THE UNIVERSITY OF
WARWICK

AUTHOR: **Robert J. B. Goudie** DEGREE: **Ph.D.**

TITLE: **Bayesian structural inference with applications in social science**

DATE OF DEPOSIT: .....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

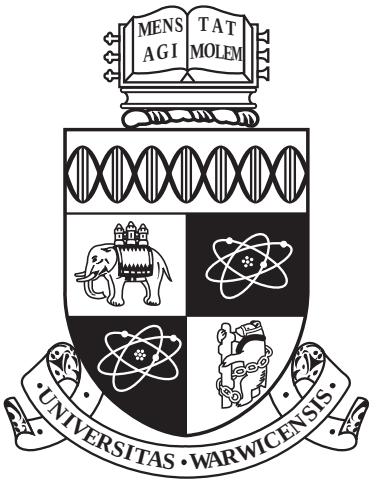
.....

.....

.....

.....

.....



Bayesian structural inference with applications in social science

by

Robert J. B. Goudie

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics
School of Health and Social Studies

October 2011

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

List of Notation

Basics

n	The number of observations	14
p	The number of variables	14
$\mathbf{y}$	A n-dimensional vector random vector	13
Θ	The parameter space of the entire model	13
θ	A vector parameter $\theta \in \Theta$	13
$\mathbf{X}$	A $n \times p$ matrix of data, or the collection of random vectors $\{X_1, \ldots, X_p\}$	20
$\mathbf{X}_A$	The columns of $\mathbf{X}$ specified by A	34
$\mathbf{X}_{-i}$	The columns of $\mathbf{X}$ except column i	20
$\perp\!\!\!\perp$	Independence of random variables	18
$\not\perp\!\!\!\perp$	Dependence of random variables	18
$\mathcal{O}$	Complexity upper bound	90

Distributions

Mult	Multinomial distribution	24
Dir	Dirichlet distribution	24

N	Normal distribution	25
MVN	Multivariate normal distribution	30
IG	Inverse-gamma distribution	25

Models

$\mathcal{M}$	A finite set of models	15
M	A model from the set of model $\mathcal{M}$	15
$\|M\|$	The number of models	15
p_M	The dimension of model M	14
Θ_M	Parameter space under model M	15
θ_M	The parameters of model M	15
$p(\mathbf{y} \mid \theta_M)$	The likelihood of model M	15
$\pi(\theta_M \mid M)$	Parameter priors under model M	15
$\pi(M)$	Prior weight for model M	15
$p(\mathbf{y} \mid M)$	The marginal likelihood of model M	16
$p(M \mid \mathbf{y})$	The posterior model distribution	15
M^{MAP}	The maximum *a posteriori* model	16

Graphs

$\mathcal{G}$	The set of all directed, acyclic graphs (DAGs) with p nodes	20
$G = (V, E)$	A mathematical graph	18
V	Vertices	18
E	Edges	18

G_{ij}	The (i, j) element of the adjacency matrix	19
p	The number of nodes in the graph G	18
ϵ	The number of edges in the graph G	90
(i, j)	Directed edge from node i to node j	18
$i \to j$	Directed edge from node i to node j	18
$j \leftarrow i$	Directed edge from node i to node j	18
G_j	The parents of node j in G	18
$\langle G_1, \ldots, G_p \rangle$	The graph G with these parent sets	19
G_A	The parent sets in G of nodes in A	19
G_{-A}	The parent sets in G of nodes not in A	19
T^G	Transitive closure matrix for G	90
C^G	A path count matrix for G	91
$\nu(G)$	The set of neighbouring graphs of G	48
G_{ij}^-	The graph G, with no edge $i \to j$	82
G_{ij}^+	The graph G, with an edge $i \to j$	82
d	The number of nodes in a path	19

Bayesian networks

G	A Bayesian network	20
$\mathbf{X}_{G_j}$	The random variables corresponding to the parents G_j of node j in G	20
$p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$	The local distribution of X_i in G	20
$p(\theta_{G,i})$	The parameter priors for the distribution of X_i, given G	36

$\pi(G)$	Prior for graph G	39
$\pi_i(G_i)$	Prior for the parents of node i in graph G	79
$p(\mathbf{X} \mid G)$	The marginal likelihood for G	39
$p(X_i \mid \mathbf{X}_{G_i})$	The local marginal likelihood	80
κ	The maximum in-degree	41

Regression models

γ	Indicator variable for the included predictors	27
M_γ	Regression model, with predictors indicated by γ	27
p_γ	The number of predictors in model M_γ	27
$\mathbf{X}_\gamma$	Matrix formed from the columns of $\mathbf{X}$ corresponding to predictors in M_γ. For normal linear regression, the matrix also includes a column of 1s	27

Discrete data

r	The number of categories (or levels) of $\mathbf{y}$	24
n_k	The number of observations in the k^{th} category	25
C_j^γ	j^{th} configuration under model M_γ	27
q_γ	Cardinality of the sample space of $\mathbf{X}_\gamma$	27

Normal data

$m, \mathbf{m}_\gamma$	The prior mean (under model M_γ)	25
$v, \mathbf{V}_\gamma$	The prior variance (under model M_γ)	25

Gibbs sampler

W	A subset of V	86

w_j	The j^{th} node in W	93
ρ	The number of nodes in W	86
$\mathfrak{F}_W$	The collection of parent sets of W that yield an acyclic graph	87
F_W	The collection of parent sets of nodes in W	86
F_{w_j}	The parents in F_W of the node w_j in W	86
$\mathfrak{F}$	The set of acyclic graphs that differ from G in the parents of nodes in W	93
$P(F_W \mid G_{-W}, \mathbf{X})$	The conditional distribution for the parents of nodes in W	87
$\mathcal{H}$	The set of all DAGs on the nodes in W	93
H	A DAG on the nodes in W	93
η	The number of nodes in $\mathcal{H}$	93
$\mathfrak{F}^{H^h}$	A component of a partition of $\mathfrak{F}$	93
G^-	The graph formed from G by removing edges directed towards a node in W	93
D_j	The descendants of node $w_j \in W$ in G^-	93
K_j	The non-descendants of node $w_j \in W$ in G^-	93
K	Nodes that are not descendants in G^- of any node in W	93
D_j^H	Descendants in G^- of nodes in H_{w_j}	93
$\mathfrak{F}_W^H$	The collection of allowed parent sets for each node $w_j \in W$, using H	98
D_{-j}^H	Descendants in G^- of nodes not in H_{w_j}	93

$\mathfrak{F}_j^H$	The set of allowed parent sets for w_j, using H	98
$\mathfrak{G}_j$	The set of possible parent sets of node j	99
$\mathfrak{G}_j^i$	The $\mathfrak{G}_j$ that contain node i	99
$\otimes$	Outer matrix product	91
$i \rightsquigarrow j$	A path from node i to node j that does not include nodes in W (except possibly i and j)	96

Acknowledgements

I am very grateful to my supervisors, Sach Mukherjee and Frances Griffiths, for their constant enthusiasm, encouragement and inspiration throughout my time at Warwick. It has also been a pleasure to collaborate with three economists (Andrew Oswald, Jan-Emmanuel De Neve and Stephen Wu) and I am particularly grateful for their efforts in helping a statistician to understand some of their field.

I have been fortunate to work amongst many other generous and friendly members of staff and students at the Department of Statistics. Thank you particularly to those in Sach's group, those who suffered me in D0.06, and to Chris Jewell who ran the departmental high performance computing facilities in his own time so pleasantly.

I am indebted to my parents and sisters for their encouragement and support. Most importantly, I thank Sarah, my fiancé and closest friend, who experienced the development and writing of this thesis the most closely, but was nevertheless supportive and encouraging throughout.

I also thank Marco Grzegorczyk, Dirk Husmeier, Karen Sachs, Amanda Goodall, Graham Loomes, and Mark Steel for an assortment of thoughts, code and data used in various parts of this thesis. Financially, I have been been supported by a joint Economic and Social Research Council (ESRC) and Engineering and Physical Sciences Research Council (EPSRC) award.

Declarations

I hereby declare that this thesis is based on my own research, except when stated otherwise. This thesis has not been submitted for a degree at another university.

Some of this work has been published, is available as a working paper, or has been submitted for publication as follows.

The material of Chapter 3 forms part of a paper '*Happiness as a driver of risk-avoiding behavior*' that is under review, co-authored with Sach Mukherjee, Jan-Emmanuel De Neve, Andrew J. Oswald and Stephen Wu. The analyses from this paper presented here are my own work, although an initial ordinary least squares analysis was conducted by Stephen Wu. An earlier version of this paper is available as CESifo Working Paper Series No. 3451.

Chapters 4 and 5 are an extension of work available as CRISM Working Paper No. 11-21, under the title '*An efficient Gibbs sampler for structural inference in Bayesian networks*'. The working paper is co-authored with Sach Mukherjee.

The material presented in Chapter 6 was published as '*Exploratory network analysis of large social science questionnaires*' in the Proceedings of Bayesian Modelling Applications Workshop (BMAW-11). This was co-authored with Sach Mukherjee and Frances Griffiths.

Abstract

Structural inference for Bayesian networks is useful in situations where the underlying relationship between the variables under study is not well understood. This is often the case in social science settings in which, whilst there are numerous theories about interdependence between factors, there is rarely a consensus view that would form a solid base upon which inference could be performed. However, there are now many social science datasets available with sample sizes large enough to allow a more exploratory structural approach, and this is the approach we investigate in this thesis.

In the first part of the thesis, we apply Bayesian model selection to address a key question in empirical economics: why do some people take unnecessary risks with their lives? We investigate this question in the setting of road safety, and demonstrate that less satisfied individuals wear seatbelts less frequently.

Bayesian model selection over restricted structures is a useful tool for exploratory analysis, but fuller structural inference is more appealing, especially when there is a considerable quantity of data available, but scant prior information. However, robust structural inference remains an open problem. Surprisingly, it is especially challenging for large n problems, which are sometimes encountered in social science. In the second part of this thesis we develop a new approach that addresses this problem—a Gibbs sampler for structural inference, which we show gives robust results in many settings in which existing methods do not.

In the final part of the thesis we use the sampler to investigate depression in adolescents in the US, using data from the Add Health survey. The result stresses the importance of adolescents not getting medical help even when they feel they should, an aspect that has been discussed previously, but not emphasised.

Chapter 1

Introduction

The aim of statistical modelling is to improve the degree of understanding of a phenomenon of interest. Statistical models can help to describe and explain many things including which factors are important; the direction and magnitude of the associated effects; and, more generally, the relationship (if any) between variables of interest. However, the level of precision that is attainable with statistical analysis is usually determined by the nature of the data that are available and the (*a priori*) assumptions one is prepared to make.

In general, more precise inferences will be possible when more data are available. The sample size is usually the most important dimension of the data. In addition, for the analysis to be useful, it will typically be important that the data are a representative sample from the larger population under study, to facilitate inference about the wider population. The second dimension of the data (the number of variables measured) is also important because of the need to minimise the possibility that a factor that was not measured performs an important role in the system under study.

The second aspect that is important in determining the precision of the analysis

is the existing level of understanding. Statistical inference is always built upon assumptions. In likelihood-based inference, many of the important assumptions are made when determining the likelihood. In some settings, these assumptions may be based upon accepted theories of the underlying system and are thus well founded. In such a case, inference is about understanding the details of a system for which the structure is already understood. In multivariate statistics, a core part of these assumptions relate to the dependencies between different variables (or components) of the system. Any assumption made about the structure of the dependency is important in statistical inference because it is built into the likelihood.

1.1 Scope of the analysis

In this thesis, we consider the situation in which high-quality data are available, but the existing accepted level of understanding of the phenomenon under study is poor. In particular, we mostly do not assume a particular structure of dependence between the components of the system. Instead, the purpose of the analysis is to make inference about dependence. Making relatively weak assumptions, such as we do here, means that we keep an open mind to unexpected relationships. Thus the analysis that we make is mostly exploratory in nature.

We also assume that only observational data are available. In such cases, without any information about the effect of interventions, it usually is very difficult to infer anything conclusive about causality. There is a large literature covering methods for analysing data collected through observational studies (Rosenbaum, 2002), but much of this avoids making causal claims. Some of the strongest claims about causality have come from researchers working with graphical models, for example, Cox and Wermuth (2004), and, most prominently, Pearl (2009). However, it remains controversial to place the emphasis on graphical approaches to causal inference, and

there are many advocates of other approaches (notably Rubin, 2005).

Here, we take the view that graphical approaches are useful tools in situations in which strong causal claims are sought, but we do not seek to construe our results in this manner. Instead, we view our work as primarily about discovering relationships that suggest interesting conjectures; these are framed in a manner that allows further work (ideally interventional) to be carried out to examine the conjectures in more detail. This point of view has been proposed previously by many authors including Williamson (2005), who views the approach as a hybrid between a hypothetico-deductive and an inductive approach to discovering causal relationships.

The cost of data collection is generally falling, and so 'large' datasets are now increasingly the norm. A considerable amount of data are now available that describe phenomena about which no consensus model is available. Datasets describing various aspects of economics, genetics, molecular and cell biology, and diverse areas of the social sciences are widely available. In many of these areas, the growth in the availability of data has exceeded the growth in theoretical understanding. This opportunity is an opening for statistical methods that improve understanding in these settings.

1.2 Statistical model selection

1.2.1 Inadequacy of the complete model

In poorly understood settings there may be many factors that could plausibly play an important role in the system under study. In this situation a model that incorporates all of these factors may seem attractive, because it incorporates all of the available information and the analysis is not prejudiced by the disregarding of potentially important factors.

The estimator associated with this COMPLETE or FULL MODEL will have many degrees of freedom, and so it is able to closely replicate features in the data. However, the 'volume' of space in which a high-dimensional probability distribution may have support (regions of positive probability) increases exponentially as its dimension increases, a phenomenon described as the 'curse of dimensionality' by Bellman (1961) in the context of dynamic programming. This effect results in the available data being sparsely dispersed across the space relative to its size.

Another example of this problem is given by Silverman (1986), who calculates the required sample size for an estimator $\hat{p}(x)$ of the density $p(x)$ at the origin of a unit multivariate normal distribution to have relative mean squared error $E\left((\hat{p}(0) - p(0))^2\right)/p(0)^2$ less than 0.1. For a univariate distribution $p(x)$, only 4 samples are required to satisfy this criterion; for a 5-dimensional distribution, 768 samples are required; and for a 10-dimensional distribution, around 842,000 samples are required. Thus even for a smooth unimodal distribution, with a simple measure of fit based around the mode of the distribution, the amount of data required rapidly becomes enormous as the dimension of the distribution grows. As a result, even with a large sample size, a single dataset in a high-dimensional setting will not exhibit all of the characteristics of the underlying probability distribution.

Thus, while on average closely matching the data will give accurate estimates, rigidly replicating the exact properties of a single dataset may be far from optimal. An estimator that does this will be particularly susceptible to small variations in the data, and so the estimator will have high variance. On the other hand, the estimator has low bias because averaging across replications of the data will give accurate estimates. Particularly in exploratory settings, the complexity of a model including all of the factors is a disadvantage. For these reasons, the complete model is often not the most useful model.

Instead, we would like to construct a model that retains the advantages of the

complete model whilst mitigating its disadvantages. The advantage that we want to keep is the small bias; the disadvantage we seek to ameliorate is its large variance. In these settings reducing variance will increase the bias, and so a trade-off exists between these properties (see, e.g. Hastie *et al.*, 2009). At the opposite end of the spectrum of model complexity to the full model, we could consider a univariate model that includes no covariates. This model will typically have the opposite problem: large bias, but low variance.

A particular example of these trade-offs is a regression model with 100 potential predictors. The ordinary least squares estimators for the regression coefficients are consistent, so as the sample size grows, the coefficients will converge to their true values. In practice, we have only a finite sample, and so the estimators will not give the true values of the coefficients. In particular, the estimates for the coefficients in the full model will have a large variance. The large variance in the estimates is intuitive because in the parameter space for the full model, the data will be sparsely dispersed, and so a small change to an individual data point may lead to a large change in the estimators. Averaging across replications of the data, however, will lead to the estimators having the correct values. Thus, the bias of the estimators is low. In contrast, a model including only one predictor will have low variance, which is intuitive because a relatively large amount of data will be used to estimate its value. However, such a simple model may not be expressive enough to capture the true form of the data, and so the bias of the estimator will be high.

1.2.2 Objectives and viewpoints

We have described why in many settings a full model may not be appropriate even if it does subsume the 'true model' (the concept of a 'true model' is discussed further below, in Sections 1.3.2 and 1.2.3). Conversely a simple, univariate model may not be sufficiently rich to represent the properties of the data. We thus aim to choose

an intermediate model that balances the competing requirements of minimising bias and variance. The models that are considered may be of differing dimension or contain different functional forms. Handling the varying dimensions of the models considered is particularly difficult. The problem of finding an appropriate model is known in general as MODEL SELECTION.

The ideal model will strike a balance of being consistent with the data without being overly complicated in such a way that over-fitting will occur. This idea has a long history and is often attributed to William of Ockham, under the name OCCAM'S RAZOR, or called the PRINCIPLE OF PARSIMONY. Each model may be associated with a particular scientific hypothesis, and so model selection may be useful in comparing the competing hypotheses.

One aim of model selection is to understand the dependence structure of the variables. The structure of the dependence within a system can be encapsulated by the likelihood function of a statistical model. Inference about the dependence structure can thus be considered as statistical model selection. The origins of this form of analysis can be traced back to the work of Sewall Wright, who developed the method of path analysis (Wright, 1921), which aims to measure the direct effect of each 'path' in a system. Another early methodology that can be viewed in this light is that of Dempster (1972), in which the covariance structure of a multivariate normal distribution is modelled with a particular focus on finding a simple description of its structure. A simple description of the structure is achieved by setting appropriate entries of the inverse covariance matrix to zero. In doing so, conditional independence, given all other variables, is implied between the corresponding variables, and the number of parameters in the model is reduced. These models can be viewed as undirected Gaussian graphical models (Lauritzen, 1996). The models considered in this thesis can be viewed as originating in similar work. However, rather than considering an undirected Gaussian graphical model, we will consider

Bayesian networks.

1.2.3 What is a statistical model?

Before turning to practical issues related to choosing the model, we discuss the meaning and role of statistical models and their relationship to 'truth'. Bernardo and Smith (1994, ch. 4, pp. 237) argue that most statisticians agree that the role of models is to provide a focused framework within which simplified representations of phenomena can be discussed. The most optimistic view is that a single statistical model can encapsulate 'truth'. Thus, if we can construct a list $\mathcal{M}$ of candidate models, we can try to determine which of these is true. This view usually seems overly-optimistic. Instead, a more appropriate view in most contexts is the pragmatic view taken by Box and Draper (1987) in a discussion of the bias-variance trade-off: "all models are false, but some are useful". Buckland *et al.* (1997) take a similar view asserting that the 'truth' is high dimensional, and effectively infinite dimensional, and so in handling model uncertainty we should seek the best approximating fit rather than the 'truth'. Another pragmatic viewpoint is taken by Fisher and Neymann (as discussed by Lehmann, 1990), who suggest that the key characteristic of models should be familiarity and simplicity. See Cox (1990) for more discussion on the role of models.

1.2.4 Implementation of model selection

In practice, choosing a model that balances bias and variance is not straightforward. For complex multivariate models, assessing the bias and variance associated with an estimator from a single, finite dataset is challenging. In particular, measuring the discrepancy between the observed data and a model is not sufficient because by this metric the full model is always selected. In addition, the traditional methods of testing the coefficients for significance using classical multivariate tests based upon

maximum likelihood estimators (MLEs) do not give sensible results in this context for several reasons.

One problem is multiple testing. Freedman (1983) examined this issue empirically. Data from 50 independent random variables were regressed against data from an entirely independent variable. Alarmingly, after dropping 35 variables which were insignificant at 0.25 level, 6 of the remaining 15 variables were judged significant at the 0.05 level. The Bonferroni correction (Bonferroni, 1936; Bland and Altman, 1995) offers a simple adjustment for this problem under an assumption of independent tests. There has been much work on multiple testing in recent years (see e.g. Benjamini and Hochberg, 1995; Dudoit and van der Laan, 2008).

Additionally, the metric by which we judge a model must account for the number of parameters that the model includes. One approach in the frequentist framework is to add a term to the likelihood function that penalises high-dimensional models. Examples include Akaike's information criteria (Akaike, 1974; Burnham and Anderson, 2002), which is known as AIC, and the Bayesian information criteria (Schwarz, 1978), which is known as BIC. Information criteria describe a general method for likelihood penalisation. Penalised likelihood approaches for model selection have a rich literature, see e.g. Claeskens and Hjort (2008). We describe AIC and BIC in more detail in Section 2.1.

In the specific context of regression, penalisation based on ℓ_1 and ℓ_2 norms of the coefficient vector are widely used (Bühlmann and van de Geer, 2011). Ridge regression (Hoerl and Kennard, 1970) uses a ℓ_2 penalty. Using this penalty allows straightforward maximisation of the (penalised) likelihood to yield a closed-form estimator. The estimates for the regression coefficients are shrunk towards zero, thereby controlling over-fitting. However, ridge regression does not set regression coefficients to exactly zero – that is, ridge regression does not result in variable selection. In contrast, the LASSO (Tibshirani, 1996) uses a ℓ_1 penalty, and can shrink

estimates for the regression coefficients to exactly zero. Here, maximisation of the penalised likelihood requires optimisation, but an efficient algorithm called LARS (Efron *et al.*, 2004) exists.

Bayesian model selection offers an alternative approach (detailed in the next section). Rather than directly using penalised likelihoods, a posterior distribution across a set of models is constructed, and comparison between pairs of models can be made using Bayes factors. Bayesian model selection and penalised likelihood methods are closely related: the log posterior distribution is given, up to a constant, by the sum of the log likelihood and the log prior. The (negative) log prior can therefore be viewed as a penalty term. This view makes clear the relationship between various penalised likelihood estimators and related Bayesian formulations.

Approaches that draw ideas from the Bayesian approach in a frequentist context are also available. For example, Buckland *et al.* (1997) propose a method for assigning weights to models, but the weights arise from functions of information criteria, rather than from a posterior distribution. The BIC also straddles both frameworks: although it takes the form of a penalised likelihood, it is also an asymptotic approximation to the Bayes factor.

1.3 Bayesian model selection

1.3.1 Basic Bayesian framing

Model selection in the Bayesian framework considers an indicator variable over models as an additional parameter, equipped with a prior and posterior distribution in the same way that all parameters do in the Bayesian framework. The usual formulation assumes that a finite collection $\mathcal{M}$ of models is being considered, and that prior mass is assigned to each of these models. The posterior distribution across

models can then be found by an application of the discrete version of Bayes' theorem. The theory of handling model uncertainty in a Bayesian framework is now well-developed; Clyde and George (2004) give a full overview.

1.3.2 Interpretations of Bayesian model selection

The interpretation of Bayesian model selection is clearest when one of the models is viewed as the 'truth'. Usually this seems unrealistic, but in practice, especially when $|\mathcal{M}|$ is large, this may be a sufficiently good approximation. Assuming one of the models in $\mathcal{M}$ is true is called $\mathcal{M}$-closed by Bernardo and Smith (1994), who also delineate two further perspectives that could be taken on the list $\mathcal{M}$ of models. In the $\mathcal{M}$-completed viewpoint none of the models in $\mathcal{M}$ is viewed as true because our true beliefs can only be represented by a separate model M_t, which is precluded from direct consideration by intractability. In this setting we need to proceed differently because it does not make sense to assign a prior to $\mathcal{M}$ when this would not represent our true prior beliefs. The final possibility, $\mathcal{M}$-open, occurs when even specifying $\mathcal{M}_t$ is not possible. For the settings considered here, an $\mathcal{M}$-open viewpoint is the most plausible, but for pragmatic reasons we will generally work in a relatively $\mathcal{M}$-closed framework.

1.3.3 Practical implementation

In the previous section, we noted the difficulty in comparing models of different dimension, because unadjusted measures of fit will invariably prefer the most complex model. In the Bayesian formulation, the relative posterior weights assigned to two models is determined by the Bayes factor, which is the relative marginal likelihood. Comparison of models of differing dimension is possible because the marginal likelihood gives a one-dimensional measure of fit.

The computation of the posterior model distribution is often challenging, and addressing this in a particular context forms a key part of this thesis. Simpler alternatives have been proposed. For example, Draper (1995) considers starting from a single model, and expanding it as suggested by context or the data. However, in the context we consider here, it is attractive to consider a fully-Bayesian approach because the high-dimensionality makes it difficult to propose a sensible starting model. Another simplification that can be often useful is BIC, which is an asymptotic approximation to the posterior distribution.

1.3.4 Summarising the posterior distribution

Once the posterior distribution over models has been evaluated, two distinct approaches can be taken to summarising its contents.

A simple approach is to find the posterior mode. The modal model (or models) is the model that is most consistent with the data. While simple and convenient, a drawback to this approach is that a level of uncertainty is ignored because it implies that the final results are made conditional on the modal model (e.g. Chatfield, 1995).

When a quantity of interest that is interpretable across all the models under consideration can be extracted from each model an alternative approach is available. In this case, it follows from Bayes' theorem that the posterior distribution for this quantity is given by taking its average across the models, weighted by the posterior mass for each model.

1.3.5 Bayesian model uncertainty in social science

Bayesian approaches to model uncertainty have not been widely adopted in social science, despite the significant model uncertainty that exists. The foremost proponent of Bayesian model selection and averaging in the context of social science

research is Raftery (1995). In econometrics, Fernández *et al.* (2001b) advocated Bayesian model averaging as a principled way to account for model uncertainty in cross-country growth regression.

