70 \\ -18 & 19 & 28 & 0 \\ 0 & 28 & 74 & -64 \\ -70 & 0 & -64 & 150 \end{pmatrix} & \begin{pmatrix} 88 & -20 & 0 & 17 \\ -20 & 111 & -2 & 0 \\ 0 & -2 & 59 & 6 \\ 17 & 0 & 6 & 213 \end{pmatrix}
 \end{array}$$

Table 4.1:  $\Omega$  matrices used for simulating data.

For each of the  $\Omega$ , data were simulated from a  $MVN(0, \Omega^{-1})$  distribution using the Cholesky decomposition of  $\Omega^{-1}$  and the R function `rnorm`. Samples of sizes  $n = 10$ , 100 and 1000 were created. The simulation process is as follows:

1. Simulate sample of size  $n$
2. Find the maximum likelihood estimator (MLE) for  $\Omega$  and thence  $\Sigma = \Omega^{-1}$  using the R-package `glasso` with a penalty zero and the model restricted to the structure of  $\Omega_{nd}$ .
3. Find the MLE for  $\Omega$  and thence  $\Sigma = \Omega^{-1}$  using the R-package `glasso` with a penalty zero and the model restricted to the structure of  $\Omega_d$ .
4. Calculate the OFI covariance matrix for both sets of parameter estimates
5. Repeat steps 1 to 4, 1000 times
6. Calculate the variance of the 1000 estimates (the empirical variance) for each non-zero off-diagonal parameter of each  $\Omega$  and for all elements of each  $\Sigma$ .

The EFI variances were also calculated for each  $\Omega$  matrix.



### 4.4.2 20 and 50 variable cases

We considered the effect of changing the triangulation of the graph that was used to make a decomposable model, as well as considering what (if any) effect the nature of the underlying  $\Omega$  matrix has for the  $p = 20$  and  $p = 50$  variable cases. Hereafter we define ‘extra edges’ to be the edges corresponding to elements which have the restriction of the parameter being zero removed, in order to make a decomposable model. Two decomposable model types were considered. The first has all extra edges radiating out from the same vertex (hereafter referred to as model type A), the second was set up to minimize the maximum number of extra edges radiating out from any one vertex (hereafter referred to as model type B). We illustrate these model types for a simpler case ( $p = 6$ ) in Figure 4.2. The edges from the non-decomposable model are labeled in an anti-clockwise direction beginning from the edge corresponding to  $\omega_{1,2}$ .

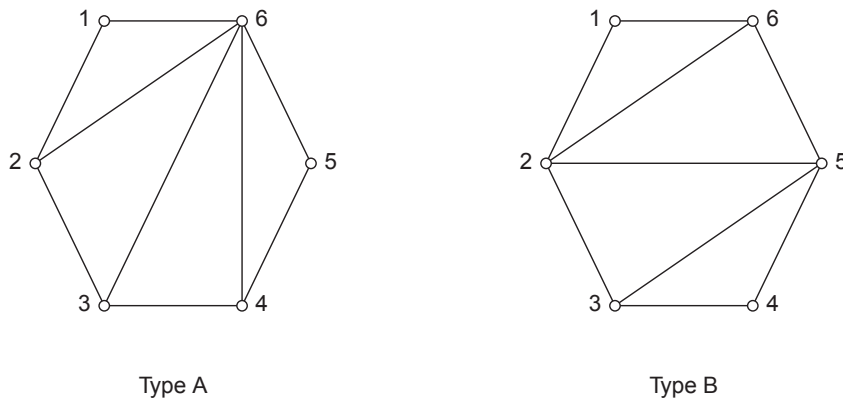


Figure 4.2: Two different decomposable models when  $p = 6$ .

The same process as for the four variable case was used to obtain three  $\Omega$  matrices for a 20-cycle and for a 50-cycle. (Tables A.6 and A.7 list the non-zero elements of  $\Omega_{same}$ ,  $\Omega_{big}$  and  $\Omega_{small}$  for the 20-cycle and 50-cycle.) In each case samples of size  $n = p + 1$  were created using the steps outlined in Section 4.4.1 with  $\Omega_d$  replaced by matrices corresponding to each of model type A and model type B.

Table 4.2:  $n = 1000$  empirical variances for cycle and percentage increase to decomposable.

	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	variance	%increase	variance	%increase	variance	%increase
$\hat{\omega}_{1,2}$	0.387	11.3%	0.489	77.3%	10.39	2.3%
$\hat{\omega}_{2,3}$	0.372	8.4%	1.031	78.0%	6.034	4.2%
$\hat{\omega}_{3,4}$	0.363	8.5%	5.635	73.1%	12.70	1.0%
$\hat{\omega}_{1,4}$	0.360	11.9%	7.141	57.3%	19.38	0.4%

## 4.5 Simulation study results

### 4.5.1 Estimating $\Omega$ - the four variable case

Figure 4.3 illustrates how variability in the three measures of variance depends upon the sample size. Variance measures for the parameter  $\omega_{1,2}$  from  $\Omega_{same}$  are shown; a similar pattern was observed for other parameters and  $\Omega$  matrices (see Figures A.1, A.2 and A.3). When fitting the non-decomposable model, the EFI variance, mean OFI variance and empirical variance of the parameter estimates were all very similar for a sample size of 1000. The empirical variance for the decomposable model is larger than that for the non-decomposable model for sample sizes of 10, 100 and 1000, as is the EFI variance. Table 4.2 shows that, when the decomposable model was used, both the variances and the percentage increase in variance varied depending on  $\Omega$ . In general the estimates for a matrix with smaller partial correlations have bigger variances, but the percentage increase in the variances when a non-decomposable model is fitted is small. The converse is true when the partial correlations are larger.

We also looked for patterns in the variance controlling for the size of the parameters by considering the relative standard deviation (RSD), as defined in equation (4.10).

$$RSD = \frac{\sqrt{Var(\hat{\omega}_{i,j})}}{|\omega_{i,j}|} \quad (4.10)$$

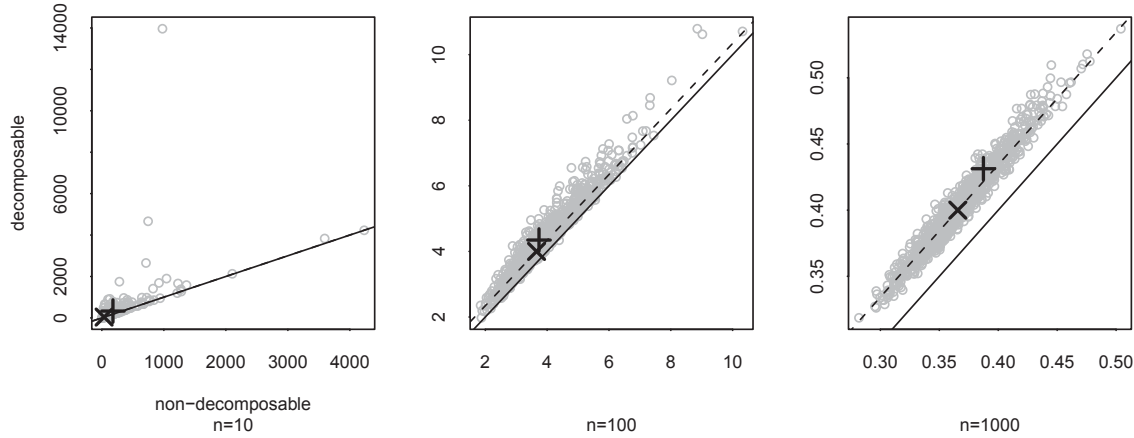


Figure 4.3: OFI variances for  $\hat{\omega}_{1,2}$  in  $\Omega_{same}$  with sample sizes 10, 100 and 1000.

(Note the radically different scales)

— represents the line  $y = x$ ;

--- represents the line  $y = x + (\text{difference in EFI variances calculated using equation(4.5)})$

× represents the expected variances;

+ represents the empirical variances.

where  $Var(\hat{\omega}_{i,j})$  is the empirical variance of  $\hat{\omega}_{i,j}$ . Here, as summarized in Table 4.3, we see that if we scale the variance relative to the parameter being estimated, then the pattern is similar with larger RSDs occurring for  $\Omega_{small}$ . However, the biggest increase when the decomposable model was used occurs with  $\Omega_{big}$ . Further investigation (see Table A.1) suggests that these differences are only relative to the value of the partial correlations in the matrix as a whole and this pattern of elements corresponding to larger partial correlations having big increases in variance is not relative to the size of individual elements within a given matrix. We also observed that if the partial correlations become very small the RSDs are very large. In this case, as seen in Table A.1, the difference in RSD between models can also become large. We note here that while the larger the RSD the more likely it is than an estimate may have the opposite sign to the true value, the increases observed here are generally not large enough to dramatically change the probability of this occurring.

The EFI variance increasingly underestimated the empirical variance as the sample size decreased, so that when the sample size reached 10, the empirical variance was of

the order of four times larger than the EFI variance for the non-decomposable model, and of the order of seven times larger than the EFI variance for the decomposable model. (See Tables A.2, A.3 and A.4.) We see this behaviour also in Figure 4.3 (and in Figures A.1, A.2 and A.3) with the empirical variances marker sitting above the line representing the difference in the EFI variances.

Figure 4.3 also illustrates the variability in the OFI variance of  $\omega_{1,2}$ . The variability increases with decreasing sample size as expected. Although mean OFI variance for the decomposable model is always larger than mean OFI variance for the non-decomposable model, as Figure 4.3 shows, for some of the small samples the reverse was true. (These are points below the  $y = x$  line.) This variation at small sample sizes was observed to some degree for all estimates. As the sample size increases the OFI variances converge towards the empirical variances. A comparison of the graphs suggests that, at each sample size, the decomposable model shows more variability than the non-decomposable model as predicted. The difference in range varies from being almost negligible (as in Figure 4.3) to around 60% more (for example see  $\omega_{1,4}$  in Figure A.2) for the large samples. The range of OFI variances for the decomposable model can be up to twice the range of OFI variances for the non-decomposable model (see Figures A.1, A.2 and A.3) for a sample of size 10. This suggests that, if we fit a decomposable model when the sample size is small, then, not only is the variance of each estimate greater but the (OFI) variance estimator itself is more variable.

Table 4.3:  $n = 1000$  Relative Standard Deviations for non-decomposable model and increase for decomposable model.

	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	RSD cycle	increase	RSD cycle	increase	RSD cycle	increase
$\omega_{1,2}$	0.069	0.004	0.039	0.013	0.159	0.002
$\omega_{2,3}$	0.068	0.003	0.036	0.012	1.222	0.006
$\omega_{3,4}$	0.067	0.003	0.037	0.012	0.593	0.001
$\omega_{1,4}$	0.067	0.004	0.038	0.010	0.254	0.006

Although our focus here is on parameter estimation after model selection we also briefly consider whether the observed changes in variance have any effect on our decision about which elements of  $\Omega$  are non-zero. We use a simple decision rule that declares an element to be non-zero if the absolute value of the estimate is greater than twice its estimated standard error ( $= \sqrt{\text{OFI variance}}$ ). True non-zero elements of  $\Omega_{same}$  and  $\Omega_{big}$  were always declared non-zero for samples of size 100 and 1000. The true values were sometimes so small that estimates would not be declared non-zero (see, for example, Table A.5) for  $\Omega_{small}$ . There was variation in which true non-zero edges were declared non-zero depending on whether the true non-decomposable model or the decomposable model was used (see Table 4.4) for small samples, and to some extent all samples using  $\Omega_{small}$ . We have already noted that the OFI variances are highly variable and so were not surprised to find that even when the true model is used an estimate for a true non-zero may not be declared non-zero. This is particularly so for small sample sizes (see Table 4.4). Furthermore there is always some variation between simulation runs so we focus our attention here on the cases where the difference between models is greater than 5%. When  $n=10$  and the decomposable model is used, for  $\Omega_{big}$  there is an increase of at least 18 in the percentage of times a true edge would be declared zero (see Table 4.4). This suggests that the increased variance already noted may affect decisions about the inclusion of edges if a decomposable model is used. The estimate of the true zero element ( $\hat{\omega}_{2,4}$ ) would be declared non-zero less than 5% of the time in each case.

### 4.5.2 Estimating $\Omega$ - 20 and 50 variable cases

We were reliant on calculating EFI variances for specific  $\Omega$  matrices as the number of variables is too large to deal with algebraically at this level. However, by comparing the EFI variances for  $\Omega$  matrices that differed only at elements corresponding to ‘extra’ edges, we determined that these elements appear explicitly in the EFI variances for some  $\Omega$  elements corresponding to true edges; this differs from the situation with  $p = 4$ . Only the EFI variances of elements corresponding to edges adjacent to the

Table 4.4: Percentage of times an element is declared zero ( $|\text{estimate}| < 2 \times \text{standard error}$ ) for 1000 simulations ( $n = 10$ ).

	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	cycle	decomp	cycle	decomp	cycle	decomp
$\hat{\omega}_{1,2}$	90.4	93.1	15.1	61.4	99.1	98.9
$\hat{\omega}_{1,4}$	88.7	90.5	1.4	25.9	99.1	99.1
$\hat{\omega}_{2,3}$	88.3	90.4	0.7	30.8	99.9	99.7
$\hat{\omega}_{2,4}$		97.8		99.3		98.3
$\hat{\omega}_{3,4}$	90.6	91.6	5.5	23.9	99.2	99.1

vertex with all extra edges radiating from it (that is  $\omega_{1,p}$  and  $\omega_{(p-1),p}$ ) were affected for type A models. The EFI variances of all elements corresponding to edges were affected for type B models.

Of more interest is the observation that even when all true partial correlations are the same the EFI variances for the various  $\Omega$  elements differ. As Figure 4.4 illustrates, the greatest percentage increase in size occurs for elements corresponding to edges which in the decomposable model are part of a triangle which contains only one added edge for both types of decomposable models. This increase declines the further the corresponding edge is from any such triangle. Figure 4.4 also shows that this pattern is visible in the empirical variances. Similar results were obtained for all three underlying  $\Omega$  matrices. (Full results for  $p=20$  and  $p=50$  can be found in Figures A.4 and A.5.)

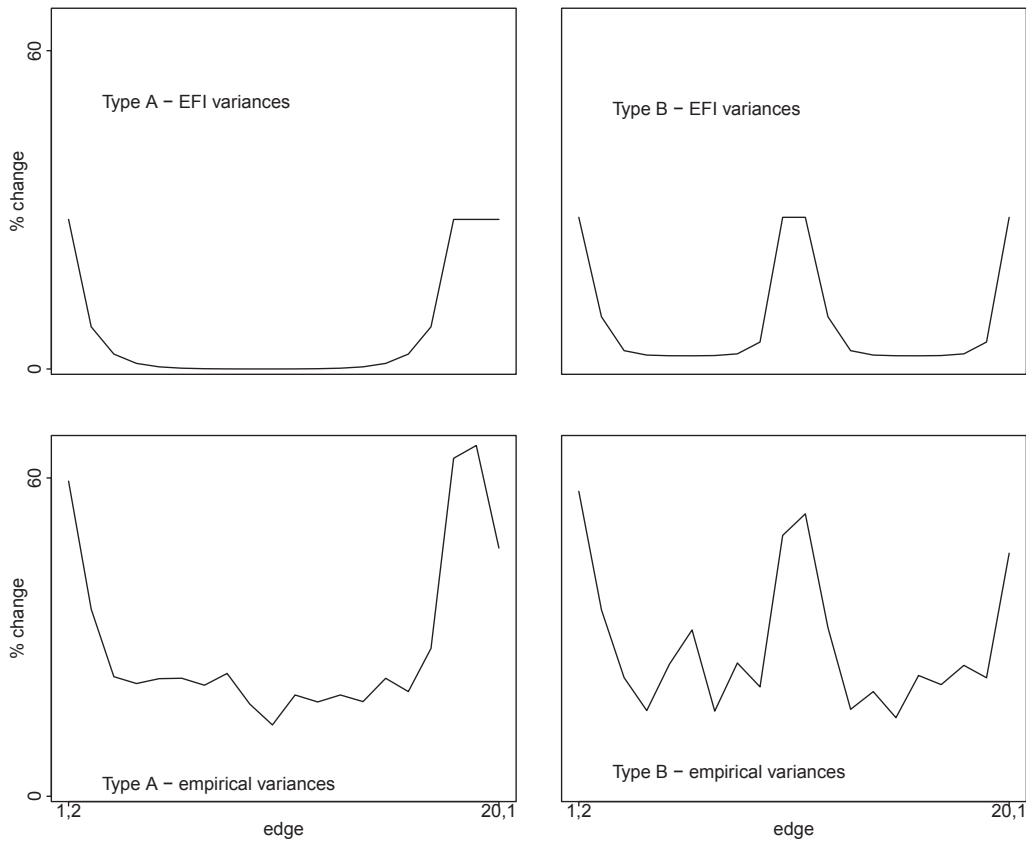


Figure 4.4: Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted.

Shown here for  $p = 20$ , when underlying  $\Omega$  matrix ( $\Omega_{same}$ ) has all partial correlations equal. Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

Figure 4.4 (and Figure A.4 ) show that for the 20 variable cases the percentage increase in the EFI variance when the decomposable model was fitted varied from almost zero to a maximum of 28% for  $\Omega_{same}$ , 150% for  $\Omega_{big}$ , and 34% for  $\Omega_{small}$ . Thus for many parameters there is a negligible change in the EFI variance of the estimates if a decomposable model is fitted. However where an increase in EFI variance is observed it has the potential to be substantial. These patterns are mirrored in the empirical variances as seen in Figure 4.4 (and in Figures A.4 and A.5 ).

We again look for patterns in the variance controlling for the size of the parameters by considering the RSDs (see equation (4.10)). Inspection of the RSDs revealed that, even when the (true) non-decomposable model was fitted, there could be a large variation in the RSDs. For example, for  $\Omega_{small}$  and  $n = 21$  the RSDs for the non-decomposable model varied from 0.63 to 17.2. In general, the larger the RSD in the non-decomposable model, the larger the RSD for the same edge when a decomposable model was fitted. However this could be accentuated or ameliorated depending upon the relationship of the edge to the extra edges, in line with the pattern observed above (see Figures A.4 and A.5 ).

The EFI variance continues to underestimate the empirical variance, although not as severely. When  $p = 50$  ( $n = 51$ ) this difference is 10%–30% for the true (non-decomposable) model and 15%–35% when a decomposable model (of either type) is fitted. Thus, although the percentage increase in many EFI variances between the true and decomposable model were almost zero, the percentage increase in empirical variances remained at least 4%, even for the  $p = 50$  ( $n = 51$ ). (see Figure A.5 ) The OFI variances remain highly variable, but this variability appears to decrease as  $n$  increases (even if it is accompanied by increasing  $p$ ). (See Figures A.6 and A.7)

Again we briefly consider whether the observed increases in variation would bring the model selected into question. As in the four variable case the elements corresponding to added edges were almost always declared different to zero less than 5% of the time (see Figures A.9 and A.11). The percentage of elements corresponding to true edges that would be declared non-zero appears to be heavily influenced by both the sample size and the size of the true partial correlations. As for the four



-variable case when the partial correlations are small ( $\Omega_{small}$ ) most, if not all, true non-zero elements would be declared zero no matter which model is fitted and there is little difference between models in the percentage declared zero (see Figures A.8 and A.10). When the partial correlations are larger ( $\Omega_{same}$  and  $\Omega_{big}$ ) there are occasions when the percentage of times a true non-zero element would be declared zero increases markedly when a decomposable model is used. These peaks correspond with the elements found to have a large percentage increase in variance. This occurs less often when  $n = 51$  than when  $n = 21$  (see Figures A.8 and A.10).

### 4.5.3 Estimating the covariance matrix ( $\Sigma$ )

In the four variable case consideration of empirical variances of the covariance matrix ( $\Sigma$ ) revealed that for elements where the corresponding element in the true  $\Omega$  matrix is zero there is always an observable difference in the variance. The variance of estimates for  $\sigma_{2,4}$  from  $\Omega_{small}$  showed extreme variation, with the variance when a decomposable model was used being up to 15 times larger than when a non-decomposable model was used. The differences in variance were tiny for elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$ . The differences only began to be seen at 4 significant figures even for  $n = 10$ . (The results are available as Table A.8.)

An increasingly extreme situation was observed for the 20 and 50 variable cases. Empirical variances for estimates of elements in the covariance matrix corresponding to an edge in the true graph showed very little (if any) difference between those obtained using the (true) non-decomposable model and those using a decomposable model. The variances were larger when the decomposable model was fitted for all other elements. As Figure 4.5(a) illustrates for  $\Omega_{same}$ , this difference is most marked for those elements in  $\hat{\Sigma}$  where the corresponding element in  $\Omega$  had the restriction to zero removed in the decomposable model. For instance, for  $\Omega_{same}$  (which had the smallest average percentage increase in variance) the variance was around 200 times larger for the element  $\sigma_{\frac{p}{2},p}$  when a type A decomposable model was used with  $p = 20$ ,  $n = 21$ . The increase was even bigger for  $\Omega_{same}$  when  $p = 50$  and  $n = 51$ . The situation is more complex, as Figure 4.5(b) and (c) show, when the elements

of  $\Omega$  vary. As with the four variable case, for both  $p = 20$  and  $p = 50$ , the largest overall percentage increase in variance was noted for the estimate of a  $\sigma$  element from  $\Omega_{small}$ . (Results when a type B decomposable model is fitted and for  $p = 50$ ,  $n = 51$  are displayed in Figures A.12, A.13 and A.14.)

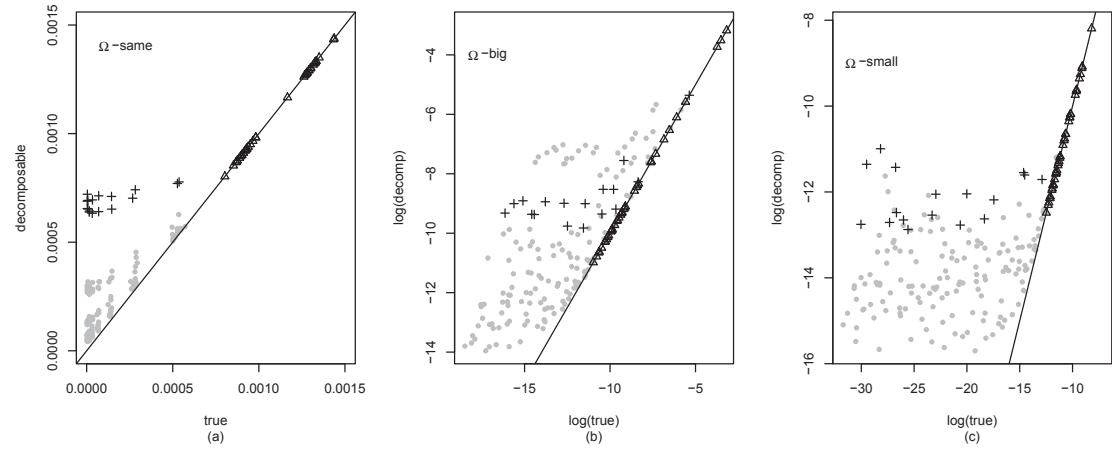


Figure 4.5: Empirical variances for elements of  $\Sigma$  when a decomposable model (Type A) vs when the true (non-decomposable) model is fitted.