1.4 Contributions of the thesis

The thesis consists of three main contributions.

The first contribution is a study of the effects of well-being on risk-taking. This question has not been considered before, although Kirkcaldy and Furnham (2000) found correlations consistent with the findings of our work. We take a Bayesian model selection approach to the question, which is unusual in empirical economics. We find evidence in support of the theory that those with higher levels of well-being are more averse to risk-taking.

We then introduce a novel Gibbs sampler for structural inference of Bayesian networks. While Gibbs samplers have been used with Bayesian networks before, they have not been used for structural inference. The Gibbs sampler introduced here explores the posterior distribution of Bayesian networks. While the general method of Gibbs sampling is well-established, the requirement of acyclicity in Bayesian networks makes designing a Gibbs sampler difficult in this context. We show empirically that the Gibbs sampler exhibits far superior performance compared to several state-of-the-art methods. Indeed, in many cases, results obtained from widely used methods are so unstable as to be unusable in practice.

The final contribution of the thesis is an explorative study of depression in adolescents. Large social science questionnaires, including the survey we use in the thesis, have not been previously studied using structural inference of Bayesian networks. Our results are consistent with earlier results, but emphasise the importance of adolescents seeing their doctor when they feel they should.

Chapter 2

Background

2.1 Model selection

A parametric statistical model does not fully describe a probability distribution. Instead, it describes a family of distributions, up to some parameters θ. For a vector valued random variable y, a model M specifies the joint distribution of y, up to unknown parameters θ. The joint probability of y can be specified conditional on both parameters θ and model M.

$$p(\mathbf{y} \mid \theta, M) \qquad \text{with } \theta \in \Theta$$

Often, dependence on the model M is left implicit and emphasis placed on the joint distribution as a function of parameters θ, i.e. the likelihood function. Statistical inference seeks to understand the relationship between these parameters, and data. In Bayesian inference, we aim to describe the posterior distribution of these parameters, given the data.

$$p(\theta \mid \mathbf{y}) \qquad \text{with } \theta \in \Theta$$

As outlined in Chapter 1, we will be considering a situation in which observations of many variables are available, but the appropriate model for the variables is not known. Therefore, we will consider the model itself as the object of interest for inference.

Suppose n samples from p variables are available. Let p_M be the dimension of model M.

In the frequentist framework, a widely used approach to model selection involves penalised likelihood methods. The most well-known of these, the AIC, was introduced by Akaike (1974) and has the following form.

$$-2\log(p(\mathbf{y} \mid \theta, M)) + 2p_M$$

The model that minimises AIC is preferred. An alternative is the Bayesian Information Criterion (BIC), introduced by Schwarz (1978).

$$-2\log(p(\mathbf{y} \mid \theta, M)) + p_M \log(n)$$

In the specific context of regression, numerous penalised estimators for the regression coefficients β have been proposed. Consider a regression model for an outcome variable $\mathbf{y}$, using a set of q predictors. Suppose we have observations $\mathbf{y} = (y_1, \ldots, y_n)$ of the outcome, and observations of the predictors arranged into the columns of a matrix $\mathbf{X}$. The LASSO (Tibshirani, 1996) penalises the regression coefficients by an ℓ_1 penalty. This penalty permits setting of some regression coefficients to exactly zero, thereby leading to variable selection. An older alternative is ridge regression (Hoerl and Kennard, 1970), which uses an ℓ_2 penalty, and yields the following estimators for the regression coefficients, with $\mathbb{I}_q$ being the $q \times q$ identity matrix.

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda \mathbb{I}_q)^{-1} X^T Y$$

While ridge regression will shrink coefficients towards zero, it will not shrink them to exactly zero in the way the LASSO does, and so does not lead to variable selection directly. Numerous other penalties have been proposed, notably the smoothly clipped absolute deviation (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005) and the adaptive LASSO (Zou, 2006).

The Bayesian approach does not use a penalised likelihood explicitly, but the log prior can be viewed as such. A particular instance in which the two approaches yield the same solution is a regression in which the prior for the regression coefficients is $\beta \sim \mathrm{N}(0, \sigma^2\lambda^{-1}\mathbb{I}_q)$. The resulting *maximum a posteriori* (MAP) estimator for β matches the ridge estimators exactly (Hoerl and Kennard, 1970; Hsiang, 1975).

2.1.1 Bayesian model selection

The Bayesian approach to model selection treats the model simply as another parameter. Suppose a finite set of models $\mathcal{M}$ is under consideration, and that a vector of observations $\mathbf{y}$ is available. Each model $M \in \mathcal{M}$ consists of a likelihood function $p(\mathbf{y} \mid M, \theta_M)$ with parameters $\theta_M \in \Theta_M$. These parameters have priors $\pi(\theta_M \mid M)$ in each model.

Since the set of models $\mathcal{M}$ under consideration is a finite set, the model prior is a discrete distribution over this set.

$$\pi(M) = \pi_M, \qquad M \in \mathcal{M} \qquad \text{where } \pi_M \geq 0 \text{ and } \sum_{M \in \mathcal{M}} \pi_M = 1$$

An expression for the posterior distribution for a model M can be written down immediately, by a simple application of Bayes Theorem. The expression depends on the marginal likelihood $p(\mathbf{y} \mid M)$ of M.

$$P(M \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M)\pi(M)}{\sum_{M \in \mathcal{M}} p(\mathbf{y} \mid M)\pi(M)} \tag{2.1}$$

The quantity $p(\mathbf{y} \mid M)$ is given by

$$p(\mathbf{y} \mid M) = \int_{\Theta_M} p(\mathbf{y} \mid M, \theta_M)\pi(\theta_M \mid M)d\theta_M, \tag{2.2}$$

and is referred to as the MARGINAL LIKELIHOOD.

Evaluation of this posterior distribution is typically difficult for two reasons. First, the integration in Equation 2.2 may be difficult to evaluate. This difficulty motivates the use of conjugate models, as described in Section 2.3.1, which enable evaluation of the integral analytically. The second difficulty is the summation over $\mathcal{M}$ in the normalising constant of Equation 2.1. When the cardinality of $\mathcal{M}$ is large, it is not possible to evaluate the summation exactly. However, Markov chain Monte Carlo methods enable the posterior distribution to be approximated without directly evaluating the normalising constant.

There are two distinct approaches for summarising the posterior distribution. A simple approach is to select a single model, and base any further inference as conditional upon this model. When choosing a single model, the maximum *a posteriori* (MAP) model is usually chosen.

$$M^{\text{MAP}} = \arg\max_M P(M \mid \mathbf{y})$$

The MAP model M^{MAP} may not be unique, and even when it is, it may not be representative of the posterior distribution. If the posterior distribution is multi-modal, with disparate models having high posterior probability, it may be unsatisfactory to choose the one model.

Alternatively, if some quantity Δ is interpretable in all models, we can average it across all of the models, weighting by the posterior model probability.

$$p(\Delta \mid \mathbf{y}) = \sum_{M \in \mathcal{M}} p(\Delta \mid \mathbf{y}, M)P(M \mid \mathbf{y})$$

This approach is called Bayesian Model Averaging (Hoeting *et al.*, 1999; Wasserman, 2000). Choosing an appropriate model prior can be challenging, and is discussed further in Sections 2.4.3 and 2.5.4.

2.2 Graphical models

Throughout this thesis, the relationship between variables is studied using graphical models. These models enable the decomposition of complex multivariate distributions into simpler local distributions. Such a decomposition can reveal a great deal about the relationships between the variables. In addition, a graphical model provides a statistical and computationally tractable description of a large joint distribution.

The decomposition is formed by the conditional independence structure, which can be represented by a graph. Thus, graphical models describe families of probability distributions using a mathematical graph. The graphical representation can ease the interpretation and clarify the structure of complex models. In addition, in some situations, the computation of particular marginal distributions can be simplified when a graphical representation is considered (see e.g. Lauritzen and Spiegelhalter, 1988).

In most graphical models, the nodes of the graph represent random variables and the edges represent the (conditional) dependence structure amongst the random variables. The conditional independence structure gives a deeper understanding of the relationships between the random variables, as we describe below.

A variety of graphical models have been developed (see e.g. Lauritzen, 1996; Smith, 2010). The two most widely used graphical models are Markov random field models, which are represented by an undirected graph, and Bayesian networks, which are represented by directed, acyclic graphs (DAGs). This thesis focuses on the latter

model.

2.2.1 Conditional independence

While a crude understanding of the relationship between random variables is provided by a simple correlation analysis, a far deeper understanding is provided by the conditional independence structure. In particular, correlation analysis gives no understanding of whether relationships between two variables are mediated by a third, whereas this is captured in the conditional independence structure. Such knowledge is generally informative, and indeed much of statistics can be considered in terms of conditional independence (Dawid, 1979). Knowledge of the conditional independence structure is particularly valuable when a loose form of causality is sought.

Two random variables A and B are conditionally independent given a third random variable C if the following property holds.

$$p(A, B \mid C) = p(A \mid C)p(B \mid C) \qquad \text{for all } C \text{ such that } p(C) > 0$$

When this property holds, we use the shorthand $A \perp\!\!\!\perp B \mid C$. We write $A \not\perp\!\!\!\perp B \mid C$ when the property does not hold.

2.2.2 Graphs

A mathematical GRAPH $G = (V, E)$ consists of a set of NODES $V = (1, \ldots, p)$, and a set of EDGES E that link pairs of nodes. We also use $v_1, \ldots, v_p$ to denote the nodes in the graph.

The edges may be DIRECTED, in which case $E \subseteq V \times V$, or UNDIRECTED, in which case E consists of unordered pairs of nodes. We will mostly consider directed graphs,

and will use three different notations to specify their edges: the collection of edges; adjacency matrices; and parent sets.

First, we can specify edges of a graph as a subset $E \subseteq V \times V$, as per the definition of a graph. Individual directed edges from node i to node j can thus be denoted by either the pair (i, j), or the symbol $i \to j$. We will refer to i as the HEAD of the edge, and j as the TAIL.

We can also specify the graph $G = (V, E)$ with an ADJACENCY MATRIX G, a $p \times p$ matrix with elements G_{ij} given by

$$G_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

The final specification of the edge set E of the graph G that we use is in terms of the parents G_j of each node j, for $j \in \{1, \ldots, p\}$. The PARENTS G_j of node j are the subset of nodes V such that $i \in G_j \Leftrightarrow (i, j) \in E$. We refer to G_j as a PARENT SET.

It will sometimes be convenient to use the collection of parent sets $\langle G_1, \ldots, G_p \rangle$ to specify a graph G. Subsets thereof are denoted by $G_A = \langle G_k : k \in A \rangle$. Thus G_A is a collection of parent sets, specifying only the parent sets of nodes in A; the parent sets of nodes not in A are not specified by G_A. The subset given by the complement $A^C = \{1, \ldots, p\} \setminus A$ of a set A is denoted by $G_{-A} = \langle G_k : k \in A^C \rangle$. Thus G_{-A} specifies the parent sets of nodes not in A, leaving the parent sets of nodes in A unspecified. The parent sets of all nodes can be specified by $\langle G_A, G_{-A} \rangle$. Thus, in particular, any graph G can be specified as $\langle G_j, G_{-j} \rangle = \langle G_1, \ldots G_p \rangle = G$ for any $j \in \{1, \ldots, p\}$.

A PATH on a graph from a node $j \in V$ to a node $k \in V$ is a sequence of nodes $j = v_0, v_1, \ldots, v_d = k$, $d \in \mathbb{N}$, such that an edge exists linking $v_{i-1} \in V$ and $v_i \in V$

for each $i = 1, \ldots, d$. In a directed graph, we usually require that $(v_{i-1}, v_i) \in E$ meaning that the path obeys the directions of the edges. However, it will occasionally be useful to consider a path that does not obey the direction of the edges on a path.

Cycles are a particular type of path that will be of key interest. A CYCLE is path v_0, $v_1, \ldots, v_d$, $d \in \mathbb{N}$, such that $v_0 = v_d$, and the path obeys the direction of the edges. A graph G in which a cycle exists is called CYCLIC; a graph without cycles is called ACYCLIC. We denote the set of all directed, acyclic graphs (DAGs) with p nodes by $\mathcal{G}$.

For undirected graphs, we denote by $i - j$ an edge between node i and node j. We define $\text{adj}(i)$ as the set of nodes j such that $i - j$. A COMPLETE UNDIRECTED GRAPH is an undirected graph in which an edge links every pair of nodes in the graph.

2.2.3 Bayesian networks

Bayesian networks are a particular type of graphical model. A Bayesian network G is a DAG with nodes $V = (1, \ldots, p)$, and directed edges $E \subset V \times V$. The nodes correspond to the components of the random variables $X_1, \ldots, X_p$. We denote by $\mathbf{X}_{G_j}$ the set of random variables that correspond to the parents G_j of node j in the graph G. It is convenient to refer to $\mathbf{X}_{G_j}$ as the PARENTS OF X_j.

A defining feature of Bayesian networks is that the joint distribution of $\mathbf{X}$ is specified in terms of $p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$, the conditional distribution of each X_i, given its parents $\mathbf{X}_{G_i}$ in the Bayesian network, with parameters θ_i. Denoting by $\mathbf{X}_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p\}$ the random vector excluding X_i, we have LOCAL MODELS (or LOCAL DISTRIBUTIONS) $p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$ that satisfy the following.

$$p(X_i \mid \mathbf{X}_{-i}, \theta_i) = p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$$

The complete joint distribution of $X_1, \ldots, X_p$, given the Bayesian network G, is the

product of these local distributions.

$$p(X_1, \ldots, X_p \mid G) = \prod_{i=1}^{p} p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$$

The conditional dependence structure of the probability distribution can be determined using the d-separation criterion (Verma and Pearl, 1990). We describe this criterion using the concept of blocked paths, which uses the concept of a path being head-to-head at a node. We say that a path $v_0, \ldots, v_d$ (not necessarily obeying edge directions) on a DAG $G = (V, E)$ is HEAD-TO-HEAD at a node v_i, for some $i \in \{1, \ldots, d-1\}$ if $(v_{i-1}, v_i) \in E$ and $(v_{i+1}, v_i) \in E$.

A blocked path can then be defined as follows. Let S be a subset of V so that S is a set of nodes in the graph. A path (not necessarily obeying edge directions) from node $v_0 \in V$ to node $v_d \in V$ in a DAG G is said to be BLOCKED by S if the path includes a node v_i, $i \in \{1, \ldots, d-1\}$, such that one of the following two criteria is satisfied.

- $v_i \in S$ and the path from node v_0 to node v_d is not head-to-head at node v_i.
- The path is head-to-head at node v_i, and neither is node v_i in S, nor does S contain any of the descendants in the graph G of node v_i.

Two subsets $A, B \subseteq V$ are d-separated if all the paths (not necessarily obeying edge directions) from A to B are blocked.

A particular conditional independence structure can be implied by multiple different Bayesian networks. However, we can define an equivalence class on the space of Bayesian networks such that Bayesian networks within the same class imply the same conditional independence structure. The definition arises from the definition of d-separation and uses the concept of the skeleton of a Bayesian network, and of v-structures. The SKELETON of a Bayesian network is the undirected graph formed by removing the directions attached to the directed edges in the network. A v-structure

is defined as an ordered triple (i, j, k) of nodes, such that $i \to j$ and $k \to j$, but no edge exists linking nodes i and k directly. Two Bayesian networks are EQUIVALENT if they share the same skeleton and v-structures (Verma and Pearl, 1990).

We can specify the equivalence class of a Bayesian network using a completed partially-directed acyclic graph (CPDAG). This name originates in Chickering (2002), but the idea has been used by other authors under a variety of names. A CPDAG is a partially-directed graph, whose directed edges do not form a cycle. For a Bayesian network G, CPDAG(G) is formed by considering all of the edges E' for which in all graphs G', such that G and G' are equivalent, that edge is oriented as in G. Then CPDAG(G) is formed by removing the direction attached to each edge not in E'. Chickering (2002) show that CPDAGs uniquely represent an equivalence class of Bayesian networks.

2.3 Univariate Bayesian models

The basic building-blocks of the models that we consider in Section 2.4 and 2.5 are simple univariate models. In this section, we first describe conjugacy, a property that characterises a class of analytically-tractable models. We then review the simplest form of the two conjugate models that are considered throughout this thesis.

We assume that $\mathbf{Y}$ is an n-dimensional random vector consisting of independent, identically distributed components. We suppose observations $\mathbf{y}$ of $\mathbf{Y}$ are available.

2.3.1 Conjugate priors

The integration required to evaluate Equation 2.2 is analytically intractable for many choices of priors for a given model. If our understanding is such that our prior needs to take a form for which the integration is intractable, numerical methods of evaluating the integral will be necessary. However, if our prior has a form close to

a prior for which the integration is straightforward, this difficulty can be avoided. For the models that we consider in this thesis, priors of the required form are well known, and are called CONJUGATE PRIORS.

Conjugate priors (Raiffa and Schlaifer, 1961) are families $\mathcal{P}$ of distributions that are closed under sampling from a distribution in a family $\mathcal{F}$ of distributions. A family $\mathcal{P}$ of prior distributions is said to be closed under sampling from a distribution $p(y \mid \theta)$ in a parametric family $\mathcal{F}$ if for every prior distribution $\pi(\theta) \in \mathcal{P}$, the posterior distribution $p(\theta \mid y) \propto \pi(\theta)p(y \mid \theta)$ is also in $\mathcal{P}$. A catalogue of many conjugate priors is given in Gelman *et al.* (2004).

Raiffa and Schlaifer (1961) list three properties that they view as desirable in a family of priors: tractability, interpretability and richness. Conjugate families are tractable, and this is the main reason for their adoption. Conjugate priors sometimes also have a simple interpretation. In exponential families we can consider the prior as constituting "virtual samples" (see, e.g. Robert, 2007), and so the relative weight implied on the prior and data can be ascertained. It is in richness, however, that conjugate families can be lacking. Ideally, a prior should exactly match a Bayesian modeller's prior beliefs, but conjugate priors are often not flexible enough to allow this to be fully achieved. Sometimes, a close approximation to prior beliefs can be constructed within the conjugate family, but often a poor approximation is accepted because of the computational advantages of conjugate priors.

In many standard Bayesian models, using non-conjugate priors is now feasible since the emergence of easily available computationally-intensive approximations. However, in the setting considered here, non-conjugate priors are not viable for the following reasons.

First, there are formidable computational challenges even when conjugate priors are used. These challenges are considerably compounded by the use of non-conjugate priors. Additionally we will be exclusively considering settings in which the sample

size of the data is large. The large sample size means that the prior will exert only a minimal effect on the posterior distribution, thus making its exact specification less important.

For these reasons, we use conjugate priors throughout.

2.3.2 Multinomial-Dirichlet

The standard Bayesian model for univariate multinomial data (e.g. O'Hagan and Forster, 2004) will form the basis of the models we consider in this thesis. Consider a random vector $\mathbf{Y}$, each component of which takes one of r discrete categories. Suppose that $\mathbf{Y}$ is distributed according to a multinomial distribution, with parameter vector $\theta = (\theta_1, \ldots, \theta_r)$, with $\theta > 0$ and $\theta_1 + \cdots + \theta_r = 1$.

$$\mathbf{Y} \sim \text{Mult}(\theta_1, \ldots, \theta_r)$$

The conjugate prior for the vector θ is Dirichlet, with hyperparameters $\alpha = (\alpha_1, \ldots, \alpha_r)$ where $\alpha_k > 0$, $k = 1, \ldots r$.

$$\theta_1, \ldots, \theta_r \sim \text{Dir}(\alpha_1, \ldots, \alpha_r) \qquad \text{with } \theta_1, \ldots, \theta_r \geq 0 \text{ and } \sum_{k=1}^{r} \theta_k = 1$$

The normalising factor in the Dirichlet likelihood is a ratio of gamma functions $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x$, for which, in particular, $\Gamma(\alpha) = (\alpha - 1)!$ for $\alpha \in \mathbb{N}$.

$$p(\theta_1, \ldots, \theta_r) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_r)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_r)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1}$$

The mean is $\alpha_k (\sum_{k=1}^{r} \alpha_k)^{-1}$ for each θ_k.

The posterior distribution of θ is parameterised in terms of a contingency table constructed from the observations $\mathbf{y}$, such that n_k is the number of observations in

the k^{th} category, $k = 1, \ldots, r$.

$$\theta_1, \ldots, \theta_r \mid \mathbf{y} \sim \text{Dir}(\alpha_1 + n_1, \ldots, \alpha_r + n_r)$$

The formulation simplifies in the natural manner for binomial data with beta priors.

In using this formulation, we are assuming that the data are independent, identically-distributed draws from a multinomial distribution. It will often be the case that some heterogeneity exists and so it is more appropriate to use a model that is conditional on some collection of covariates; we consider this possibility in Section 2.4.1.

2.3.3 Normal inverse-gamma

The models for normally-distributed data that we consider will similarly build upon standard univariate models (e.g. Gelman *et al.*, 2004). Suppose we have a random vector $\mathbf{Y}$, components of which are independent random variables distributed according to a normal distribution, with mean μ and variance σ^2.

$$\mathbf{Y} \sim N(\mu, \sigma^2)$$

When both μ and σ^2 are unknown, the conjugate priors for μ and σ^2 are normal and inverse-gamma respectively.

$$\mu \mid \sigma^2 \sim \text{N}(m, v^{-1}\sigma^2)$$
$$\sigma^2 \sim \text{IG}(a, b)$$

The hyperparameters a and b are respectively the shape and scale parameters of the inverse-gamma distribution. The hyperparameters m can be interpreted as the prior mean, and v is inversely proportional to the prior variance. The inverse-gamma

distribution has density

$$\pi(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp(-b/\sigma^2).$$

The joint prior for (μ, σ^2) is thus normal inverse-gamma $\text{NIG}(m, v, a, b)$.

$$\pi(\mu, \sigma^2) = \frac{\sqrt{v}}{\sigma\sqrt{2\pi}}\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp\left(-\frac{2b + v(\mu - m)^2}{2\sigma^2}\right)$$

By conjugacy, the joint posterior distribution for (μ, σ^2) is also normal inverse-gamma.

$$\mu, \sigma^2 \mid \mathbf{y} \sim \text{NIG}(m^\star, v^\star, a^\star, b^\star)$$

where, with $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2$, the parameters are

$$\begin{aligned} m^\star &= \frac{mv + n\bar{y}}{n + v} \\ v^\star &= \frac{1}{n + v} \\ a^\star &= a + \frac{n}{2} \\ b^\star &= b + \frac{1}{2}\left(s^2 + \frac{nv(\bar{x} - m)^2}{n + v}\right). \end{aligned}$$

2.4 Model selection for Bayesian regression models

Regression models aim to characterise the relationship between a response variable and a collection of predictor variables. The model for the response is specified conditionally on the predictor variables. We consider situations in which the parametric form of the conditional distribution is known, up to the choice of predictor variables. Model uncertainty in this context is therefore uncertainty about which set of predictor variables should be used. The problem is known as VARIABLE SELECTION.

We assume that the observations consist of a n-dimensional random vector $\mathbf{y}$ of 'out-

come' values, and a $n \times p$ random matrix $\mathbf{x}$ of observations of p predictor variables. These observations come from a random vector $\mathbf{Y}$, and a $n \times p$ random matrix $\mathbf{X}$, the columns of which are random vectors $X_1, \ldots, X_p$, respectively.

We aim to determine which subset of the predictors $\{X_1, \ldots, X_p\}$ is best suited to predicting $\mathbf{Y}$, the outcome variable. There are 2^p subsets of the p predictors, each of which corresponds to a possible model for $\mathbf{Y}$. A regression model M_γ is specified using a p-dimensional indicator vector $\gamma = (\gamma_1, \ldots, \gamma_p)$, the i^{th} component of which takes the value 1 when the i^{th} variable is included in the model, for $i = 1, \ldots p$. Let $\mathbf{X}_\gamma = \{X_i : \gamma_i = 1\}$ be the set of predictor variables included in model M_γ, and $p_\gamma = \sum_{i=1}^{p} \gamma_i$ be the number of predictors included in the model. We use $\mathcal{M}_\gamma$ to refer to the set of all models. Note that the predictor variables are assumed to be observed without error.

2.4.1 Multinomial-Dirichlet

Suppose each component of the response vector $\mathbf{Y}$ has r levels, or categories. We will be considering models for the response $\mathbf{Y}$ specified to be conditional on a subset $\mathbf{X}_\gamma$ of the set of discrete variables $\{X_1, \ldots, X_p\}$. Each component of X_i, for $i = 1, \ldots, p$, has r_i levels. We define the CONFIGURATIONS of $\mathbf{X}_\gamma$ to be the components of its sample space, for which $q_\gamma = \prod_{i=1}^{p} r_i^{\gamma_i}$ is the cardinality. We label the configurations C_j^γ for $j = 1, \ldots, q_\gamma$. We assume that observations $\mathbf{y}$ and $\mathbf{x}$ for the outcome random vector $\mathbf{Y}$ and the predictor random variables $\mathbf{X}$ respectively are available; we denote by $\mathbf{x}_\gamma$ the observations for the predictors included in model M_γ.

For a particular model M_γ, we assume that the distribution of $\mathbf{Y}$ is independently parameterised for different configurations of $\mathbf{X}_\gamma$, the predictors in the model. Thus the parameter space Θ_γ of the distribution of $\mathbf{Y}$ under model M_γ can be broken into

smaller parameter spaces $\Theta_{\gamma,j}$ corresponding to the configurations of the predictors.

$$\Theta_\gamma = \bigtimes_{j=1,\ldots,q_\gamma} \Theta_{\gamma,j}$$

Thus the likelihood factorises across configurations.

$$p(\mathbf{Y} \mid \mathbf{X}_\gamma, \theta_{\gamma,j}, M_\gamma) = \prod_{j=1}^{q_\gamma} p(\mathbf{Y} \mid \mathbf{X}_\gamma = C_j^\gamma, \theta_{\gamma,j}, M_\gamma) \qquad \text{with } \theta_{\gamma,j} \in \Theta_{\gamma,j}$$

This independence assumption means that no information is 'shared' about the distribution of Y between cases in which the configuration of the predictors differ, and the model may be entirely different for different configurations. In particular, linearity in the predictors is not a requirement for the fitted model. This unstructured form of model means that availability of a large sample size is important for useful inference to be possible.

The distribution of $\mathbf{Y}$ conditional on the configuration C_j^γ of the predictors is specified to be multinomial for each configuration, with $j = 1, \ldots, q_\gamma$, and with an r-dimensional parameter vector $\theta_{\gamma,j} \in \Theta_\gamma$, each component of which corresponds to a category of $\mathbf{Y}$.

$$\mathbf{Y} \mid \mathbf{X}_\gamma, \theta_{\gamma,j}, M_\gamma \sim \text{Mult}(\theta_{\gamma,j,1}, \ldots, \theta_{\gamma,j,r}) \qquad \text{for } j = 1, \ldots, q_\gamma$$

The likelihood for $\mathbf{y}$ under a model M_γ is a function of the random variable $N_{\gamma,j,k}$ given by the number of times that the predictors take the j^{th} configuration C_j^γ and the outcome variable has the k^{th} category, for $j = 1, \ldots, q_\gamma$ and $k = 1, \ldots, r$.

$$p(\mathbf{y} \mid \mathbf{X}_\gamma, \theta_\gamma, M_\gamma) = \prod_{j=1}^{q_\gamma} \prod_{k=1}^{r} \theta_{\gamma,j,k}^{N_{\gamma,j,k}}$$

The conjugate prior distribution for $\theta_{\gamma,j,1}, \ldots, \theta_{\gamma,j,r} \mid M_\gamma$ is Dirichlet, for each $j =$

$1, \ldots, q_\gamma$.

$$\theta_{\gamma,j,1}, \ldots, \theta_{\gamma,j,r} \sim \mathrm{Dir}(\alpha_{\gamma,j,1}, \ldots, \alpha_{\gamma,j,r}) \qquad \text{with } \theta_{\gamma,j,1}, \ldots, \theta_{\gamma,j,r} \geq 0, \sum_{k=1}^{r} \theta_{\gamma,j,k} = 1$$

We assume that $\theta_{\gamma,j}$ are *a priori* independent. This assumption, when taken with the assumption that the distribution of $\mathbf{y}$ is independently parameterised, is called LOCAL INDEPENDENCE (Spiegelhalter and Lauritzen, 1990). The joint prior for $(\theta_{\gamma,1}, \ldots, \theta_{\gamma,q_\gamma})$ is thus the product of $\mathrm{Dir}(\alpha_{\gamma,j,1}, \ldots, \alpha_{\gamma,j,r})$ distributions.

$$\pi(\theta_{\gamma,1}, \ldots, \theta_{\gamma,q_\gamma} \mid M_\gamma) = \prod_{j=1}^{q_\gamma} \frac{\Gamma(\alpha_{\gamma,j,1} + \cdots + \alpha_{\gamma,j,r})}{\Gamma(\alpha_{\gamma,j,1}) \ldots \Gamma(\alpha_{\gamma,j,r})} \prod_{k=1}^{r} \theta_{\gamma,j,k}^{\alpha_{\gamma,j,k}-1}$$

For each $j = 1, \ldots, q_\gamma$, unless otherwise stated, we take the hyperparameters $\alpha_{\gamma,j,k} = (r_i q_\gamma)^{-1}$ for all $k = 1, \ldots, r$, following Buntine (1991) and Heckerman *et al.* (1995).

Given observations $n_{\gamma,j,k}$ of the contingency table random variables $N_{\gamma,j,k}$, formed from observations $\mathbf{y}$ and $\mathbf{x}$, the posterior distribution for each $\theta_{\gamma,j}$ is $\mathrm{Dir}(\alpha_{j,1} + n_{\gamma,j,1}, \ldots, \alpha_{j,r} + n_{\gamma,j,r})$, for each $j = 1, \ldots, q_\gamma$. Defining the collection of counts $n_\gamma = \{n_{\gamma,j,k} : j = 1, \ldots, q_\gamma \text{ and } k = 1, \ldots, r\}$ under a model M_γ, and the corresponding collection of hyperparameters $\alpha_\gamma = \{\alpha_{j,k} : j = 1, \ldots, q_\gamma \text{ and } k = 1, \ldots, r\}$, the marginal likelihood can be written in closed-form.

$$p(\mathbf{y} \mid M_\gamma, n_\gamma, \alpha_\gamma) = \prod_{j=1}^{q_\gamma} \frac{\Gamma(\alpha_{\gamma,j,1} + \cdots + \alpha_{\gamma,j,r})}{\Gamma(\sum_{k=1}^{r} n_{\gamma,j,k} + \sum_{k=1}^{r} \alpha_{\gamma,j,k})} \prod_{k=1}^{r} \frac{\Gamma(n_{\gamma,j,k} + \alpha_{\gamma,j,k})}{\Gamma(\alpha_{\gamma,j,k})}$$

2.4.2 Linear regression

The second model we consider is for a normally-distributed random variable $\mathbf{Y}$ taking values in $\mathbb{R}$. We again assume that $\mathbf{Y}$ is dependent on a subset $\mathbf{X}_\gamma$ of random variables $\{X_1, \ldots, X_p\}$ that defines the model M_γ, but we now assume that these variables are continuous. The strong independence assumptions between the

different configurations of $\mathbf{X}_\gamma$ that we made in the previous section do not translate sensibly into a continuous setting. Instead we make the usual assumption that $p(\mathbf{Y} \mid \mathbf{X}_\gamma)$ is a smooth function of $\mathbf{X}_\gamma$, and in particular assume that the expectation of $\mathbf{Y}$ is a linear function in the model parameters.

In linear regression settings, it is convenient to redefine $\mathbf{X}_\gamma$ to be the $n \times (p_\gamma + 1)$ design matrix. All of the linear regressions that we consider include an intercept term, and so we include a column of 1s in the design matrix.

$$\mathbf{X}_\gamma = \left[\begin{array}{cccc} \mathbf{1} & \mathbf{X}_{\gamma_1} & \ldots & \mathbf{X}_{\gamma_{p_\gamma}} \end{array} \right]$$

We assume the rank of $\mathbf{X}_\gamma$ is $p_\gamma + 1$.

The normal linear regression model for $\mathbf{Y}$ is specified conditional on $\mathbf{X}_\gamma$ for a model M_γ.

$$\mathbf{Y} \mid \beta, \sigma, \mathbf{X}_\gamma, M_\gamma \sim \mathrm{MVN}(\mathbf{X}_\gamma \beta, \sigma^2 \mathbb{I}_{p_\gamma})$$

Normal inverse-gamma

The general joint conjugate prior for $\beta, \sigma \mid M_\gamma$ is normal inverse-gamma (e.g. O'Hagan and Forster, 2004). Let $\mathbf{m}_\gamma$ be the prior mean for the regression coefficients β, and $\sigma^2 \mathbf{V}_\gamma$ their prior variance.

$$\beta \mid \sigma^2, \mathbf{X}_\gamma, M_\gamma \sim \mathrm{MVN}(\mathbf{m}_\gamma, \sigma^2 \mathbf{V}_\gamma) \qquad \sigma^2 > 0$$

$$\sigma^2 \sim \mathrm{IG}(a, b) \qquad a, b > 0$$

As before, a and b are the shape and scale parameters of an inverse-gamma distribution. The joint prior for $\beta, \sigma^2 \mid M_\gamma$ is thus NIG$(\mathbf{m}_\gamma, \sigma^2\mathbf{V}_\gamma, a_\gamma, b_\gamma)$.

$$\begin{aligned}\pi(\beta, \sigma^2 \mid \mathbf{X}_\gamma, M_\gamma) = &\frac{1}{(2\pi\sigma^2)^{(p_\gamma+1)/2}\,|\mathbf{V}_\gamma|^{1/2}} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mathbf{m}_\gamma)^T\mathbf{V}_\gamma^{-1}(\beta - \mathbf{m}_\gamma)\right\} \\ &\times \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)}(\sigma^2)^{-(a_\gamma+1)}\exp(-b_\gamma/\sigma^2)\end{aligned}$$

The posterior distribution for $\beta, \sigma^2 \mid M_\gamma$ is NIG$(\mathbf{m}_\gamma^\star, \mathbf{V}_\gamma^\star, a_\gamma^\star, b_\gamma^\star)$, with parameters defined as follows.

$$\begin{aligned}\mathbf{V}_\gamma^\star &= (\mathbf{V}_\gamma^{-1} + \mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1} \\ \mathbf{m}_\gamma^\star &= (\mathbf{V}_\gamma^{-1} + \mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}(\mathbf{V}_\gamma^{-1}\mathbf{m}_\gamma + \mathbf{X}_\gamma^T\mathbf{y}) \\ a_\gamma^\star &= a_\gamma + n/2 \\ b_\gamma^\star &= b_\gamma + \{\mathbf{m}_\gamma^T\mathbf{V}_\gamma^{-1}\mathbf{m}_\gamma + \mathbf{y}^T\mathbf{y} - (\mathbf{m}_\gamma^\star)^T(\mathbf{V}_\gamma^\star)^{-1}\mathbf{m}_\gamma^\star\}/2\end{aligned} \tag{2.3}$$

The marginal likelihood takes the following form.

$$p(\mathbf{y} \mid M_\gamma) = \frac{\left|\mathbf{V}_\gamma^\star\right|^{1/2} b_\gamma^{a_\gamma}\Gamma(a_\gamma^\star)}{|\mathbf{V}_\gamma|^{1/2}\,\pi^{n/2}\Gamma(a_\gamma)}(b_\gamma^\star)^{-a_\gamma^\star}$$

It can be difficult to specify the hyperparameters of the normal inverse-gamma formulation, particularly the matrix $\mathbf{V}_\gamma$ of prior variances between the coefficients, and so we turn to a special form of the normal inverse-gamma formulation.

Zellner g-prior

The Zellner g-prior (Zellner, 1986) specification for a Bayesian linear model is a special case of the normal inverse-gamma formulation that is easier to specify. The specification has been widely used (Smith and Kohn, 1996; Fernández *et al.*, 2001b)

and discussed (e.g. Laud and Ibrahim, 1995; Fernández *et al.*, 2001a; Clyde and George, 2004). The g-prior takes the following form.

$$\beta \mid \sigma^2, \mathbf{m}_\gamma, \mathbf{X}_\gamma, M_\gamma \sim \text{MVN}(\mathbf{m}_\gamma, g\sigma^2(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1})$$
$$\pi(\sigma^2) \propto \sigma^{-2}$$

The prior for σ^2 is the improper Jeffrey's prior.

The prior for $\beta \mid \sigma^2, M_\gamma$ has mean $\mathbf{m}_\gamma$, and variance that depends on the predictor variables $\mathbf{X}$. At first glance, this suggests that the prior is dependent on the data. However, because we assume that the predictor variables are observed without error, the dependence is only on a part of the structure of the data that we assume is 'fixed'. The procedure is thus not an empirical Bayes estimator.

The term $\mathbf{X}^T\mathbf{X}$ is the sample second moment of the predictors, and so the prior variance of β is greater if the observations are close together, as is natural. Dividing by the second moment also has the advantage of making the prior invariant to scale.

Using a prior variance of this form has the additional benefit of combining naturally with the posterior variance (2.3), which enables the parameter g to be interpreted as determining the relative weight assigned to the information in the prior and the sample. For example, taking $g = 1$ assigns equal weight to the prior information and the sample, whereas $g = n$ assigns equal weight to the prior information and a single unit of the sample. The latter specification is particularly widely used (e.g. Kass and Wasserman, 1995). Alternatives include calibrating g by an empirical Bayes procedure (George and Foster, 2000). These and other choices are reviewed and empirically compared by Liang *et al.* (2008).

The relevant posterior distributions and the marginal likelihood are defined in terms

of MLE $\hat{\beta}$ for the regression coefficients and the residual sum of squares s^2.

$$\hat{\beta} = (\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^T \mathbf{y}$$
$$s^2 = (\mathbf{y} - \mathbf{X}_\gamma \hat{\beta})^T (\mathbf{y} - \mathbf{X}_\gamma \hat{\beta})$$

Given these definitions, the posterior distributions for β and σ^2 are as follows under a g-prior specification.

$$p(\beta \mid \mathbf{y}, \mathbf{m}_\gamma, \mathbf{X}_\gamma, M_\gamma) \sim \text{MVN}\left(\frac{g}{g+1}(\mathbf{m}_\gamma / g + \hat{\beta}),\ \frac{\sigma^2 g}{g+1}(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}\right)$$
$$p(\sigma \mid \mathbf{y}, \mathbf{m}_\gamma, \mathbf{X}_\gamma, M_\gamma) \sim \text{IG}\left(\frac{n}{2},\ \frac{s^2}{2} + \frac{1}{2(g+1)}(\mathbf{m}_\gamma - \hat{\beta})^T \mathbf{X}_\gamma^T \mathbf{X}_\gamma (\mathbf{m}_\gamma - \hat{\beta})\right)$$

The marginal likelihood for a model M_γ takes the following form when a g-prior is chosen.

$$p(\mathbf{y} \mid M_\gamma) \propto (g+1)^{-(p_\gamma+1)/2} \times \left(\mathbf{y}^T \mathbf{y} - \frac{g}{g+1}\mathbf{y}^T \mathbf{X}_\gamma (\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^T \mathbf{y} - \frac{1}{g+1}\mathbf{m}_\gamma^T \mathbf{X}_\gamma^T \mathbf{X}_\gamma \mathbf{m}_\gamma\right)^{-n/2}$$

This marginal likelihood has a closed-form and is straightforward to compute.

2.4.3 Model priors

The model prior should be a discrete distribution over the 2^p possible regression models. A simple choice is a uniform distribution over the models.

$$\pi(M_\gamma) = \frac{1}{|\mathcal{M}_\gamma|} \qquad \text{for } M_\gamma \in \mathcal{M}_\gamma$$

This choice is widely used (e.g. Smith and Kohn, 1996; Raftery *et al.*, 1997), but has some unsatisfactory attributes that we discuss in Chapter 7.

2.4.4 Posterior distribution over models

The posterior distribution $P(M_\gamma \mid \mathbf{X})$ is a discrete distribution over these models.

$$P(M_\gamma \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid M_\gamma)\pi(M_\gamma)}{\sum_{M_\gamma \in \mathcal{M}_\gamma} p(\mathbf{X} \mid M_\gamma)\pi(M_\gamma)}$$

The maximum *a posteriori* model is a single model M_γ^{MAP} from $\mathcal{M}_\gamma$. In contrast, model averaging reflects an aspect of the complete posterior distribution. For example, we might consider the inclusion probabilities, defined for each predictor X_i as the posterior probability that X_i is in the model for $\mathbf{y}$.

2.5 Model selection for Bayesian networks

The aim of model selection for Bayesian networks is to understand the dependence structure of the random variables. We will consider on an equal footing all of the random variables for which we have observations. We thus assume we simply have a $n \times p$ matrix of independent observations of p variables, which are from a random vector $\mathbf{X} = (X_1, \ldots, X_p)$. Note that we refer to a particular subset of the random vector $\mathbf{X}$ by $\mathbf{X}_A$ for a set $A \subseteq \{1, \ldots, p\}$.

We wish to determine which Bayesian network G best describes the joint distribution of the random variables $X_1, \ldots, X_p$. Recall that we refer to the set of possible Bayesian networks with p nodes by $\mathcal{G}$, which is a finite set, each member of which identifies a family of models for $X_1, \ldots, X_p$.

The models that we consider for the local distributions in the Bayesian networks are straightforward generalisations of those that we consider for Bayesian variable

selection. Rather than regarding a single random variable as the response and the remainder as predictors, we now consider all of the random variables $X_1, \ldots X_p$ as responses, and, for each random variable, we consider which of the remaining variables should be selected as predictors. The local model for a variable X_i is thus specified conditionally on its parents $\mathbf{X}_{G_i}$ in the Bayesian network. The complete model must be a Bayesian network; in particular, it must correspond to an acyclic directed graph.

We aim to make inference about this using statistical model selection. These methods have been widely adopted in molecular biology (Husmeier, 2003; Friedman, 2004; Needham *et al.*, 2007; Mukherjee and Speed, 2008), and have been used in some areas of medical sciences (Acid *et al.*, 2004). In this chapter, we focus on the structure of the model, as given by the graph. The structure of the model suggests how the different components of the system interact, which may be helpful in understanding the system as a whole.

2.5.1 Independence assumptions

We assume that the parameters $\theta_G \in \Theta_G$ for the complete Bayesian network G can be broken into components $\theta_{G,i} \in \Theta_G$ corresponding to the individual random variables.

$$\Theta_G = \bigtimes_{i=1}^{p} \Theta_{G,i}$$

In addition, we assume that the $\theta_{G,i}$, $i = 1, \ldots, p$, are *a priori* independent for a particular Bayesian network G. This assumption is called GLOBAL INDEPENDENCE (Spiegelhalter and Lauritzen, 1990).

We also assume PARAMETER MODULARITY. This assumptions states that if a random variable X_i has the same parents in two Bayesian networks $G^{(1)}$ and $G^{(2)}$, then

the priors for the parameters $\theta_{G^{(1)},i}$ and $\theta_{G^{(2)},i}$ are the same.

$$p\left(\theta_{G^{(1)},i}\right) = p\left(\theta_{G^{(2)},i}\right)$$

This assumption is normally sensible, except in settings in which detailed prior information is available that suggests that dependence exists between the parameter prior of a node and the structure of the graph beyond the immediate parents of the node. This scenario is different to the focus of this thesis: if such detailed information is available, the analysis is considerably less exploratory than the settings we consider. In addition, typically only when there are a small number of variables under consideration is it practical to elicit such a prior. To make no assumption about the equality of prior parameters would necessitate specifying the prior parameters for each parameter under every possible Bayesian network; with a large number of variables this is impractical.

2.5.2 Multinomial-Dirichlet

Consider a particular Bayesian network G. We label the configurations of $\mathbf{X}_{G_i}$ for a random variable X_i by $C_j^{G,i}$, with $j \in \{1, \ldots, q_{G,i}\}$, where $q_{G,i}$ is the number of configurations of the parents in G of X_i.

As in multinomial regression, we assume that the parameters of the multinomial models for different configurations are independent. In addition to this local independence assumption, we make the global independence assumption described in Section 2.5.1. Together these mean that we can describe separately the models for each node, and for each configuration of the parents of that node. Each of these models, we assume, has the following form, with $j = 1, \ldots, q_{G,i}$.

$$X_i \mid \theta_{G,i,j}, \mathbf{X}_{G_i} = C_j^{G,i} \sim \text{Mult}(\theta_{G,i,j,1}, \ldots, \theta_{G,i,j,r_i})$$

Thus, the conditional distribution of each X_i has the following form.

$$p\left(X_i \mid \theta_{G,i,j}, \mathbf{X}_{G_i} = C_j^{G,i}\right) = \theta_{G,i,j,k} \qquad \text{for } k = 1, \ldots, r_i$$

We denote by θ_G the collection of all $\theta_{G,i,j,k}$ for $i = 1, \ldots, p$, $j = 1, \ldots, q_{G,i}$ and $k = 1, \ldots, r_i$. Let $N_{G,i,j,k}$ be the cells of a contingency table for $\mathbf{X}$ that counts the number of X_i in the i^{th} category when $\mathbf{X}_{G_i} = C_j^{G,i}$, for $i = 1, \ldots, p$, $j = 1, \ldots, q_{G,i}$, $k = 1, \ldots, r_i$. The joint distribution of $\mathbf{X}$ is thus

$$p(\mathbf{X} \mid G, \theta_G) = \prod_{i=1}^{p} \prod_{j=1}^{q_{G,i}} \prod_{k=1}^{r_i} \theta_{G,i,j,k}^{N_{G,i,j,k}}.$$

The assumptions of local and global independence mean that $\theta_{G,i,j}$ are assumed *a priori* independent. For each $i = 1, \ldots, p$ and $j = 1, \ldots, q_{G,i}$, the prior for $\theta_{G,i,j}$ is the conjugate Dirichlet prior.

$$\theta_{G,i,j,1}, \ldots, \theta_{G,i,j,r_i} \mid G \sim \text{Dir}(\alpha_{G,i,j,1}, \ldots, \alpha_{G,i,j,r_i})$$

The joint prior for θ_G is thus the product of Dirichlet distributions, with $\theta_{G,i,j,k} \geq 0$ and $\sum_{k=1}^{r_i} \theta_{G,i,j,k} = 1$ for all $i = 1, \ldots, p$ and $j = 1, \ldots, q_{G,i}$.

$$p\left(\theta_G\right) = \prod_{i=1}^{p} \prod_{j=1}^{q_{G,i}} \frac{\Gamma(\alpha_{G,i,j,k} + \cdots + \alpha_{G,i,j,r_i})}{\Gamma(\alpha_{G,i,j,k}) \ldots \Gamma(\alpha_{G,i,j,r_i})} \prod_{k=1}^{r_i} \theta_{G,i,j,k}^{\alpha_{G,i,j,k}-1}$$

For each $i = 1, \ldots, p$ and $j = 1, \ldots, q_{G,i}$, unless otherwise stated, we take the hyperparameters $\alpha_{G,i,j,k} = (r_i q_{G,i})^{-1}$ for all $k = 1, \ldots, r_i$, following Buntine (1991) and Heckerman *et al.* (1995). The choice of hyperparameters can be important: the maximum a posteriori graph is very sensitive to the specification of the hyperparameters (Silander *et al.*, 2007). The most satisfactory way to reduce this sensitivity is to treat the effective sample size α as an unknown parameter, and choose for it a suitable prior distribution. Our choice of hyperparameter corresponds to $\alpha = 1$

(Heckerman *et al.*, 1995). Silander *et al.* (2007) propose a discrete, uniform prior distribution on the range 1 to 100, and are able to implement this approach for networks with a small number of nodes ($p < 15$). For larger networks, however, the considerable extra computational effort required precludes this approach, and so we do not investigate this further. Instead, we take the approach described by Silander *et al.* (2007) as "being Bayesian about the structure", and use model averaging approaches (Section 2.1.1), for which the sensitivity to the hyperparameters is likely to be not as strong as it is for MAP estimation, since the focus on not on a single model.

Suppose we have an observation $\mathbf{x}$ of $\mathbf{X}$, and from this we form the cells of the contingency table $n_{G,i,j,k}$. Then the marginal likelihood can be written in closed-form.

$$p(\mathbf{X} \mid G, n_{G,i,j,k}) = \prod_{i=1}^{p} \prod_{j=1}^{q_{G,i}} \frac{\Gamma(\sum_{k=1}^{r_i} \alpha_{G,i,j,k})}{\Gamma(\sum_{k=1}^{r_i} n_{G,i,j,k} + \sum_{k=1}^{r_i} \alpha_{G,i,j,k})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{G,i,j,k} + \alpha_{G,i,j,k})}{\Gamma(\alpha_{G,i,j,k})}$$

2.5.3 Normal linear regression

The generalisation of the normal linear regression model to Bayesian networks is straightforward when global independence (Section 2.5.1) is assumed. Consider a particular Bayesian network G. For its computational convenience, we will use a g-prior for the regressions at each variable.

$$\beta_i \mid \sigma_i^2, \mathbf{X}_{G_i}, G \sim \mathrm{MVN}(\mathbf{m}_{G,i}, g\sigma_i^2 (\mathbf{X}_{G_i}^T \mathbf{X}_{G_i})^{-1})$$
$$\pi(\sigma_i^2) \propto \sigma_i^{-2}$$

The form of these priors is given in Section 2.4.2 above. The marginal likelihood for a Bayesian network G is

$$p(\mathbf{X} \mid G) \propto \prod_{i=1}^{p} (g+1)^{-(q_{G,i}+1)/2} \times \left(X_i^T X_i - \frac{g}{g+1} X_i^T \mathbf{X}_{G_i} (\mathbf{X}_{G_i}^T \mathbf{X}_{G_i})^{-1} \mathbf{X}_{G_i}^T X_i - \frac{1}{g+1} \mathbf{m}_{G,i}^T \mathbf{X}_{G_i}^T \mathbf{X}_{G_i} \mathbf{m}_{G,i} \right)^{-n/2}.$$

When the variables are centred, so that $\mathbf{m} = 0$, the final term of the marginal likelihood is zero.

2.5.4 Model priors

The simplest prior $\pi(G)$ for Bayesian networks is a uniform prior over the space of DAGs.

$$\pi(G) = \frac{1}{|\mathcal{G}|} \qquad G \in \mathcal{G}$$

This prior is used in most discussions of structural inference for Bayesian networks (e.g. Cooper and Herskovits, 1992; Madigan and Raftery, 1994). However, the unsatisfactory aspects of uniform priors for regression models (Chapter 7) may well also apply to these priors. In addition, a uniform prior such as this may not match prior beliefs, and so other priors have been proposed. In molecular biology applications informative priors are popular (Mukherjee and Speed, 2008; Werhli and Husmeier, 2007) because useful prior information is often available. Their use has also been proposed in other contexts (e.g. Angelopoulos and Cussens, 2008).