Shown here for  $p=20$ , when underlying  $\Omega$  matrix has (a) all partial correlations equal, (b) large partial correlations and (c) small partial correlations. Note that log scales are used for (b) and (c).

$\triangle$  = elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$

$+$  = elements of  $\Sigma$  corresponding to elements only non-zero in the decomposable estimate of  $\Omega$

## 4.6 Case studies: Fisher’s iris data and 12 node data

### 4.6.1 Fisher’s iris data

We now consider Fisher’s *Iris virginica* dataset. It consists of  $n = 50$  measurements of the the sepal length, sepal width, petal width and petal length of 50 flowers. Previous analyses by Roverato (2002) and Atay-Kayis and Massam (2005) suggest that a non-decomposable 4-cycle is the most probable model for the data. We estimate parameters using the non-decomposable model and two different decomposable models: first, with the restriction that  $\omega_{1,3} = 0$  removed; second, with the restriction that  $\omega_{2,4} = 0$  removed. Estimates for both  $\Omega$  and  $\Sigma$  were obtained. We expected OFI variances for estimates of the parameters of  $\Omega$  obtained using the decomposable models to be larger than those obtained when the non-decomposable model was used and  $\hat{\sigma}_{1,3}$  and  $\hat{\sigma}_{2,4}$  to be the only elements in the inferred  $\Sigma$  matrices to vary between the three models.

Estimates for most of the edges in the decomposable graphs were larger than the estimates using the non-decomposable model, as shown in Table 4.5. The estimates for  $\omega_{1,2}$  and  $\omega_{3,4}$  showed the greatest percentage change when the decomposable models were used both in terms of the actual estimates and the OFI variance of the estimates. Most of the OFI variances were larger when the decomposable model was used (see Table 4.6). However in only one case ( $\hat{\omega}_{1,2}$  in the include  $\omega_{2,4}$  model) was this increase sufficient to cause a change in the decision to declare an element non-zero. Element  $\omega_{2,3}$  in the model with the restriction that  $\omega_{2,4} = 0$  removed was an exception in that both the estimate and the OFI variance decreased. Given the results in Section 4.5.1 we believe that this is due to the variability of the OFI variances, rather than a reason to prefer the decomposable model estimate of  $\hat{\omega}_{2,3}$ . As expected the only elements of  $\Sigma$  to show any variation between models, at 4dp accuracy, were  $\sigma_{1,3}$  and  $\sigma_{2,4}$ . In both cases this was most marked when the edge was estimated as non-zero in  $\Omega$ .

Table 4.5: Estimated  $\Omega$  and  $\Sigma$  matrices for *Iris virginica* dataset.

	$\hat{\Omega}$	$\hat{\Sigma}$
cycle	$\begin{pmatrix} 10.27 & -2.47 & 0 & -9.63 \\ -2.47 & 15.45 & -7.88 & 0 \\ 0 & -7.88 & 18.97 & -1.14 \\ -9.63 & 0 & -1.14 & 13.06 \end{pmatrix}$	$\begin{pmatrix} 0.4043 & 0.0938 & 0.0572 & 0.3033 \\ 0.0938 & 0.1040 & 0.0476 & 0.0733 \\ 0.0572 & 0.0476 & 0.0754 & 0.0488 \\ 0.3033 & 0.0733 & 0.0488 & 0.3046 \end{pmatrix}$
include $\omega_{1,3}$	$\begin{pmatrix} 10.22 & -2.66 & 1.32 & -9.71 \\ -2.66 & 15.78 & -8.24 & 0 \\ 1.32 & -8.24 & 19.09 & -2.30 \\ -9.71 & 0 & -2.30 & 13.33 \end{pmatrix}$	$\begin{pmatrix} 0.4043 & 0.0938 & 0.0491 & 0.3033 \\ 0.0938 & 0.1040 & 0.0476 & 0.0766 \\ 0.0491 & 0.0476 & 0.0754 & 0.0488 \\ 0.3033 & 0.0766 & 0.0488 & 0.3046 \end{pmatrix}$
include $\omega_{2,4}$	$\begin{pmatrix} 10.37 & -2.70 & 0 & -9.69 \\ -2.70 & 15.43 & -7.88 & 0.33 \\ 0 & -7.88 & 19.01 & -1.20 \\ -9.63 & 0.33 & -1.20 & 13.05 \end{pmatrix}$	$\begin{pmatrix} 0.4043 & 0.0938 & 0.0580 & 0.3033 \\ 0.0938 & 0.1040 & 0.0476 & 0.0714 \\ 0.0580 & 0.0476 & 0.0754 & 0.0488 \\ 0.3033 & 0.0714 & 0.0488 & 0.3046 \end{pmatrix}$

Table 4.6: Estimates and standard error ( $\sqrt{\text{OFI variance}}$ ) for *Iris virginica* dataset.

	cycle		include $\hat{\omega}_{1,3}$		include $\hat{\omega}_{2,4}$	
	estimate	std error	estimate	std error	estimate	std error
$\hat{\omega}_{1,2}$	-2.47	1.018	-2.66	1.063	-2.70	1.633
$\hat{\omega}_{2,3}$	-7.88	2.623	-8.24	2.694	-7.88	2.616
$\hat{\omega}_{3,4}$	-1.14	1.206	-2.30	2.012	-1.20	1.243
$\hat{\omega}_{1,4}$	-9.63	2.106	-9.72	2.120	-9.69	2.138
$\hat{\omega}_{1,3}$			1.32	1.797		
$\hat{\omega}_{2,4}$					0.33	1.840

## 4.6.2 12 node case

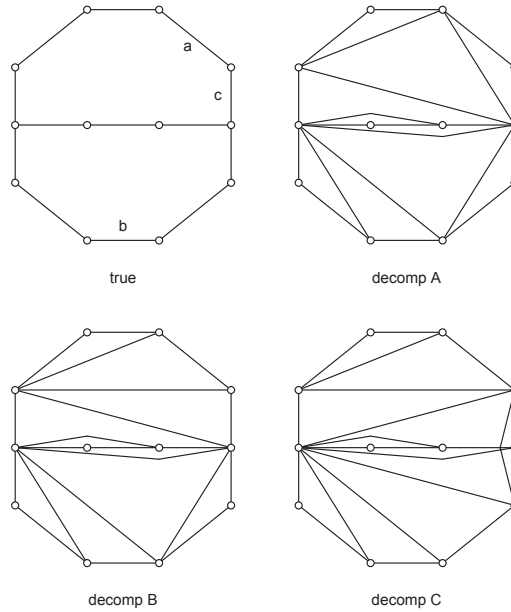


Figure 4.6: The 12 variable models.

We also consider a larger sample ( $n = 250$ ) of data simulated for the 12 node case illustrated in Figure 4.6, with all diagonal elements of  $\Omega$  being 20 and all non-zero off-diagonal elements being -8. We fitted both the true underlying non-decomposable graph and the top three graphs found using the feature inclusion stochastic search (Scott and Carvalho, 2008) which restricts its search to decomposable graphs. We note here that the third graph, model decomp C, is not a minimal superset. That is we can triangulate the graph by adding fewer edges than this triangulation requires.

We focus our interest on the estimates for  $\omega_{1,2}$ ,  $\omega_{10,11}$  and  $\omega_{1,8}$ . These elements correspond to the edges labeled  $a$ ,  $b$  and  $c$  respectively in Figure 4.6. We expected to see higher OFI variances when the decomposable models were fitted. In particular we expected that the increase in the OFI variance for  $\hat{\omega}_{1,2}$  would be smaller for a decomposable model where  $a$  is part of a triangle with two extra edges (models decomp B and decomp C). Edge  $b$  is always part of a triangle with two extra edges so we expected the OFI variances for  $\hat{\omega}_{10,11}$  to all be very similar, with some minor differences caused by the different configurations. Edge  $c$  is part of two triangles, (a

Table 4.7: Estimates and standard error ( $\sqrt{\text{OFI variance}}$ ) for  $\hat{\omega}_{1,2}$ ,  $\hat{\omega}_{10,11}$  and  $\hat{\omega}_{1,8}$ .

	$\hat{\omega}_{1,2}$		$\hat{\omega}_{10,11}$		$\hat{\omega}_{1,8}$	
	estimate	std error	estimate	std error	estimate	std error
true	-9.736	1.213	-7.012	1.003	-7.176	0.954
decomp A	-10.327	1.359	-6.912	1.047	-7.730	1.110
decomp B	-9.787	1.214	-6.912	1.047	-7.235	0.956
decomp C	-9.787	1.214	-6.912	1.047	-8.632	1.180

one extra edge triangle and a two extra edge triangle) in model decomp C. We were interested to observe the effect of this on the OFI variance for  $\hat{\omega}_{1,8}$ .

Results for the three elements of focus are given in Table 4.7. (Results for all elements are given in Table A.4.) As expected when edge  $a$  was part of a triangle with two extra edges (model decomp A) the OFI variance for  $\hat{\omega}_{1,2}$  was greater. Both the estimated value of  $\hat{\omega}_{1,2}$  and its variance were close to the values obtained using the true non-decomposable model when it was part of a triangle with one extra edge. We expected the OFI variances for  $\hat{\omega}_{10,11}$  to be similar in all decomposable models, in fact the estimates and OFI variances were the same (to 4 and 6dp respectively) in every case. Here it appeared that although there was some minor variation in the configuration of nearby triangles there was almost no effect on both the estimates and their OFI variances. Unsurprisingly  $\hat{\omega}_{1,8}$  showed the most variation in both estimates and OFI variances. When edge  $c$  was part of a two extra edge triangle (model decomp B) the increase in OFI variance (compared to the true decomposable model) was minimal. The OFI variance was greatest when edge  $c$  was part of two triangles, suggesting that here the increases were compounded. However even for this case the sample was large enough to ensure that the variance was still small compared to the estimate itself.

In the simulation studies we found that the variance of estimates for elements of  $\Sigma$  corresponding to non-zero edges in  $\Omega$  showed almost no variation between models, we therefore expected that estimates for these elements would be similar across all

four models, which was the case. Estimates for all other elements showed some variation between models; some estimates were up to 50% bigger or smaller when a decomposable model was used. In this more complex graph there was no discernible pattern to the placement of these elements (see Figure A.15).

## 4.7 Discussion

We have examined three ways of estimating the variance of the inverse covariance elements. The EFI and empirical variances provide ways of examining how changes in a known  $\Omega$  affect the difficulty of estimation. The empirical variance is a gold standard, but requires a separate simulation for each different  $\Omega$ . For small samples, as we would expect, the EFI variance underestimates the empirical variance. However, it preserves the pattern of variance changes between models, and for small models yields analytical formulae for the difference in variance between two models. In a data analysis context, the OFI variance is the most readily available variance estimate. It is highly variable both in absolute terms and in its mean difference to the empirical variance when the sample size is small. We also considered calculating a bootstrap variance, this was found to mostly overestimate the empirical variance by up to 20% even when  $n = 1000$ .

The different variance estimates all point to the fact that estimates of  $\Omega$  obtained by fitting a decomposable model always have larger variance than those fitted using the (true) non-decomposable model. This is not surprising, given that for each decomposable model we are using the same amount of data to estimate more parameters. However, the differences in the variance of the  $\Omega$  elements are typically small, and the absolute (though not percentage) difference between the model types decreases with increasing sample size. In this regard the size of the sample appears to be more important than the size relative to the number of variables. The structure of the decomposable model chosen has an effect on the inference quality for particular elements: if a particular element of  $\Omega$  is of interest, we should avoid placing the corresponding edge in a triangle which in the decomposable model contains only

one extra edge.

Our focus in this chapter has been on parameter estimation, that is we assume model selection has already occurred. We do, however, note here that a comparison of estimates for elements of  $\Omega$  with their standard error ( $= \sqrt{\text{OFI variance}}$ ) does give some pointers to which elements may in fact have a true value of zero. Correspondingly in all but the very small sample sizes, the increased OFI variances for true non-zero elements suggests only increased variability in the estimates, rather than the possibility of a completely different parameter value.

Estimation of  $\Sigma$  is more problematic in that the penalty for fitting a decomposable model when the true model is non-decomposable has the potential to be substantial. The only estimates for elements of  $\Sigma$  which can be relied upon are those which correspond to non-zero edges in the true  $\Omega$ . Further in our simulation studies, for other elements, we noted that the empirical variances for the estimates derived from the decomposable models could be many times greater than the actual parameter values. The detrimental effect in our experiments of additional edges associated with small but non-zero estimates in the inverse covariance matrix suggests that procedures that estimate many zero parameters as small but non-zero (e.g. the graphical lasso Friedman *et al.* (2008b)), should be scrutinized carefully when the goal is estimation of the covariance matrix.

We note again that all decomposable models we used were ‘true’ models where the edges were a superset of the edges in the non-decomposable model. Output from a model selection procedure may not have the superset property. For example Roverato (2002) and Atay-Kayis and Massam (2005) show the decomposable model with highest *a posteriori* probability for the Iris data set is a chain of the 4 variables, i.e. a subset of the best non-decomposable model. The role of model averaging may also avert the consequences of restricting to decomposable models: Scott and Carvalho (2008) report good estimates of  $\Sigma$  from such a procedure. We leave the behavior of model selection and model averaging procedures restricted to decomposable models for a future investigation.





# Chapter 5

## The performance of covariance selection methods that consider decomposable models only

### 5.1 Introduction

Gaussian graphical models (Dempster, 1972; Lauritzen, 1996; Whittaker, 2008) are a powerful tool for both exploring the partial independence structure of data and for regularization of the covariance matrix. The  $p$ -node graph for  $p$ -dimensional data will have edges corresponding to non-zero off-diagonal elements in the inverse covariance matrix  $\Omega$ . The inverse covariance matrix thus yields the partial independence structure of the data. In high dimensional problems, especially when the sample size ( $n$ ) is similar to, or less than, the number of variables ( $p$ ), regularization of the covariance matrix leads to improved estimation. This regularization can be achieved via covariance selection to achieve a sparse inverse covariance matrix. In both the above situations we require models that distinguish relevant edges from irrelevant ones. This introduces computational challenges, particularly in high-dimensions.

In this chapter we compare two commonly used approaches to estimating the in-

verse covariance matrix, Bayesian model selection and graphical lasso. Bayesian models usually restrict consideration to decomposable models for computational convenience (Scott and Carvalho, 2008; Armstrong *et al.*, 2009; Jones *et al.*, 2005). Few authors who have studied Bayesian methods without the restricting the class of models (Dellaportas *et al.*, 2003; Wong *et al.*, 2003; Moghaddam *et al.*, 2009), and only one (Moghaddam *et al.*, 2009) presents a method that appears to be scalable to high dimensions. Here we use feature-inclusion stochastic search (FINCS) (Scott and Carvalho, 2008), a decomposable restricted method, demonstrably better than other comparable Bayesian methods (Scott and Carvalho, 2008). Graphical lasso approaches (Friedman *et al.*, 2008b; Fan *et al.*, 2009; Ambroise *et al.*, 2009), on the other hand, do not make any restrictions on the structure of the graph but the resulting graphs are usually not as sparse as graphs estimated using Bayesian methods. The motivation for this chapter arises from consideration of the implications of restricting model selection to decomposable models in cases where the true model is non-decomposable.

We begin by examining the behaviour of FINCS under model misspecification, that is, when the true model is non-decomposable. Here we consider the expected asymptotic behaviour. We then simulate data from a very sparse non-decomposable inverse covariance matrix and use FINCS to fit a model. When the sample size is large ( $> 12p$ ), the non-zero elements of the top models, as expected, include all the non-zero elements from the true model and the (minimum) extra non-zero elements needed to make a decomposable model. On the other hand when the sample size is smaller, the non-zero elements in the top models frequently miss some of the true non-zero elements. Furthermore we also found that the magnitude of all true partial correlations needed to be greater than 0.2 for the top models to find all true non-zero elements. FINCS (and indeed graphical lasso) rarely distinguish between a true zero and a very small partial correlation.

In their paper Scott and Carvalho (2008) compare the performance of FINCS models in prediction to those obtained using lasso regression of each variable on the remaining variables, as in Meinhausen and Bühlmann (2006), to obtain a sparse graph. We were also interested in ascertaining whether Bayesian methods still out

perform lasso methods when more recent adaptations and methods of selecting the penalty parameter are used, namely graphical lasso (Friedman *et al.*, 2008b) and adaptive graphical lasso (Fan *et al.*, 2009). In our comparisons graphical lasso was always the least sparse and adaptive graphical lasso, while increasing in sparsity, was only sparser than the FINCS top model for a few cases at very low dimensions ( $p = 4$ ). We also compare parameter estimation by using the Kullback-Leibler divergence as a measure of the nearness of the estimates to the true parameters. Here the results were less clear cut with the graphical lasso methods (especially adaptive graphical lasso) often being ‘closer’ to the true model (smaller Kullback-Leibler divergence) than the FINCS based estimate.

Finally we consider predictive performance. Here we found that, in terms of predictive ability, while graphical lasso was clearly the worst performer, at low dimensions there is no clear winner between the FINCS top graph and adaptive graphical lasso. However, as the dimension increased FINCS began to outperform adaptive graphical lasso. Model averaging is often promoted as useful tool for improving the predictive ability of Bayesian models. However here the gains in predictive accuracy by using model averaging were so slight that the extra computational time involved makes it not worthwhile.

The rest of this chapter is organized as follows. In Section 2 we review the general properties of graphs used in this chapter and detail the algorithms used for the two approaches to model selection. Section 3 details data used, how each method was implemented and how comparisons were quantified. Results are presented in Section 4 and we conclude in Section 5 with a summary of our findings and suggestions for future work.

## 5.2 Background

### 5.2.1 General properties of graphs

Without loss of generality we assume that our Gaussian data is centred with mean zero and  $p$ -dimensional covariance matrix  $\Sigma$ . Using Gaussian graphical models (GGMs) focuses attention on the inverse covariance matrix  $\Omega = \Sigma^{-1}$ . If  $\mathcal{G}=(V,E)$  is a graph with edges  $E$  and ( $p$ ) vertices  $V$ , then  $\mathcal{G}$  is a *Gaussian graphical model* if  $e_{i,j} \in E \Leftrightarrow \omega_{i,j} \neq 0, \forall i \neq j, i, j \in V$ . If all  $\omega_{ij}$  are non-zero, then the graph has all possible edges and is said to be *complete*. The symmetry of  $\Omega$  ensures that this is an undirected graph, that is  $e_{i,j} \in E \Leftrightarrow e_{j,i} \in E$ . A sample of  $n$  elements from our  $MVN(0,\Sigma)$  distribution will have a sample covariance matrix  $S$ . In general the graph for  $S^{-1}$  will be complete even if the graph for  $\Omega$  is not. Furthermore if  $p > n$  we cannot invert  $S$ . Dempster (1972) first identified the issue of deciding which elements of  $\hat{\Omega}$  should be zero as the covariance selection problem.

If  $A$ ,  $B$  and  $C$  are disjoint subsets of  $V$  and  $A \cup B \cup C = V$  then  $C$  *separates*  $A$  and  $B$  if all paths from  $A$  to  $B$  must pass through  $C$ . If furthermore  $C$  is complete then  $(A, B, C)$  is a *decomposition* of  $\mathcal{G}$ . If we iteratively decompose the graph until no further decompositions can be found, then the subgraphs so found are the set of *prime components* (Jones *et al.*, 2005). If the prime components are all complete they are called *cliques* and the graph is a (fully) *decomposable graph* (Lauritzen, 1996). We emphasize here that the existence of a decomposition does not imply that a graph is decomposable. If any of the prime components found by iterative decomposition are not complete and cannot be further decomposed then that component is non-decomposable and therefore so is the whole graph.

In this thesis we define a *superset graph* to be one which includes all the true edges plus at least one other edge which is not part of the true edge set. A minimal superset graph of a non-decomposable graph includes only the minimal number of extra edges needed to achieve a triangulation. In a similar vein we define a *subset graph* to be one which includes no extra edges, and also fails to include at least one of the true edges.

In the GGM setting decomposability means that the density function can be written as a product of the marginal densities of the cliques and reciprocal of the marginal densities of the separators (Lauritzen, 1996). This allows the marginal likelihoods to be calculated analytically. We use the fractional-Bayes approach of Carvalho and Scott (2009) which specifies the hyper-inverse Wishart scale parameter of the prior in terms of the sums of squares matrix (see Section 2.2.2). Thus if  $X^T X$  is the sum of squares matrix for the data matrix  $X$  then

$$p(X|\mathcal{G}) = (2\pi)^{-np/2} \frac{h(\mathcal{G}, gn, gX^T X)}{h(\mathcal{G}, n, X^T X)} \quad (5.1)$$

where we set  $g$  as  $1/n$  (see Section 2.2.2). If  $\mathcal{P}$  represents the set of all prime components and  $\mathcal{S}$  the set of all separators in  $\mathcal{G}$  then the normalizing constant  $h$  is calculated as:

$$h(\mathcal{G}, b, D) = \frac{\prod_{P \in \mathcal{P}} |\frac{1}{2} D_P|^{\frac{(b+|P|-1)}{2}} \Gamma_{|P|} \left( \frac{(b+|P|-1)}{2} \right)^{-1}}{\prod_{S \in \mathcal{S}} |\frac{1}{2} D_S|^{\frac{(b+|S|-1)}{2}} \Gamma_{|S|} \left( \frac{(b+|S|-1)}{2} \right)^{-1}} \quad (5.2)$$

and

$$\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(x + (1-j)/2)$$

is the multivariate gamma function.