2.5.5 Posterior distribution over models

The posterior distribution $P(G \mid \mathbf{X})$ is a discrete distribution over these models.

$$p(G \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid G)\pi(G)}{\sum_{G \in \mathcal{G}} p(\mathbf{X} \mid G)\pi(G)} \tag{2.4}$$

As with regression, we may consider either the MAP model or consider averaging over all (or some) of the models. A drawback using the MAP graph is that it is very sensitive to the specification of the hyperparameters (Silander *et al.*, 2007). However, evaluating the posterior distribution in Equation 2.4 is often not straightforward, and so in the next section we focus on evaluation and approximation for the posterior.

2.6 Posterior distribution computation

Evaluating the posterior distribution over models (Equation 2.4) is in principle straightforward, but in practice is extremely challenging when many random variables are under consideration. When conjugate local models are used, the marginal likelihood is straightforward to compute. The challenge arises when the cardinality of $\mathcal{M}$ is large. A large cardinality makes the summation in the denominator of Equation 2.1 intractable. Instead, we seek an approximation to the posterior distribution. This thesis focuses on approximations that use Markov chain Monte Carlo, the details of which in this context are described in Section 2.6.4. Before describing this, we describe the procedure for evaluating the posterior exactly when this is tractable.

As shown by Robinson (1973), the cardinality of $\mathcal{G}$ grows super-exponentially in p, and so when $p > 6$, say, direct enumeration is not practical.

2.6.1 Exact evaluation of the posterior distribution

When the cardinality of the model space is small, we can evaluate the posterior distribution (Equation 2.1) exactly by exhaustive enumeration. The procedure is the straightforward in principle.

1. Make a list of all DAGs $G \in \mathcal{G}$ with correct number of nodes
2. Evaluate $p(G \mid \mathbf{X})$ for each $G \in \mathcal{G}$.

For Bayesian networks, this algorithm runs without difficulty on modern computers when $p \leq 6$. Adding even a couple of extra variables vastly increases the computational burden of the algorithm, but the maximum of $p = 6$ can be exceeded slightly when a large cluster is available because the algorithm is trivially parallelisable. Another method that reduces the computational burden is to apply an IN-DEGREE RESTRICTION κ that specifies that only Bayesian networks $G = \langle G_1, \ldots, G_p \rangle$ with $|G_i| \leq \kappa$ for all $i = 1, \ldots, p$ are allowed.

2.6.2 MAP-finding methods

The difficulty in approximating the full posterior distribution led many authors to focus on finding the MAP Bayesian network. Greedy (local) searches (Heckerman *et al.*, 1995), or transforming the problem into a MAX-SAT (Cussens, 2008) or into a linear programming problem (Jaakkola *et al.*, 2010; Cussens, 2011) are among the many proposals for finding the MAP Bayesian network. However, in this thesis we focus on methods that allow model averaging, which is not directly possible using these methods.

2.6.3 Markov chain Monte Carlo

Computational methods, including Markov chain Monte Carlo (MCMC), have in the last 20 years transformed Bayesian statistics (see e.g. Gilks *et al.*, 1996; Brooks *et al.*, 2011). While MCMC methods had existed for at least 25 years before, the full potential was not realised until the early 1990s when the generality of the methods in Bayesian statistics was highlighted by Gelfand and Smith (1990) and others.

The aim of MCMC is to estimate properties of a probability distribution that is not easily analytically tractable. Most of the key properties of interest of a probability distribution can be estimated by obtaining a large sample from the distribution, even if the samples are not independent. The idea is to construct a Markov chain on the sample space of the target distribution in such a manner that the equilibrium distribution of the Markov chain is the target distribution. Constructing a Markov chain with the correct equilibrium distribution is remarkably straightforward because generic frameworks are available.

Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) is one such framework that enables a vast range of distributions to be approximated. Suppose we wish to approximate a distribution $p(x)$, which we call the target distribution, and in which $x \in \mathcal{X}$ may be a vector. To do this, we draw samples $\{x^{(t)} : t = 1, \ldots, N\}$ from a Markov chain with each sample drawn conditional on the previous sample according to a transition kernel $K(x' \mid x)$, samples from which are drawn as follows. Given a current state $x^{(t-1)}$, the algorithm uses a proposal distribution $q(x' \mid x^{(t-1)})$ to draw a proposal for the next state. Then, either the proposal is accepted so that $x^{(t)} = x'$, or the proposal is rejected and the current state retained so that $x^{(t)} = x^{(t-1)}$. The decision whether to accept the proposal is

Algorithm 1 A Metropolis-Hasting sampler (Metropolis *et al.*, 1953; Hastings, 1970)

Initialise at an arbitrary starting point $x^{(0)}$

for t in 1 to N **do**

Draw $x' \sim q(x' \mid x^{(t-1)})$

Set

$$x^{(t)} = \begin{cases} x' & \text{with probability } \alpha(x', x^{(t-1)}) \\ x^{(t-1)} & \text{with probability } 1 - \alpha(x', x^{(t-1)}), \end{cases}$$

where $\alpha(x', x^{(t-1)}) = \min\left\{1, \frac{p(x')}{p(x^{(t-1)})} \frac{q(x^{(t-1)}|x')}{q(x'|x^{(t-1)})}\right\}$.

end for

made probabilistically in the manner detailed in Algorithm 1.

The original algorithm by Metropolis *et al.* (1953) used a symmetric proposal, which means the HASTINGS FACTOR $\frac{q(x|x')}{q(x'|x)}$ is unity. In this case, the ability to evaluate the ratio of $q(x^{(t-1)} \mid x')$ and $q(x' \mid x^{(t-1)})$ is not required.

Gibbs sampling

Gibbs sampling (Ripley, 1979; Geman and Geman, 1984) is a particular form of the Metropolis-Hastings algorithm, but leads to rather different samplers. A Gibbs sampler uses samples from the conditional distribution of the components of a multivariate distribution to approximate the complete joint distribution.

We again consider $p(x) = p(x_1, \ldots, x_p)$ to be the target distribution for which we seek an approximation. The Gibbs sampler forms its transition kernel from the full conditional distributions $p(x'_k \mid x_{-k})$, where $x_{-k} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)$ for $k = 1, \ldots, p$. Given a current state $x^{(t-1)} = (x_1^{(t-1)}, \ldots, x_p^{(t-1)})$, the algorithm draws a sample x'_k from $p(x'_k \mid x_{-k})$, for some $k \in \{1, \ldots, p\}$. Then $x^{(t)} = x'$ where $x' = (x_1^{(t-1)}, \ldots, x_{k-1}^{(t-1)}, x'_k, x_{k+1}^{(t-1)}, \ldots, x_p^{(t-1)})$ resulting in $x^{(t-1)}$ and $x^{(t)}$ differing in only

Algorithm 2 A Gibbs sampler (Ripley, 1979; Geman and Geman, 1984)

Initialise at an arbitrary starting value $x = (x_1^{(0)}, \ldots, x_p^{(0)})$.

for t in 1 to N **do**

  for k in 1 to p **do**

    Draw $x' \sim p(x_k \mid x_{-k}^{(t-1)})$

    Set $x^{(t)} = (x_1^{(t-1)}, \ldots, x_{k-1}^{(t-1)}, x'_k, x_{k+1}^{(t-1)}, \ldots, x_p^{(t-1)})$

  end for

end for

the k^{th} component. There is no accept-reject decision; all draws from $p(x'_k \mid x_{-k})$ are used.

Algorithm 2 describes a systematic scan Gibbs sampler, in which each component of x is sampled in turn. An alternative is a random-scan sampler, in which k is drawn from $\{1, \ldots, p\}$ at random, typically uniformly. There are few theoretical results to guide the choice between random- and systematic-scan Gibbs samplers (Roberts and Sahu, 1997). In this thesis, random-scan Gibbs samplers are used throughout.

A positivity condition is required for the Gibbs sampler to be useful in the manner described in the next section; we discuss this in a particular context in Section 4.3.3.

Using the samples

The samples $\{x^{(t)} : t = 1, \ldots, N\}$ are useful because, under weak conditions, the probability that the Markov chain is at a particular state will match that probability of that state in the target distribution.

We first consider the STATIONARY (or INVARIANT or EQUILIBRIUM) distribution of the Markov chain. A stationary distribution of a Markov chain is a distribution such that if the current value of the Markov chain is a draw from the stationary

distribution, the subsequent values of the Markov chain retain this distribution. We want the stationary distribution to be the target distribution. We can establish that this holds by checking that the detail balance condition holds. The transition kernel K of the Markov chain and target distribution p are said to be in DETAILED BALANCE if

$$p(x)K(x \mid y) = p(y)K(y \mid x) \qquad \text{for all } x, y \in \mathcal{X}.$$

The construction of the Markov chain associated with the Metropolis-Hastings algorithm or the Gibbs sampler means that detailed balance will hold.

We also require that the Markov chain is irreducible and aperiodic. A Markov chain is IRREDUCIBLE if the probability of making a transition between any two states is positive. A state in a Markov chain has PERIOD k if any return to that state must occur in multiples of k time steps. If the period of a state is 1, the state is said to be APERIODIC. If all states are aperiodic, then the Markov chain is said to be APERIODIC.

When the Markov chain is irreducible and aperiodic, even if we start the Markov chain at an arbitrary initial value, the samples drawn from the Markov chain will tend to draws from the stationary distribution as the number of samples $N \to \infty$. Thus, after a suitable 'burn-in' period of, say, T iterations, the points $\{x^{(t)} : t \geq T\}$ can be regarded as dependent samples from the target distribution.

In Bayesian inference, the quantities of interest (means, variances, quantiles etc) can all be expressed in terms of a posterior expectation. Thus, in general, we seek an approximation to $E(\mathcal{S}(X))$, for some function $\mathcal{S}(x)$ of the distribution p, that uses the MCMC samples $\{x^{(t)} : t \geq T\}$. An appropriate estimator is the sample mean.

$$E(\mathcal{S}(X)) = \frac{1}{N} \sum_{t=1}^{N} \mathcal{S}(x^{(t)}) \tag{2.5}$$

Recall that the MCMC samples will be dependent, and so the usual law of large

numbers does not apply to this estimator. However, the ergodic theorem for Markov chains reassures us that, for an irreducible Markov chain with stationary distribution p, the sample average converges to $E(\mathcal{S}(X))$ as $N \to \infty$, when $E(\mathcal{S}(X)) < \infty$.

A less dependent set of samples is given by using only every k^{th} sample in Equation 2.5. This is called THINNING. Usually thinning is not beneficial because the associated estimator has a larger variance than the estimator that uses all of the samples (Geyer, 1992; MacEachern and Berliner, 1994). However thinning can be necessary when equilibrium is only reached with N large, but storing all N samples is not feasible for computational reasons.

Convergence diagnostics

Assessing when the dynamics of the MCMC sampler follow those of the equilibrium regime is not straightforward. Although some theoretical results are available (e.g. Roberts and Rosenthal, 1998) the results mostly relate to settings that are elementary relative to typical practical uses of MCMC, and so in practice convergence is assessed using statistical methods based on the samples themselves. Numerous methods for assessing convergence statistically have been proposed; a review is given by Brooks and Roberts (1998).

The most straightforward method for assessing convergence is to examine the sequential 'trace-plot' of a statistic of the samples, against iteration. Yu and Mykland (1998) argue however, that a plot of cumulative sums of statistics enable a better test of convergence: when the sampler is mixing well, there will be regular excursions around the mean in each statistic, which will be represented by a 'spiky' cumulative sum plot. Brooks (1998) develop the idea further with a related quantitative measure of convergence.

Convergence diagnostics are only necessary to assess how long the burn-in period

should be. However, burn-in is simply a method for finding a starting point for the Markov chain, and so if we have an alternative method that gives a starting point that is representative of the target distribution then using a burn-in period is wasteful of samples, a point emphasised by Geyer (2011). Indeed, the ergodic theorem shows that the sample averages will converge to the true expectation regardless of the starting value. In the setting considered here, it would be possible to choose a starting point from the result of a MAP estimation method (Section 2.6.2), which may give a representative point of the target distribution.

We prefer the alternative approach in which multiple chains are run, initialised at disparate, over-dispersed points in the sample space. Using multiple runs provides more reassurance that the asymptotic regime has been reached, rather than the sampler simply being 'stuck' in a local mode. Thus, using multiple runs reduces the chances of being unaware of significant areas of mass in the target distribution. If convergence has been reached in all of the runs, then all statistics of the samples should be similar across the runs. When using this approach it is clear that a burn-in period is required. For example, it is clear from Figure B.1 that in all cases the first part of each run is not representative of the target distribution.

The advantage of this approach is that these multiple runs can be run simultaneously in parallel, and if sufficient computing resources are available, in the same amount of time we can be more confident in our result. There is not an accepted answer as to how many independent chains should be run. Gelman and Shirley (2011) recommend running at least three chains in parallel. A formal numerical diagnostic using multiple chains has been proposed by Gelman and Rubin (1992).

Mixing rate

The mixing rate (or time) of a Markov chain is informally the rate at which (or time until) dependence on the initial conditional is forgotten, and the rate at which

passage between the areas of significant posterior mass occurs. The mixing times of an MCMC sampler depend on the relationship between the target distribution and the transition kernel. When the transition kernel takes a form that allows the sampler to move freely in and between the areas of mass in the target distribution, the sampler will mix quickly. The transition kernels given by the most simple MCMC samplers, however, may corner themselves into unrepresentative local modes, especially in high dimensions. Disastrous results from MCMC samplers are certainly not limited to the contrived textbook examples of bad behaviour.

2.6.4 Approximations for the posterior distribution

In the case of approximating the posterior distribution over the space of models, Markov chain Monte Carlo is required because the cardinality of the space of models $|\mathcal{G}|$ is large.

The standard form of MCMC that is used for structural inference for Bayesian networks is MC^3 (Madigan and York, 1995), a simple Metropolis-Hastings sampler. This sampler moves through the space of DAGs $\mathcal{G}$ by drawing proposals from the NEIGHBOURHOOD of a graph G, defined as the DAGs that can be formed by adding or removing a single edge from G.

$$\nu(G) = \{G' : G \text{ and } G' \text{ differ by a single edge}\}$$

The size of the neighbourhood can vary because certain edge additions may introduce a cycle, and so are not allowed.

Given a current state $G^{(t-1)}$, which is a DAG, a proposal G' is drawn uniformly at random from $\nu(G^{(t-1)})$, the neighbourhood of the current state. The proposal distribution is thus

$$q(G' \mid G) = \frac{1}{|\nu(G)|}.$$

Algorithm 3 MC^3 (Madigan and York, 1995)

Initialise initial Bayesian network $G^{(0)}$

for t in 1 to N **do**

Evaluate $\nu(G^{(t-1)})$

Draw a proposal G' uniformly at random from $\nu(G^{(t-1)})$

Evaluate the acceptance probability $\alpha(G', G)$

Draw $u \in [0, 1]$ uniformly at random

if $u < \alpha(G', G)$ **then**

Set $G^{(t)} = G'$

else

Set $G^{(t)} = G^{(t-1)}$

end if

end for

The proposal is accepted according to the usual Metropolis-Hastings acceptance probability.

$$\alpha(G', G) = \min \left\{ 1, \frac{P(G' \mid \mathbf{X})\pi(G')}{P(G^{(t-1)} \mid \mathbf{X})\pi(G^{(t-1)})} \frac{\frac{1}{|\nu(G')|}}{\frac{1}{|\nu(G^{(t-1)})|}} \right\}$$

The complete algorithm is detailed in Algorithm 3.

In many situations, MC^3 works surprisingly well, but if the posterior distribution is not unimodal, the local moves may fail to explore the space fully because the sampler may become 'trapped' in one mode. This issue becomes more severe as the sample size increases because the posterior distribution becomes more concentrated. We examine this issue in more detail in Chapter 4.

2.7 Constraint-based methods

An alternative class of methods for structural inference is constraint-based approaches. These methods determine the structure of the Bayesian network by making firm decisions about the structure of the Bayesian network through a series of tests of conditional independence. The conditional independence structure is returned as a CPDAG (Section 2.2.3), which specifies an equivalence class of Bayesian networks which with the data are consistent.

In this section, we survey some constraint-based methods, and then describe in detail one such method, the PC-algorithm (Spirtes and Glymour, 1991).

2.7.1 Survey of available methods

Several constraint-based methods have been proposed. The earliest methods, such as the IC algorithm (Verma and Pearl, 1990) and the SGS algorithm (Spirtes *et al.*, 2000), test the independence of each pair of random variables, conditional on each set of other variables. More recent proposals, such as the PC-algorithm (Spirtes and Glymour, 1991) and the recursive method of Xie and Geng (2008), are more selective in the conditional independencies that they consider and are thus more efficient.

The frequentist constraint-based methods have mostly been developed separately from the Bayesian methods that we focus on, but recently Tsamardinos *et al.* (2006) proposed a method that combines constraint-based methods with the score-based methods, such as the Bayesian posterior.

2.7.2 PC-algorithm

The PC-algorithm (Spirtes and Glymour, 1991) works in two stages. First, an undirected graphical model, called the SKELETON is constructed. Then as many of the edges on the graph as possible are assigned directions ('oriented').

The first part of the procedure constructs an undirected model by initially assuming that there are no independencies or conditional independencies between any of the variables. This assumption corresponds to initialising G as the complete undirected graph. Standard frequentist tests of independence are then made; when the p-value of the tests suggest an independence, the relevant edge of the undirected graph is removed. For example, if variable i is discovered to be independent of variable j, given some variables in a set S, then the edge $i - j$ will be removed. The tests are made in increasing order of cardinality of S. Testing in this order reduces the number of independence tests that are required, because if, for example, $i \perp\!\!\!\perp j$ then we do not need to test whether $i \perp\!\!\!\perp j \mid S$ for any $S \subseteq V$.

In this thesis, we use the default cut-off p-value 0.05 that is commonly used with the PC-algorithm (e.g. Tsamardinos *et al.*, 2006; Buhlmann *et al.*, 2010) so that our comparisons with alternative methods correspond to common practice. Meinshausen and Bühlmann (2010) describe an approach that may lead to a more principled choice of cut-off parameter, but we do not investigate this approach in this thesis.

In the second stage of the algorithm, the undirected edges that can be unambiguously assigned a direction are oriented. First, we identify v-structures by considering all triples (i, j, k) such that $i - j$, $j - k$ but with i and k not linked. Then the v-structure $i \rightarrow j \leftarrow k$ is present if and only if there is no conditional independence $i \not\perp\!\!\!\perp k \mid j$, which we ascertained in the first stage. Further edges can be oriented if a particular orientation would induce a cycle in the graph, or if a particular orientation would introduce a v-structure that had been rejected in the previous stage. The

final section of the algorithm follows the method introduced by Verma and Pearl (1992). Correctness was proved by Meek (1995).

The full outline is described in Algorithm 4. The PC-algorithm has been shown to be asymptotically consistent (Kalisch and Bühlmann, 2007).

The conditional independence tests in the algorithm (line 7) are made using standard likelihood ratio tests. For example, the G^2 statistic is used in the case of discrete data, with $\mathfrak{n}_{ijk}^{abc}$ denoting the number of occurrences of $X_i = a$, $X_j = b$, $\mathbf{X}_k = c$; $\mathfrak{n}_{ij}^{ab}$ denoting the number of occurrences of $X_i = a$, $X_j = b$; and $\mathfrak{n}_s^d$ is the number of occurrences of $X_s = d$.

$$G^2 = 2 \sum_{a,b,c} \mathfrak{n}_{ijk}^{abc} \log \left(\frac{\mathfrak{n}_{ijk}^{abc} \mathfrak{n}_k^c}{\mathfrak{n}_{ik}^{ac} \mathfrak{n}_{jk}^{bc}} \right)$$

This statistic is asymptotically χ^2-distributed with $r_i r_j \prod_{l \in k} r_l$ degrees of freedom. The null hypothesis is $i \perp\!\!\!\perp j \mid k$, indicating conditional independence between i and j given k.

Algorithm 4 PC-algorithm (Spirtes and Glymour, 1991)

Initialise initial graph G as the complete undirected graph.

Initialise $a = 1$

while pairs (i, j) with neighbours of large enough cardinality remain **do**

  while suitable pairs (i, j) remain **do**

    Choose an ordered pair of nodes (i, j) such that $i - j$ and $|\text{adj}(i) \setminus \{j\}| \geq a$

    Choose a set $S \subseteq \text{adj}(i)$ such that $|S| = a$

    if $i \perp\!\!\!\perp j \mid S$ **then**

      Set G to the graph G with the edge $i - j$ removed.

      Add S to SepSet(i, j) and SepSet(j, i)

    end if

  end while

  $a = a + 1$

end while

for triples (i, j, k) in which $i - j - k$ but i not adjacent to k **do**

  if $j \notin \text{SepSet}(i, k)$ **then**

    Orient $i \to j \leftarrow k$

  end if

end for

while orientable edges remain **do**

  Rule 1 Orient $j - k$ as $j \to k$ if there exists an edge $i \to j$, but no edge links nodes i and k.

  Rule 2 Orient $i - j$ as $i \to j$ if there a node k such that $i \to k \to j$

  Rule 3 Orient $i - j$ as $i \to j$ if there exist nodes k and l such that $i - k \to j$ and $i - l \to j$

  Rule 4 Orient $i - j$ as $i \to j$ if there exist nodes k and l such that $i - k \to l$ and $k \to l \to j$, and no edge links nodes j and k, but an edge does link nodes i and l.

end while

Chapter 3

Subjective well-being and risk-avoiding behaviour

Measures of subjective well-being aim to encapsulate the human experience of 'happiness', 'well-being', and 'satisfaction with life'. These terms are thus often used interchangeably. Subjective well-being has been discussed since at least the 1970s (e.g. Easterlin, 1974), but the subject has developed considerably in recent years (e.g. Easterlin, 2003; Oswald and Wu, 2010). A particular focus of recent work has been on identifying factors that affect subjective well-being (e.g. Diener *et al.*, 1995; Fowler and Christakis, 2008). This work has motivated, and in part been motivated by, recent political adoption of the aim of increasing 'gross national happiness', which proponents argue is a more relevant indicator of the success of a country (or policy) than gross domestic product (GDP). Oswald (1997) argues that economic measures of performances are only relevant as a means to an end, and that end is well-being. In particular, nobody has any real concern with the standard economic indicators (inflation, growth, unemployment, etc) except as proxies for the well-being of the population.

While factors that influence subjective well-being have been widely-studied, the effects of happiness are relatively unstudied. In this chapter, we propose and provide empirical evidence that supports the idea that subjective well-being influences the risk taking characteristics of individuals. We investigate this by considering seatbelt-wearing as a proxy for avoidable risk taking. We find that individuals who describe themselves as happier are more likely to wear a seatbelt.

We use data on reported well-being and seatbelt use in a sample of 300,000 Americans, and find evidence strongly consistent with this theory. That is, the less satisfied people are with life, the less conscientious they are in taking action to preserve their life by the wearing of a seatbelt. The result is obtained with various methodological approaches, with an emphasis on Bayesian model-selection. We find evidence that none of the confounders, either singly or jointly, can explain the observed connection between seatbelt use and subjective well-being (even after accounting for non-linear effects). To the best of our knowledge, the principal finding has not been established in this manner before, although simple correlations consistent with the result have been reported by Kirkcaldy and Furnham (2000), in the psychology literature.

The remainder of this chapter is organised as follows. We first detail the background to the study. We then present details of the data and methods used in the study, before presenting the main results. Finally, we discuss shortcomings and implications, as well as directions for further work.

3.1 Background

3.1.1 Risky behaviour

Understanding the reasons why individuals take risks, particularly avoidable risks, is an important open question in economics (Barsky *et al.*, 1997; Dohmen *et al.*,

2011). Some researchers argue that in the industrialised world—where affluence has become the norm—the key question for policy-making has become that of how to understand risky health behaviours (Offer, 2006; Offer *et al.*, 2010).

Decision processes involving risk are complex. They are affected by a wide range of factors—including underlying risk preferences, perceptions, framing, level of involvement in the outcome-generating process, previous outcomes, and biological factors (Kahneman and Tversky, 1979; Zeckhauser and Viscusi, 1990; Thaler and Johnson, 1990; Kimball, 1993; Fong and McCabe, 1999; Sapienza *et al.*, 2009).

We use the use of seatbelts as an indicator of risk-taking because it represents an interesting indicator of self-preserving behaviour. In a modern industrialised nation, there are few widespread activities in which people are at risk of instantaneous death or serious injury. However, driving is one activity that carries with it the risk of serious physical harm. The wearing of seatbelts is a demonstrably effective measure in reducing this risk (Wild *et al.*, 1985). There is little cost associated with seatbelt use and so, rationally, the wearing of seatbelts should be universal.

Yet seatbelt use in the United States is far from universal. Only 83 percent of individuals in the data used in this study state they always use a seatbelt. This figure is corroborated by the National Occupant Protection Use Survey by National Highway Traffic Safety Administration (Pickrell and Ye, 2008), which directly also observed that 83 percent of individuals actually used a seatbelt. Thus, there remain as yet unexplained patterns of variation in this key risk behaviour.