## 5.2.2 Feature-inclusion stochastic search

Feature-inclusion stochastic search (FINCS)(Scott and Carvalho, 2008) has been proposed as a method superior to both other Bayesian methods and lasso based methods for model selection. FINCS is a serial procedure that combines three types of moves through the space of all possible graphs. Most moves are local moves which exploit the computational advantages of adding or deleting only one edge at a time. The decision to add or delete is made randomly with the actual edges to add (or delete) being chosen in proportion to the relative inclusion probabilities (see equation (5.4)), and to maintain decomposability. As each new model is found the model score (log of the non-normalized posterior probability) is calculated using:

$$\text{model score} = \log P(X|\mathcal{G}_k) + \log mc(\mathcal{G}_k) \quad (5.3)$$

where  $P(X|\mathcal{G}_k)$  is calculated using equation (5.1) and  $mc(\mathcal{G}_k)$ , the multiplicity correction prior over the graphs, using equation (5.5).

Global moves are used to move to another part of the graph space by generating a randomized median triangulation pair, in order to avoid missing regions that are not easily found in stepwise moves. This is done by starting with an empty graph and adding edges in proportion to their current estimated inclusion probability (see equation (5.4)). The graph so formed ( $\mathcal{G}_N$ ) is usually not decomposable so a minimal decomposable supergraph ( $\mathcal{G}^+$ ) and a maximal decomposable subgraph ( $\mathcal{G}^-$ ) are found. Model scores are then calculated for both  $\mathcal{G}^+$  and  $\mathcal{G}^-$  and the one with the highest model score chosen.

Finally resampling moves revisit graphs in proportion to their model score and thereby ensure that the global moves do not irretrievably direct the search away from ‘good’ graphs. A blend of 80-90% local moves and 10-15% resampling moves with the balance being global moves is recommended by Scott and Carvalho (2008). We implement FINCS with a global move every 20 iterations and a resampling move every 10 iterations.

For each edge  $e_{i,j}$  the inclusion probability at step  $t$  is estimated by the relative inclusion probability:

$$\hat{q}_{ij}(t) = \frac{\sum_{k=1}^{k=t} 1_{(i,j) \in \mathcal{G}_k} P(X|\mathcal{G}_k) mc(\mathcal{G}_k)}{\sum_{k=1}^{k=t} P(X|\mathcal{G}_k) mc(\mathcal{G}_k)} \quad (5.4)$$

$P(X|\mathcal{G}_k)$  is calculated using equation (5.1) and

$$mc(\mathcal{G}_k!) = \frac{\kappa(m - \kappa)!}{(m + 1)(m!)} \quad (5.5)$$

where  $\mathcal{G}_k$  has  $\kappa$  edges out of the  $m = p(p - 1)/2$  total possible edges. This is the multiplicity correction prior over the graphs (Scott and Carvalho, 2008) which places a conjugate beta prior on the success probability  $r$  of the standard binomial prior. (See Section 2.3 for more details) As the relative inclusion probabilities are only based on the graphs visited they do not converge to the true inclusion probabilities except in the trivial sense of all models eventually being enumerated. They do however give useful pointers as to the importance of an edge. We use the C++ implementation of FINCS described in Scott and Carvalho (2008).

In a similar manner, the exponentiated scores for the final list of retained models can be normalized to compute ‘relative posterior probabilities’. These are true probabilities only relative the restricted list of models retained, but again are useful indicators of importance within this set.

We found that stable results were obtained with as few as 100 iterations for the  $p = 4$  cases, where there are only 61 possible decomposable graphs. We define stable results here to mean that the model score of the top graph did not change and correspondingly there was little or no change in the relative inclusion probabilities when the number of iterations was increased to 3 million. In all other cases we ran FINCS for 3 million iterations and found there was little or no change in the relative inclusion probabilities when the number of iterations was increased to 5 million.

### 5.2.3 Kullback-Leibler divergence

White (1982) showed that, in the event of model misspecification, the maximum likelihood estimator should converge to the model that minimizes the Kullback-Leibler divergence (Kullback and Leibler, 1951). The Kullback-Leibler divergence between two density functions  $f$  and  $g$  is  $E[\log f(X)/g(X)]$  where the expectation is with respect to  $f$ . If  $f$  and  $g$  are both multivariate normal with the same mean then Whittaker (2008, p168) gives a formula for calculating the divergence between their variances. We set  $f$  as the true model and  $g$  the estimate so that the Kullback-Leibler divergence ( $KL$ ) is calculated as

$$KL = \frac{1}{2}tr(\Sigma\hat{\Omega} - I_k) - \frac{1}{2}\log \det(\Sigma\hat{\Omega}) \quad (5.6)$$

where  $I_k$  is the  $k$  by  $k$  identity matrix,  $\Sigma = \Omega^{-1}$  is the true covariance matrix and  $\hat{\Omega}$  is the estimated  $\Omega$  matrix .

The posterior mode will (asymptotically) behave in a similar manner to the likelihood, because of the central role of the likelihood in all Bayesian approaches (Berger and Wolpert, 1984). Restricting model selection to decomposable models when the



true model is non-decomposable is a special type of misspecification, in that we can get arbitrarily close to our true model by selecting a decomposable model with close-to-zero estimates for many elements. Since there are, in fact, many such models we would expect that, with a large amount of data, any well mixing posterior search or sampler will visit several modes that are supersets of the true model. Shalizi (2009, page 11) show that even when the model is misspecified the posterior concentrates on the divergence minimising part of the search space. Where the likelihood and the prior have the same support, the likelihood will always dominate the prior asymptotically, thus we expect that these models will have essentially equal Kullback-Leibler divergence. If there is a penalty on the inclusion of edges in these models, then they will be minimal decomposable supersets of the true model edges. Furthermore since all the supersets contain the true model edges, and there are many possible combinations from which edges to add to create a decomposable graph, the relative inclusion probability should be a good indicator of the edges in the true model.

Model averaging is often promoted as useful tool for improving the predictive ability of Bayesian models. When the true model is sparse there are many supersets, even with  $p$  moderate. Thus if we restrict consideration to (say) the top 500 models, then most will be supersets. The estimated partial correlations for edges other than true edges will be close to zero, meaning that all these supersets represent essentially the same model. We therefore expected to see minimal benefit to model averaging, which was the case.

#### 5.2.4 Graphical lasso and adaptive graphical lasso

In their paper Scott and Carvalho (2008) compare the performance of FINCS based models in prediction to those obtained using lasso regression of each variable on the remaining variables to obtain a sparse graph in the manner of Meinhausen and Bühlmann (2006). The more recent graphical lasso (Friedman *et al.*, 2008b) applies an  $L_1$  penalty directly to the inverse covariance matrix with superior performance.

Thus the objective function is

$$\log \det \Omega - \text{tr}(\Omega S) - \lambda \sum_{i=1}^p \sum_{j=1}^p |\omega_{ij}| \quad (5.7)$$

where  $\Omega$  is a positive definite matrix,  $S$  is the sample covariance matrix and  $\lambda > 0$  is the penalty.

As with other lasso-based methods selecting the penalty is an important first step. Friedman *et al.* (2008b) recommend using cross-validation. We use the R-package `glasso`, (R Development Core Team, 2009; Friedman *et al.*, 2008a), with the penalty selected by 5-fold cross validation and the sample covariance estimated with an  $n$  divisor to obtain our graphical lasso estimate. In this section we use 5-fold rather than 10-fold cross validation due to the smaller size of many of our samples. The graphical lasso algorithm as implemented in `glasso` yields an estimated inverse covariance matrix that is not perfectly symmetric (at 3-4 significant figures). We used an inverse covariance matrix made exactly symmetric by using the average of the  $i, j^{\text{th}}$  and  $j, i^{\text{th}}$  elements.

Fan *et al.* (2009) apply an adaptive lasso penalty and a Smoothly Clipped Absolute Deviation (SCAD) penalty to graphical lasso as a method of solving the issue of all estimates being biased towards zero and also of obtaining a sparser graph. Here we focus on the adaptive graphical lasso as it gave better estimates in initial experiments. The adaptive graphical lasso is implemented using a penalty matrix ( $\zeta$ ) rather than the scalar penalty term of graphical lasso. The elements of  $\zeta$  are  $\zeta_{i,j} = 1/|\tilde{\omega}_{i,j}|^\gamma$ , where  $\tilde{\Omega} = (\tilde{\omega}_{i,j})_{1 \leq i,j \leq p}$  is any consistent estimate of  $\Omega$  and  $\gamma > 0$ . Thus for adaptive graphical lasso the objective function becomes

$$\log \det \Omega - \text{tr}(\Omega S) - \lambda \sum_{i=1}^p \sum_{j=1}^p \zeta_{i,j} |\omega_{ij}| \quad (5.8)$$

We implemented adaptive graphical lasso using the symmetric graphical lasso estimated inverse covariance matrix as  $\tilde{\Omega}$ ,  $\gamma = 0.5$  and selecting the penalty by 5-fold cross-validation. We again used the R-package `glasso`, making the estimate exactly symmetric in the same manner as for the graphical lasso estimate.

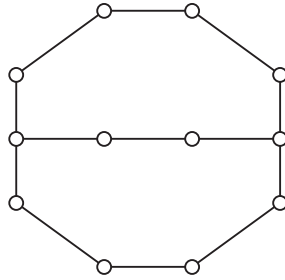


Figure 5.1: The 12 variable model.

## 5.3 Data and methods

### 5.3.1 Data

We simulate large ( $n = 1000$ ) datasets in order to explore how FINCS treats data where the true underlying model is non-decomposable. We use two different model structures for  $\Omega$  to simulate our data, a cycle (with 4, 6, or 20 nodes) and the 12-node structure (see Figure 5.1) used in examples in Jones *et al.* (2005) and in Chapter 4. In each case all diagonals of  $\Omega$  are 20 with non-zero off-diagonals for the cycles being -9, and in the 12-node case -8, making all partial correlations 0.45 (or 0.4 for the 12-node case). Henceforth we refer to these matrices as  $\Omega_{same}$  (or the 12-node  $\Omega_{same}$ ). This was done in order to focus attention on the ability (or otherwise) of FINCS to point to the true edges without the distraction of small partial correlations. For each  $\Omega$  data was simulated from a  $MVN(0, \Omega^{-1})$  distribution using the Cholesky decomposition of  $\Omega^{-1}$  and the R function `rnorm`. Smaller datasets ( $n = 50$ ) were also generated to observe the behaviour when  $n$  is small. We also generated datasets of other varying sizes to try and gauge at what point ‘large’  $n$  behaviour begins. In order to explore the ability of each method to distinguish true

small partial correlations from true zero partial correlations we also simulated data using  $\Omega$  matrices with the same structures and smaller off diagonals giving partial correlations between 0.05 and 0.4. We also used  $\Omega$  matrices with varying sized partial correlations:  $\Omega_{big}$ , and  $\Omega_{small}$  for the 4 and 20 node cycles as in Chapter 4 (Tables 4.1 and A.6); and for the 12-node case a matrix ( $\Omega_{twelve}$ ) with partial correlations ranging from approximately 0.01 to 0.7 which is given in Table B.1.

We used three  $\Omega$  matrices for simulating data to compare model selection and prediction by FINCS and the two graphical lasso-based methods. These were the 4-node and 20-node  $\Omega_{same}$  matrices and the  $\Omega_{twelve}$  (with varying partial correlations). Each matrix was used to simulate five different datasets of size  $n = 50$  and five of size  $n = 1000$ . For prediction purposes a prediction dataset (of size  $n = 50$ ) was also simulated in the same manner from each of the three  $\Omega$ .

We also explored the behaviour of each method with real data with a 59-node mutual-funds dataset (Scott and Carvalho, 2008). We split the 86-month sample into a 60 month training set (the first 60 months) and a 26 month prediction set (the remaining 26 months) which enabled us to compare predictions using FINCS derived estimates of the covariance matrix with predictions using lasso derived estimates of the covariance matrix.

### 5.3.2 Model selection

To assess how FINCS treats non-decomposable graphs we consider the graphs with the highest model score comparing them to the true model to ascertain whether or not they are a superset graph. We use the Kullback-Liebler divergence (see Section 5.2.3) as a measure of distance from the the true model to assess whether FINCS behaves as expected. In each case  $\hat{\Omega} = \bar{\Omega}$  is the posterior mean for that graph (see Section 2.2.2 ).

We also consider the ability of relative inclusion probabilities to point towards the true (non-decomposable) graph by identifying the graph obtained by specifying as edges, those with an relative inclusion probability of at least 0.8. In this case  $\Omega$

is estimated as the maximum likelihood estimate (MLE), calculated using the R package `glasso` (with shrinkage penalty `rho = 0`, and zero elements specified as those associated with edges with an relative inclusion probability less than 0.8). FINCS models (the posterior mean of the top graph and the inclusion probabilities graph MLE) are compared to the graphical lasso and adaptive graphical lasso models using the Kullback-Leibler divergence as above.

We quantify the accuracy of edge selection using precision and recall. For each model

$$\text{precision} = \frac{T_E}{T_E + F_E} \quad \text{and} \quad \text{recall} = \frac{T_E}{T_E + F_0}$$

where  $T_E$  is the number of true edges found,  $F_E$  is the number of edges found that are not true edges and  $F_0$  is the number of true edges that were not found. Thus precision is the proportion of edges in the model that are true edges and recall is the proportion of true edges found by the current model. A superset graph as defined in Section 5.2.1 is thus a model with a recall of one and precision of less than one. A subset graph has a precision of one and a recall of less than one.

As a worst case scenario, we also calculate the Kullback-Leibler divergence for the unregularized maximum likelihood estimate (that is the inverse of the sample covariance matrix). We replicate all analyses for each of the 10 simulated datasets (five with  $n = 50$  and five with  $n = 1000$ ) and for each of the three different  $\Omega$ .

### 5.3.3 Prediction

We compute the posterior mean ( $\bar{\Omega}$ ) for each of the top 500 models found using FINCS, and then set  $\hat{\Sigma} = \bar{\Omega}^{-1}$  for prediction purposes. A model averaged prediction was then obtained by calculating a weighted mean of predictions using each of the top 500 models (where the weight is the normalized model score). We also calculate  $\hat{\Sigma}$  ( $= \hat{\Omega}^{-1}$ ) from the symmetric estimates of  $\Omega$  obtained from graphical lasso, adaptive graphical lasso and the MLE of the FINCS inclusion probability graph (as in Section 5.3.2). Comparisons are made between these four predictions and that made using the inverse of the posterior mean ( $\bar{\Omega}^{-1}$ ) of the top FINCS graph.

We initially use the mutual-funds data to assess the performance of each method in predicting unknown values. We in turn compute, for each month in the prediction dataset, the conditional expectation of each return based on the ‘observed’ values of the remaining 58 returns and calculate the squared error. The five methods detailed above are compared using the total sum of the squared errors. We use the same process to obtain the total sum of squared errors for the simulated datasets. As a worst case scenario we also compute the total sum of squared errors using the sample covariance matrix (the maximum likelihood estimate) as  $\hat{\Sigma}$ . Again we replicate the analyses for all the simulated datasets.

## 5.4 Results

### 5.4.1 Feature-inclusion stochastic search (FINCS) treatment of non-decomposable graphs

The graphs with the highest model score were all supersets of the true graph for large ( $n = 1000$ ) samples simulated from distributions where all true non-zero partial correlations are 0.45. There are only three such supersets when  $p = 4$ . We observed that, when  $n=1000$ , these three superset graphs had both a larger log-posterior and a smaller Kullback-Leibler divergence than all the other graphs (see Figure 5.2(a)). In almost all cases, when  $n$  was smaller, we found that the Kullback-Leibler divergence was smaller for superset graphs. This occurred even though a superset may not have the largest log-posterior. Figure 5.2(b) illustrates this for the 50 graphs with the highest model score as found by FINCS for a sample of 50 simulated from the 6-node  $\Omega_{same}$  as specified in Section 5.3.1. Here we observe that, even though the top graph is not a superset, there is only one non-superset graph with a Kullback-Leibler divergence as small as those for the superset graphs.

Varying the sample size (only) we found that the graphs with the highest model scores were always superset graphs when  $n \gg 12p$ . It should be noted here that, apart from the  $p = 4$  case, not all superset graphs are visited and as  $n$  decreases and

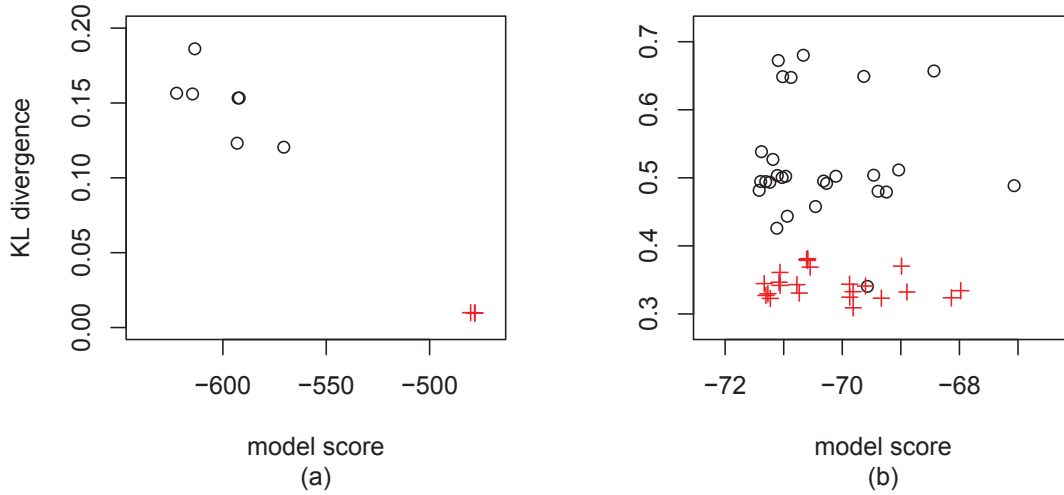


Figure 5.2: Model score vs Kullback-Leibler divergence for:

(a) the top 10 graphs found by FINCS for a 4 variable cycle ( $n=1000$ ) and

(b) the top 50 graphs found by FINCS for a 6 variable cycle ( $n = 50$ ).

+ = superset graphs; o = non-superset graphs

gets close to  $12p$  not all superset graphs will have model scores larger than other graphs, as observed in Figure 5.2 (b).

We now turn to the effect of the size of the partial correlations. We kept  $n$  large (1000) and decreased the value of off diagonal elements of  $\Omega_{same}$  and thus the partial correlations ( $\tilde{\rho}_{ij}$ ). We observed here that there appears to be a threshold ( $|\tilde{\rho}_{ij}| > 0.1$  when  $p = 4$  and  $|\tilde{\rho}_{ij}| > 0.3$  when  $p = 20$ ) above which the top graphs are all supersets. The top graph was a subset of the the true graph when all the non-zero  $|\tilde{\rho}_{ij}|$  were below 0.1. Subsequent graphs both missed true edges and included incorrect ones (see Figures B.1 and B.2). We also note that the subset graphs tended to have a smaller Kullback-Leibler divergence than graphs which also included incorrect edges.

We also considered the situation where data was simulated from a distribution where the value of the true partial correlations varied. The top graph was a superset only for  $\Omega_{big}$ ,  $p = 4$  which had all  $|\tilde{\rho}_{ij}| > 0.5$  (see Appendix Figure B.3) . In other cases edges corresponding to small ( $|\tilde{\rho}_{ij}| < 0.1$ ) were almost always omitted. In these

situations the subset graphs did not necessarily have the smallest Kullback-Leibler divergence. (Detailed results are presented in Figure B.4 and Tables B.2, B.3, B.4, B.5 and B.6.)

As the relative inclusion probabilities point to the importance of edges we also considered whether the relative inclusion probabilities gave any pointers as to the true non-decomposable nature of the graph. With  $n = 1000$ ,  $p$  very small (4 or 6) and all partial correlations 0.45 ( $\Omega_{same}$ ), the relative inclusion probabilities for the true edges (only) were all 1.000 (see Table B.7). When  $n = 1000$ , for the 12-node  $\Omega_{same}$  and the 20-node  $\Omega_{same}$ , not only did the true edges have an relative inclusion probability of 1 (to 3dp) but so too did at least one other edge. (Table B.8 shows this.) The relative inclusion probabilities for all true edges were more than 0.8 under the same condition that led to the top graphs being supersets. In most cases (as we see in Table B.9) there was also at least one other edge that had an relative inclusion probability in the same range, thus making it not clear exactly which were the true edges. There is however a definite indication that the true model is non-decomposable and is sparser than any of the top graphs. The high relative inclusion probabilities often miss a true edge and may include other edges for smaller sample sizes, where the top graphs may not be supersets (see Table B.10). Edges associated with  $|\tilde{\rho}_{ij}| < 0.1$  in the true partial correlation matrix have low ( $< 0.5$ ) relative inclusion probabilities (see Table B.11).