3.1.2 Subjective well-being

In recent years, an increasing number of authors (e.g. Easterlin, 1974; Oswald, 1997; Frey and Stuzer, 2002) have argued that subjective well-being should play an important role in the study of human behaviour. However, while the concept of evaluating

policy by its effect on well-being may be an incontestably worthy aim, using this idea in practice invariably involves relying on self-assessed measures of subjective well-being. There has been considerable debate in the literature about whether self-reported measures of well-being are meaningful (Argyle, 2001; Bertrand and Mullainathan, 2001), specifically whether they accurately reflect the true state of a respondent's well-being. It may seem that relying on self-assessed measures leads to a lack of scientific objectivity, but as Easterlin (1974) notes, "If one is interested in how happy people are—in their subjective satisfaction—why not let each person set his own standard and decide how closely he approaches it". In fact, substantial new evidence suggests that these measures are correlated with biological and other indicators (Udry, 1998; Steptoe and Wardle, 2005; Fliessbach *et al.*, 2007), and thus do provide meaningful information in an objective sense. It has also recently been demonstrated that there is a close spatial match between U.S. life satisfaction scores and objective well-being indicators (Oswald and Wu, 2010).

A diverse literature is emerging on the determinants of human happiness (e.g. Diener, 1984; Diener *et al.*, 1995; Oswald, 1997; Radcliff, 2001; Clark, 2003; Easterlin, 2003; Di Tella and MacCulloch, 2005; Layard, 2005; Luttmer, 2005; Dolan and White, 2007; Dolan and Kahneman, 2008; Fowler and Christakis, 2008; Stevenson and Wolfers, 2008; Pittau *et al.*, 2009), how its changes over a lifespan (Blanchflower and Oswald, 2004, 2008; Pischke, 2011), and its relationship to utility (Kimball and Willis, 2006; Benjamin *et al.*, 2010). One of the notable claims is that subjective well-being is 'U-shaped' over the course of a life (Blanchflower and Oswald, 2008); that is, individuals are satisfied in youth, but become less satisfied in middle-age, and then recover satisfaction in old age. Another interesting claim is that subjective well-being at a country level is disconnected from economic growth; in particular, subjective well-being in the U.S. has not increased as it has become richer (Oswald, 1997; Stevenson and Wolfers, 2008). Less is known, however, about the influence of people's well-being on their actions: that is, on what happiness 'does', rather than

the factors that shape it.

3.2 Data and methods

This section describes the two data sources and briefly outlines Bayesian variable selection and joint confounding methods. Importantly, these Bayesian techniques allow a relaxation of the assumption of linearity.

3.2.1 Behavioural Risk Factor Surveillance System Survey

We draw data from the publicly available Behavioural Risk Factor Surveillance System Survey (BRFSS). This is a household-level random-digit telephone survey, collected by the U.S. Government's National Center for Chronic Disease Prevention and Health, that has been conducted throughout the United States since 1984. Seatbelt-use statistics were collected in 2006 and 2008, but to avoid a discontinuous time-period, we use only 2008 data (results using 2006 data are similar). Following previous work (Oswald and Wu, 2010), we restrict our analyses to those between 18 and 85 years old, not residing in unincorporated U.S. territories, and exclude respondents who refused or were unsure of their response, or whose response is missing, for any of the 19 variables included in our analyses (Tables 3.1 and 3.2). The resulting sample size is 313,354.

Our measure of life satisfaction is the response, on a 4-point scale ranging from 'Very satisfied' to 'Very dissatisfied', to the question, "In general, how satisfied are you with your life?". Seatbelt use is recorded as self-reported frequency of use when driving or riding in a car, on a 5-point scale. Respondents were also able to declare that they do not use a car. These questions were separated in the survey by at least 4 other questions. Table 3.3 (page 72) lists the questions from which the covariates

Table 3.1: The main covariates used from BRFSS in Chapter 3. The discretisation in Column 2 ('Levels') is used in the linear analyses, while the analyses based upon model selection use the discretisation in Column 3 ('Collapsed Levels'). (The additional covariates used in the model selection analyses are detailed in Table 3.2 on page 71.)

Variable	Levels	Collapsed categories
Seatbelt	Always (coded 5)	Always
	Nearly always (4)	Not always
	Sometimes (3)	Not always
	Seldom (2)	Not always
	Never (1)	Not always
Subjective well-being	Very satisfied (4)	Very satisfied
	Satisfied (3)	Not very satisfied
	Dissatisfied (2)	Not very satisfied
	Very dissatisfied (1)	Not very satisfied
Gender	Male	Male
	Female	Female
Race	White only, non-Hispanic	White only, non-Hispanic
	Black only, non-Hispanic	Black only, non-Hispanic
	Asian only, non-Hispanic	Asian only, non-Hispanic
	Other/Multiracial, non-Hispanic	Other/Multiracial, non-Hispanic
	Hispanic	Hispanic
Age	(Age in years)	Young (18–34 years)
		Middle-aged (35–64 years)
		Old (65 years or older)
Marital Status	Never Married	Never Married
	Married	In couple
	Divorced	Formerly in couple
	Separated	Formerly in couple
	Widowed	Widowed
	Unmarried couple	In couple
Education	No high school	Not a high school graduate
	Some high school	Not a high school graduate
	High school graduate	High school graduate
	Some college/technical school	High school graduate
	College graduate	College graduate
Employment	Employed for wages	Employed
	Self-employed	Employed
	Unemployed	Unemployed
	Homemaker	Not in workforce
	Student	Not in workforce
	Retired	Not in workforce
	Unable to work	Not in workforce
Annual Income	$10,000 or less	Low income
	$10,000 – $15,000	Low income
	$15,000 – $20,000	Low income
	$20,000 – $25,000	Medium income
	$25,000 – $35,000	Medium income
	$35,000 – $50,000	Medium income
	$50,000 – $75,000	High income
	$75,000 or more	High income
State of residence	(State of residence)	
Month of interview	(Month of interview)	
Number of children	(Number of children in household)	No children
		1 child
		2 or more children

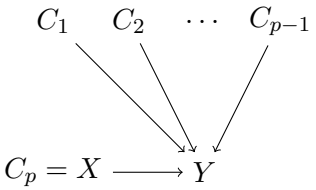


Figure 3.1: A graphical representation of the form of the models used in variable selection for joint effects of multiple covariates. The variable selection formulation explores subsets of $\{C_1, \ldots, C_{p-1}, X\}$ as joint explanatory factors for response Y.

are derived.

3.2.2 Bayesian methods

Bayesian variable selection

We fit standard regression models to the data. We additionally consider Bayesian variable selection, which is useful in this context because it accounts for the possibility of non-linearity and interactions. This framework provides a more rigorous test of the importance of a covariate because a larger number of possible alternative explanations are considered, including interaction effects that are sometimes key (e.g. in Gelman *et al.*, 2007) and yet are often overlooked.

The models M_γ for seatbelt use that we consider are a particular form of the multinomial-Dirichlet Bayesian variable selection introduced in Section 2.4.1. The models describe the distribution of seatbelt use Y in terms of a collection of potential predictors $C_1, \ldots, C_{p-1}$ and well-being X, which we refer to collectively as the covariates $C_1, \ldots, C_p$ for simplicity. Recall that the models are defined by an indicator variable $\gamma = (\gamma_1, \ldots, \gamma_p)$, so that $\{C_i : \gamma_i = 1\}$ is the subset of covariates included in model M_γ. We let the number of covariates $p_\gamma = \mathbb{I}^T\gamma$ included in the model be such that $p_\gamma \leq 9$ (Figure 3.1). Suppose each of the p covariates consists of r_i categories, $1 \leq i \leq p$. For a model M_γ, let $\mathcal{C} = \{C_1^\gamma, \ldots, C_q^\gamma\}$ be the set

containing all $q_\gamma = \prod_{i=1}^{p} r_i^{\gamma_i}$ combinations of values of the covariates included in the model. Note that this is equivalent to defining $\mathcal{C}$ as the sample space of the included covariates. To control complexity in this setting, we simplify the data by reducing the levels of some variables with many categories, as shown in Tables 3.1 and 3.2 (on pages 59 and 71), and binarise the response, enabling a simple contrast between those who always wear seatbelts with those who do not. For each of the n individuals, let y_i be the indicator of whether individual i always uses a seatbelt, and c_i be the p-dimensional vector of covariates that incorporates an indicator of well-being. We use a binomial model for the responses, with parameter θ_j dependent on the configuration $C_j^\gamma \in \mathcal{C}$ of the covariates. Thus the joint probability for vector of responses $\boldsymbol{y}$ depends on n_j, the number of observed individuals who have covariates C_j^γ, and m_j, the number of these individuals who use a seatbelt.

The posterior distribution over models M_γ, given the data, provides a measure of the fit of each model that incorporates a preference for simpler models of lower dimension. The posterior, up to proportionality, is given by the product of the model prior $\pi(M_\gamma)$, and, using the standard assumption of independent beta(α, β) parameter priors (Cooper and Herskovits, 1992), the closed-form marginal likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{c}_1, \ldots, \boldsymbol{c}_p, M_\gamma) = \prod_{j=1}^{q_\gamma} \frac{\Gamma(m_j + \alpha)\Gamma(n_j - m_j + \beta)\Gamma(\alpha + \beta)}{\Gamma(n_j + \alpha + \beta)\Gamma(\alpha)\Gamma(\beta)},$$

where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_p$ are the vectors of observations of the covariates.

As usual and following previous authors (Heckerman *et al.*, 1995), we set the hyperparameters $\alpha = \beta = (2q_\gamma)^{-1}$ for each θ_j. We choose a flat prior $\pi(M_\gamma) \propto 1$, but the large sample results in insensitivity to this choice. Penalised likelihood approaches offer an alternative to the Bayesian approach taken here: indeed, here we find that a BIC-based analysis (with $p_\gamma \leq 5$, for computational reasons) in this setting selected the same model.

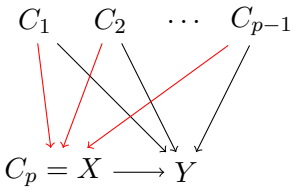


Figure 3.2: The form of the model used in model selection for joint confounding by multiple factors. A graphical representation of family of models for considering the influence of conjectured explanatory variable X on response Y with potential observed confounders $C_1, \ldots, C_{p-1}$. A model selection approach is used to explore evidence in favour of a direct link from X to Y in light of subsets of $\{C_1, \ldots, C_{p-1}\}$ which may jointly explain both X and Y.

Joint confounding

An alternative to regression approaches, which model risk-taking behaviour conditional on the observed covariates and life-satisfaction, is additionally to model life-satisfaction conditional on the observed covariates (Robins *et al.*, 1992; Senn *et al.*, 2007). This approach has the advantage of explicitly modelling the unbalanced distribution of subjective well-being among individuals, for which we must account to compare meaningfully how seatbelt-use varies with life-satisfaction. We can restore balance by identifying covariates that explain both subjective well-being and seatbelt use, and examining the effect of life-satisfaction within particular values of these covariates.

We take a model selection approach to discovering such covariates (Robins and Greenland, 1986) that is similar to Bayesian variable selection, but as shown in Figure 3.2 we now mirror dependences between covariates C_i and seatbelt use (Y) with corresponding direct dependences, for $i \leq p-1$, between C_i and subjective well-being (X). This approach can be thought of as exploring different stratifications for a model of the effect of X on Y. Any residual relationship after stratification between subjective well-being and seatbelt use represents the controlled effect (Rosenbaum, 2002). The approach taken here can also be regarded as a special case of structural

inference in Bayesian networks (Heckerman *et al.*, 1995; Madigan and York, 1995; Mukherjee and Speed, 2008).

Each model $M_{\gamma,\delta}$ is defined by a set of confounders (a subset of the covariates $C_1, \ldots, C_{p-1}$ defined by γ, excluding subjective well-being X, and with $p_\gamma \leq 9$ for computational tractability) and an indicator variable δ for whether the direct dependence between X and Y is present. We redefine $\mathcal{C}$ to be the set containing all combinations of values of the confounders alone (i.e. excluding subjective well-being) in $M_{\gamma,\delta}$, and, with $q'_\gamma = q_\gamma r_p^\delta$, denote by $\mathcal{D} = \{D_1^\gamma, \ldots, D_{q'_\gamma}^\gamma\}$, the corresponding set including subjective well-being. We denote the number of observed individuals with confounding variables $C_j^\gamma \in \mathcal{C}$ by w_j, and number of these individuals who are 'very satisfied' by v_j. Similarly defining n_l to be number of observed individuals with covariates $D_l^\gamma \in \mathcal{D}$ and the number of these who always use a seatbelt by m_l, we have the following marginal likelihood for seatbelt use $\boldsymbol{y}$, subjective well-being $\boldsymbol{x}$, and confounders $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p-1}$.

$$p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p-1}, M_{\gamma,\delta}) = \prod_{l=1}^{q'} \frac{\Gamma(m_l + \alpha)\Gamma(n_l - m_l + \beta)\Gamma(\alpha + \beta)}{\Gamma(n_l + \alpha + \beta)\Gamma(\alpha)\Gamma(\beta)} \times \prod_{j=1}^{q} \frac{\Gamma(v_j + \alpha)\Gamma(w_j - v_j + \beta)\Gamma(\alpha + \beta)}{\Gamma(w_j + \alpha + \beta)\Gamma(\alpha)\Gamma(\beta)}$$

We again choose beta priors for α, β, with $\alpha = \beta = (2q_\gamma)^{-1}$ for X, and $\alpha = \beta = (2q'_\gamma)^{-1}$ for Y. Note that the result of adding extra dependencies is simply an additional term in the marginal likelihood, and so the computation time is identical to variable selection.

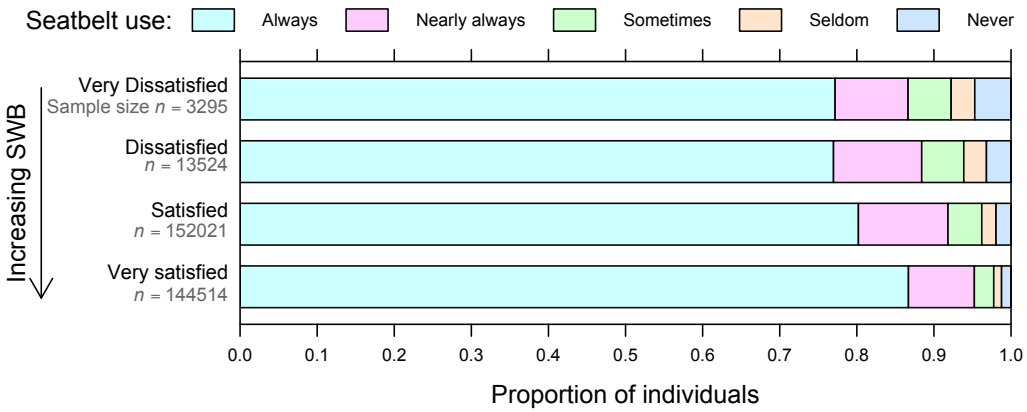


Figure 3.3: Frequency of seatbelt use cross-tabulated by subjective well-being (SWB). Each category contains at least 101 individuals. Pearson's chi-squared statistic is 3242 (p-value $p < 2.2 \times 10^{-16}$).

3.3 Results

3.3.1 Raw data

By analysing the 313,354 individuals with complete relevant data in a random sample in the United States, the study finds evidence that an individual's life-satisfaction (subjective well-being) is an important determinant of their attitude to taking risks, even when a wide range of other factors are accounted for. Figure 3.3 shows that, in raw data, subjective well-being and seatbelt use are strongly associated.

The main idea of the chapter is visible in the raw uncorrected data. Across the entire sample of $n = 313,354$ U.S. residents used here we find that, while 86.7 percent of individuals who are 'very satisfied' with their life report always using their seatbelt, only 77.2 percent of adults who are 'very dissatisfied' do so. Moreover, 4.7 percent of individuals who are 'very dissatisfied' with their life report never using their seatbelt, whereas only 1.2 percent of adults who are 'very satisfied' do so. The differences across all the levels in this large sample corresponds to a statistically highly significant association (Figure 3.3), yielding a Chi-squared p-

value with $p < 2.2 \times 10^{-16}$.

3.3.2 Regression for seatbelt use

To try to investigate this more fully, and to understand the influence of other explanatory factors, we employed a range of analyses. First, we carried out a logistic regression that predicts whether an individual always wears a seatbelt. This regression includes sex, age, race, marital status, educational achievement, employment status, income, month of interview, and state of residence as independent variables. The resulting fitted odds ratio for always wearing a seatbelt in favour of very satisfied individuals is large at 1.383 (Table 3.4 on page 73). This result shows that subjective well-being remains a quantitatively important determinant of seatbelt use after inclusion of a wide range of social, economic and demographic factors. The same conclusion, that subjective well-being is substantively important, is given when predicting the level of seatbelt use by OLS, as shown in Table 3.5 on page 74.

After allowing for a range of covariates, an increase of one level (out of four) in subjective well-being is associated with an increase by a factor of 1.383 in the odds ratio of wearing a seatbelt.

3.3.3 Bayesian variable selection

A more rigorous test of the hypothesis can be performed by allowing non-linearity and interactions into the model, as detailed in Section 3.2.2 above, to check that the result is robust to such deviations in the modelling assumptions. This approach addresses the possibility that in combination, and potentially through a non-linear relationship, other covariates may adequately describe seatbelt use, without any dependence on subjective well-being. To consider this possibility, we use a vari-

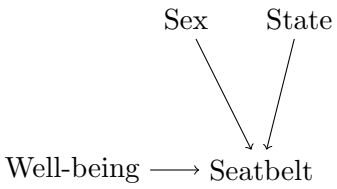


Figure 3.4: The model selected by variable selection for seatbelt use for joint effects of multiple covariates, with selection occurring from 19 covariates, including subjective well-being (Tables 3.1 and 3.2 on pages 59 and 71). The approach accounts for interactions and non-linear effects, and so provides a more stringent test of the influence of subject well-being on seatbelt use. The (posterior) probability of the model shown was close to unity: this shows that subjective well-being appears as a salient influence on seatbelt use even when interactions and non-linear effects of other explanatory factors are allowed.

able selection framework to explore all possible subsets S_γ of covariates (up to and including 9 covariates jointly, for computational tractability) to quantify the joint explanatory ability of those subsets in terms of probability scores. We find that, with probability 0.99, the subset of predictors that jointly best describe seatbelt use are state of residence, sex and life satisfaction (Figure 3.4). Fitted posterior probabilities from this model are shown in Figure 3.5 by state, arranged into groups defined by seatbelt legislation. It can be seen in Figure 3.5 that seatbelt-wearing rates vary widely across U.S. states and that differing legislation at the state-level explains some of this variation. Females are more likely to use a seatbelt than males. These patterns are expected and fairly well-known, but it is the high rate of seatbelt use in very satisfied individuals that, to the best of our knowledge, is a new one in social science. This model estimates that the probability of an individual who is very satisfied always wearing their seatbelt is 0.067 higher.

3.3.4 Joint confounding

The regression approaches described above focus on factors associated with seatbelt use. However, it is factors that explain, possibly in combination, both subjective

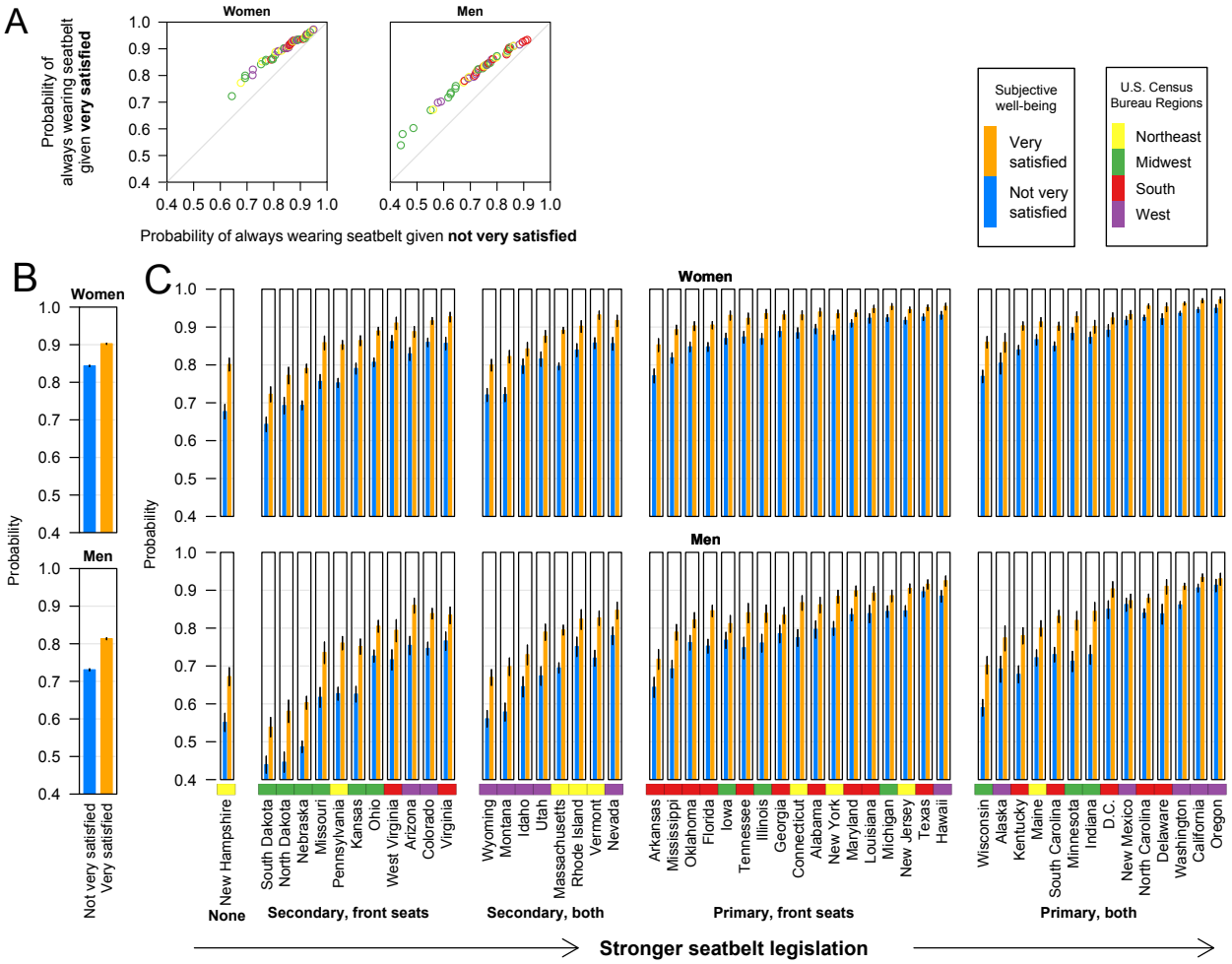


Figure 3.5: Fitted (posterior) probabilities of always wearing a seatbelt given subjective well-being. (A) For each state, the probability of always wearing a seatbelt for very satisfied residents against the probability of always wearing a seatbelt for residents who are not very satisfied. The colours denote U.S. Census Bureau Regions. (B) Probability of always wearing a seatbelt (Bayesian posterior probabilities, with bars indicating 95 percent highest probability density region), given subjective well-being, stratified by gender. (C) As (A), but stratified by state of residence and gender (these covariates were identified as influential by a variable selection approach; see the main text for details and Figure 3.1). States are grouped by legislation type, and the adjacent colours denote U.S. Census Bureau Regions. Both state/legislation and gender effects are important, but the association between subjective well-being and seatbelt use remains clear under stratification.

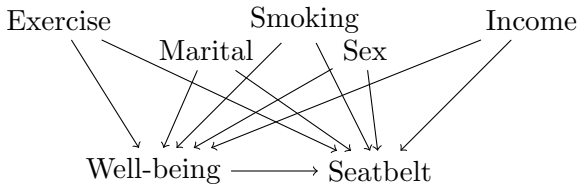


Figure 3.6: The model selected for joint confounding by multiple factors of the relationship between well-being and seatbelt use, treating seatbelt as Y, subjective well-being as X and selecting potential confounders C_i from Tables 3.1 and 3.2 on pages 59 and 71. The model shown was selected with high confidence (posterior probability of model was close to unity); it includes five factors, but retains the link from subjective well-being to seatbelt use, showing that well-being remains an important influence on seatbelt use even when all possible joint stratifications are considered in a fully general non-linear model.

well-being and seatbelt use that may bias the result; this can happen through the unbalancing of the distribution of subjective well-being. We consider this problem explicitly with models of form shown in Figure 3.2, so that the covariates explain *both* subjective well-being and seatbelt use. This approach makes it possible to isolate the fully controlled relationship between subjective well-being and seatbelt use.

The best model (Figure 3.6), in which the Bayesian posterior probability of the model is close to unity, retains the link from subjective well-being to seatbelt use. This model is preferred to the corresponding model—without such a link—with high confidence (Bayes factor $\approx 10^{33}$). Applying the back-door theorem (Pearl, 2009), which here implies taking the weighted average of the effect over the strata defined by the model, the probability of always wearing a seatbelt is estimated to be 0.053 higher in individuals who report themselves very satisfied with their life.

3.4 Discussion

Economists and behavioural scientists currently lack a full understanding of why some people take extreme risks with their lives. This chapter provides some of the first evidence of a powerful link between life-satisfaction and risk-avoiding behaviour. The study finds that the less happy an individual is with life, the less conscientious that person is in taking action to preserve their life by the wearing of a seatbelt.

Goudie *et al.* (2011) provide further evidence: using widowhood at 60 years old or younger as an instrument, they show that an exogenous increase of one class of subjective well-being category increases seatbelt use by 0.188 categories. In addition, longitudinal data from Add Health shows that the less happy an individual is with life, the more likely they are to be involved in a motor vehicle accident later in life.

Our results are consistent with a rational-choice account of extreme risk-taking. It can be shown that standard expected-utility theory predicts that 'happier' people will be more cautious in their risk-taking and invest more in safety (Goudie *et al.*, 2011). Put informally, this is because humans who greatly enjoy life have a lot to lose. By contrast, people who gain only a small utility premium from life have less to lose; thus, on an expected-utility calculation, they will rationally take greater risks (with their lives), in the sense that they are less willing to pay the costs associated with safety-seeking.

We have used seatbelt use as an indicator of individual propensity for risky behaviour. Although relatively little-studied by economists and social scientists, driving is one of the few mainstream activities that even in developed countries remains potentially life-threatening. In contrast to behaviours like smoking and drug-taking, seatbelt use is probably habitual rather than addictive. For this reason, it is less likely that current seatbelt-wearing behaviour is strongly affected by long-past attitudes to risk. In contrast, current smoking status, for example, may relate to

decision-making processes of an individual some decades previously. Additionally, the 'passive' effects on others brought about by the non-use of seatbelts are arguably smaller, or at least less well appreciated, than for smoking, and so seatbelt use may reflect a more personal indication of propensity for risk than other measures. Seatbelt use has in addition been demonstrated to be associated with risk preference as elicited by a lottery choice experiment (Anderson and Mellor, 2008).

The chapter's conceptual account potentially has implications for science and policy. If a government wants to alter the dangerous actions chosen by citizens, it may need to change its citizens' intrinsic happiness with their lives rather than, as now, concentrating policy upon detailed behavioural symptoms themselves.

Table 3.2: Additional covariates from BRFSS used in model selection analyses in Chapter 3.

Variable	Raw categories	Collapsed categories
Body Mass Index (BMI)	(Height and weight)	
	BMI < 2500	Neither overweight or obese
	$2500 <$ BMI < 3000	Overweight
	BMI > 3000	Obese
Heavy alcohol	(Number drinks of drinks/month)	
	Men $>$ 2 drinks/day	Heavy drinker
	Women $>$ 1 drinks/day	Heavy drinker
	Men $\leq$ 2 drinks/day	Not heavy drinker
	Women $\leq$ 1 drinks/day	Not heavy drinker
Physical Activity	Do exercise	Do exercise
	Don't exercise	Don't exercise
Diabetes	Have diabetes	Have diabetes
	Had diabetes when pregnant	Had diabetes when pregnant
	No diabetes	No diabetes
	Only pre- or borderline	Only pre- or borderline
Heart Attack	Had heart attack	Had heart attack
	Not had heart attack	Not had heart attack
Special Equipment	Use special equipment	Use special equipment
	Don't use special equipment	Don't use special equipment
Current Smoker	Current smoker	Current smoker
	Not current smoker	Not current smoker
Asthma	Currently have asthma	Currently have asthma
	Do not currently have asthma	Do not currently have asthma

Table 3.3: Questions used in the study from BRFSS in Chapter 3.

Variable	Question
Seatbelt	How often do you use seat belts when you drive or ride in a car?
Life Satisfaction	In general, how satisfied are you with your life?
Gender	(Noted by interviewer)
Race	Are you Hispanic or Latino? Which one or more of the following would you say is your race? [Mark all that apply.] (from White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Other.)
Age	What is your age?
Marital Status	Are you: Married, Divorced, Widowed, Separated, Never married, A member of an unmarried couple?
Education	What is the highest grade or year of school you completed?
Employment	Are you currently: Employed for wages, Self-employed, Out of work for more than 1 year, Out of work for less that 1 year, A homemaker, A student, Retired, Unable to work
Income	Is your annual household income from all sources: (from Less than \$25,000, \$10,000 – \$15,000, \$15,000 – \$20,000, \$20,000 – \$25,000, \$25,000 – \$35,000, \$35,000 – \$50,000, \$50,000 – \$75,000, \$75,000 or more)
Number of children	How many children less than 18 years of age live in your household?
Body Mass Index	About how much do you weigh without shoes? About how tall are you without shoes?
Heavy alcohol	One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. During the past 30 days, on the days when you drank, about how many drinks did you drink on the average? [A 40 ounce beer would count as 3 drinks, or a cocktail drink with 2 shots would count as 2 drinks.]
Physical Activity	During the past month, other than your regular job, did you participate in a activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
Diabetes	Have you ever been told by a doctor that you have diabetes?
Heart Attack	Has a doctor, nurse, or other health professional ever told you that you had a heart attack, also called a myocardial infarction?
Special Equipment	Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone? (Include occasional use or use in certain circumstances.)
Current Smoker	Do you now smoke cigarettes every day, some days, or not at all?
Current Asthma	Have you ever been told by a doctor, nurse, or other health professional that you had asthma? Do you still have asthma?

Table 3.4: Logistic regression equations for seatbelt use. The model predicts seatbelt use from a panel of covariates (Table 3.1 on page 59), including subjective well-being. We show the estimated coefficients β, and their standard errors and p-values, and the odds ratios (OR), for the model as fitted to data from $n = 313,354$ individuals from the BRFSS in 2008. Subjective well-being has p-value $p < 2 \times 10^{-16}$. All estimates have controlled for state of residence and interview month.

Effect	Coefficient, β	Std. err.	p value	Odds ratio, $\exp(\beta)$
Subjective well-being	**0.324**	**0.008**	**< 0.001**	**1.383**
Gender (baseline Male)				
Female	0.716	0.011	< 0.001	2.047
Race (baseline White)				
Black	-0.009	0.021	0.668	0.991
Asian	0.593	0.060	< 0.001	1.809
Hispanic	-0.038	0.026	0.149	0.963
Other race	0.353	0.026	< 0.001	1.424
Age	0.032	0.002	< 0.001	1.032
Age^2	0.000	0.000	< 0.001	1.000
Marital Status (baseline Never Married)				
Married	0.230	0.018	< 0.001	1.259
Divorced	0.110	0.020	< 0.001	1.116
Widowed	0.182	0.025	< 0.001	1.200
Separated	0.159	0.037	< 0.001	1.173
Unmarried couple	0.006	0.034	0.855	1.006
Educational achievement (baseline No High School)				
Attended High School	-0.090	0.038	0.017	0.914
Graduated High School	-0.033	0.034	0.325	0.967
Attended College	0.100	0.034	0.004	1.105
Graduated college	0.410	0.035	< 0.001	1.506
Employment status (baseline Employed)				
Self-employed	-0.477	0.016	< 0.001	0.620
Unemployed	0.023	0.025	0.374	1.023
Homemaker	0.219	0.025	< 0.001	1.245
Student	0.172	0.042	< 0.001	1.187
Retired	0.198	0.019	< 0.001	1.219
Unable to work	0.177	0.023	< 0.001	1.193
Income (baseline Less than $10,000)				
$10,000 – $15,000	-0.047	0.031	0.125	0.954
$15,000 – $20,000	-0.022	0.029	0.460	0.978
$20,000 – $25,000	0.007	0.029	0.795	1.007
$25,000 – $35,000	-0.054	0.028	0.054	0.947
$35,000 – $50,000	-0.064	0.028	0.022	0.938
$50,000 – $75,000	-0.004	0.029	0.895	0.996
More than $75,000	0.158	0.029	< 0.001	1.171
Number of children	0.001	0.001	0.262	1.001
Constant	-0.873	0.086	< 0.001	0.418

Table 3.5: Ordinary Least Squares (OLS) equations for seatbelt use. The model predicts seatbelt use from a panel of covariates (Table 3.1 on page 59), including subjective well-being (shown in bold). We show the estimated coefficients β, the standard error and the p-value for the model as fitted to data from $n = 313,354$ individuals from the 2008 Behavioral Risk Factor Surveillance System Survey (BRFSS). Subjective well-being has p-value $p < 2 \times 10^{-16}$. All estimates have controlled for state of residence and interview month.

Effect	Coefficient, β	Standard error	p value
Subjective well-being	**0.081**	**0.002**	**< 0.001**
Gender (baseline Male)			
Female	0.196	0.003	< 0.001
Race (baseline White)			
Black	0.016	0.005	0.003
Asian	0.059	0.008	< 0.001
Hispanic	-0.032	0.008	< 0.001
Other race	0.084	0.006	< 0.001
Age			
Age	0.007	0.001	< 0.001
Age^2	-4.4×10^{-5}	<0.001	< 0.001
Marital Status (baseline Never married)			
Married	0.086	0.005	< 0.001
Divorced	0.028	0.006	< 0.001
Widowed	0.064	0.007	< 0.001
Separated	0.050	0.011	< 0.001
Unmarried couple	0.025	0.010	0.015
Educational achievement (baseline No High School)			
Attended High School	-0.016	0.012	0.193
Graduated High School	0.016	0.011	0.138
Attended College	0.077	0.011	< 0.001
Graduated college	0.160	0.011	< 0.001
Employment status (baseline Employed)			
Self-employed	-0.144	0.005	< 0.001
Unemployed	-0.008	0.008	0.276
Homemaker	0.024	0.005	< 0.001
Student	0.070	0.011	< 0.001
Retired	0.023	0.004	< 0.001
Unable to work	0.003	0.007	0.670
Income (baseline Less than $10,000)			
$10,000 – $15,000	-0.002	0.010	0.871
$15,000 – $20,000	0.007	0.009	0.473
$20,000 – $25,000	0.019	0.009	0.034
$25,000 – $35,000	0.005	0.009	0.538
$35,000 – $50,000	0.010	0.009	0.239
$50,000 – $75,000	0.026	0.009	0.004
More than $75,000	0.051	0.009	< 0.001
Children			
Number of children	-0.001	0.000	0.016
Constant			
Constant	3.997	0.023	< 0.001

Chapter 4

An efficient Gibbs sampler for structural inference

In this chapter we propose a Gibbs sampler for structural inference in Bayesian networks. The standard Markov chain Monte Carlo (MCMC) algorithms used for this problem are random-walk Metropolis-Hastings samplers, but for problems of even moderate dimension, these samplers often exhibit slow mixing. The Gibbs sampler proposed here conditionally samples the complete set of parents of a set of nodes in a single move, by blocking together particular components. The resulting MCMC algorithm mixes more rapidly.

In Chapter 5, we will examine the performance of the sampler using data simulated from the ALARM network, and on real datasets from a social science survey and a multi-variable single-cell molecular assay. We find that the existing approaches are unsatisfactory because they give results that are highly unstable across Monte Carlo replications, and across bootstrap replications of the data. In contrast, the proposed approach permits robust structural inference across a wide range of settings.

In this chapter, we introduce the Gibbs sampler, and describe how it can be im-

plemented efficiently. We start by introducing structural inference for Bayesian networks. We then describe a naïve Gibbs Sampler for this problem, which motivates the development of an improved Gibbs sampler, which makes larger moves by considering multiple parent sets together. We then describe how these algorithms can be implemented efficiently.

4.1 Introduction

4.1.1 Problems with small local moves

Most MCMC algorithms for model selection rely on small 'local' moves, based on the heuristic that models that are 'close' to each other will be similar and that the target distribution is at least very loosely locally monotonic. In many settings such algorithms converge to their target distribution rapidly and mix freely.

The standard random-walk Metropolis-Hastings algorithms used for structural inference of Bayesian networks, including MC^3 (Madigan and York, 1995) and variants that improve its efficiency (Giudici and Castelo, 2003), propose small, local changes to the current state. These proposals are accepted according to the usual acceptance probability. In some settings, in which the sample size of the observations is small, and the number of variables p in the Bayesian network is small, such samplers work well. Unfortunately, there are many settings in which using samplers that make small local moves of this kind will yield a sample with undesirable properties. This occurs particularly when the target distribution is 'peakier' than anticipated.

4.1.2 Methods for improving mixing

The fundamental shortcoming with making only small changes is that it leaves the algorithms incapable of 'escaping' local modes. Particularly in high dimensions,

this means that the MCMC sampler converges slowly to its target distribution, and does not mix well. A variety of general techniques is available to improve mixing. For example, methods that introduce an auxiliary variable can often improve convergence. However, the most successful methods, for example Hybrid/Hamiltonian Monte Carlo (Duane *et al.*, 1987), the Metropolis adapted Langevin algorithm (Roberts and Tweedie, 1996) and further developments (Girolami and Calderhead, 2011), use knowledge of the derivatives of the log target distribution. However, useful derivatives are clearly not directly available for discrete distributions.

A simple idea for improving mixing in these settings is to consider larger moves. However, to do this, we need to be able to identify large moves that focus on areas of significant posterior mass. This is often not straightforward. Even when it is, it is not clear precisely how large the moves should be. Some guidance in the discrete case for Metropolis-Hastings algorithms is given by Roberts (1998), who shows that the usual optimal acceptance rate of 0.234 applies. However this is only proved for a very specific example.

4.1.3 A Gibbs sampler

In this chapter we propose a method for constructing Gibbs samplers for structural inference of Bayesian networks. Gibbs samplers make moves that are tailored to the local form of the distribution by using the conditional distribution, and thus identify areas of significant posterior mass. The Gibbs sampler we consider here is also able to make large moves by using 'blocking'.

Specifically, the Gibbs sampler proposed here considers the parents of a set of nodes as a single component, and conditionally samples parents of each node in the set from the appropriate joint distribution. These moves are formed by 'blocking' together the parents of the set of nodes. Blocking allows the sampler to make 'large' moves that are sampled exactly from the local conditional posterior distribution, enabling

the sampler to locate and explore the areas of significant posterior mass efficiently. The method exploits the simple heuristic that the parents of a node are similar to the independent variables chosen in Bayesian variable selection, with the node as the dependent variable. The deficiency in the heuristic is that the acyclicity requirement of Bayesian networks is ignored. The Gibbs sampler is constructed around this idea, but exactly accounts for acyclicity so that the target distribution is indeed the true posterior distribution over Bayesian networks.

4.1.4 Constraints on in-degree

Typically it is not useful in applications to consider Bayesian networks in which random variables have many parents (large in-degree) because there is usually not enough data to estimate the parameters of extremely complex models, especially for discrete models. In addition, if a model averaging approach is taken, a constraint on in-degree is not as restrictive as it may at first seem; we discuss this further in Section 5.6 (page 127).

For these reasons, most methodologies for structural inference (e.g. Friedman and Koller, 2003; Koivisto and Sood, 2004) take advantage of the reduction in the cardinality of the space of Bayesian networks that is given by imposing a maximum in-degree (fan-in) restriction on the Bayesian networks. We adopt this restriction for the sampler introduced here, enabling dramatic improvements in mixing for this class of problems. As we show below, better mixing has clear practical consequences because it means that in finite compute time there is a much reduced chance of seeing extreme Monte Carlo artefacts (which might otherwise be reported as the output of inference).

4.2 Background and notation

4.2.1 Graphs and Bayesian networks

We first recall the definition of Bayesian networks (introduced in Section 2.2.3), and the related notation. A Bayesian network G is a directed, acyclic graph (DAG) with nodes $V = (1, \ldots, p)$, and directed edges $E \subset V \times V$.

Particularly in this chapter, we will make use of the specification of the edge set E of the graph G in terms of the parents G_j of each node j, for $j \in \{1, \ldots, p\}$. The parents G_j of node j are the subset of nodes V such that $i \in G_j \Leftrightarrow (i, j) \in E$. We refer to G_j as a parent set and use X_{G_j} to refer to the set of random variables that correspond to the parents G_j of node j in the graph G.

We will use the collection of parent sets $\langle G_1, \ldots, G_p \rangle$ to specify a graph G. Subsets thereof are denoted by $G_A = \langle G_k : k \in A \rangle$, and the subset given by the complement $A^C = \{1, \ldots, p\} \setminus A$ of a set A is denoted by $G_{-A} = \langle G_k : k \in A^C \rangle$. In particular, note that any graph G can be specified as $\langle G_i, G_{-i} \rangle = \langle G_1, \ldots G_p \rangle = G$ for any $i \in \{1, \ldots, p\}$.

4.2.2 Joint distribution and priors

The joint distribution of $\mathbf{X}$ is specified in terms of $p(X_i \mid \mathbf{X}_{G_i}, \theta_i)$, the conditional distribution with parameters θ_i of each X_i, given the parents $\mathbf{X}_{G_i}$ of node i in the Bayesian network. For structural inference our interest focuses on the posterior distribution on Bayesian networks $P(G \mid \mathbf{X})$, which is proportional to the product of the marginal likelihood $p(\mathbf{X} \mid G)$, and a prior $\pi(G)$ for the Bayesian network structure.

In principle we do not need to assume that the graph prior $\pi(G)$ takes any particular form, but the required computation is simplified if we assume the graph prior

factorises as $\pi(G) = \prod_{i=1}^{p} \pi_i(G_i)$ across the nodes of the graph. A prior satisfying this condition is called MODULAR (Friedman and Koller, 2003). Both flat priors over graph space, and informative priors that penalise or reward the presence or absence of particular edges (Werhli and Husmeier, 2007; Mukherjee and Speed, 2008) can be formulated in this manner. Note that the prior is not specified over the space of orders and so does not suffer from the difficulties involved in doing so (Ellis and Wong, 2008; Eaton and Murphy, 2007), in contrast to the methods used in Friedman and Koller (2003) and Koivisto and Sood (2004).

We will assume that conjugate priors for the parameters $\theta_i \mid G$ have been chosen, and assume that local parameter independence and modularity (Heckerman *et al.*, 1995) holds. Under the assumptions we have made, we can obtain a closed-form marginal likelihood. In addition, the marginal likelihood factorises across the nodes of the graph, and the posterior distribution on Bayesian networks is

$$P(G \mid \mathbf{X}) \propto \prod_{i=1}^{p} p(X_i \mid \mathbf{X}_{G_i}) \pi_i(G_i),$$

where $p(X_i \mid \mathbf{X}_{G_i})$ is the marginal likelihood for node i given the graph $G = \langle G_1, \ldots, G_p \rangle$. This is the target distribution for our sampler.

4.3 Preliminaries

To introduce the Gibbs sampler, we first recall the standard MC^3 sampler, and an analogous naïve Gibbs sampler. Usually convergence of Gibbs samplers follows from the Hammersley-Clifford theorem (Besag, 1974), but this does not apply in this context. An alternative argument is outlined.

4.3.1 MC^3 sampler

The standard sampler for structural inference for Bayesian networks is MC^3 (Madigan and York, 1995), which is a Metropolis-Hastings sampler that explores $\mathcal{G}$ by proposing to add or remove a single edge from the current graph G, subject to acyclicity. Each proposal G' is drawn uniformly at random from the neighbourhood $\nu(G)$ of the current graph, defined as the set of DAGs that differ from G by the addition or removal of a single edge. The proposal G' is accepted with probability $\min(1, \alpha(G', G))$, where

$$\alpha(G', G) = \min\left\{1, \frac{P(G' \mid \mathbf{X})\,|\nu(G')|^{-1}}{P(G \mid \mathbf{X})\,|\nu(G)|^{-1})}\right\}.$$

4.3.2 A naïve Gibbs sampler

Constructing a Gibbs sampler that is analogous to MC^3 is straightforward. To do this, we consider the posterior distribution on Bayesian networks to be a joint distribution for the off-diagonal entries in the adjacency matrix, which is a $p \times p$ matrix whose elements G_{ij} are indicator variables for whether G includes an edge from i to j, and whose diagonal elements $G_{ii} = 0$ for all i. We thus have $p(p-1)$ random variables G_{ij}, each of which takes the value 1 or 0. The proposal distribution of MC^3 can be viewed as proposing to toggle the value of G_{ij} of the adjacency matrix for some $i \neq j$, subject to the restriction that the proposal must be acyclic. A simple Gibbs sampler works in a similar way. At each step of the Gibbs sampler a sample from the conditional distribution of G_{ij} is drawn, for some $i, j \in \{1, \ldots, p\}, i \neq j$, given the rest of the graph $G_{ij}^C = \{G_{uv} : 1 \leq u \leq p, 1 \leq v \leq q\} \setminus \{G_{ij}\}$. Define G_{ij}^+ as the graph G with an edge from i to j, and G_{ij}^- as the graph G with no edge from i to j. If G_{ij}^+ is cyclic, G_{ij}^- is sampled with probability 1. If G_{ij}^+ is acyclic, the

conditional distribution of G_{ij} is Bernoulli.

$$P(G'_{ij} = g \mid G^C_{ij}) = \begin{cases} 1 & g = 0,\ G^+_{ij} \text{ cyclic} \\ 0 & g = 1,\ G^+_{ij} \text{ cyclic} \\ \dfrac{P(G^-_{ij} \mid \mathbf{X})}{P(G^-_{ij} \mid \mathbf{X}) + P(G^+_{ij} \mid \mathbf{X})} & g = 0,\ G^+_{ij} \text{ acyclic} \\ \dfrac{P(G^+_{ij} \mid \mathbf{X})}{P(G^-_{ij} \mid \mathbf{X}) + P(G^+_{ij} \mid \mathbf{X})} & g = 1,\ G^+_{ij} \text{ acyclic} \end{cases} \tag{4.1}$$

The choice of i and j can either be made sequentially (systematically) or randomly. There are few theoretical results to guide the choice of random- and systematic-scan Gibbs samplers (Roberts and Sahu, 1997); here, random-scan Gibbs samplers are used throughout.

This naïve Gibbs sampler offers no advantages over MC^3. However, thinking of structural inference from a Gibbs sampling perspective opens up the possibility of drawing on ideas from the Gibbs sampling literature to improve the mixing rate of the MCMC algorithm, which we discuss in Section 4.4.

4.3.3 Convergence conditions for Gibbs samplers

Convergence of a Gibbs sampler for Bayesian networks does not follow from the usual justification of Gibbs sampling that relies on the Hammersley-Clifford theorem (Besag, 1974). The theorem gives a positivity condition that is sufficient to prove that the univariate conditional distributions, used by the Gibbs sampler, uniquely define the joint distribution. The required condition is that the support of the joint distribution is given by the Cartesian product of the supports of the marginal distributions. An example of when this condition does not hold is the density $p(x, y)$ with support only on $[0, 1] \times [0, 1]$ and $[2, 3] \times [2, 3]$. Clearly $p(x)$ and $p(y)$ are both positive on $[0, 1]$ and $[2, 3]$ but neither $[0, 1] \times [2, 3]$ or $[2, 3] \times [0, 1]$ are in the support

of the joint distribution (Hobert *et al.*, 1997; O'Hagan and Forster, 2004).

The acyclicity requirement of Bayesian networks means that this positivity condition is not satisfied. Consider a Bayesian network consisting of two correlated random variables X_1 and X_2. The correlation means that both the graph with a single edge $1 \to 2$ and the graph with a single edge $2 \to 1$ have positive probability. Thus $P(G_{12} = 1) > 0$ and $P(G_{21} = 1) > 0$ in the marginal distributions. However, the joint distribution $P(G_{12} = 1 \text{ and } G_{21} = 1) = 0$ because the corresponding graph (the complete graph) is cyclic. The complete graph is thus not in the support of the joint distribution but is in the Cartesian product of the supports of the marginal distributions.

An alternative sufficient condition for uniqueness of the joint distribution and convergence of the Gibbs sampler when positivity is not satisfied is given by Besag (1994) in a discussion of Tierney (1994), which was expanded upon in continuous settings by Hobert *et al.* (1997). The condition requires that for every $G^{(0)} \in \mathcal{G}$ and $G \in \mathcal{G}$ there exists a finite sequence $G^{(1)}, \ldots, G^{(d)}$, with $G^{(d)} = G$ and $d \in \mathbb{N}$, such that $G^{(i)}$ and $G^{(i-1)}$ differ in only a single component, and that the joint distribution $P(G^{(i)}) > 0$ for all $i = 1, \ldots d$. When the graph prior $\pi(G) > 0$ for all G, this condition is clearly satisfied: one such finite sequence removes every edge of $G^{(0)}$, one at a time, and then adds every edge of G, one at a time. Each graph in the sequence is clearly acyclic, since the sequence is composed of subgraphs of the acyclic $G^{(0)}$ and G, and so has positive probability in the joint distribution when the graph prior is positive everywhere in $\mathcal{G}$. A similar proof follows if the graph prior has support on all subgraphs of graphs with support in the graph prior, as is true for most widely used priors.

4.4 Optimising Gibbs samplers

The mixing of Metropolis-Hastings algorithms depends strongly upon the choice of proposal distribution, which is often chosen for convenience to be a local random-walk. Gibbs samplers make moves according to the full conditional distributions. These distributions have the attractive property that they exactly reflect some local structure of the target distribution.

Nonetheless, Gibbs sampling is not always efficient. Inefficiency occurs when there is strong correlation between the components of the random vector. To see this, consider a Gibbs sampler for a multivariate continuous distribution with highly correlated components. At each step, a single component of the random vector is sampled according to its conditional distribution, but since this component is strongly correlated with another component, the conditional distribution is concentrated on only a small part of its support. This means that the sampler is likely to make only small moves, and thus explore the sample space slowly. The same issue arises with discrete distributions.

For Bayesian networks, there is strong dependence between the edge indicator variables G_{ij}, particularly for the collections $\{G_{ij} : i \in \{1, \ldots, p\}\}$ for each $j \in \{1, \ldots, p\}$ that correspond to parent sets. For example, there may be random variables X_r and X_s that do not individually predict X_j well, but do when taken in combination. In this case, G_{rj} and G_{rs} will be correlated. Another possibility is of two pairs of random variables X_r, X_s and X_u, X_v that in combination both predict X_j well, but such that any of the four random variables individually do not. In this case, the probability of transitioning from a graph in which the parents of X_j are X_r and X_s to a graph in which the parents of X_j are X_u and X_v may be extremely low when using a sampler that only makes single edge changes, such as MC^3.