#### 5.4.2 Model selection comparison of FINCS with graphical lasso methods.

Figure 5.3 gives the results for the two sets of five datasets simulated from the 4-node  $\Omega_{same}$ . All methods always find all the true edges when  $n$  is large for this four variable cycle (recall=1 in Figure 5.3 for  $n = 1000$ ). There were two instances where the inclusion probabilities missed one true edge when  $n = 50$  (see Figure 5.3). Adaptive graphical lasso and the inclusion probabilities were the only methods which ever correctly identified the model. However, the model selected varied from sample to sample, even when  $n$  was large. With so few variables the addition or deletion of



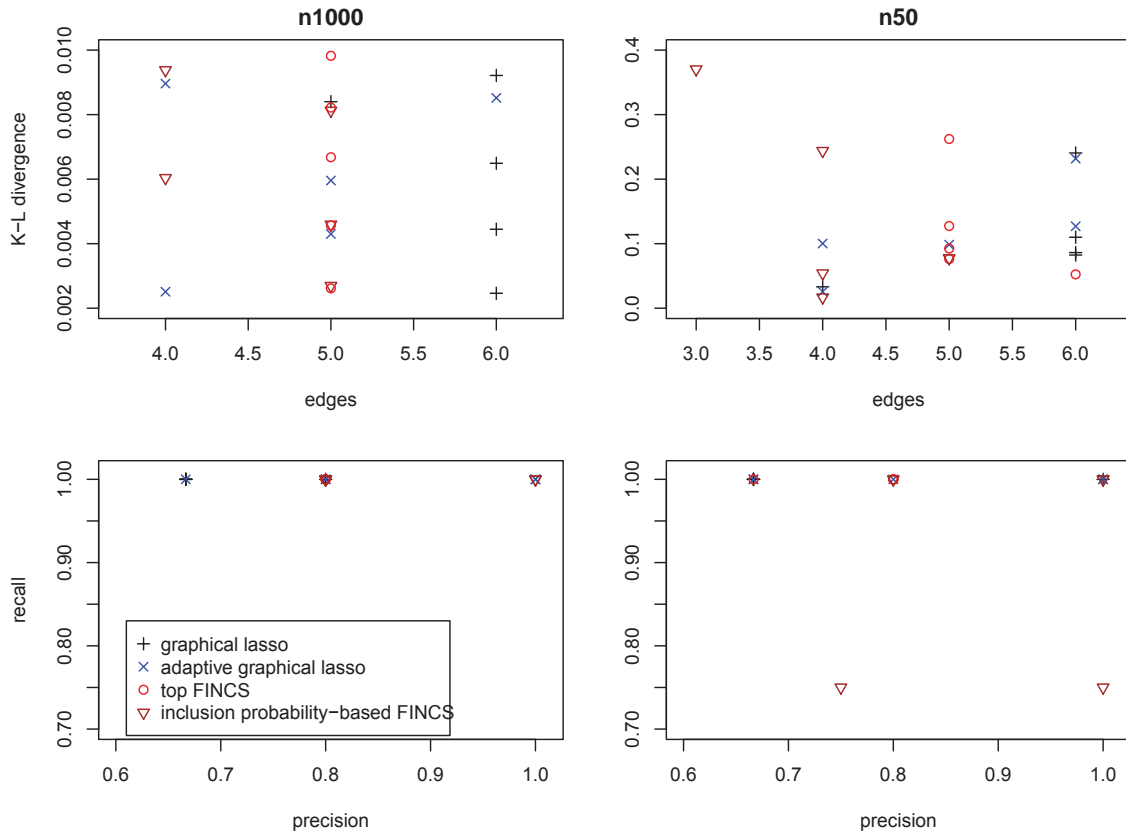


Figure 5.3: Model comparison measures for 5 samples of  $n = 50$  and 5 of  $n = 1000$  from a  $p = 4$  cycle

one edge or minor changes to the parameter estimates can have observable effects on the Kullback-Leibler divergence. This point is best illustrated by an extreme case where the model selected by adaptive lasso was correct but the divergence was greater than the divergence of the (unregularised) maximum likelihood estimator! All methods displayed considerable variation, with the (obvious) exception of the maximum likelihood estimate. Each method had the smallest divergence for at least one of the five  $n = 50$  samples. When  $n = 1000$  the Kullback-Leibler divergence was quite similar across all model selection methods. Nevertheless, the combination of shrunk estimates and the ability to select the correct model means that the graphical lasso methods tended to have a slightly smaller Kullback-Leibler divergence.

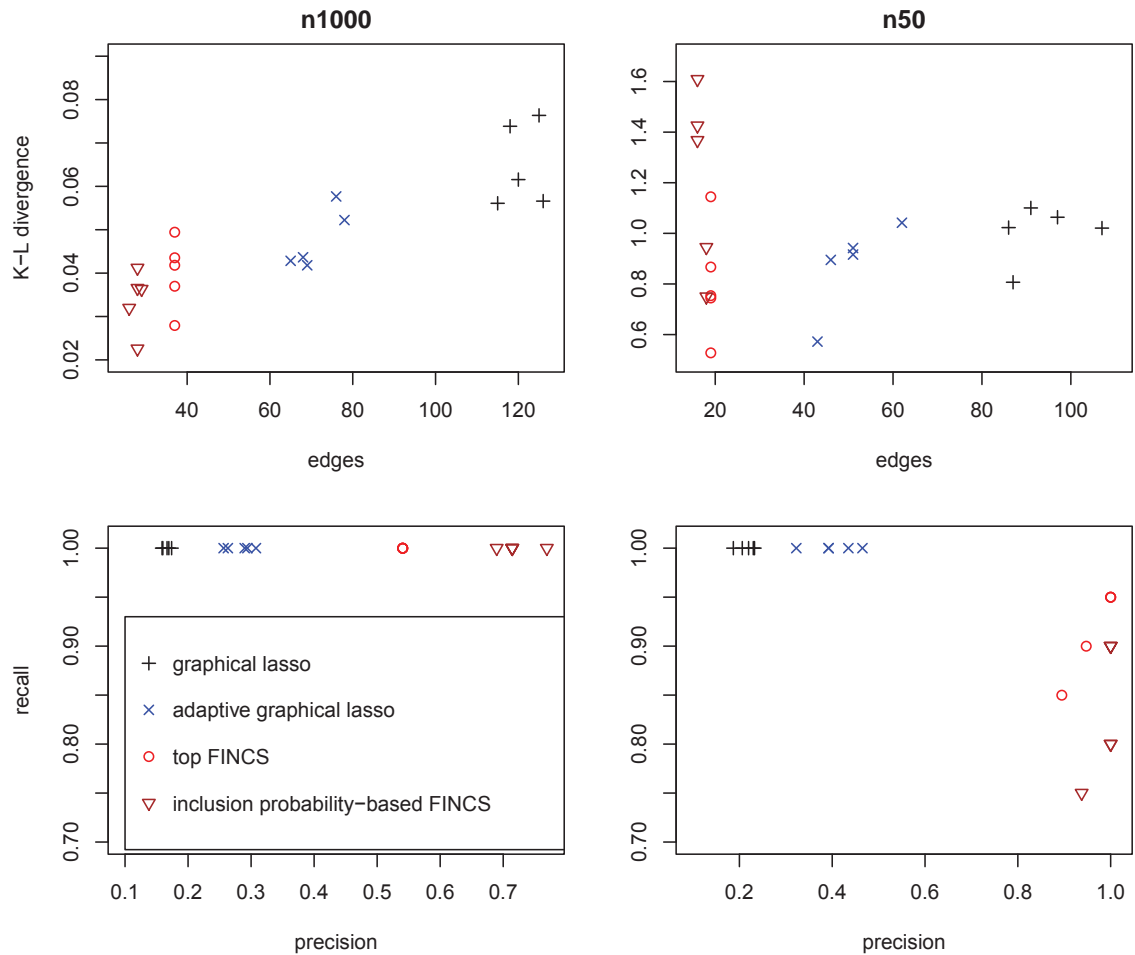


Figure 5.4: Model comparison measures for 5 samples of  $n = 50$  and 5 samples of  $n = 1000$  from a  $p = 20$  cycle.

Figure 5.4 shows that for the 20 variable cycle by all measures the estimates based on FINCS graphs tend to outperform the two graphical lasso graphs. MLE results are not shown as the number of edges and Kullback-Leibler divergence were so much greater their inclusion would have distorted the scale of the figures. The MLE results are given in Table B.12. The FINCS top graph is sparser than both the graphical lasso-based graphs, and the inclusion probabilities point to an even sparser superset graph for the large ( $n = 1000$ ) samples. Here the advantages of sparsity are also apparent in parameter estimation with both the sparser FINCS based estimates always having a smaller Kullback-Leibler divergence. The only possible issue with FINCS is that for small samples (seen for  $n = 50$ ) the recall is less than 1 meaning that the top graphs are not superset graphs (see Figure 5.4). This in turn means that

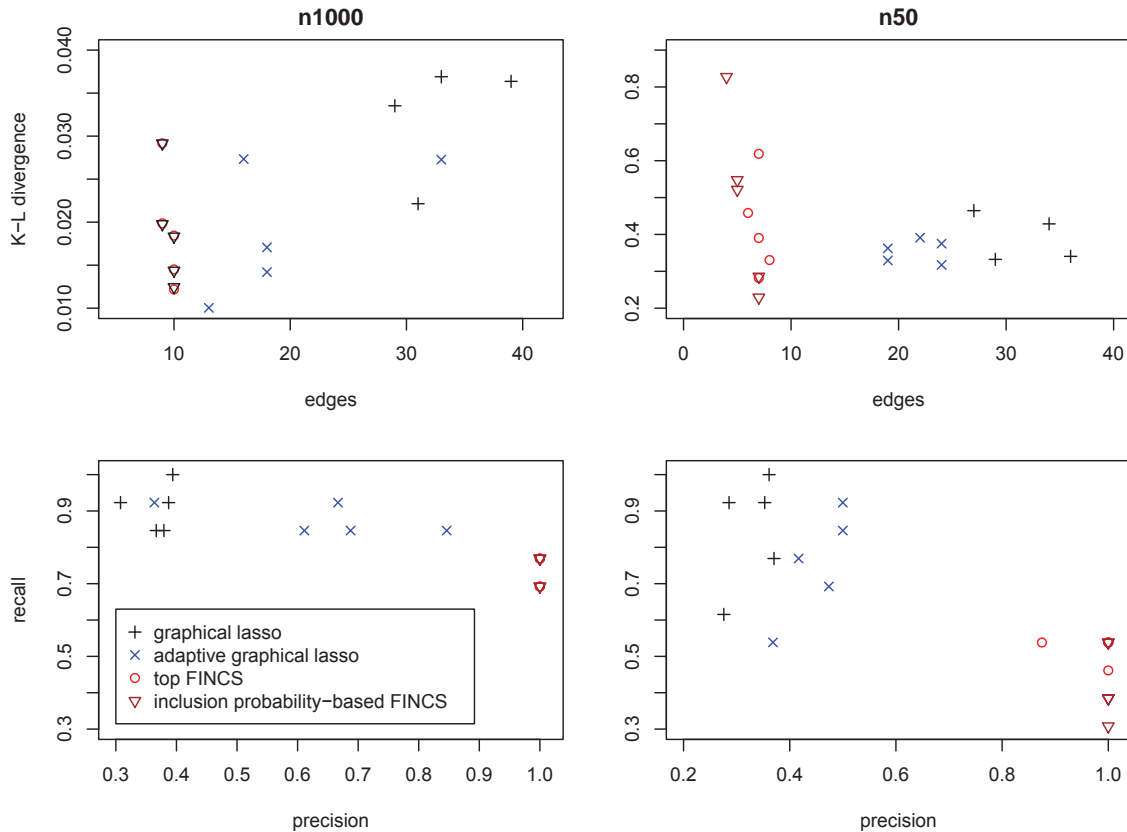


Figure 5.5: Model comparison measures for simulated data, where the true model includes some very small partial correlations ( $p = 12$ ;  $n = 50$  and  $n1000$ ).

the increased sparsity of the inclusion probabilities graph yields worse rather than better estimates. Despite this in all but one case the Kullback-Leibler divergence for the estimates based on the FINCS top graph were the smallest.

We saw in Section 5.4.1 that, when true partial correlations are small, FINCS may fail to find those edges. We now consider the models selected when the true model includes some small partial correlations. In this situation Figure 5.5 reveals that all methods miss some of the edges (recall  $< 1$ ). While the two graphical lasso methods do identify more of the true edges (higher recall) they also include incorrect edges (low precision). The sparser FINCS graphs only ever included one incorrect edge and this only occurred once (for the top graph when  $n = 50$ ). While adaptive graphical lasso and FINCS (using the top graph) clearly perform better than graphical lasso (and to some extent using FINCS inclusion probabilities) the choice between these two is not clear. FINCS top graph gains in sparsity and precision but the adaptive

graphical lasso often has a smaller Kullback-Leibler divergence and has greater recall. It may depend upon the situation which is the preferred option.

### 5.4.3 Comparison of predictions using FINCS and graphical lasso derived graphs

Mutual-funds data

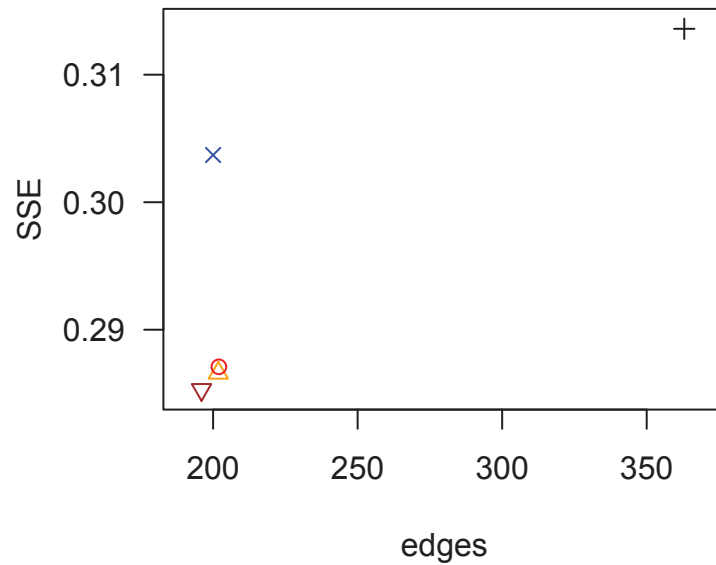


Figure 5.6: Mutual-funds data: Number of edges in graph and total sum of squared errors.

+ = graphical lasso; × = adaptive graphical lasso;

o = top FINCS graph; △ = FINCS model averaged predictions; ▽ = inclusion probability FINCS graph

The three FINCS based methods selected similar edges resulting in a similar sum of squared errors (see Figure 5.6). The most interesting aspect of this is that, as expected, model averaging appears to make only a very small improvement to the overall accuracy of predictions. As seen already in Section 4.2 the FINCS derived

graphs are much sparser than the graphical lasso graph. Both graphical lasso and adaptive graphical lasso have sum of square errors 2-3% greater than the FINCS based models. The adaptive graphical lasso while similar in sparsity to the FINCS models has a sum of squared errors closer to the graphical lasso. We initially suspected that was due to shrinkage in the elements of  $\hat{\Omega}$ . While this may be a partial explanation a comparison of the actual edges found by the two methods reveals that although the number of edges is similar, the actual edges found vary considerably, with only 30% of the edges found being common to both models. Again as a worst case scenario we also computed the maximum likelihood estimate. This is a complete graph of 1711 edges and gave a total sum of squared errors of 5.6.

### Simulated data

We also considered the ability of the various models from Section 5.4.2 to predict data in a prediction set of  $n = 50$  simulated in the same manner as the original datasets. Here, as Figure 5.7 shows, the situations is not so clear. In most cases graphical lasso was the worst performer, apart from using the (unregularised) Maximum Likelihood Estimator (see Table B.13 for MLE sum of squared errors). However which of the other methods performed best varied between samples for all four scenarios.

We then simulated a data set of size  $n = 51$  from a  $p = 50$  cycle in order to assess the performance of each method in a situation with dimensions closer to the mutual funds data. We observed that for each of these five simulated datasets the top FINCS graph had the lowest sum of squared error. Although there was some overlap between methods in the range of the sum of squared errors, the ‘extra edges’ in both graphical lasso methods appear to be having a more detrimental effect here. However, the superiority of FINCS is less impressive than for the mutual-funds data.

We have no way of knowing the true model for the mutual-funds data, but it is feasible that there are groups of closely related funds. This suggests that the true model may be decomposable or at least have a decomposition where many of the prime

components are cliques. This could lead to the dramatically superior performance of FINCS with the mutual-funds data.

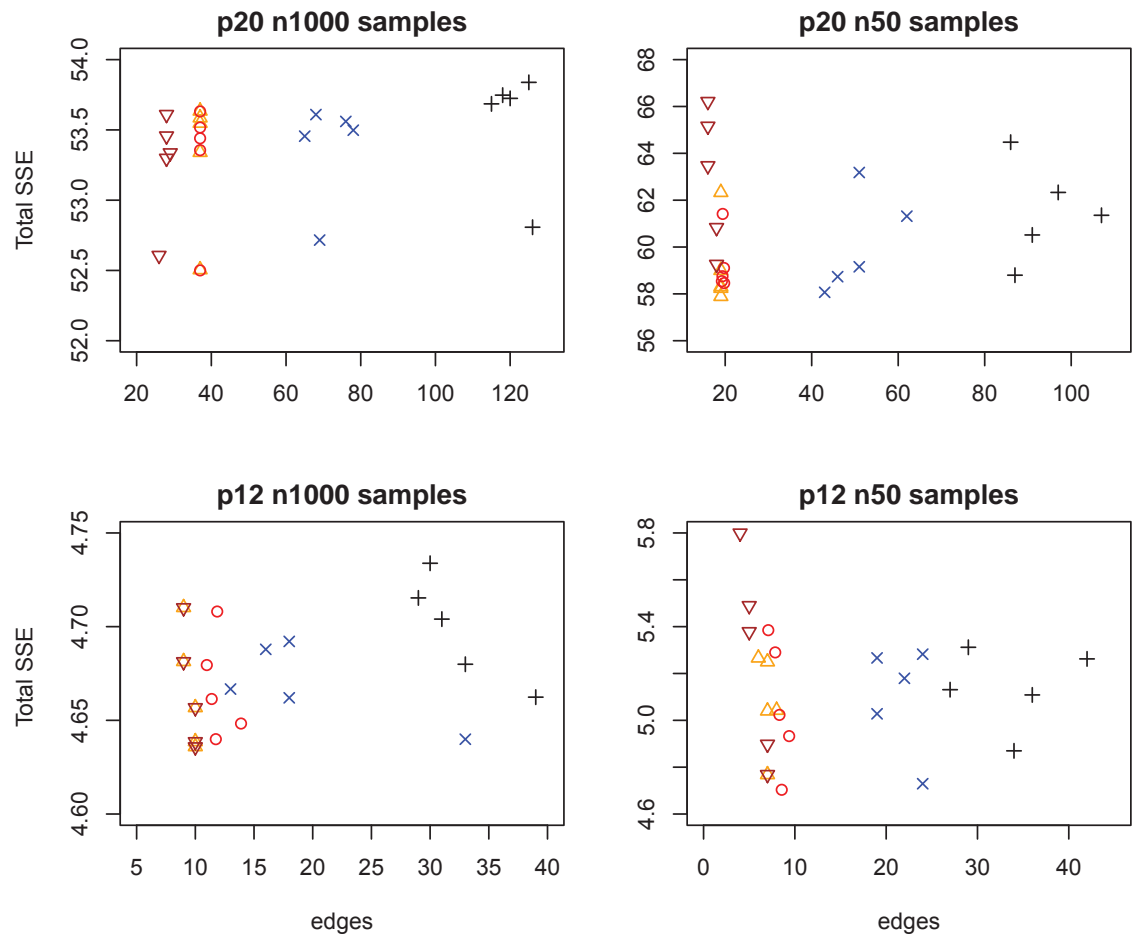


Figure 5.7: Total sum of squared errors and number of edges for 5 simulated samples of each  $n$  and  $p$  as specified.

+ = graphical lasso;  $\times$  = adaptive graphical lasso;

$\circ$  = top FINCS graph;  $\triangle$  = FINCS model averaged predictions;  $\nabla$  = inclusion probability FINCS graph

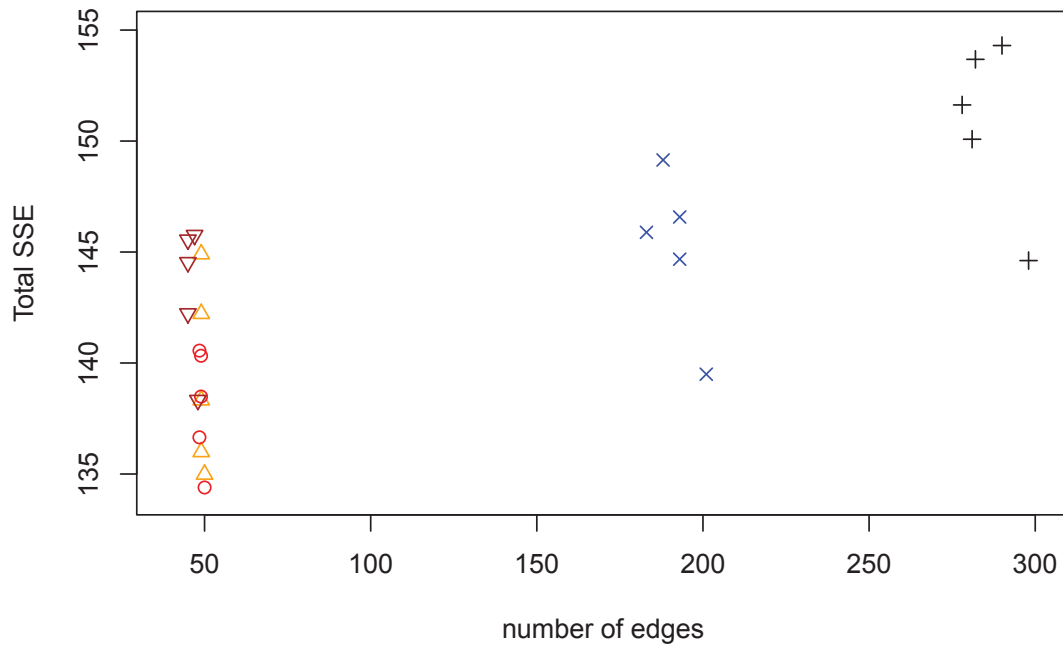


Figure 5.8: Total sum of squared errors and number of edges for 5 simulated samples:  $n = 51$  and  $p = 50$ .

+ = graphical lasso; × = adaptive graphical lasso;

o = top FINCS graph; Δ = FINCS model averaged predictions; ∇ = inclusion probability FINCS graph

## 5.5 Discussion

FINCS behaved as expected for large samples ( $n > 12p$ ). Minimal superset graphs had a smaller Kullback-Leibler divergence than others and as  $n$  increased they also had higher posterior probability. The caveat on this is that FINCS (as with both graphical lasso methods) often failed to pick up edges corresponding to elements with a small ( $< 0.1$ ) true partial correlation. Further simulation studies are needed to ascertain whether these thresholds are unchanged with a non-decomposable graph built from many components like the ones studied here.

A consideration of the edges in the top 500 graphs reveals that the edges in the different models are combinations of only a subset of all possible edges. This suggests that for large samples the top graphs are circulating around the decomposable supersets of the true non-decomposable model. Thus top FINCS graphs all essentially represent the same model and hence model averaging does not improve predictions. At very small dimensions we were able to observe that the Kullback-Leibler divergence is very similar for all superset graphs (only). Requiring retained graphs to exceed a minimum Kullback-Leibler divergence from already-retained graphs, as well as having a high model score, could ensure that truly different models (graphs) are available for model averaging purposes. We leave exploration and implementation of these ideas to future work.

Graphical lasso, while an improvement on early lasso regression-based models, performed the worst of the methods we considered, both in terms of precision of edge selection and prediction. Adaptive graphical lasso which allows for a different penalty on each edge shows substantial improvement in terms of sparsity, reduced parameter divergence and prediction accuracy.

FINCS gave good results even though we were working with datasets where the true model is non-decomposable, which were often better than the adaptive graphical lasso model. This is despite the inability of FINCS to choose the correct model. Using the top graph under FINCS gave results on a par with model averaging. While the relative inclusion probabilities may point to a non-decomposable model,



for large samples there was not much to be gained in using the maximum likelihood estimator of the graph based on relative inclusion probabilities  $> 0.8$  and for smaller samples the estimate was often worse.

Some studies suggest that graphical lasso methods can be improved by using methods other than cross validation to select the penalty. Information criteria (AIC, BIC) are used by Ambroise *et al.* (2009) to choose the penalty parameter in graphical lasso and Gao *et al.* (2009) to choose the penalty parameter in adaptive graphical lasso. In each case the penalty chosen yields a sparser graph than the cross validation penalty. However as Liu *et al.* (2010) point out, AIC and BIC tend to perform poorly when the dimension ( $p$ ) is large relative to the sample size ( $n$ ). To overcome this Liu *et al.* (2010) introduce a stability approach (StARS) to choosing the penalty parameter which they claim outperforms cross validation, AIC and BIC, but only for high-dimensional problems. We wanted to use a method for selecting the penalty parameter which could be used for all variations in  $n$  and  $p$ , hence we used cross-validation.

FINCS gave good results even though we were working with datasets where the true model is non-decomposable. These results were often better than the adaptive graphical lasso model. This is despite the inability of FINCS to choose the correct model. Notably, although FINCS often included ‘extra’ edges to make the model decomposable, in moderate to high dimensions it had better precision than the adaptive graphical lasso.

# Chapter 6

## Concluding discussion

In this concluding chapter we summarize our findings, further highlighting the contributions made by this thesis. As with any research, questions arise naturally from our conclusions and so we also make suggestions for further research.

Our focus has been on computationally tractable methods for fitting Gaussian models. We have considered two penalized likelihood approaches, the graphical lasso and the adaptive graphical lasso, and also a decomposable Bayesian approach feature-inclusion stochastic search. An attraction of using penalized likelihood approaches is that they are faster than Bayesian methods, even when the Bayesian search is restricted to decomposable models for computational convenience. Bayesian models on the other hand are sparser and separate model selection from parameter estimation. While the overriding theme to emerge on the relative costs and benefits of the two approaches is that which is ‘best’ depends on the application, we offer some guidelines and quantify the costs and benefits of each approach.

### 6.1 Estimating the inverse covariance matrix

In many situations, as in Chapter 3, our aim in estimating the inverse covariance matrix is to gain insights from the conditional independence pattern and so we focus the first part of our discussion on model selection and estimation of the inverse

covariance matrix ( $\Omega$ ).

Using partial correlations gave demonstrably better results both in terms of the number of genes able to be classified and the accuracy of classification. In terms of using different methods to obtain the estimated partial correlations we found that the results using graphical lasso were on a par with high-dimensional Bayesian covariance selection, an approach which places a less stringent restriction on possible models than FINCS. Given that very small estimated partial correlations will naturally be excluded by being on a long path we would expect that both adaptive graphical lasso and FINCS would give similar results. Although the restriction to decomposable models does lead to significant savings in computational time, FINCS is still more computationally expensive than adaptive graphical lasso even allowing for cross validation.

In Chapter 5 we saw that, when using FINCS, the sample size needed to be at least  $12p$  before a superset is the top model. Thus our work in Chapter 4 suggests that, in the cases where the FINCS restriction to decomposable models results in a superset model being used for parameter estimation, the practical effect of increased variability in the estimates will be small. In this situation, the sparser FINCS derived models perform better simply because they are sparser and therefore closer to the true model. Not answered here, however, is whether the  $12p$  ‘superset threshold’ is actually  $12p$  or whether when the true model has more than one prime component it may actually be smaller. It is possible that this ‘superset threshold’ is only related to the relative size of the number of variables in the largest non-decomposable prime component and the sample. We leave verification of that to future research.

The situation is not so clear cut when the sample size is smaller. In Chapter 4 we saw that there is always increased variability if a superset model is used and that, with small samples, this variability can be substantial. However in Chapter 5 we found that only the penalized likelihood methods select a superset model when the sample size is small. Furthermore the approach which yielded the model with the smallest Kullback-Leibler divergence varied between samples. Thus the decision as to which method to use will vary according to context. If missing true edges is of

higher concern than adaptive graphical lasso is more likely to include all the true edges plus a minimum of extras. If reducing the number of incorrect edges is more important than the top FINCS graph has fewer incorrect edges while missing the least true ones.

Finally in this section we comment on small partial correlations. In Chapter 5 we saw that edges corresponding to small partial correlations are often missed by all methods, no matter what the sample size. If they are picked up, particularly if the true partial correlations all tend to be small, then our work in Chapter 4 suggests that, the presence of incorrect edges in the model is unlikely to have much affect on the variability of the estimates.

## 6.2 Regularizing the estimate of the covariance matrix

In other situations we have an interest in obtaining a ‘good’ estimate of the inverse covariance matrix is so that we can obtain a regularized estimate of the covariance matrix. We saw in Chapter 4 that the presence of incorrect edges in the inverse covariance matrix graph can markedly increase the variance of some elements of the covariance matrix. The greater the difference in sparsity between the true model and the estimate of the inverse covariance matrix, the greater this increase in the variance of some elements of the covariance matrix. In Chapter 5 we used a regularized estimate of  $\Sigma$  for prediction purposes. We were therefore not surprised to see that graphical lasso, which always included the highest number of incorrect edges, gave the least accurate predictions.

As observed in Chapter 5 the difference between adaptive graphical lasso and estimation derived from the FINCS top graph is not so clear cut. The best estimates will be obtained by using the sparser FINCS top graph in large sample situations when all approaches point to a superset model. We found that when a smaller sample meant that only the adaptive graphical lasso models were a superset, and also when the presence of small partial correlation in  $\Omega$  meant that both methods missed

identifying these elements as non-zero, the errors were quite similar. The approach which gave the smallest errors varied between samples. The increased variability in estimates of elements of  $\Sigma$  corresponding to extra edges does not appear to be reflected in increased variability in the total sum of squared errors. Whether this would show up with more replications, or whether missed edges are also affecting the variability, is something we leave to future research.

Finally we note again here that the adaptive graphical lasso and top FINCS graphs for the mutual-funds dataset (see Section 5.4.3) were very different although containing the same number of edges. Furthermore in this more complex situation FINCS clearly outperformed adaptive graphical lasso.

### 6.3 Concluding remarks

Gaussian graphical models are a useful tool for eliciting and understanding relationships in high-dimensional data. In Chapter 3 we saw that using partial correlations enables both more, and more accurate, classifications of genes than using a correlation graph. The question then is, what method should one use to select the model and estimate the parameters. We have considered the latest methods for two common approaches: penalized likelihood and decomposable Bayesian. If time is of the essence, then penalized likelihood approaches are faster. In all cases adaptive graphical lasso gave superior results to graphical lasso. While the decision to restrict to decomposable models is made purely for computational convenience our research suggests that in most high-dimensional settings the results will be comparable with or even better than those obtained using adaptive graphical lasso. This may not be the case at low-dimensions, particularly with small samples as is discussed above.

As with all research in the process of answering our initial questions more have arisen. Model averaging gives superior results in most Bayesian situations. This was not the case here. As noted earlier we suggest including consideration of the Kullback -Leibler divergence in the search algorithm as one way to ensure the top

graphs actually represent different models. More investigation could also be done to confirm that our results do hold at even higher-dimensions and for more complex graphs with many components.



# Appendix A

## Supplementary Tables and Figures for Chapter 4

### A.1 Tables and Figures for the four variable case

Matrices with small partial correlations were obtained in the following manner:

- $small\Omega_{same}$  was obtained by dividing all off diagonals elements of  $\Omega_{same}$  by 100;
- $small\Omega_{big}$  was obtained by dividing all off diagonals elements of  $\Omega_{big}$  by 10;
- $small\Omega_{small}$  was obtained by dividing all off diagonals elements of  $\Omega_{small}$  by 1000.



Table A.1:  $n = 1000$  Relative Standard Deviations for non-decomposable model and increase for decomposable model.

	$small\Omega_{same}$		$small\Omega_{big}$		$small\Omega_{small}$	
	RSD	cycle increase	RSD	cycle increase	RSD	cycle increase
$\omega_{1,2}$	7.008	0.001	0.624	0.006	155.80	0.033
$\omega_{2,3}$	6.822	0.001	0.410	0.008	1241.92	0.024
$\omega_{3,4}$	7.100	0.002	0.523	0.131	596.98	0.049
$\omega_{1,4}$	7.127	0.001	0.459	0.041	258.27	0.067

Table A.2: Four variable empirical and EFI variances for elements of  $\widehat{\Omega}_{same}$  for cycle and decomposable, when  $n=10$ .

	cycle		decomposable	
	EFI variance	Empirical variance	EFI variance	Empirical variance
$\hat{\omega}_{1,2}$	36.57	178.96	40	329.81
$\hat{\omega}_{2,3}$	36.57	291.92	40	495.78
$\hat{\omega}_{3,4}$	36.57	272.89	40	446.35
$\hat{\omega}_{1,4}$	36.57	168.57	40	320.03

Table A.3: Four variable empirical and EFI variances for elements of  $\widehat{\Omega}_{big}$  for cycle and decomposable, when  $n=10$ .

	cycle		decomposable	
	EFI variance	Empirical variance	EFI variance	Empirical variance
$\hat{\omega}_{1,2}$	50.12	221.23	88.72	632.27
$\hat{\omega}_{2,3}$	107.66	635.85	183.21	1965.61
$\hat{\omega}_{3,4}$	564.12	2681.58	978.41	11173.80
$\hat{\omega}_{1,4}$	699.18	3344.64	1124.15	7681.47

Table A.4: Four variable empirical and EFI variances for elements of  $\hat{\Omega}_{small}$  for cycle and decomposable, when  $n=10$ .

	cycle		decomposable	
	EFI variance	Empirical variance	EFI variance	Empirical variance
$\hat{\omega}_{1,2}$	1001.22	4664.80	1016.20	7930.63
$\hat{\omega}_{2,3}$	626.80	3305.47	628.48	5416.36
$\hat{\omega}_{3,4}$	1240.26	5267.63	1240.92	9185.29
$\hat{\omega}_{1,4}$	1821.72	7789.67	1897.93	14514.84

Table A.5: Percentage of times an element is declared zero ( $|\text{estimate}| < 2 \times \text{standard error}$ ) for 1000 simulations ( $\Omega_{small}$ ).

	$n=10$		$n=100$		$n=1000$	
	cycle	decomp	cycle	decomp	cycle	decomp
$\hat{\omega}_{1,2}$	99.1	98.9	49.7	50.1	0	0
$\hat{\omega}_{1,4}$	99.1	99.1	95.7	95.5	89.6	89.7
$\hat{\omega}_{2,3}$	99.9	99.7	94.9	94.9	60.4	60.9
$\hat{\omega}_{2,4}$		98.3		95.4		94.4
$\hat{\omega}_{3,4}$	99.2	99.1	76.8	77.3	2.6	2.9

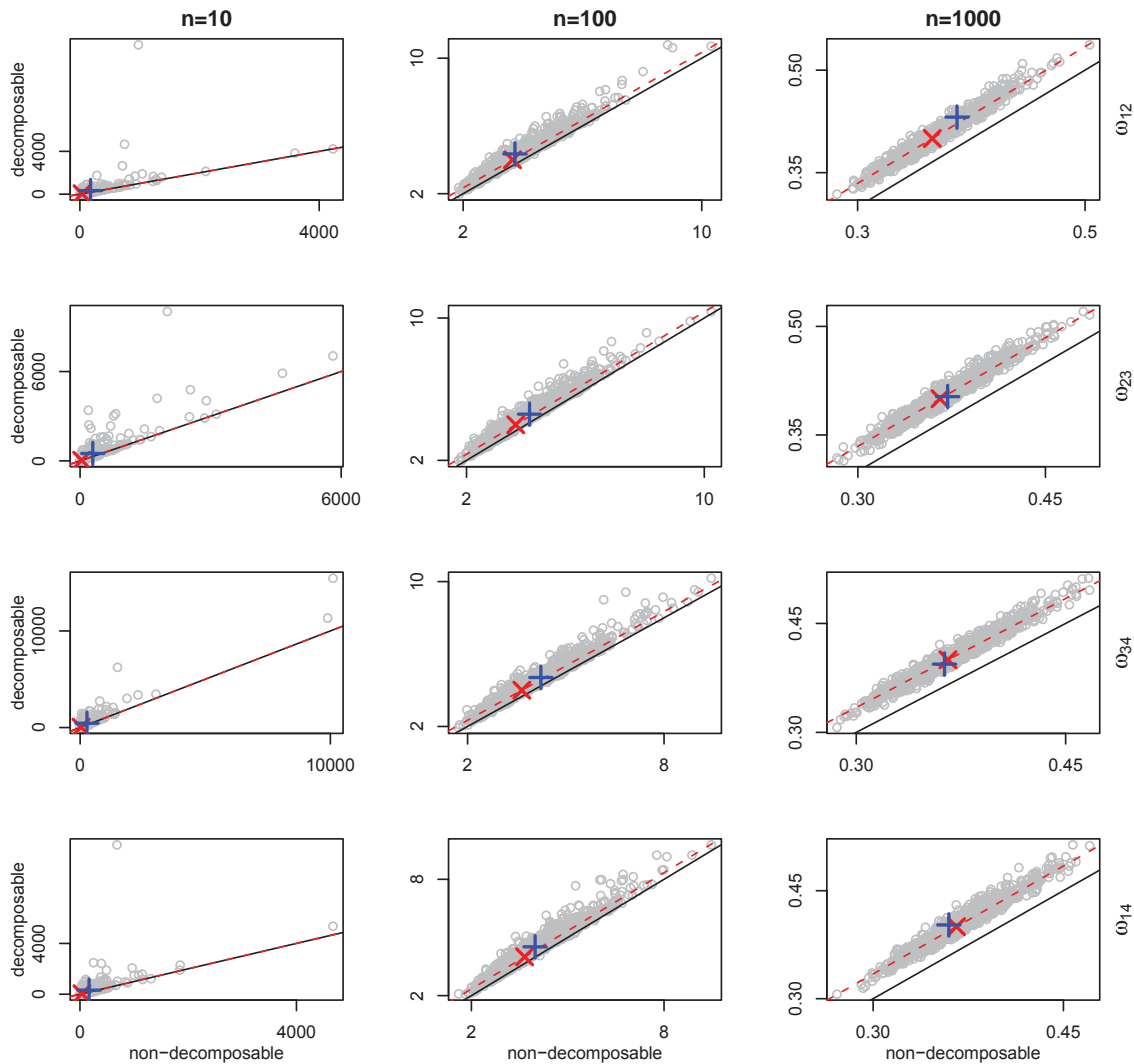


Figure A.1: OFI variances for  $\widehat{\Omega}_{same}$  with sample sizes 10, 100 and 1000. (Note the radically different scales.)

— represents the line  $y = x$ ;

- - - represents the line  $y = x + \text{difference in EFI variances calculated using equation(5)}$

× represents the expected variances;

+ represents the empirical variances.

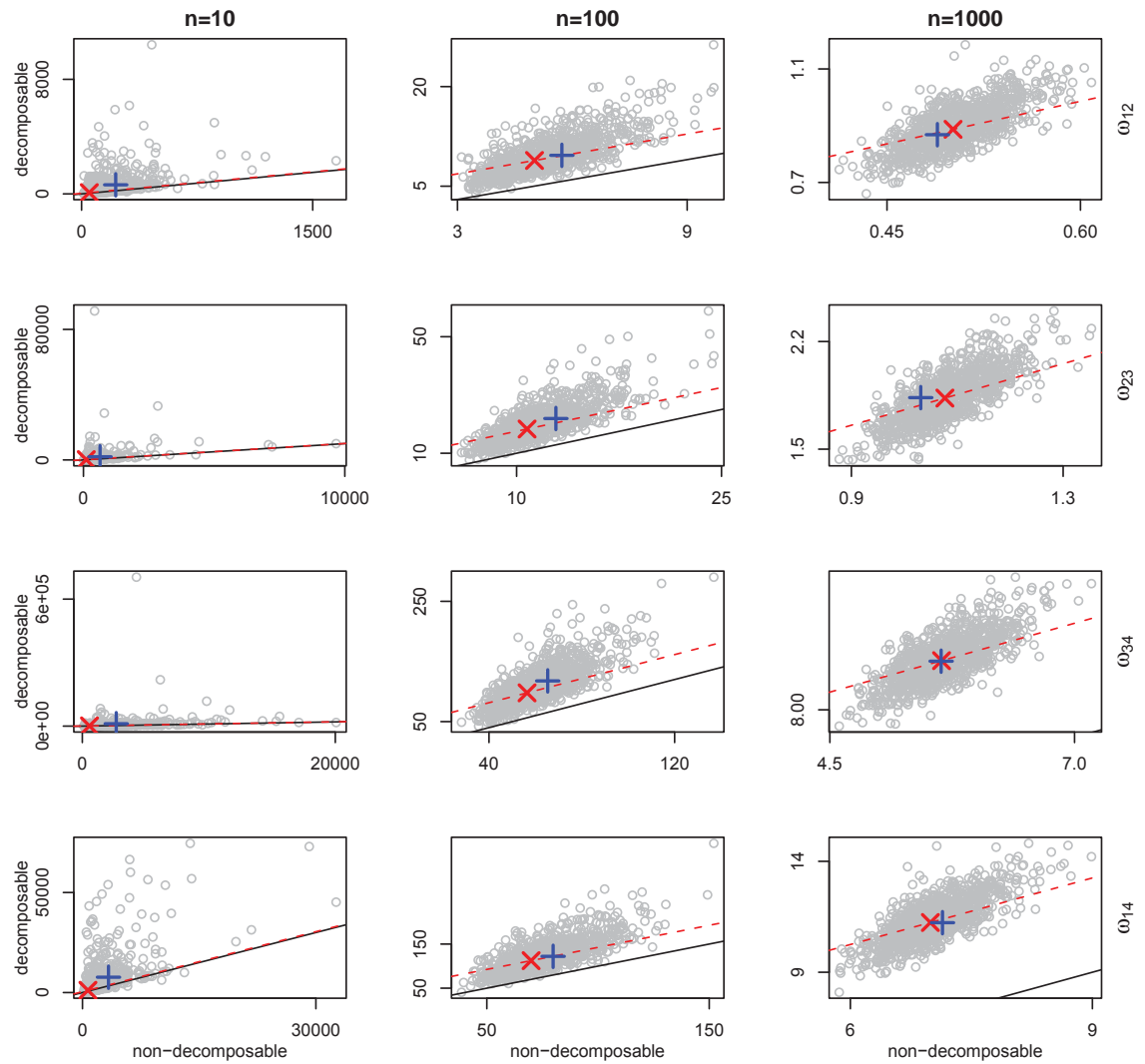


Figure A.2: OFI variances for  $\widehat{\Omega}_{big}$  with sample sizes 10, 100 and 1000. (Note the radically different scales.)

— represents the line  $y = x$ ;

- - - represents the line  $y = x + \text{difference in EFI variances calculated using equation(5)}$

× represents the expected variances;

+ represents the empirical variances.

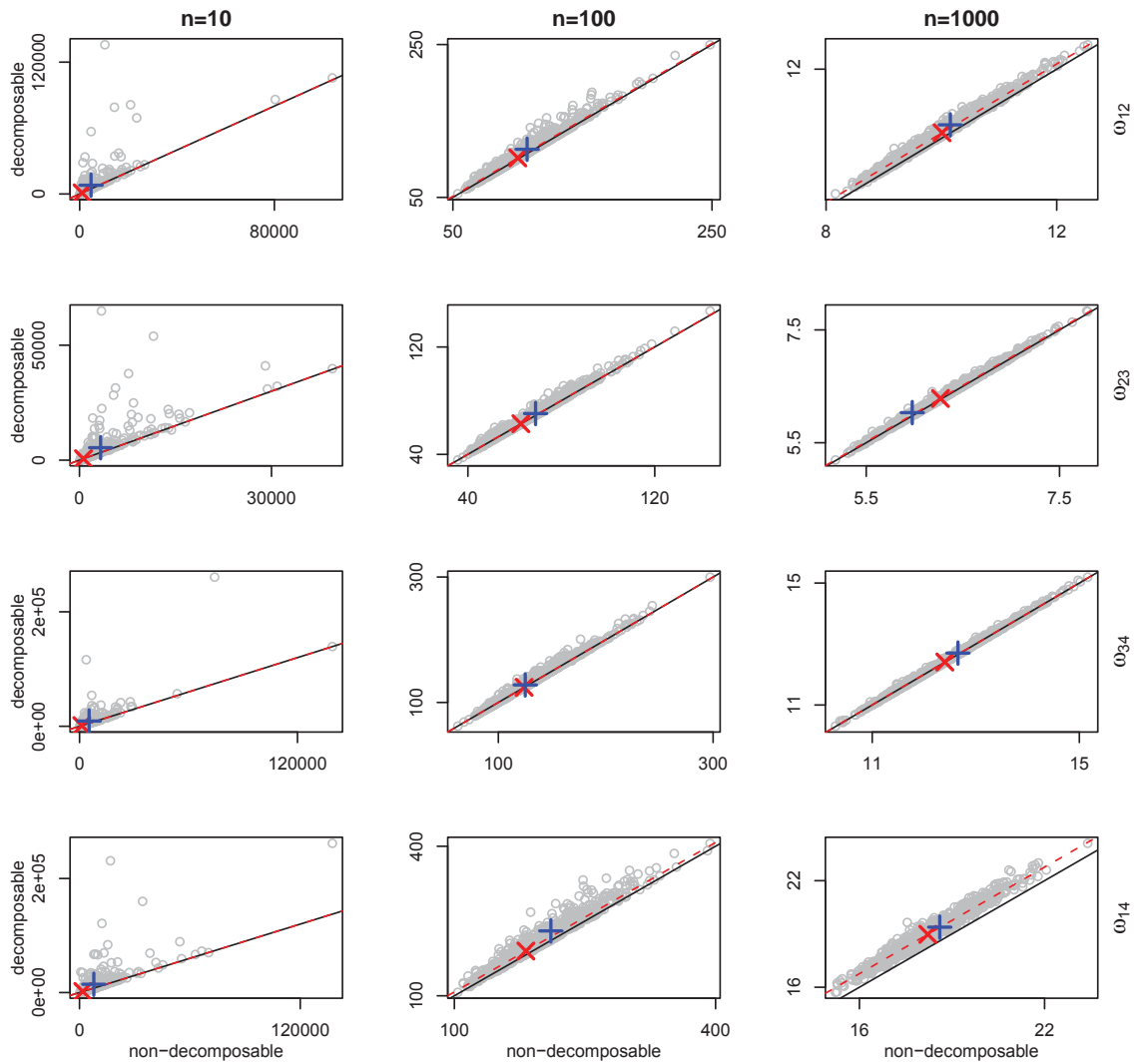


Figure A.3: OFI variances for  $\widehat{\Omega}_{small}$  with sample sizes 10, 100 and 1000. (Note the radically different scales)

— represents the line  $y = x$ ;

- - - represents the line  $y = x + \text{difference in EFI variances calculated using equation(5)}$

× represents the expected variances;

+ represents the empirical variances.