In addition to this local form of dependence, the acyclicity restriction creates strong

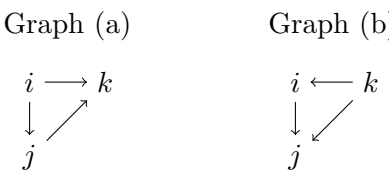


Figure 4.1: Illustrative graphs of when small local moves may fail to enable transitions between two regions of high probability. If both (a) and (b) have high probability, the near-cyclic nature of the graphs makes transitions between (a) and (b) difficult.

dependence between the parents of separate nodes. For example, suppose two random variables X_i and X_j are strongly correlated so that both the edge (i, j) and the reversed edge (j, i) have high probability. If the edge (i, j) is present, the probability of it being removed is low, but its presence precludes the reversed edge (j, i) from ever being added. It is this possibility that motivates the 'edge reversal' move that is commonly used in variants of the MC^3 algorithm.

More complex dependence is also possible. If three nodes are strongly correlated then many of the 'almost cyclic' graphs will have high probability. For example, suppose both graph (a) and (b) in Figure 4.1 have high probability. Since reversing the edge $i \to k$ forms a cycle in (a), moves that consider only the parents of a pair of nodes i and k at the same time will not move between graphs (a) and (b) easily. Samplers that alter only a single edge indicator, such as MC^3, will also fail. However, if the parents of all three nodes are sampled jointly, the sampler is able to move between graphs (a) and (b) easily.

One method for alleviating this problem is to transform the distribution so that the components of the random variable are not correlated. In general, finding a suitable transformation can be very difficult, and for Bayesian networks would need to encapsulate the requirement for acyclicity.

Instead we propose to group a number of the components together and sample from their joint conditional distribution. In Gibbs sampling, this is known as 'block-

ing'. The method is widely thought to be beneficial in settings such as this in which there is strong correlation between components of the random variable. In the case of multivariate normal distributions, Roberts and Sahu (1997) have shown that for random-scan Gibbs sampling, convergence improves when components of the random vector are sampled as blocks. By sampling from the joint conditional distribution of a group of components we avoid the issues caused by any correlation between these components because the joint conditional distribution naturally incorporates the correlation structure, and so can account for it.

4.5 A Gibbs sampler for Bayesian networks

As we noted above, the efficiency of a Gibbs sampler can be improved by blocking together a group of components, and sampling from their joint conditional distribution. In theory, any group of components can be taken as a block, but sampling from their joint conditional distribution needs to be possible, and ideally simple. The blocks that we consider correspond to the parent sets of a set of nodes, so that the parent sets of several nodes to be considered simultaneously, ameliorating the problems caused by the correlations described in Section 4.4.

Let $W \subseteq V$ denote a subset of $\rho = |W|$ nodes whose parent sets are sampled together as a block. Let $F_W = \langle F_{w_1}, \ldots, F_{w_\rho} \rangle$ denote the collection of parent sets for the nodes in W. We propose to group these nodes and sample $F_{w_1}, \ldots, F_{w_\rho}$ jointly so that in each Gibbs step the algorithm selects the set $W = (w_1, \ldots, w_\rho)$ of nodes uniformly at random, and samples new parents for all nodes in W. Each block $\{G_{ij} : i \in \{1, \ldots, p\}, j \in \{w_1, \ldots, w_\rho\}, i \neq j\}$ consists of the indicator variables that determine the parents of the nodes in W. Note that for computational reasons $|W|$ must be small.

It is natural that each block is a collection of parent sets because we can parameterise

both G and its marginal likelihood $p(\mathbf{X} \mid G)$ by parent sets $G_1, \ldots, G_p$. The marginal likelihood factorises across nodes into conditionals $p(X_j \mid X_{G_j})$, each of which is a function of parents G_j of that node.

$$P(G_1, \ldots, G_p \mid \mathbf{X}) \propto \prod_{i=1}^{p} p(X_i \mid \mathbf{X}_{G_i}) \pi(G_i)$$

At each Gibbs step, new parent sets F_W for nodes in W are sampled, whereas the parent sets G_{-W} for nodes not in W remain the same. The new graph $G' = \langle F_W, G_{-W} \rangle$ is thus formed by changing the parents of the nodes in W to $F_{w_1}, \ldots, F_{w_\rho}$, and leaving the parents of nodes not in W unchanged.

To be able to construct a Gibbs sampler using these blocks, we need to find the conditional distribution on F_W, given the parent sets G_{-W} of nodes not in W. For parent sets F_W such that $G' = \langle F_W, G_{-W} \rangle$ is cyclic, the conditional probability is 0. Let $\mathfrak{F}_W$ be the set of collections F_W of parent sets $F_{w_1}, \ldots, F_{w_\rho}$ such that $G' = \langle F_W, G_{-W} \rangle$ is acyclic. For $F_W \in \mathfrak{F}_W$, the conditional posterior distribution is multinomial, with weights given by the posterior distribution of the graph $G = \langle F_W, G_{-W} \rangle$.

$$\begin{aligned} P(F_W \mid G_{-W}, \mathbf{X}) &= \frac{P(F_W, G_{-W} \mid \mathbf{X})}{P(G_{-W} \mid \mathbf{X})} \\ &= \frac{P(G' \mid \mathbf{X})}{\sum_{F_W \in \mathfrak{F}_W} P(F_W, G_{-W} \mid \mathbf{X})} \end{aligned} \tag{4.2}$$

Algorithm 5 (below) outlines the algorithm. The correctness of the sampler can be easily proved using the condition given by Besag (1994) in the same way that correctness of the naïve Gibbs sampler is proved in Section 4.3.3, and in fact the requirements on the graph prior will be weaker than for the naïve sampler.

We need to be able to sample from $P(F_W \mid G_{-W}, \mathbf{X})$ and so its normalising constant poses a problem. Since Bayesian networks with large in-degree are rarely of interest in applications (see Section 4.1.4 on page 78), we can reduce this problem by intro-

ducing a restriction on the maximum in-degree κ of each node. We constrain the in-degree to a maximum of $\kappa = 3$ in the examples in Chapter 5. We view this as only a minor restriction, because even higher order interactions are visible when a Bayesian model averaging approach is taken (see further discussion in Section 5.6). We investigate the computational aspects of sampling from $P(F_W \mid G_{-W}, \mathbf{X})$ in Section 4.6, where we describe a two-stage approach to sampling.

Any choice of $|W|$ is in principle possible. However, when $|W|$ is large, the computational requirement for evaluating the conditional distribution in Equation 4.2 is unmanageable. The parameter can be used to tune the algorithm. Throughout Chapters 4 and 4, we use $|W| = 3$, which ensures that all scenarios described in Section 4.4 are avoided.

When $|W| = 1$, it is interesting to note that if all choices of parent set do not induce a cycle, $\mathfrak{F}_W$ equals the power set $\wp(V \setminus \{w_1\})$, with $W = w_1$. When this occurs the conditional distribution (Equation 4.2) can be viewed as the posterior distribution of a standard Bayesian variable selection problem with dependent variable w_1, and the other variables as independent variables. If the addition of particular nodes would introduce a cycle, we have a constrained Bayesian variable selection problem. Suppose that a cycle would be created by adding nodes $b_1, \ldots, b_k \in V$, $k \in \{1, \ldots, p\}$ as parents of node w_1. In this case $\mathfrak{F}_W = \wp(V \setminus \{w_1, b_1, \ldots, b_k\})$ and the conditional distribution in Equation 4.2 is a constrained Bayesian variable selection in which the variables corresponding to the nodes $b_1, \ldots, b_k$ are excluded.

4.6 Computational aspects

Up to this point we have not discussed the computational aspects of the algorithm. In this section, we describe how the algorithm described above can be implemented efficiently. A key bottleneck for many MCMC algorithms for structural inference for

Algorithm 5 A Gibbs sampler, with blocks

Initialise starting point $G^{(0)} = \langle G_1^{(0)}, \ldots, G_p^{(0)} \rangle$

for t in 1 to N **do**

Sample W from V

Draw $F_W \sim P(F_W \mid G_{-W}^{(t-1)}, \mathbf{X})$

Set $G^{(t)} \leftarrow G = \langle F_W, G_{-W}^{(t-1)} \rangle$

end for

Bayesian networks is checking for cycles, and so we first describe how these checks can be made quickly and efficiently. Sampling efficiently from the conditional distribution in Equation 4.2 is also not straightforward because $\mathfrak{F}_W$ has large cardinality. We introduce a two-stage approach to sampling that reduces this problem. We use these methods together to efficiently implement the Gibbs sampler described above.

4.6.1 Online cyclicity checking

Bayesian networks are described by DAGs, and so any algorithm that explores the space of Bayesian networks must ensure that each graph considered does not include a cycle. This constraint must be considered at each step of the algorithm, and so this is often a key bottleneck in algorithms for structural inference. There are various methods of checking for cycles. In the following section, we describe how such checks can be made using the transitive closure. We then describe an online algorithm for updating the transitive closure. An online algorithm greatly improves efficiency because, at each MCMC iteration, many parts of the Bayesian network do not change.

Checking for cycles

The most straightforward method for checking for cycles is depth-first search, which takes $\mathcal{O}(p+\epsilon)$ time, where ϵ is the number of edges in the graph. For MC^3, to evaluate the acceptance probability we must consider all possible single-edge changes to a directed graph of which there are $\mathcal{O}(p^2)$, and so checking for cycles at each iteration takes $\mathcal{O}(p^3)$ time in the worst case.

Using a $\mathcal{O}(p^3)$ algorithm at each step creates a bottleneck in the algorithm, but we can avoid this by using ideas first proposed in this context by Giudici and Castelo (2003). We describe an alternative method that was proposed in the dynamic algorithms literature by King and Sagert (2002). Let T^G be the transitive closure of the current state of the sampler, which for a graph $G = (V, E)$ is defined as the directed graph $T^G = (V, E^T)$, where $(i, j) \in E^T$ if and only if a path (obeying edge directions) from i to j exists in G. Knowing the transitive closure is of use because its adjacency matrix $T^G = (T^G_{ij})$ immediately reveals which alterations can be made to G without introducing a cycle. The addition of an edge (i, j) introduces a cycle if and only if $T^G_{ji} = 1$. Removing an edge from G never introduces a cycle. The adjacency matrix of the transitive closure therefore enables graphs created by single-edge additions to be screened for cycles in $\mathcal{O}(1)$ time.

Online transitive closure updates

The transitive closure for an arbitrary directed graph can be determined in $\mathcal{O}(p^\omega)$ time (Munro, 1971), where ω is the best known exponent for matrix multiplication (Coppersmith and Winograd, 1990, show $\omega < 2.376$). However, only incremental changes are made to the current state G of the sampler, so a dynamic algorithm can be used to compute the transitive closure more efficiently. We need a fully dynamic transitive closure algorithm, so that both insertion and deletion of edges

are supported. This problem has been the subject of significant interest in the dynamic algorithms literature; for an overview see Demetrescu *et al.* (2010).

Algorithms for this problem provide a procedure for querying the transitive closure, and procedures that update the transitive closure when an edge is added or removed from the graph. A trade-off exists between the performance of these two operations (Demetrescu and Italiano, 2005). We choose to implement the algorithm introduced by King and Sagert (2002), which allows queries to be performed in $\mathcal{O}(1)$ time, and updates in $\mathcal{O}(p^2)$ worst-case time, assuming a word size of $\mathcal{O}(\log p)$. This bound is thought to be the best bound possible for updates that retains $\mathcal{O}(1)$ queries (Demetrescu and Italiano, 2005), yet the algorithm is simple to implement.

The algorithm maintains a path count matrix $C^G = (C^G_{ij})$, where C^G_{ij} is the number of distinct paths from node i to node j in G. Clearly, $T^G_{ij} = 1$ if and only if $C^G_{ij} > 0$, and so query operations are performed in $\mathcal{O}(1)$ by simply checking whether the relevant component of C^G is positive.

The routines for updating C^G when an edge is added or removed are also straightforward. We first consider adding an edge (i, j) to a graph G to form a graph G'. Denote the i^{th} column of C^G by $C^G_{\bullet i}$, and the j^{th} row by $C^G_{j\bullet}$. The increase in the number of distinct paths between any two nodes a and b is given by the (a, b) element of the outer product of $C^G_{\bullet i}$ and $C^G_{j\bullet}$. The path count matrix for G' is thus formed by adding this outer product, denoted by $\otimes$, to the existing path count matrix.

$$C^{G'} = C^G + C^G_{\bullet i} \otimes C^G_{j\bullet}$$

Updating C^G when an edge (i, j) is removed from the graph is performed analogously.

$$C^{G'} = C^G - C^G_{\bullet i} \otimes C^G_{j\bullet}$$

This algorithm is simple to implement, and provides a fast method for determining

which edges can be added to a DAG without introducing a cycle.

4.6.2 Efficient implementation of a Gibbs sampler

The key part of an implementation of the Gibbs sampler in Algorithm 5 (above) is the method of sampling from the conditional distribution $P(F_W \mid G_{-W}, \mathbf{X})$ in Equation 4.2. To do this exactly, we need to be able to evaluate its normalising constant. This is not straightforward for two reasons. First, we need to be able to identify the collection of parent sets $F_W \in \mathfrak{F}_W$ for which the graph $\langle F_W, G_{-W} \rangle$ is acyclic. Second, the cardinality of $\mathfrak{F}_W$ may be large. The methods described in Section 4.6.1 enable identification of the collection of parent sets $F_W \in \mathfrak{F}_W$ that form acyclic graphs.

In this section, we describe the details of how the difficulty in sampling from a distribution with large cardinality can be managed. First, the scale of the problem is reduced by enforcing a maximum in-degree, as described in Section 4.1. Then we use a two-stage approach that first samples a component of a partition of $\mathfrak{F}_W$, and then samples a member of $\mathfrak{F}_W$ in that component. In the remainder of this section we detail the partition of $\mathfrak{F}_W$ used, and then describe how this can be used in a two-stage sampling procedure.

The key idea is to choose the partition of F_W so that, conditional on a component of the partition, the parents of each node are independent. This enables the efficient two-stage sampling method described in Section 4.6.2. The partition is specified through a DAG on the nodes in W. Membership of a particular component of the partition of F_W is specified through separate conditions on the parent set of each node in W. This gives the desired property: that conditional on a component of the partition, any choice of parent sets (allowed by the component of the partition) for each node yields an acyclic graph and the parents of each node are independent.

Forming the partition

It is convenient to consider partitioning $\mathfrak{F} = \{\langle F_W, G_{-W}\rangle : F_W \in \mathfrak{F}_W\}$, rather than partitioning $\mathfrak{F}_W$ directly. The set $\mathfrak{F}$ consists of all the acyclic graphs that can be formed from the set of collections of parent sets $\mathfrak{F}_W$. The partition of $\mathfrak{F}$ will take the form $\mathfrak{F} = \{\mathfrak{F}^{H^1}, \ldots, \mathfrak{F}^{H^\eta}\}$. We describe the form of the components $\mathfrak{F}^{H^h}$, $h = 1, \ldots, \eta$, in the following.

The partition of $\mathfrak{F}$ is formed by considering DAGs $H = (W, F)$ on the set of nodes W, with edges $F \subset W \times W$ defined by the parent sets $\langle H_{w_1}, \ldots, H_{w_\rho}\rangle$. Let $\mathcal{H} = \{H^1, \ldots, H^\eta\}$, with cardinality η, be the set of all DAGs on the nodes in W. Each graph H is associated with a set $\mathfrak{F}^H$ of graphs in $\mathfrak{F}$ and the components of the partition $\{\mathfrak{F}^{H^1}, \ldots, \mathfrak{F}^{H^\eta}\}$ are these sets. We will show that $\{\mathfrak{F}^{H^1}, \ldots, \mathfrak{F}^{H^\eta}\}$ is a partition of $\mathfrak{F}$ in Lemma 1 below.

We now describe the relation between a DAG $H \in \mathcal{H}$ and the associated set $\mathfrak{F}^H$ of graphs. For a particular H, the set $\mathfrak{F}^H$ is formed in the following manner, starting from a graph G.

First, we form the reduced graph $G^- = \langle G^-_{w_1}, \ldots, G^-_{w_\rho}, G^-_{-W}\rangle$ by removing any edges that are directed into W, so that $G^-_{w_j} = \emptyset$, for all $j \in 1, \ldots, \rho$, and $G^-_i = G_i$ for all nodes i such that $i \notin W$.

To define $\mathfrak{F}^H$ we will require notation for the following sets. We define $D_j = \{i : T^{G^-}_{w_j i} = 1\}$ for each node $w_j \in W$ to be the nodes that are descendants in G^- of node w_j, and $K_j = \{i : T^{G^-}_{w_j i} = 0\}$ to be the nodes that are not descendants (non-descendants) in G^- of node $w_j \in W$. Note that node $w_j \in D_j$ but that $w_j \notin K_j$ by definition. In addition we make the following definitions, in which we use the definition that the edges of H are specified by the parent sets $\langle H_{w_1}, \ldots, H_{w_\rho}\rangle$ of each node in W.

- The set of nodes $K = \bigcap_{k=1,\ldots,\rho} K_k$ that are not descendants in G^- of any node

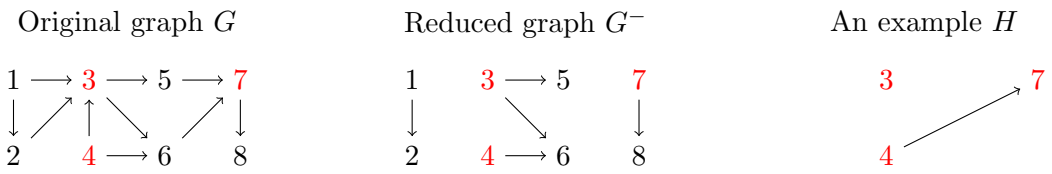


Figure 4.2: An illustrative example of the relevant graphs and sets, with $W = \{w_1, w_2, w_3\} = \{3, 4, 7\}$ shown in red. From the original graph G, the edges into W are removed to form G^-. If we choose H as shown, we get $K = \{1, 2\}$, and for $w_3 = 7$ we get $D_3^H = \{4, 6\}$ and $D_{-3}^H = \{3, 5, 6, 7, 8\}$.

in W.

- The descendants $D_j^H = \bigcup_{\{w_k \,:\, w_k \in H_{w_j}\}} D_k$ in G^- of the parents in H of node w_j.
- The descendants $D_{-j}^H = \bigcup_{\{w_k \,:\, w_k \notin H_{w_j}\}} D_k$ in G^- of nodes in W that are not parents in H of node w_j.

Figure 4.2 illustrates the notation.

We now describe the conditions that define membership of the set $\mathfrak{F}^H$ of graphs, for some $H \in \mathcal{H}$. A graph $F = \langle F_W, G_{-W} \rangle$ is a member of $\mathfrak{F}^H$, where $H = \langle H_{w_1}, \ldots, H_{w_\rho} \rangle$, if and only if the collection F_W of parent sets $F_{w_1}, \ldots, F_{w_\rho}$ satisfies the following two conditions for all $j = 1, \ldots, \rho$.

(A) $F_{w_j} \subseteq \left(K \cup D_j^H \right) \setminus D_{-j}^H$

(B) $F_{w_j} \cap \left(D_k \setminus D_{-j}^H \right) \neq \emptyset$ for all nodes $w_k \in H_{w_j}$

Note that (B) depends on D_k not D_k^H.

The condition (A) ensures that no cycle is formed in the graph. The condition prevents cycles because each parent of a node $w_j \in W$ must either be a non-descendant in G^- of any node in W, or a node whose ancestors in G^- are all parents in H of w_j. In particular, no descendant of w_j is added as an ancestor of w_j, which would allow a cycle to be formed.

The condition (B) ensures that there is a unique $H \in \mathcal{H}$ such that $F \in \mathfrak{F}^H$. Uniqueness is required for $\{\mathfrak{F}^{H^1}, \ldots, \mathfrak{F}^{H^\eta}\}$ to be a partition of $\mathfrak{F}$. The condition enforces uniqueness by ensuring each edge in H is 'used', by checking that for a node $w_j \in W$, at least one descendant $v_l \in V$ of each of its parents in H is in F_{w_j} and that v_l is not a descendant in G^- of a node $w_k \in W$ that is not a parent in H of w_j. An example of the need for this condition is the graph $F = \langle F_W, G_W \rangle$ with F_W such that $F_{w_j} = \emptyset$ for all $j = 1, \ldots, \rho$. Without condition (B), F would be in $\mathfrak{F}^H$ for all $H \in \mathcal{H}$, and thus $\{\mathfrak{F}^{H^1}, \ldots, \mathfrak{F}^{H^\eta}\}$ would not be a partition of $\mathfrak{F}$.

Lemma $\{\mathfrak{F}^{H^1}, \ldots \mathfrak{F}^{H^\eta}\}$ *form a partition of* $\mathfrak{F}$.

Proof. We show that the graphs form a partition by showing

(i) $\bigcup_{h=1,\ldots\eta} \mathfrak{F}^{H^h} = \mathfrak{F}$, and

(ii) $\mathfrak{F}^{H^{h_1}} \cap \mathfrak{F}^{H^{h_2}} = \emptyset \quad$ for $H^{h_1} \neq H^{h_2} \quad$ with $\quad H^{h_1}, H^{h_2} \in \mathcal{H}$.

(i) $\underline{\bigcup_{h=1,\ldots,\eta} \mathfrak{F}^{H^h} \subseteq \mathfrak{F}}$

We proceed by showing that $\mathfrak{F}^{H^h} \subseteq \mathfrak{F}$ for all $h = 1, \ldots, \eta$. To show this, starting from a DAG G, we need, for each $h = 1, \ldots, \eta$, that each $F = \langle F_W, G_{-W} \rangle \in \mathfrak{F}^{H^h}$ is such that

(a) The parents in F of nodes not in W match those in G, and

(b) F is acyclic.

By definition of $\mathfrak{F}^{H^h}$, (a) is true.

To prove (b), first note that G^- is acyclic because G^- is a subgraph of the acyclic G. We proceed by contradiction.

Suppose some graph $F \in \mathfrak{F}^{H^h}$ is cyclic. Since F differs from the acyclic G^- only in the parents of nodes in W, any cycle in F must include at least one node in W. Let $c_1, \ldots, c_d \in W$, $d \in \{1, \ldots, \rho\}$, be the (minimal) complete set

of nodes in W included in some cycle in F. Denote the existence of a path (that obeys the edge directions) in F from node $w_a \in W$ to node $w_b \in W$ that does not include any nodes in W (except w_a and w_b) by $w_a \rightsquigarrow w_b$, and without loss of generality suppose that $c_1 \rightsquigarrow c_2 \rightsquigarrow \cdots \rightsquigarrow c_d$ in F. Note that since $c_1, \ldots, c_d$ is the complete set of nodes in W in the cycle, no node between c_i and c_{i+1} in the path can be in W, $i \in \{1, \ldots, d-1\}$.

We now show that for $w_a, w_b \in W$, $w_a \rightsquigarrow w_b$ only if an edge $w_a \to w_b$ links node w_a to w_b in H^h. Since $w_a \rightsquigarrow w_b$, there must exist a node $v_l \in V$ that is a parent of w_b in F such that v_l is a descendant in F of w_a. Note that in some cases $v_l = w_a$. Since v_l is a parent of node w_b in the graph F, $v_l \in \left(K \cup D_b^{H^h}\right) \setminus D_{-b}^{H^h}$, since $w_b \in W$. Also since $w_a \rightsquigarrow w_b$ does not include any nodes in W, v_l is also a descendant in G^- of w_a. We proceed by contradiction. Suppose no edge $w_a \to w_b$ exists in H^h. Then v_l is a descendant of w_a in the graph G^-, but w_a is not a parent of w_b in the graph H^h. So $v_l \in D_{-b}^{H^h}$, which is a contradiction. Thus $w_a \rightsquigarrow w_b$ only if $w_a \to w_b$ in H^h, for $w_a, w_b \in W$.

Now, recall that $c_1 \rightsquigarrow c_2 \rightsquigarrow \cdots \rightsquigarrow c_d$. Since a cycle is formed we must in addition have a path in F from node c_d to node c_1. Since $c_1, \ldots, c_d$ is the complete set of nodes in W involved in the cycle, no node on the path from c_d to c_1 can be in W. Thus $c_d \rightsquigarrow c_1$. However, this implies that $c_1 \to c_2 \to \cdots \to c_d \to c_1$ in H^h, which implies H^h is cyclic. But H^h is acyclic by assumption, and so we have a contradiction. Thus F is acyclic.

$\mathfrak{F} \subseteq \bigcup_{h=1,\ldots,\eta} \mathfrak{F}^{H^h}$

Suppose we start from a graph $G = \langle G_W, G_{-W} \rangle$. We want to show that for each DAG G' that is identical to G in its parents of nodes in W, there is some $H \in \mathcal{H}$ such that $G' \in \mathfrak{F}^H$. Thus consider $G' = \langle G'_W, G'_{-W} \rangle$, with the collection of parent sets $G'_{-W} = G_{-W}$ and with G'_W any collection of parent

sets such that G' is a DAG.