## **A.2 Tables and Figures for 20 and 50 variable cases**

Table A.6: Non-zero elements of  $\Omega$  matrices for  $p=20$ .

i	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$
1	20	-9	36	15	52	-20
2	20	-9	15	-7	52	-12
3	20	-9	50	-25	40	-10
4	20	-9	79	-23	69	2
5	20	-9	55	-65	55	-14
6	20	-9	209	-46	110	-34
7	20	-9	127	-132	153	13
8	20	-9	230	-7	36	-15
9	20	-9	90	-91	160	41
10	20	-9	169	-41	54	-19
11	20	-9	101	-63	29	-9
12	20	-9	84	-4	149	60
13	20	-9	49	-15	134	-44
14	20	-9	65	-22	132	15
15	20	-9	46	-10	61	-2
16	20	-9	69	-35	132	33
17	20	-9	38	-37	105	-1
18	20	-9	121	-4	39	-3
19	20	-9	14	-8	51	-3
20	20	-9*	125	-46*	168	27*

\* is element  $\omega_{20,1}$

Table A.7: Non-zero elements of  $\Omega$  matrices for  $p=50$ .

i	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$
1	20	-9	80	-103	22	-9
2	20	-9	206	-13	93	-5
3	20	-9	64	-19	82	-11
4	20	-9	30	-17	62	-1
5	20	-9	65	-88	113	-4
6	20	-9	197	-69	118	18
7	20	-9	119	-9	101	47
8	20	-9	42	15	113	-57
9	20	-9	28	-15	73	24
10	20	-9	29	-10	65	13
11	20	-9	8	-11	131	-20
12	20	-9	127	-76	98	-5
13	20	-9	138	-24	47	-27
14	20	-9	202	-200	53	-23
15	20	-9	240	-25	121	40
16	20	-9	40	15	63	8
17	20	-9	20	-5	132	5
18	20	-9	22	-12	138	-19
19	20	-9	44	-27	84	47
20	20	-9	63	-38	98	-13
21	20	-9	140	-119	11	3
22	20	-9	224	-56	89	5
23	20	-9	137	-105	28	-16
24	20	-9	123	-9	128	-41
25	20	-9	89	-30	86	-7
26	20	-9	117	-86	19	-13
27	20	-9	113	-77	164	22

*Continued on next page*



i	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$	$\omega_{i,i}$	$\omega_{i,(i+1)}$
28	20	-9	200	28	132	-58
29	20	-9	52	-24	88	30
30	20	-9	18	-11	68	-20
31	20	-9	45	13	85	10
32	20	-9	68	-37	151	7
33	20	-9	96	26	77	10
34	20	-9	60	-67	123	-24
35	20	-9	174	-68	134	18
36	20	-9	126	-19	35	-36
37	20	-9	160	-144	79	25
38	20	-9	223	-131	84	-19
39	20	-9	229	-42	70	41
40	20	-9	124	-33	91	-5
41	20	-9	31	-4	79	-34
42	20	-9	72	-29	313	-66
43	20	-9	144	-115	261	10
44	20	-9	115	-10	24	-3
45	20	-9	26	-5	86	33
46	20	-9	33	10	95	-69
47	20	-9	36	-7	381	-8
48	20	-9	30	12	25	-5
49	20	-9	27	-18	82	-46
50	20	-9*	47	-20*	207	-1*

\* is element  $\omega_{50,1}$

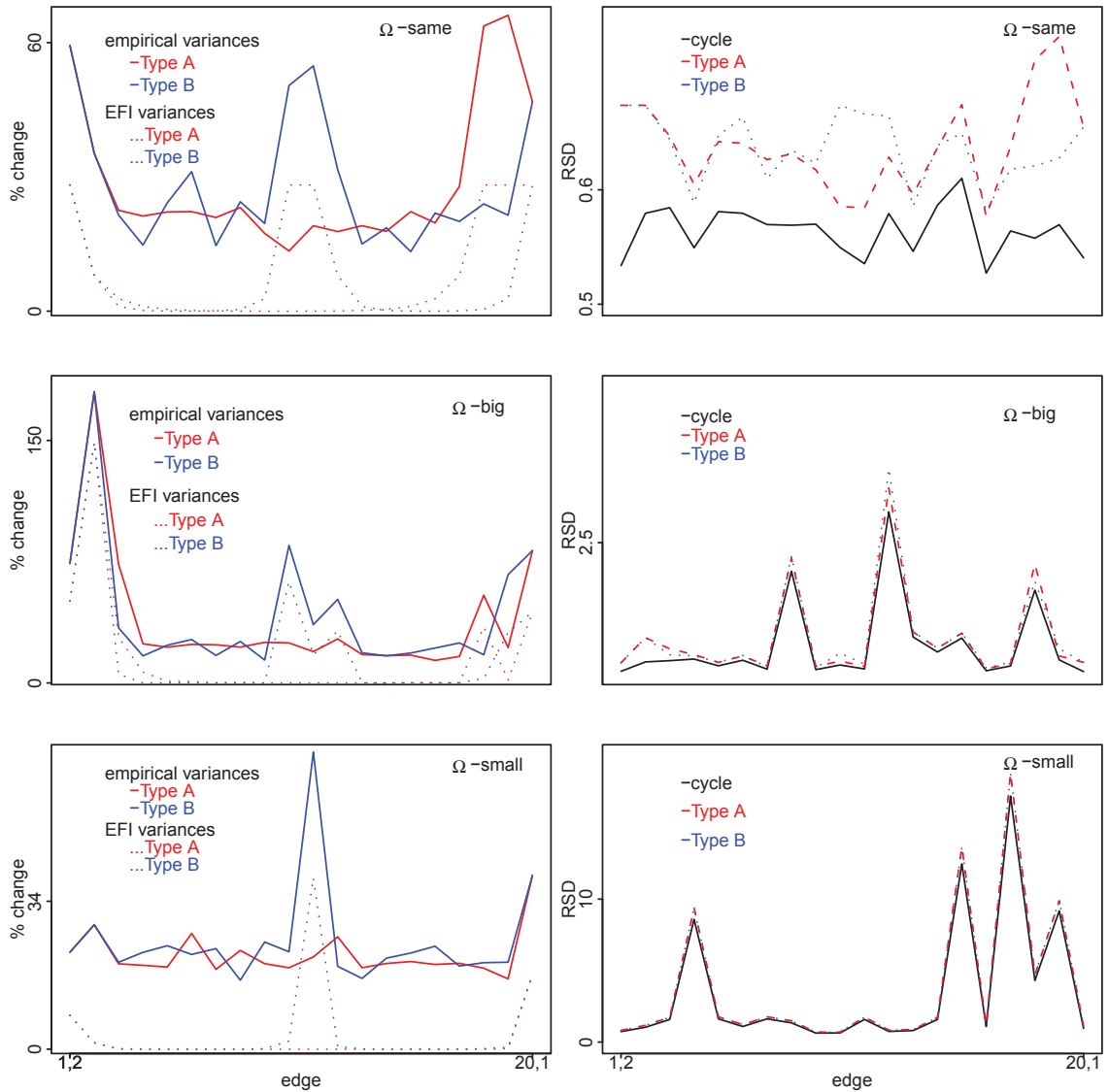


Figure A.4: Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted and relative standard deviation (RSD) when fitting true (cycle) and decomposable models.

Shown here for  $p=20$ ,  $n=21$  three different  $\Omega$  matrices, and two decomposable models. Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

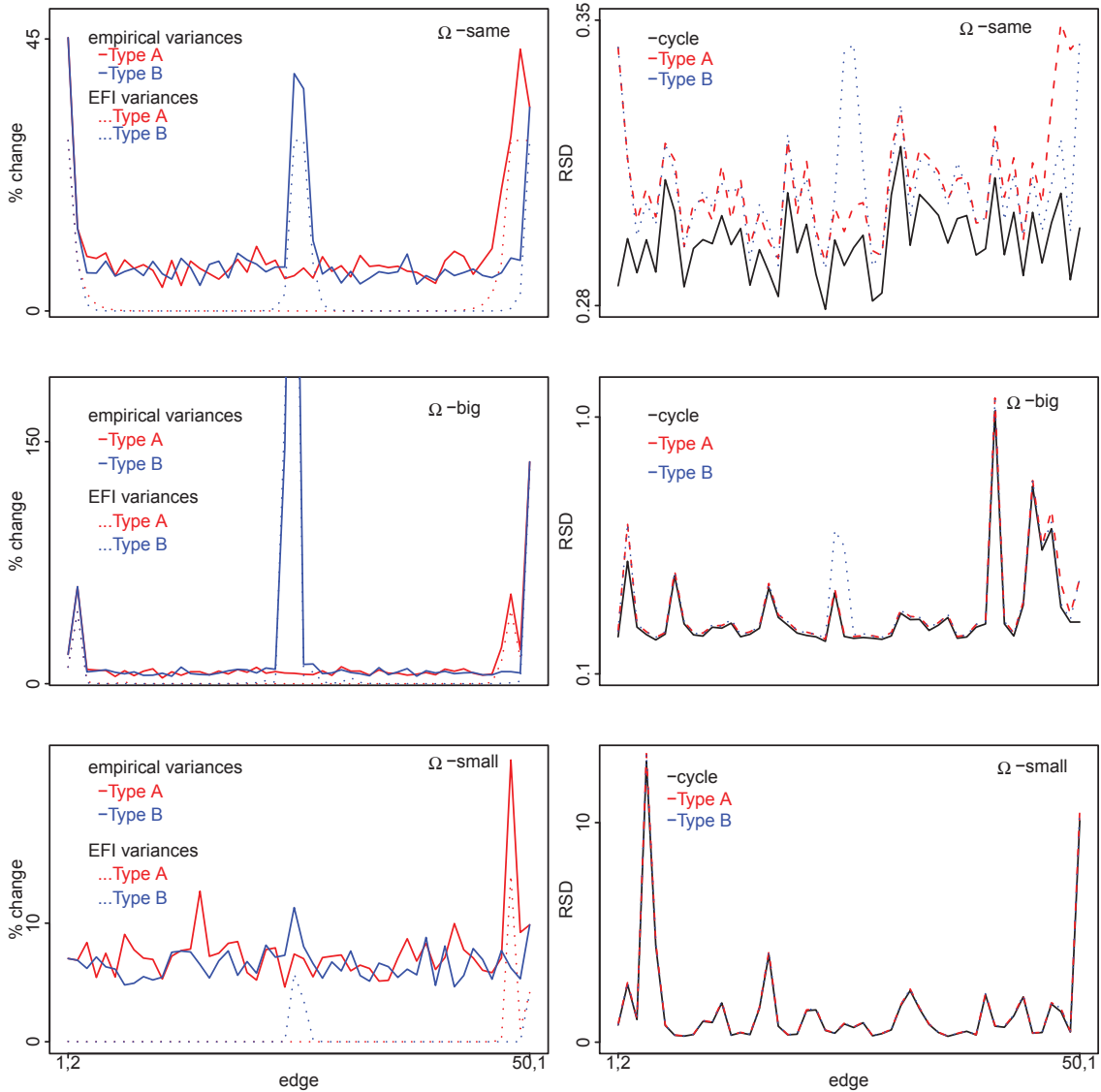


Figure A.5: Percentage change in expected (EFI) and empirical variances when a decomposable model is fitted and relative standard deviation (RSD) when fitting true and decomposable models.

Shown here for  $p=50$ ,  $n=51$  three different  $\Omega$  matrices, and two decomposable models. Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

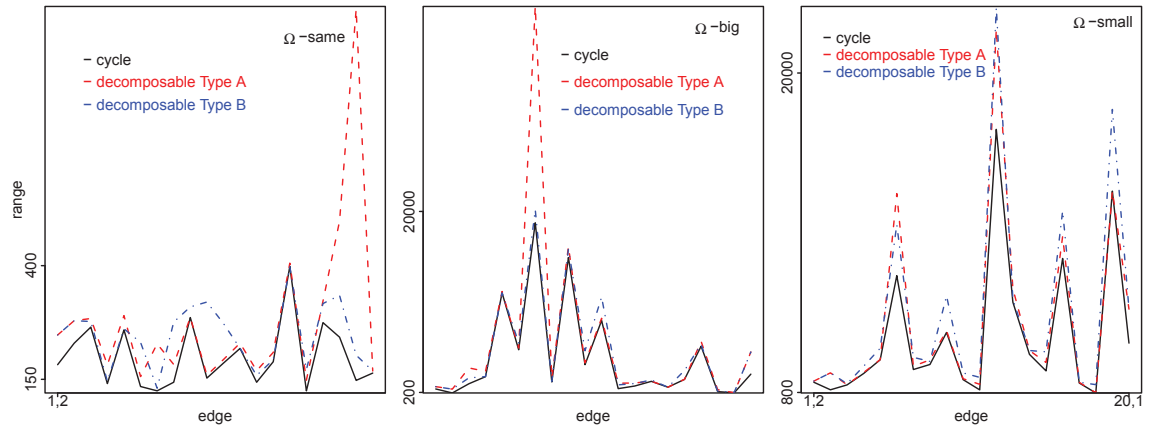


Figure A.6: Range of OFI variances when a cycle and two different decomposable models are fitted.

Shown here for  $p=20$ ,  $n=21$ . Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

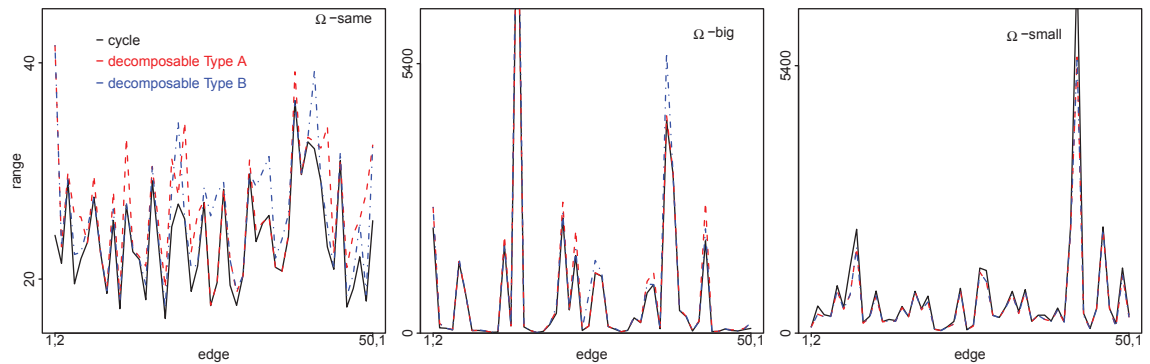


Figure A.7: Range of OFI variances when a cycle and two different decomposable models are fitted.

Shown here for  $p=50$ ,  $n=51$ . Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

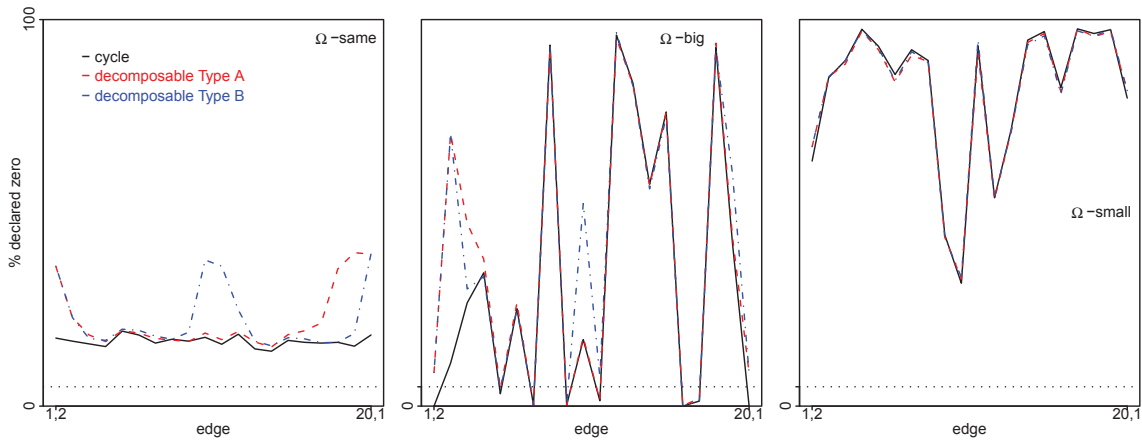


Figure A.8: Percentage of true non-zero elements declared zero when a cycle and two different decomposable models are fitted.

Shown here for  $p=20$ ,  $n=21$ . Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

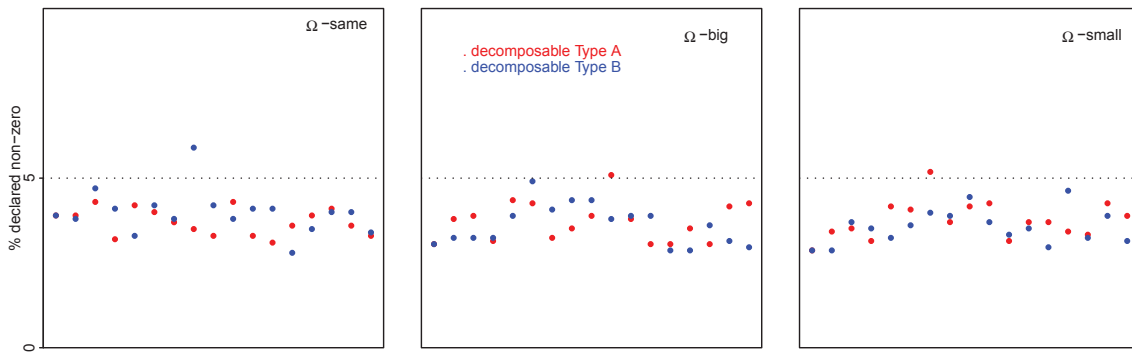


Figure A.9: Percentage of elements corresponding to 'extra edges' which are declared non-zero. Shown for two different decomposable models, for  $p=20$ ,  $n=21$ .

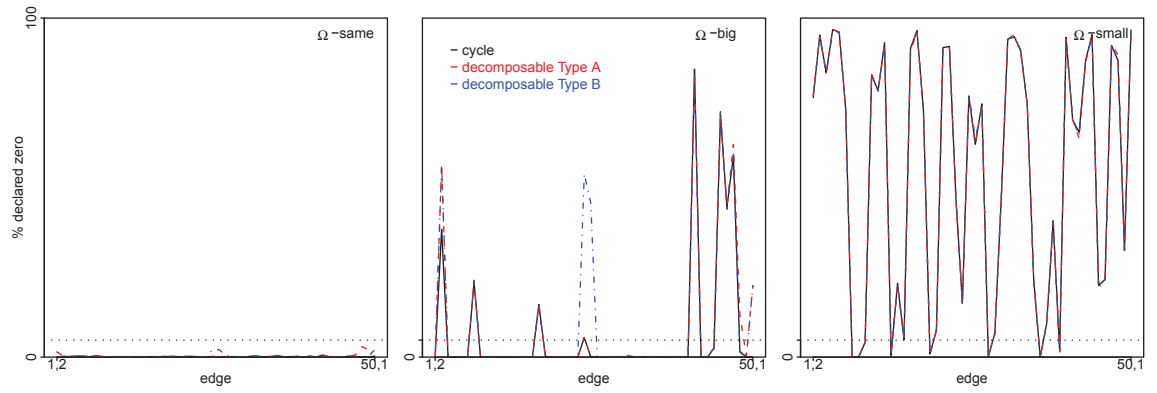


Figure A.10: Percentage of true non-zero elements declared zero when a cycle and two different decomposable models are fitted.

Shown here for  $p=50$ ,  $n=51$ . Edges are labeled in an anticlockwise direction beginning with the edge corresponding to  $\omega_{1,2}$ .

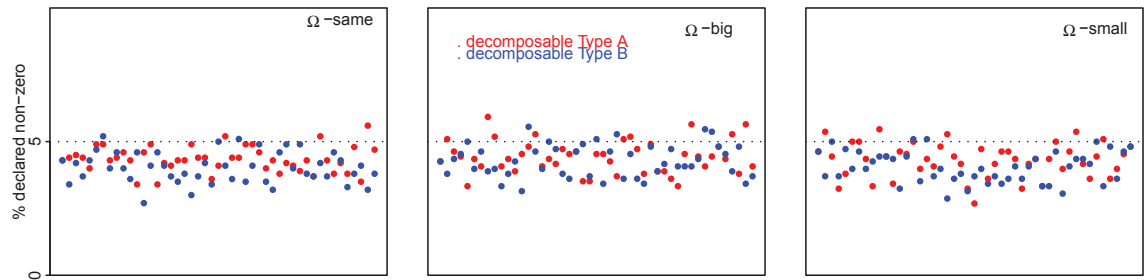


Figure A.11: Percentage of elements corresponding to ‘extra edges’ which are declared non-zero. Shown for two different decomposable models, for  $p=50$ ,  $n=51$ .

### A.3 Tables and Figures for estimations of $\Sigma$

Table A.8: Four variable empirical variances for elements of  $\hat{\Sigma}$  for cycle and percentage increase to decomposable, when  $n=10$ .

	$\Omega_{same}$		$\Omega_{big}$		$\Omega_{small}$	
	variance	%increase	variance	%increase	variance	%increase
$\hat{\sigma}_{1,1}$	0.005236	<1%	0.000921	<1%	0.0000310	<1%
$\hat{\sigma}_{1,2}$	0.004146	<1%	0.003202	<1%	0.0000127	<1%
$\hat{\sigma}_{1,3}$	0.003697	<1%	0.001628	<1%	0.0000048	18.75%
$\hat{\sigma}_{1,4}$	0.004150	<1%	0.000773	<1%	0.0000059	<1%
$\hat{\sigma}_{2,2}$	0.005422	<1%	0.045355	<1%	0.0000200	<1%
$\hat{\sigma}_{2,3}$	0.004191	<1%	0.017674	<1%	0.0000188	<1%
$\hat{\sigma}_{2,4}$	0.003636	6.57%	0.004875	1.17%	0.0000012	300%
$\hat{\sigma}_{3,3}$	0.005619	<1%	0.008990	<1%	0.0000753	<1%
$\hat{\sigma}_{3,4}$	0.004453	<1%	0.002844	<1%	0.0000085	<1%
$\hat{\sigma}_{4,4}$	0.005407	<1 %	0.001195	<1%	0.0000048	<1%

Note: that the parameter values are themselves small hence the small variances.

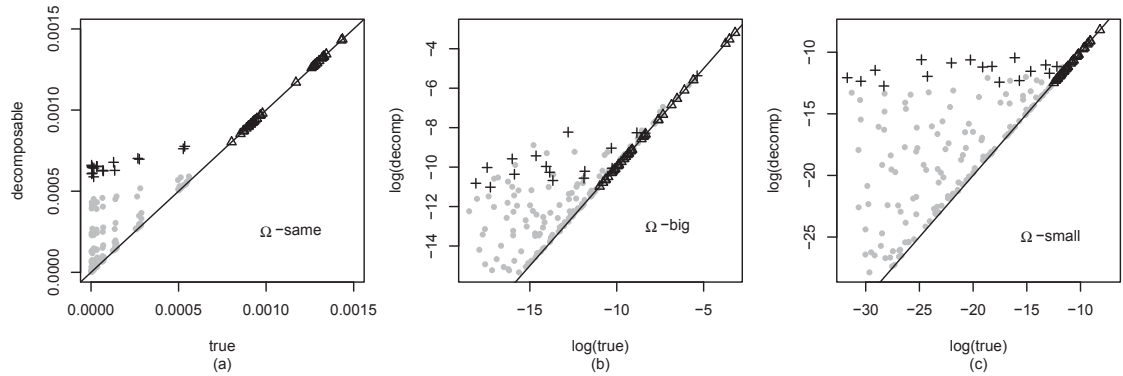


Figure A.12: Empirical variances for elements of  $\Sigma$  when a decomposable model (type B) vs when the true (non-decomposable) model is fitted.

Shown here for  $p=20$  and  $n=21$ , when underlying  $\Omega$  matrix has (a) all partial correlations equal, (b) large partial correlations and (c) small partial correlations. Note that log scales are used for (b) and (c).

$\Delta$ =elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$

$+$  = elements of  $\Sigma$  corresponding to elements only non-zero in the decomposable estimate of  $\Omega$

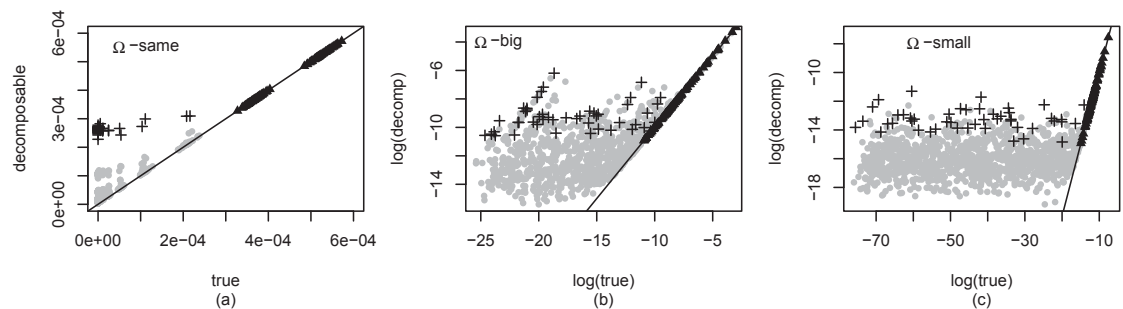


Figure A.13: Empirical variances for elements of  $\Sigma$  when a decomposable model (type A) vs when the true (non-decomposable) model is fitted.

Shown here for  $p=50$  and  $n=51$ , when underlying  $\Omega$  matrix has (a) all partial correlations equal, (b) large partial correlations and (c) small partial correlations. Note that log scales are used for (b) and (c).

$\Delta$ =elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$

$+$  = elements of  $\Sigma$  corresponding to elements only non-zero in the decomposable estimate of  $\Omega$



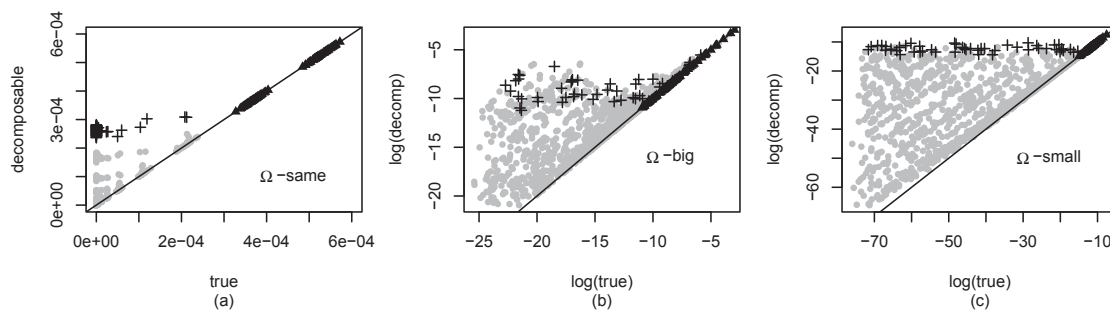


Figure A.14: Empirical variances for elements of  $\Sigma$  when a decomposable model (type B) vs when the true (non-decomposable) model is fitted.

Shown here for  $p=50$  and  $n=51$ , when underlying  $\Omega$  matrix has (a) all partial correlations equal, (b) large partial correlations and (c) small partial correlations. Note that log scales are used for (b) and (c).

$\Delta$  = elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$

$+$  = elements of  $\Sigma$  corresponding to elements only non-zero in the decomposable estimate of  $\Omega$

## A.4 Tables and Figures for 12 node example

Table A.9: Estimates and OFI standard deviations for elements of  $\hat{\Omega}$  for cycle and three decomposable models, when  $p=12$  and  $n=250$ .

	cycle(sd)	decomp A(sd)	decomp B(sd)	decomC(sd)
$\hat{\omega}_{1,1}$	20.42(1.7149)	21.11(1.8883)	20.58(1.7284)	20.97(1.7532)
$\hat{\omega}_{1,2}$	-9.74(1.2125)	-10.33(1.3594)	-9.79(1.2144)	-9.79(1.2144)
$\hat{\omega}_{1,3}$	0	0	-0.65(1.0007)	-0.63(1.0135)
$\hat{\omega}_{1,5}$	0	0	0	-0.46(0.7277)
$\hat{\omega}_{1,8}$	-7.18(0.9538)	-7.73(1.1102)	-7.23(0.9558)	-8.63(1.1796)
$\hat{\omega}_{1,9}$	0	0	0	2.41(1.0827)
$\hat{\omega}_{2,2}$	19.91(1.6967)	20.18(1.7135)	20.07(1.7091)	20.07(1.7091)
$\hat{\omega}_{2,3}$	-6.89(1.0284)	-7.87(1.1630)	-7.87(1.1630)	-7.87(1.1630)
$\hat{\omega}_{2,4}$	0	1.81(0.9656)	2.16(1.0799)	2.16(1.0799)
$\hat{\omega}_{2,8}$	0	0.79(0.8954)	0	0
$\hat{\omega}_{3,3}$	18.55(1.6047)	19.54(1.7478)	19.54(1.7478)	19.54(1.7478)
$\hat{\omega}_{3,4}$	-7.13(1.0715)	-8.16(1.2146)	-8.16(1.2146)	-8.16(1.2146)
$\hat{\omega}_{4,4}$	20.99(1.7953)	21.37(1.8204)	21.38(1.8209)	21.28(1.8163)
$\hat{\omega}_{4,5}$	-7.35(0.9466)	-7.41(0.9480)	-7.41(0.9480)	-7.40(0.9483)
$\hat{\omega}_{4,8}$	0	-0.70(0.7970)	-0.44(0.8745)	0
$\hat{\omega}_{5,5}$	16.79(1.2995)	16.99(1.3096)	16.99(1.3096)	17.06(1.3122)
$\hat{\omega}_{5,6}$	-7.26(0.9053)	-7.31(0.9700)	-7.31(0.9700)	-7.31(0.9700)
$\hat{\omega}_{5,7}$	0	-0.87(0.8739)	-0.87(0.8739)	-0.87(0.8739)
$\hat{\omega}_{5,8}$	0	1.71(0.6993)	1.71(0.6993)	1.62(0.7828)
$\hat{\omega}_{5,9}$	0	0	0	0.59(0.7937)
$\hat{\omega}_{5,10}$	0	-0.56(0.7022)	-0.56(0.7022)	-0.85(0.7715)
$\hat{\omega}_{5,11}$	0	-0.57(0.8576)	-0.57(0.8576)	-0.57(0.8576)
$\hat{\omega}_{5,12}$	-7.22(0.8989)	-6.86(0.9943)	-6.86(0.9943)	-6.86(0.9943)

*Continued on next page*

	cycle(sd)	decomp A(sd)	decomp B(sd)	decompC(sd)
$\hat{\omega}_{6;6}$	17.82(1.5093)	17.89(1.6001)	17.89(1.6001)	17.89(1.6001)
$\hat{\omega}_{6;7}$	-6.73(0.9594)	-6.81(1.0973)	-6.81(1.0973)	-6.81(1.0973)
$\hat{\omega}_{7;7}$	19.21(1.6343)	20.07(1.7261)	20.07(1.7261)	20.07(1.7261)
$\hat{\omega}_{7;8}$	-7.84(1.0176)	-8.50(1.0748)	-8.50(1.0748)	-8.50(1.0748)
$\hat{\omega}_{8;8}$	20.53(1.6055)	20.86(1.6217)	20.79(1.6191)	22.11(1.8075)
$\hat{\omega}_{8;9}$	-9.31(1.1496)	-9.99(1.2793)	-9.99(1.2793)	-10.81(1.3508)
$\hat{\omega}_{8;10}$	0	1.01(0.9138)	1.01(0.9138)	0
$\hat{\omega}_{9;9}$	22.72(1.9179)	23.72(2.1213)	23.72(2.1213)	23.26(1.9558)
$\hat{\omega}_{9;10}$	-8.12(1.1291)	-9.03(1.3396)	-9.03(1.3396)	-8.47(1.1576)
$\hat{\omega}_{10;10}$	19.15(1.6405)	19.43(1.6606)	19.43(1.6606)	19.58(1.6847)
$\hat{\omega}_{10;11}$	-7.01(1.0032)	-6.91(1.0465)	-6.91(1.0465)	-6.91(1.0465)
$\hat{\omega}_{11;11}$	19.12(1.6172)	19.09(1.6454)	19.09(1.6454)	19.09(1.6454)
$\hat{\omega}_{11;12}$	-9.49(1.1568)	-9.06(1.2621)	-9.06(1.2621)	-9.06(1.2621)
$\hat{\omega}_{12;12}$	20.37(1.6878)	19.84(1.7744)	19.84(1.7744)	19.84(1.7744)

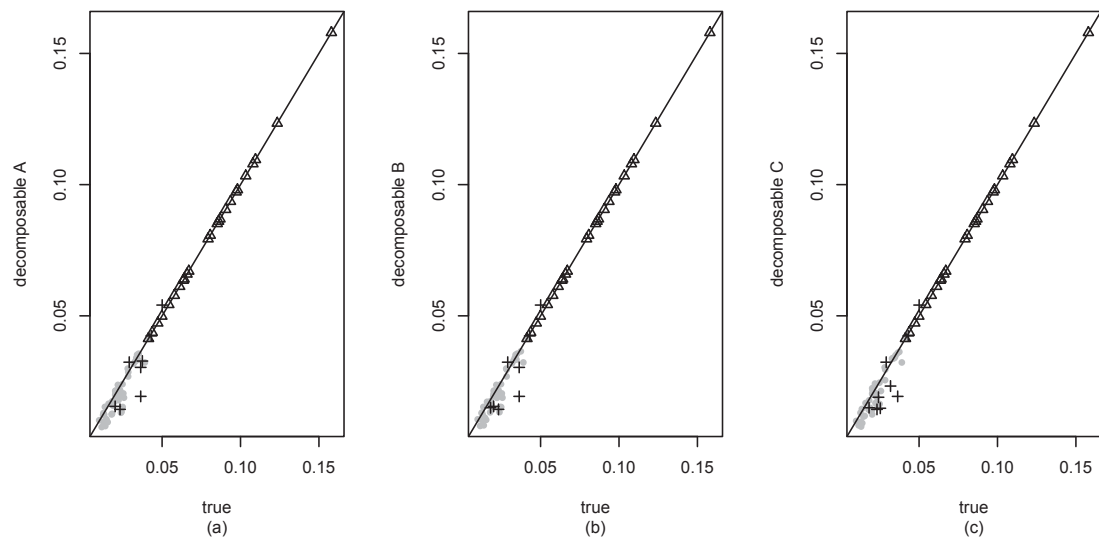


Figure A.15: Elements of  $\Sigma$  when a decomposable model vs when the true (non-decomposable) model is fitted.

Shown here for decomposable models (a) A, (b) B and (c) C.

$\Delta$  = elements of  $\Sigma$  corresponding to non-zero elements in  $\Omega$

$+$  = elements of  $\Sigma$  corresponding to elements only non-zero in the decomposable estimate of  $\Omega$



# Appendix B

## Supplementary Tables and Figures for Chapter 5

### B.1 $\Omega$ matrices

Table B.1:  $\Omega_{twelve}$

$$\begin{pmatrix} 63.7 & 1.1 & 0 & 0 & 0 & 0 & 0 & 7.6 & 0 & 0 & 0 & 0 \\ 1.1 & 247.8 & 150.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 150.9 & 188.3 & 23.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 23.2 & 85.2 & -9.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -9.0 & 197.2 & -99.9 & 0 & 0 & 0 & 0 & 0 & -43.4 \\ 0 & 0 & 0 & 0 & -99.9 & 119.8 & -37.4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -37.4 & 137.1 & 41.0 & 0 & 0 & 0 & 0 \\ 7.6 & 0 & 0 & 0 & 0 & 0 & 41.0 & 144.1 & 42.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 42.8 & 188.0 & -89.9 & 0 & 70 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -89.9 & 240.5 & -9.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -9.3 & 234.4 & -54.6 \\ 0 & 0 & 0 & 0 & -43.4 & 0 & 0 & 0 & 0 & 0 & -54.6 & 59.89 \end{pmatrix}$$

## B.2 Graphs and Tables for Section 5.4.1

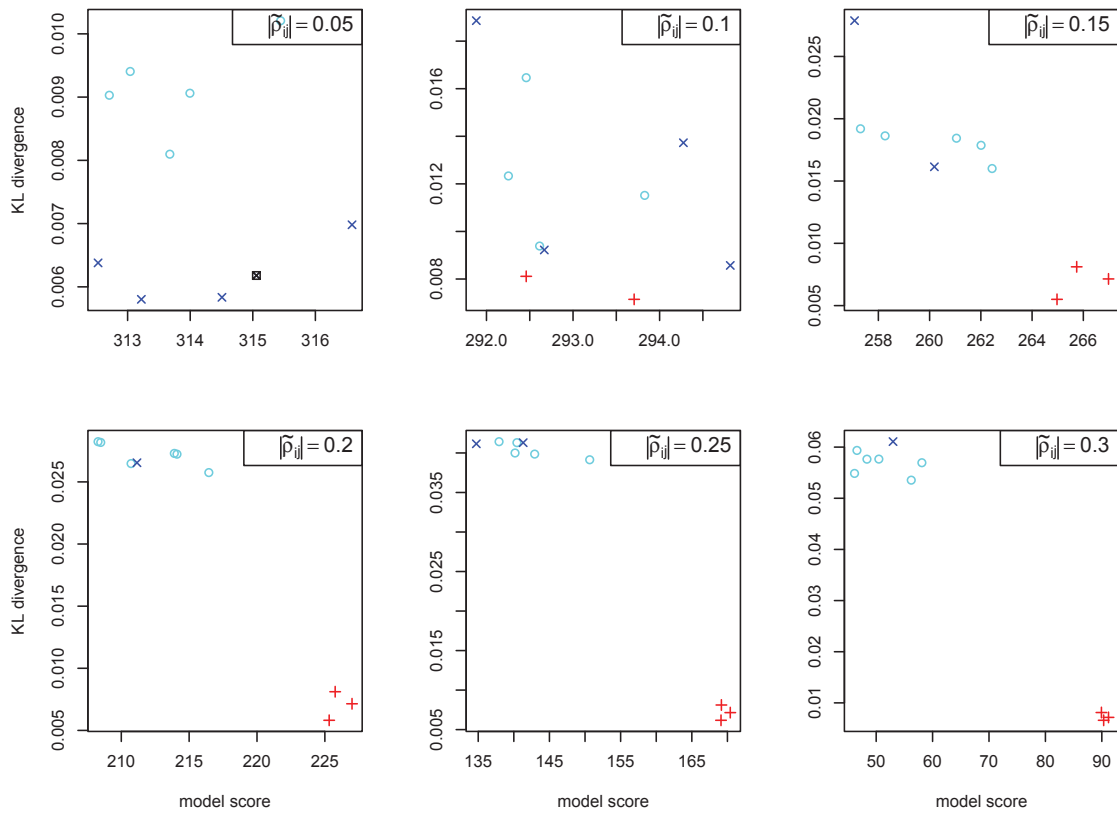


Figure B.1: Top 10 graphs found by FINCS for  $p=4$ , samples of  $n=1000$  for different  $|\tilde{\rho}_{ij}|$ .

+ = superset graphs; × = subset graphs; o = subset plus incorrect edges;

⊠ = empty graph

*note there are only 3 possible superset graphs*

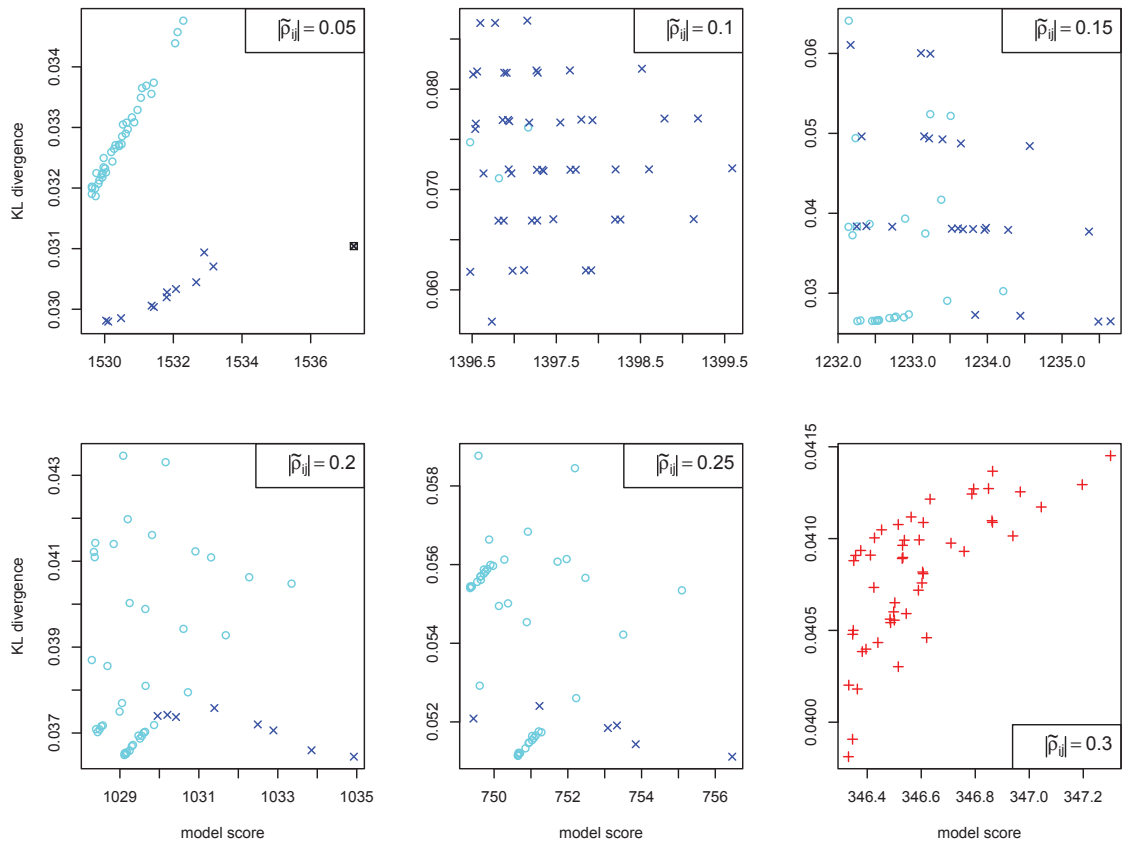


Figure B.2: Top 50 graphs found by FINCS for  $p=20$ , samples of  $n=1000$  for different  $|\tilde{\rho}_{ij}|$ .

$+$  = superset graphs;  $\times$  = subset graphs;  $\circ$  = subset plus incorrect edges;

$\boxtimes$  = empty graph



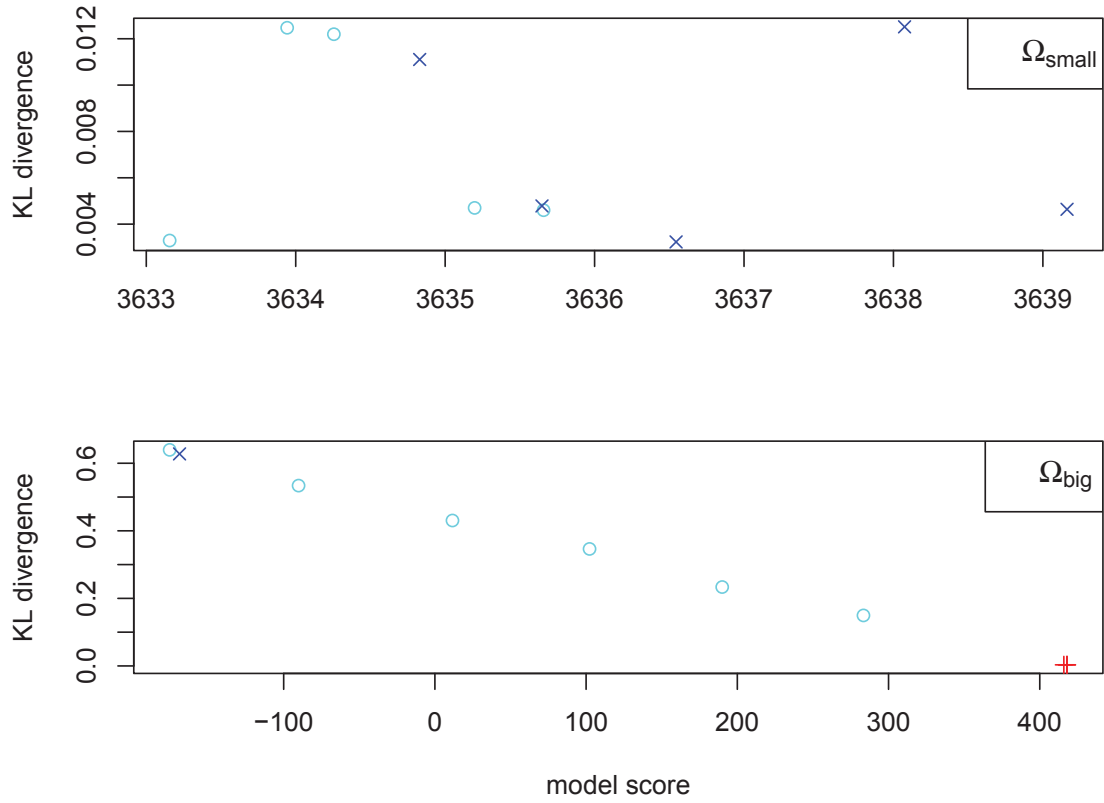


Figure B.3: Top 10 graphs found by FINCS for  $p=4$ , samples of  $n=1000$  for  $\Omega_{small}$  and  $\Omega_{big}$ .

+ = superset graphs; x = subset graphs; o = subset plus incorrect edges;

⊠ = empty graph

*note there are only 3 possible superset graphs*

Table B.2: True edges missing in top 10 graphs found by FINCS for  $\Omega_{small}$  when  $p=4$ .

(x indicates the edge is missing in the graph.)

edge	true $\tilde{\rho}$	graph rank									
		1	2	3	4	5	6	7	8	9	10
1,2	-0.20										
2,3	-0.02	x	x	x	x		x	x	x	x	x
3,4	0.05	x	x		x	x	x		x	x	
4,1	0.12		x						x	x	x

Table B.3: True edges missing in top 10 graphs found by FINCS for  $\Omega_{big}$  when  $p=4$ .

(x indicates the edge is missing in the graph.)

edge	true $\tilde{\rho}$	graph rank									
		1	2	3	4	5	6	7	8	9	10
1,2	-0.50				x		x	x		x	
2,3	0.75										
3,4	-0.61					x					x
4,1	-0.70								x		

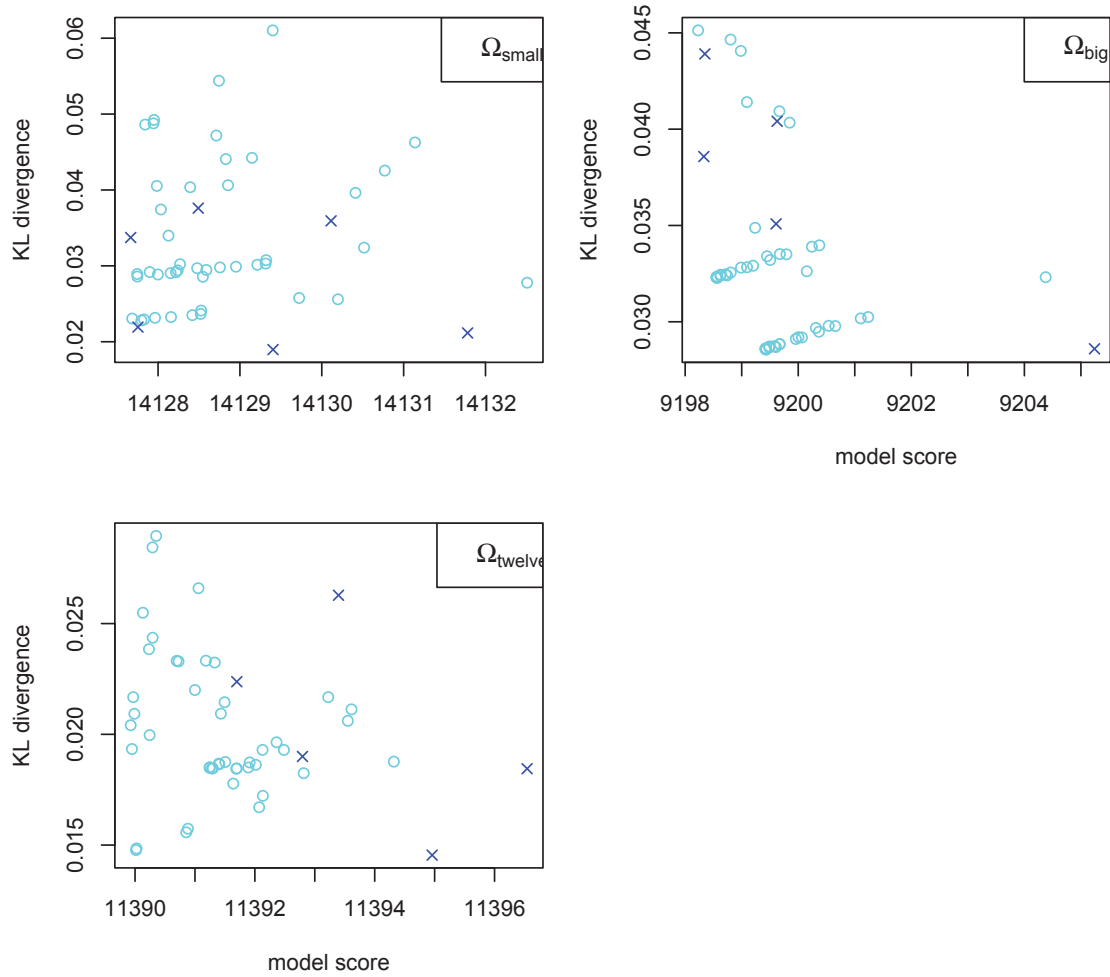


Figure B.4: Top 50 graphs found by FINCS for samples of  $n=1000$  for for  $\Omega_{small}$  and  $\Omega_{big}$  (both  $p=20$ ) and for  $\Omega_{twelve}$  ( $p=12$ ).

+ = superset graphs; x = subset graphs; o = subset plus incorrect edges;

⊠ = empty graph







### B.3 Relative inclusion probabilities

Table B.7: Relative inclusion probability matrices for  $n=1000$  and  $\tilde{\rho}_{ij}=-0.45$ .

$$\begin{pmatrix} * & \mathbf{1.000} & 0.351 & \mathbf{1.000} \\ * & * & \mathbf{1.000} & 0.724 \\ * & * & * & \mathbf{1.000} \\ * & * & * & * \end{pmatrix} \quad \begin{pmatrix} * & \mathbf{1.000} & 0.188 & 0.320 & 0.513 & \mathbf{1.000} \\ * & * & \mathbf{1.000} & 0.602 & 0.446 & 0.369 \\ * & * & * & \mathbf{1.000} & 0.251 & 0.122 \\ * & * & * & * & \mathbf{1.000} & 0.284 \\ * & * & * & * & * & \mathbf{1.000} \\ * & * & * & * & * & * \end{pmatrix}$$

**Note:** Relative inclusion probabilities associated with true edges are in **red**.

Table B.8: Relative inclusion probability matrix for 12 node case when  $n=1000$  and  $\tilde{\rho}_{ij}=-0.4$  and for 20-node cycle when  $n=1000$  and  $\tilde{\rho}_{ij}=-0.45$ .

$$\begin{pmatrix} * & \mathbf{1} & 0.6 & 0.46 & 0.4 & 0.0 & 0.0 & \mathbf{1} & 0.0 & 0 & 0.0 & 0.0 \\ * & * & \mathbf{1} & 0.2 & 0.1 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 \\ * & * & * & \mathbf{1} & 0.3 & 0.0 & 0.0 & 0.3 & 0.0 & 0 & 0 & 0 \\ * & * & * & * & \mathbf{1} & 0.0 & 0.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & \mathbf{1} & 0.5 & \mathbf{1} & 0.4 & 0.3 & 0.5 & \mathbf{1} \\ * & * & * & * & * & * & \mathbf{1} & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & \mathbf{1} & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & \mathbf{1} & 0.3 & 0.3 & 0.3 \\ * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.4 & 0.3 \\ * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.3 \\ * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} \\ * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$$

$$\begin{pmatrix} * & \mathbf{1} & 0.6 & 0.0 & 0.5 & 0 & 0 & 0 & 0.0 & 0 & 0.3 & 1.0 & 0.2 & 0.0 & 0 & 0 & 0 & 0.4 & \mathbf{1} \\ * & * & \mathbf{1} & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0 \\ * & * & * & * & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & \mathbf{1} & 0.5 & 0.0 & 0.4 & 0.0 & 1.0 & 0.6 & 0.0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 \\ * & * & * & * & * & * & \mathbf{1} & 0.0 & 0.3 & 0.0 & 0.3 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & \mathbf{1} & 0.9 & 0.1 & 0.4 & 0.00 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & \mathbf{1} & 0.0 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.9 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.0 & 0.0 & 0.0 & 0 & 0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.2 & 0.0 & 0 & 0 & 0 & 0.6 & 0.5 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.0 & 0 & 0 & 0 & 0.7 & 0.3 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.0 & 0 & 0.0 & 1.0 & 0.1 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.3 & 0.4 & 1.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.5 & 0.5 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0.3 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & \mathbf{1} \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$$

Note:

- Relative inclusion probabilities associated with true edges are in red;
- Relative inclusion probabilities that are 1.000 and are associated with other edges are in cyan;
- Relative inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.



Table B.9: Relative inclusion probability matrices when  $n=1000$  and top graphs are supersets.

$$\begin{matrix}
 \tilde{\rho}_{ij} \text{ big} & & \tilde{\rho}_{ij} = -0.15 \\
 \left( \begin{array}{cccc}
 * & 1.000 & 0.823 & 1.000 \\
 * & * & 0.987 & 0.361 \\
 * & * & * & 0.999 \\
 * & * & * & *
 \end{array} \right) & & \left( \begin{array}{cccc}
 * & 1.000 & 0.346 & 1.000 \\
 * & * & 1.000 & 0.694 \\
 * & * & * & 1.000 \\
 * & * & * & *
 \end{array} \right)
 \end{matrix}$$

$$\begin{matrix}
 \tilde{\rho}_{ij} = -0.3 \\
 \left( \begin{array}{cccccccccccccccccccccccc}
 * & 1 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0.3 & 1 \\
 * & * & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.8 & 0.6 \\
 * & * & * & 1 & 0.9 & 0.0 & 0.0 & 0.0 & 0.1 & 0 & 0 & 0 & 0 & 0.0 & 0.8 & 0.8 & 1.0 & 0.9 & 0.2 \\
 * & * & * & * & * & 1 & 0 & 0.0 & 0.0 & 0.0 & 0 & 0 & 0 & 0.0 & 0.1 & 0.0 & 0 & 0 & 0 \\
 * & * & * & * & * & * & 1 & 0.6 & 0.3 & 0.3 & 0 & 0 & 0 & 0.0 & 0.1 & 0.9 & 0.2 & 0.0 & 0 \\
 * & * & * & * & * & * & * & 1 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0.0 & 0.2 & 0 & 0 & 0 \\
 * & * & * & * & * & * & * & * & 1 & 0.4 & 0 & 0.0 & 0 & 0.0 & 0 & 0.3 & 0.5 & 0.0 & 0 \\
 * & * & * & * & * & * & * & * & * & 1 & 0 & 0 & 0 & 0.0 & 0 & 0.2 & 0.3 & 0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & 1 & 0.8 & 0.3 & 1.0 & 0.0 & 0.9 & 0.6 & 0.0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.1 & 0 & 0 & 0 & 0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.7 & 0 & 0.0 & 0.0 & 0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0 & 0.0 & 0.0 & 0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 1.0 & 0.1 & 0.0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 \\
 * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1
 \end{array} \right)
 \end{matrix}$$

Note:

- Relative inclusion probabilities associated with true edges are in red;
- Relative inclusion probabilities that are  $> 0.8$  and are associated with other edges are in cyan;
- Relative inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.

Table B.10: Relative inclusion probability matrices when  $n=50$ .

$\tilde{\rho}_{ij} = -0.45$	$\tilde{\rho}_{ij} = -0.4$
$\begin{pmatrix} * & 1.000 & 0.890 & 0.768 \\ * & * & 0.902 & 0.597 \\ * & * & * & 0.993 \\ * & * & * & * \end{pmatrix}$	$\begin{pmatrix} * & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0 & 0.0 \\ * & * & * & * & 1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & 1.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.8 \\ * & * & * & * & * & * & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 \\ * & * & * & * & * & * & * & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & 0.8 & 0.1 & 0.0 & 0.3 \\ * & * & * & * & * & * & * & * & * & 0.5 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & 1.0 & 0.1 \\ * & * & * & * & * & * & * & * & * & * & * & 0.4 \\ * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$
$\tilde{\rho}_{ij} = -0.45$	
$\begin{pmatrix} * & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0 & 0.0 & 0 & 0.0 & 0 & 0.0 & 0.0 & 0.0 & 1 \\ * & * & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & 0.8 & 0.3 & 0.0 & 0.0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & 1.0 & 0.0 & 0.0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & 0.9 & 0.2 & 0.0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & 0.9 & 0.0 & 0.0 & 0.0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & 1.0 & 0.0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & 1.0 & 0.0 & 0 & 0.0 & 0 & 0 & 0.0 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & 1.0 & 0.0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & 1.0 & 0.0 & 0.0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.7 & 0.3 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.9 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$	

Note:

- Relative inclusion probabilities associated with true edges are in red;
- Relative inclusion probabilities that are  $> 0.8$  and are associated with other edges are in cyan;
- Relative inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.

Table B.11: Relative inclusion probability matrices when  $n=1000$  and partial correlations are small.

$\begin{pmatrix} * & 1.000 & 0.016 & 0.856 \\ * & * & 0.013 & 0.012 \\ * & * & * & 0.036 \\ * & * & * & * \end{pmatrix}$	$\begin{pmatrix} * & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & 1 & 0.0 & 0.0 & 0.0 & 0 & 0 & 0.0 & 0 & 0.0 & 0.0 \\ * & * & * & 1 & 0.0 & 0.0 & 0.0 & 0 & 0 & 0.0 & 0 & 0.0 \\ * & * & * & * & 0.9 & 0.0 & 0.0 & 0 & 0.0 & 0.0 & 0 & 0.0 \\ * & * & * & * & * & 1 & 0.1 & 0.0 & 0 & 0 & 0.0 & 1 \\ * & * & * & * & * & * & 1 & 0.0 & 0 & 0 & 0.0 & 0.1 \\ * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 & 0.0 \\ * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & 1 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * & * & 1 \\ * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$
$\begin{pmatrix} * & 1 & 0.0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1 \\ * & * & 1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & 1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0 \\ * & * & * & * & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0 \\ * & * & * & * & * & 1 & 0.0 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & 1 & 0.3 & 0.0 & 0 & 0 & 0.0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & 0.8 & 0.0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 1 & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.0 & 0.0 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.1 & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & 0.0 \\ * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * & * \end{pmatrix}$	

Note:

- Relative inclusion probabilities associated with true edges are in red;
- Relative inclusion probabilities that  $> 0.8$  and are associated with other edges are in cyan;
- Relative inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.

## B.4 Maximum likelihood estimate (MLE) results

Table B.12: Model comparison measures for MLE results.

$p$	$n$	K-L divergence for simulated sample				
		1	2	3	4	5
20	50	6.4636	5.5157	4.5946	4.9008	5.5059
20	1000	0.1148	0.1114	0.1076	0.1010	0.1268
12	50	1.7893	0.8728	0.8285	1.5253	1.3253
12	1000	0.0563	0.0389	0.0364	0.0369	0.0308

When  $p=20$  there were 190 edges giving a precision of 0.1053 and recall of 1.

When  $p=12$  there were 66 edges giving a precision of 0.1970 and recall of 1.

Table B.13: Sum of squared errors using the MLE.

$p$	$n$	SSE for simulated sample				
		1	2	3	4	5
20	50	97.48	106.70	76.70	119.62	83.89
20	1000	54.41	54.47	53.78	54.67	55.82
12	50	5.837	5.352	5.234	6.159	5.772
12	1000	4.655	4.756	4.750	4.675	4.762



# Appendix C

## DRC forms

The signed statements of contribution to a doctoral thesis containing publications are attached immediately following this page. Although bound into the thesis they constitute additional pages and are thus not numbered.



DRC 16



**MASSEY UNIVERSITY**  
GRADUATE RESEARCH SCHOOL

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: ANNE MARIE FITCH

Name/Title of Principal Supervisor: Dr Beatrix Jones

Name of Published Paper: Shortest path analysis using


partial correlations for classifying gene functions from  
gene expression data

In which Chapter is the Published Work: Chapter 3

What percentage of the Published Work was contributed by the candidate: 90%

  
Candidate's Signature

25/8/2011  
Date

  
Principal Supervisor's signature

26/8/2011  
Date



DRC 16



MASSEY UNIVERSITY  
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: ANNE MARIE FITCH

Name/Title of Principal Supervisor: Dr Beatrix Jones

Name of Published Paper: The cost of using decomposable Gaussian graphical models for computational convenience

In which Chapter is the Published Work: Chapter 4

What percentage of the Published Work was contributed by the candidate: 90%

  
Candidate's Signature

25/8/2011  
Date

  
Principal Supervisor's signature

26/8/2011  
Date

# References

- Aburatani, S., Kuhara, S., Toh, H., and Horimoto, K. (2003). Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling. *Signal Processing*, **83**, 777–788.
- Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, **3**, 205–238.
- Armstrong, H., Carter, C. K., Wang, K. F. K., and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, **19**(3), 303–316.
- Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, **92**(2), 317–335.
- Banjeree, O., Ghaoui, L. E., and d’Apremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, **9**, 485–516.
- Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA.
- Berry, A., Blair, J. R., Heggernes, P., and Peyton, B. (2004). Maximum cardinality search for computing minimal triangulation of graphs. *Algorithmica*, **39**(4), 287–298.
- Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, **96**(3), 497–512.
- Cox, D. R. and Wermuth, N. (2000). On the generation of the chordless four-cycle. *Biometrika*, **87**(1), 206–212.
- Dahl, J., Vandenberghe, L., and Roychowdbury, V. (2008). Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods and Software*, **23**(4), 501–520.
- Dawid, A. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**(3), 1272–1317.
- de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlations coefficients. *Bioinformatics*, **20**(18), 3565–3574.
- Dellaportas, P., Giudici, P., and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā*, **65**(1), 43–55.
- Dempster, A. (1972). Covariance selection. *Biometrics*, **28**(1), 157–175.
- Dijkstra, E. (1959). A note on two problems in connection with graphs. *Numerische Mathematik*, **1**, 269–271.
- Dobra, A. and Fienberg, S. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences, USA*, **97**(22), 11885–1192.
- Dobra, A. and West, M. (2004). Bayesian covariance selection. Technical report, Duke University.

- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196–212.
- Drton, M. and Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, **33**(2), 247–257.
- Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**(3), 457–482.
- Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, **3**(2), 521–541.
- Fitch, A. M. and Jones, B. (2009). Shortest path analysis using partial correlations for classifying gene functions from gene expression data. *Bioinformatics*, **25**(1), 42–47.
- Fitch, A. M. and Jones, B. (2012). The cost of using decomposable gaussian graphical models for computational convenience. *The Journal of Computational Statistics and Data Analysis*, <http://dx.doi.org/10.1016/j.csda.2012.01.020>.
- Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**(2), 302–332.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008a). *glasso: Graphical lasso- estimation of Gaussian graphical models*. R package version 1.2.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008b). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.
- Gao, X., Ou, D. Q., and Xu, H. (2009). Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *ArXiv e-prints*, pages 1–20.
- Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**(4), 785–801.
- Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Association*, **102**(478), 507–516.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Ed 2*. Springer, USA.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennet, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P., Stepaniants, S. B., Shoemaker, D., Gachotte, D., Chakraburtt, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**(1), 109–126.
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, **20**(4), 388–400.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.
- Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.
- Liu, H., Roeder, K., and Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems(NIPS)*, **23**, 14.
- Matusno, T., Tominga, N., Arizono, K., Iguchi, T., and Kohara, Y. (2006). Graphical Gaussian modeling for gene association structures based on expression deviation patterns induced by various chemical stimuli. *IEICE Transactions on Information and Systems*, **E89-D**(4), 1563–1574.
- Meinhausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

- Meinhausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high dimensional data. Technical report, Department of Statistics, UC Berkeley.
- Moghaddam, B., Marlin, B. M., Khan, M. E., and Murphy, K. P. (2009). Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. *Proceedings of the 23rd Neural Information Processing Systems Conference*, pages 1285–1293.
- O’Hagen, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B*, **57**(1), 99–138.
- Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, **104**(486), 735–746.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *The Annals of Statistics*, **36**(6), 2818–2849.
- Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). *Biometrika*, **90**(3), 491–515.
- Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, **29**, 391–411.
- Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**(1), 1–32.
- Scott, J. G. and Carvalho, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, **17**(4), 790–808.
- Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, **3**, 1039–1059.
- Shimamura, T., Imoto, S., Yamaguchi, R., and Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, **19**, 142–153.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, **58**(1), 267–288.
- Tibshirani, R., M.Saunders, Rosset, S., J.Zhu, and K.Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B*, **67**(1), 91–108.
- Toh, H. and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**(2), 287–297.
- Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *Journal of the American Statistical Association*, **75**(372), 963–972.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**(1), 1–25.
- Whittaker, J. (2008). *Graphical Models in Applied Multivariate Statistics*. John Wiley and sons, UK.
- Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, **5**(11), R92.
- Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, **90**(4), 809–830.

- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, **68**(1), 49–67.
- Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, **31**(1), 585–603.
- Zhou, X., Kao, M. C. J., and Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences, USA*, **99**, 12783–12788.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, **67**(2), 301–320.