We will show that $G' \in \mathfrak{F}^{H'}$, where $H' = \langle H'_1, \ldots, H'_\rho \rangle \in \mathcal{H}$ is a DAG on nodes in W. For each node $w_j \in W$, the parents H'_{w_j} of w_j in H' are defined as follows.

$$H'_{w_j} = \{w_k \in W : \text{ there exists some } v_l \in G'_{w_j} \text{ such that } v_l \in D_k\}$$

As usual, G'_{w_j} is the parent set in the graph G' of the node w_j; and D_k is the descendants in $(G')^-$ of the node w_k.

Note that H' is a subgraph (on the nodes in W) of the transitive closure $T^{G'}$. By definition, G' is a DAG, so $T^{G'}$ is also a DAG, and thus H' is a DAG.

We show that $G' = \langle G'_1, \ldots, G'_p \rangle \in \mathfrak{F}^{H'}$ by showing that for each node $w_j \in W$, both conditions (A) and (B) that specify membership of $\mathfrak{F}^{H'}$ are satisfied.

(a) $G'_{w_j} \subseteq \left(K \cup D_j^{H'}\right) \setminus D_{-j}^{H'}$

Let $v_l \in G'_{w_j}$ meaning that v_l is a parent of the node w_j in the graph G'. First we show that $v_l \notin D_{-j}^{H'}$, and then show that $v_l \in K \cup D_j^{H'}$.

To see that $v_l \notin D_{-j}^{H'}$, note that if $v_l \in D_{-j}^{H'}$ then v_l must be a descendant in $(G')^-$ of some node $w_k \in W$ that is not in H'_{w_j}. However, every such w_k is in H'_{w_j} by the definition of H'_{w_j}, thus $v_l \notin D_{-j}^{H'}$.

To see that $v_l \in K \cup D_j^{H'}$, we suppose $v_l \notin K$ and show this implies that $v_l \in D_j^{H'}$. This follows because if $v_l \notin K$ then it must be the descendant in $(G')^-$ of some node $w_k \in W$. Then $w_k \in H'_{w_j}$ by definition of H'. Therefore $v_l \in D_j^{H'}$, as required. Thus $v_l \in K \cup D_j^{H'}$.

(b) $G'_{w_j} \cap \left(D_k \setminus D_{-j}^{H'}\right) \neq \emptyset$ for all $w_k \in H'_j$

Consider $w_k \in H'_{w_j}$. By the definition of H'_{w_j}, this means that there exists some node $v_l \in G'_{w_j}$ such that $v_l \in D_k$.

Additionally, any $v_{k'} \in W$ such that $v_l \in D_{k'}$ is such that $v_{k'} \in H'_{w_j}$, by definition of H'_{w_j}. Thus v_l is not a descendant in $(G')^-$ of any node in W that is not in H'_{w_j}. Then $v_l \in G'_{w_j} \cap (D_k \setminus D^{H'}_{-j})$, and thus the condition is satisfied.

(ii) $\underline{\mathfrak{F}^{H^{h_1}} \cap \mathfrak{F}^{H^{h_2}} = \emptyset}$

Since $H^{h_1} \neq H^{h_2}$, there must be at least one node that has a different parent in H. Suppose that the node w_j is such a node, and that w_k is a parent of w_j in H^{h_1} but not in H^{h_2}.

Consider a graph $G^{(1)} = \langle G^{(1)}_{w_j}, G^{(1)}_{-w_j} \rangle \in \mathfrak{F}^{H^{h_1}}$. Recall that, in particular,

$$G^{(1)}_{w_j} \cap \left(D_k \setminus D^H_{-j}\right) \neq \emptyset \qquad \text{for } w_k \in H^{h_1}_{w_j}.$$

Since this condition must be satisfied, there must exists some $v_l \in D_k \setminus D^H_{-j}$ such that $v_l \in G^{(1)}_{w_j}$.

We will show that for every graph $G^{(2)} = \langle G^{(2)}_{w_j}, G^{(2)}_{-w_j} \rangle \in \mathfrak{F}^{H^{h_2}}$, it is the case that $v_l \notin G^{(2)}_{w_j}$, meaning that no graph is in both $\mathfrak{F}^{H^{h_1}}$ and $\mathfrak{F}^{H^{h_2}}$.

This follows because $v_l \in D_k$ and so is a descendant in $(G')^-$ of w_k, which is not a parent of w_j in H^{h_2}. Thus $v_l \in D^{H^{h_2}}_{-j}$, and so $v_l \notin \left(K \cup D^H_j\right) \setminus D^H_{-j}$. Therefore v_l cannot be a parent of w_j in $G^{(2)}$.

■

We need to be able to find $\mathfrak{F}^H$ easily so that we can draw samples easily. First define $\mathfrak{F}^H_W$, for a graph $H \in \mathcal{H}$, to be the collection of parent sets of nodes $w_j \in W$ such that for all $F_W \in \mathfrak{F}^H_W$ graphs $\langle F_W, F_{-W} \rangle \in \mathfrak{F}^H$. Then, for a given $H \in \mathcal{H}$ and for each node $w_j \in W$, define $\mathfrak{F}^H_j$ as the set of parent sets that satisfy both (A) and (B). Note that, since membership of $\mathfrak{F}^H$ is defined by a property of the each parent set of nodes in W, $\mathfrak{F}^H_W$ is simply the Cartesian product of the parent sets $\mathfrak{F}^H_j$ for

$w_j \in W$.

$$\mathfrak{F}_W^H = \bigtimes_{w_j \in W} \mathfrak{F}_j^H$$

This simplicity to the structure of $\mathfrak{F}_W^H$ is the key to the efficiency of the method.

It is straightforward to find $\mathfrak{F}_j^H$ directly from (A) and (B), by noting that nodes in D_{-j}^H may not be parents of w_j, but that at least one node in $R_k = D_k \setminus D_{-j}^H$ for each $w_k \in H_{w_j}$ must be a parent. Let $\mathfrak{G}_j$ be the complete set of all parent sets of node j (subject to a maximum in-degree κ), and create look-up tables $\mathfrak{G}_j^i$, $i \in \{1, \ldots, p\}$ that list the parent sets that contain i.

The set $\mathfrak{F}_j^H$ can be found by considering the set of parent sets that include nodes that cannot be parents of w_j

$$Q_j = \bigcup_{l \in D_{-j}^H} \mathfrak{G}_j^l,$$

and the set of parent sets that include at least one descendant of all parents of w_j in H

$$S_j = \bigcap_{w_k \in H_{w_j}} \bigcup_{r \in R_k} \mathfrak{G}_j^r.$$

Satisfying (A) requires that $\mathfrak{F}_j^H \subseteq \mathfrak{G}_j \setminus Q_j$ and, when the parents $H_{w_j} \neq \emptyset$, we need $\mathfrak{F}_j^H \subseteq S_j$ to satisfy condition (B). These conditions give the following expression for $\mathfrak{F}_j^H$, which can be evaluated efficiently.

$$\mathfrak{F}_j^H = \begin{cases} S_j \setminus Q_j & \text{if } H_{w_j} \neq \emptyset \\ \mathfrak{G}_j \setminus Q_j & \text{if } H_{w_j} = \emptyset \end{cases}$$

We fix $|W|$ to be a small constant for all p so that $\left|H_{w_j}\right|$ does not increase and enforce a maximum in-degree κ. In this setting $\mathfrak{F}_j^H$ can be evaluated in $\mathcal{O}(p^{\kappa+1})$ time by storing the lookup-tables $\mathfrak{G}_j^i$ as a bit map.

Two-stage sampling of new parent sets

The partition is used in the two-stage approach to sampling in the following manner. We first sample a component $\mathfrak{F}^H$ of the partition from $P(\mathfrak{F}^H \mid G_{-W})$, and then sample new parents F_W from $P(F_W \mid \mathfrak{F}^H, G_{-W})$.

We can sample a component of the partition using the following identity.

$$\begin{aligned} P(\mathfrak{F}^{H^h} \mid G_{-W}, \mathbf{X}) &= \frac{P(\mathfrak{F}^{H^h}, G_{-W} \mid \mathbf{X})}{P(G_{-W} \mid \mathbf{X})} \\ &= \frac{\sum_{F_W \in \mathfrak{F}_W^{H^h}} \prod_{w_j \in W} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j})}{\sum_{H \in \mathcal{H}} \sum_{F_W \in \mathfrak{F}_W^H} \prod_{w_j \in W} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j})} \end{aligned} \tag{4.3}$$

The structure of the partition of $\mathfrak{F}$ means that we are able to interchange the sum and products in Equation 4.3, in a similar way to the interchange used in Friedman and Koller (2003).

Lemma *The following identity holds.*

$$\sum_{F_W \in \mathfrak{F}_W^{H^h}} \prod_{w_j \in W} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j} = \prod_{w_j \in W} \sum_{F_{w_j} \in \mathfrak{F}_j^{H^h}} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}$$

Proof. To show this, we first simplify notation. Define

$$p_{w_j}^{(i)} = p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j}^{(i)}), \qquad i \in \{1, \ldots \mathbb{F}\},$$

where $\mathbb{F}$ is the cardinality of $\mathfrak{F}_W^{H^h}$, and where $F_{w_j}^{(i)}$ is the parent set of node w_j for the i^{th} member of $\mathfrak{F}_W^{H^h}$.

$$\mathfrak{F}_W^{H^h} = \left\{ \langle F_{w_1}^{(1)}, \ldots, F_{w_\rho}^{(1)} \rangle, \ldots, \langle F_{w_1}^{(\mathbb{F})}, \ldots, F_{w_\rho}^{(\mathbb{F})} \rangle \right\}$$

We similarly introduce notation for each member of $\mathfrak{F}_j^{H^h}$. We let $\mathbb{F}_j$ denote the

cardinality of this set.

$$\mathfrak{F}_j^{H^h} = \left\{ F_{w_j}^{(1)}, \ldots, F_{w_j}^{(\mathbb{F}_j)} \right\}$$

As mentioned previously, the key observation is that $\mathfrak{F}^{H^h}$ is the Cartesian product of the parent sets $\mathfrak{F}_j^{H^h}$ for $w_j \in W$.

$$\mathfrak{F}_W^{H^h} = \bigtimes_{j=1,\ldots,\rho} \mathfrak{F}_j^{H^h}$$

Thus,

$$\begin{aligned}
\prod_{w_j \in W} \sum_{F_{w_j}^{(i)} \in \mathfrak{F}_j^{H^h}} p_{w_j}^{(i)} &= \prod_{w_j \in W} \left(p_{w_j}^{(1)} + \cdots + p_{w_j}^{(\mathbb{F}_j)} \right) \\
&= \sum_{i_1 \in \{1,\ldots,\mathbb{F}_1\}, \ldots, i_\rho \in \{1,\ldots,\mathbb{F}_\rho\}} p_{w_1}^{(i_1)} \ldots p_{w_\rho}^{(i_\rho)} \\
&= \sum_{\langle F_{w_1}^{(i)}, \ldots, F_{w_\rho}^{(i)} \rangle \in \mathfrak{F}_W^{H^h}} p_{w_1}^{(i)} \ldots p_{w_\rho}^{(i)} \\
&= \sum_{\langle F_{w_1}^{(i)}, \ldots, F_{w_\rho}^{(i)} \rangle \in \mathfrak{F}_W^{H^h}} \prod_{w_j \in W} p_{w_j}^{(i)} \, .
\end{aligned}$$

■

We can thus sample a partition using the following expression.

$$P(\mathfrak{F}^{H^h} \mid G_{-W}, \mathbf{X}) = \frac{\prod_{w_j \in W} \sum_{F_{w_j} \in \mathfrak{F}_j^{H^h}} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j})}{\sum_{H \in \mathcal{H}} \prod_{w_j \in W} \sum_{F_{w_j} \in \mathfrak{F}_j^{H}} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j})} \tag{4.4}$$

The inner sums in Equation 4.4 can thus be evaluated separately for each node, for each graph $H \in \mathcal{H}$. This makes evaluation of the expression more efficient.

Once we have sampled a component of the partition, we can sample parent sets for each node in W in the following manner. The parents of each node, conditional on

H^h, are independent, and so can be sampled separately using the following identity.

$$P(F_{w_j} \mid \mathfrak{F}^{H^h}, G_{-W}, \mathbf{X}) = \frac{p(X_{w_j} \mid \mathbf{X}_{F_{w_j}})\pi_{w_j}(F_{w_j})}{\sum_{F_{w_j} \in \mathfrak{F}_j^{H^h}} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}})\pi_{w_j}(F_{w_j})}$$

Sampling new parent sets F_W given $\mathfrak{F}^{H^h}$ is thus straightforward because this density is simply the posterior distribution of a constrained Bayesian variable selection with response w_j and with $\mathfrak{F}_j^{H^h}$ as the set of possible predictor sets.

Complete algorithm

The complete algorithm is described in Algorithm 6 (below). The algorithm uses the two-stage sampling procedure (Section 4.6.2), which depends on the descendants D_j and non-descendants K_j. Fast access to these sets is maintained by updating the path count matrix C^G as described in Section 4.6.1.

The run-time of the algorithm depends on the number of nodes p, the maximum in-degree κ of each node, and the number of nodes in W. We fix $|W|$ to be a small constant for all p, and so the run-time is determined by the evaluation of $\mathfrak{F}_j^H$. As described earlier, this is $\mathcal{O}(p^{\kappa+1})$.

Algorithm 6 An efficient Gibbs sampler, with general blocks

Initialise starting point $G^{(0)} = \langle G_1^{(0)}, \ldots, G_p^{(0)} \rangle$

Compute initial path count matrix $C^{G^{(0)}}$

for t in 1 to N **do**

Sample W from V

Generate G^-, update C^{G^-}

for $w_j \in W$ **do**

Retrieve $D_j = \{k : C_{jk}^{G^-} \geq 1\}$

Retrieve $K_j = \{k : C_{jk}^{G^-} = 0\}$

end for

for $H \in \mathcal{H}$ **do**

for $w_j \in W$ **do**

Evaluate $\mathfrak{F}_{w_j}^H$

$Z_j^H = \sum_{F_{w_j} \in \mathfrak{F}_j^H} p(X_{w_j} \mid \mathbf{X}_{F_{w_j}}) \pi_{w_j}(F_{w_j})$

end for

$Z^H = \prod_{w_j \in W} Z_j^H$

end for

Sample H, according to $P(H) = \frac{Z^H}{\sum_{H \in \mathcal{H}} Z^H}$

for $w_j \in W$ **do**

Sample F_{w_j} from $P(F_{w_j} \mid \mathfrak{F}^{H^h})$

end for

Set $G^{(t)} \leftarrow G = \langle F_W, G_{-W}^{(t-1)} \rangle$

Update $C^{G^{(t)}}$

end for

Chapter 5

Evaluation of the Gibbs sampler

In Chapter 4, we introduced a Gibbs sampler for structural inference of Bayesian networks. In this chapter, we present empirical results comparing the Gibbs sampler to several widely used existing methods using simulated data and two real datasets: a social science survey (Centers for Disease Control and Prevention, 2008) and single-cell molecular data from a study of immune responses (Bendall *et al.*, 2011).

Accuracy and stability are two key characteristics by which an algorithm for structural inference of Bayesian networks can be assessed. A good algorithm provides accurate results that are consistent with the true underlying system, and its results are stable in the sense of not being overly sensitive to perturbations in the initial conditions of the algorithm, or the dataset. Badly mixed MCMC applications are not stable because the results depend on initial conditions.

We will investigate the performance of the Gibbs sampler according to both of these aspects, and compare its performance to some existing methods. We start by considering synthetic data generated from the widely-studied ALARM network. We then consider two recent real datasets, the first from a large social science survey and the second from a molecular biology study in which multiple variables were measured

in thousands of individual cells. Both datasets are from areas of current scientific interest and enjoy relatively large sample sizes, facilitating objective comparison of results, as detailed below.

5.1 Setup

In this section, we first outline the alternative methods that we compare to the Gibbs sampler, and then describe the simulation setting in which we compare the methods.

5.1.1 Alternative methods

We compare the performance of our Gibbs sampler with MC^3 (Section 2.6.4) and the REV sampler of Grzegorczyk and Husmeier (2008), which is a variant of MC^3 that uses a more extensive edge reversal move.

We also provide a comparison with two constraint-based methods: the PC-algorithm (Spirtes *et al.*, 2000), which we described in Section 2.7.2, and another constraint-based approach introduced by Xie and Geng (2008), who demonstrate it can outperform the PC-algorithm.

REV sampler

The REV sampler (Grzegorczyk and Husmeier, 2008) augments the simple moves in MC^3 with a more extensive edge reversal move. The algorithm chooses an edge uniformly at random, and then samples new parents for both the node at the head and then, conditionally, the node at the tail. We describe the REV sampler in more detail in the discussion of Chapter 5.

Xie-Geng algorithm

The method introduced by Xie and Geng (2008) is a constraint-based approach that resembles the PC-algorithm. It utilises a clever decomposition that means that the inference problem can be split recursively into smaller problems. Suppose that $A \perp\!\!\!\perp B \mid C$. Then Xie and Geng (2008) show that the local skeleton can be constructed by amalgamating the local skeletons in the following manner. Let $G_{A\cup B} = (V_{A\cup B}, E_{A\cup B})$ and $G_{B\cup C} = (V_{B\cup C}, E_{B\cup C})$ be the local skeletons of $A \cup B$ and $B \cup C$ respectively. Then the local skeleton $G_{A\cup B\cup C}$ of $A \cup B \cup C$ has edge set $E_{A\cup B\cup C}$, where

$$E_{A\cup B\cup C} = E_{A\cup B} \cup E_{B\cup C} \setminus \{(u, v) : u, v \in C, (u, v) \notin E_{A\cup B} \cap E_{B\cup C}\}.$$

The algorithm uses this decomposition to break the problem into smaller problems, and so starts by seeking a decomposition $V = A\cup B\cup C$ such that $A \perp\!\!\!\perp B \mid C$. Local skeletons are then constructed for both $A\cup B$ and $B\cup C$, by seeking a decomposition of $A\cup B$ and $B\cup C$, and then combining the local skeletons using the formula above.

Orientation of the edges of the graph is performed using the same procedure used by the PC-algorithm. The full algorithm is described in Algorithm 7.

5.1.2 Simulation setup

In all of the examples, we use the default settings for all of the methods. This means that comparisons with alternative methods correspond with common practice. In particular, we use the default significance level $\alpha = 0.05$ for the constraint-based methods. Meinshausen and Bühlmann (2010) describe an approach that may lead to a more principled choice of cut-off parameter, but we do not investigate this approach here.

Algorithm 7 Recursive decomposition (Xie and Geng, 2008)

Initialise initial graph G as the complete undirected graph.

Seek a decomposition (A, B, C) such that $A \perp\!\!\!\perp B \mid C$

if a decomposition exists **then**

Save C to SepSet(a, b) for all $a \in A, b \in B$

$L_{A\cup C} \leftarrow$ recursively decompose $A \cup C$

$L_{B\cup C} \leftarrow$ recursively decompose $B \cup C$

$L_{A\cup B\cup C} \leftarrow$ combine $L_{A\cup C}$ and $L_{B\cup C}$

else

Construct $L_{A\cup B\cup C}$ using IC- or PC-algorithm.

end if

The Gibbs sampler we use is a random-scan sampler, with $|W| = 3$ so that the parent sets of three nodes are sampled jointly at each step. To provide a fair comparison, our implementation of MC^3 implements the fast updating of the transitive closure described in Section 4.6.1, and the pre-computation and caching of local marginal likelihoods used in our Gibbs sampler.

We use a flat graph prior $\pi(G) \propto 1$ and constrain all of the MCMC samplers to graphs with in-degree $\kappa \leq 3$.

5.2 Evaluation metrics

In this section, we introduce the metrics by which we evaluate the accuracy and stability of the methods of structural inference. We will focus on different aspects of accuracy and stability for the experiments using synthetic and using real data. To make structural comparisons, we use completed partially directed acyclic graphs (e.g. Chickering, 2002) to make comparisons, so that these are on a common scale across the various methods.

5.2.1 Synthetic data

When using synthetic data, the true graph is known and so the accuracy of structural learning algorithms can be assessed using ROC (receiver-operating characteristic) curves. ROC curves compare the graphs given by each method to the true graph.

While, from a Bayesian perspective, the main aim is to approximate the posterior distribution accurately, it is informative to compare the regions of high posterior probability to the data-generating graph for two reasons. First, the true graph should be close to the regions of high posterior probability in large sample size settings, such as many of the scenarios considered here. This means that, in the settings considered here, the data-generating graph may provide a reasonable proxy for the true posterior distribution. Second, because the exact distribution is intractable it is simply not possible to compare the results to the true distribution except in trivial examples with $p < 6$, say. Such trivial examples are not especially interesting because there is little reason to assume that MCMC samplers that perform well when p is small will also perform well when p is large.

The Bayesian (MCMC) and frequentist (constraint-based) methods return different forms of result and so their representation on the ROC plot differs. The Bayesian methods return estimates for the posterior distribution on Bayesian networks $P(G \mid \mathbf{X})$, from which the posterior probability of any edge $P(e)$, $e \in E$ can be computed. With a threshold τ, $0 \leq \tau \leq 1$, we define $E_\tau \subset V \times V$ as the set of edges with posterior probability $P(e) \geq \tau$. Since we know the true graph, we can compare the true graph to the edges E_τ and in particular count true and false positive edges. These counts can be placed on a standard scale by defining the true and false positive rates as the proportion of true and false edges present in E_τ, compared to the true graph. An ROC curve for an MCMC algorithm is then given by plotting the true positive rate against the false positive rate, for a range of values of thresholds $0 \leq \tau \leq 1$. The frequentist constraint-based methods return a point estimate, so

these appear as a single point on the ROC plane.

Naturally, we seek to maximise the number of true positives for a given number of false positives, and so the algorithms with the greatest area under the curve are preferred. For this reason, we will additionally compare the areas under the ROC curves directly. We will particularly focus on the region of the ROC curve corresponding to a small false positive rate because in many applications, for example in molecular biology, high-scoring edges may be used to design validation experiments. In such settings, it is important that the methods return almost no false positives so that validation of artifactual edges is not attempted.

In addition to considering ROC curves, we can also examine the accuracy of the methods by computing the distance between the graph returned by each method and the true graph. To assess the distance between the graphs we use the structural Hamming distance (SHD; Hamming, 1950), defined as the number of edge additions and removals needed to change one graph into the other. We will particularly focus on the number of true edges that are not detected. For the Bayesian methods, we define an edge as 'not detected' if it has a posterior edge probability of less than 0.5. For the frequentist constraint-based methods we assess this directly by counting true edges that are not returned.

The threshold 0.5 corresponds to selecting edges that are *a posteriori* more likely to be present than not present. This threshold gives the 'median probability model'. For selection among normal linear models, in some settings it can be shown this is the optimal model for prediction (Barbieri and Berger, 2004). This gives the choice an appeal even when optimality is not known.

The aspect of stability that we focus on with synthetic data is Monte Carlo stability. A good MCMC sampler gives consistent results regardless of the initial conditions of the sampler. We will assess this by comparing the posterior edge probabilities of 10 independent runs of the samplers, initialised at disparate initial graphs. The

consistency of the runs can be examined using a convergence diagnostic plot, in which the posterior edge probabilities of each edge in two independent runs are plotted against each other. When the edge probabilities of the two runs agree, all of the points in the scatter plot will lie on the $y = x$ line.

We will consider two different convergence diagnostic plots. In this first of these, we compare all 10 independent runs of each sampler. We do this by plotting two panels, each of which consists of a 10-by-10 matrix of plots, in which each cell compares the edge probabilities between the corresponding pair of runs. The lower triangle of both panels shows the Gibbs sampler runs, which are to be contrasted with the MC^3 and REV runs shown in the upper half of top and bottom panels respectively. Each point plotted is an individual edge probability. The colour represents the distance of the point from the line $y = x$. The orange points are the furthest from the $y = x$ line.

We additionally consider a convergence diagnostic plot in which the points are binned into hexagonal areas, to avoid over-plotting. This plot makes clear the number of edges that have, for example, posterior probability of 1 in one run and 0 in another. We also consider such edges by plotting the number of 'major discrepancies' between independent runs for each MCMC sampler, at a range of sample sizes. A major discrepancy is defined as an edge with posterior probability greater than 0.9 in one run, and less than 0.1 in another run.

5.2.2 Real data

For the real data, we focus on assessing the stability of the methods. We first consider Monte Carlo stability, using the same diagnostics as for the synthetic data.

We then consider the sensitivity of the methods to small perturbations in the data. One method for assessing this property is to consider bootstrap samples of the data.

Since bootstrap replicates of a dataset when n is large are similar, the estimate of the Bayesian network associated with each replicate should be similar. We measure the similarity of two Bayesian networks with the structural Hamming distance (SHD), which measures the number of edges that are present in one network and absent in the other. The result of the MCMC methods is an edge probability matrix rather than a point estimate of the Bayesian network that is given by the constraint-based methods, and so we will consider three methods for choosing the graph to compare with the constraint-based methods: thresholding the edge probabilities to match the number of edges given by the PC-algorithm, by the Xie-Geng method, and thresholding at 0.5 posterior edge probability (giving the 'median probability model').

5.3 Synthetic data

We first analysed the performance of the methods using synthetic data, generated from the ALARM network (Beinlich *et al.*, 1989). This network is widely used to examine the performance of methods of structural learning (e.g. Friedman and Koller, 2003; Grzegorczyk and Husmeier, 2008). In this section, we describe the details of the simulations, and then assess the performance of the methods. We consider the accuracy, Monte Carlo stability and finally the trace plots of the MCMC runs.

5.3.1 Simulation setup

There are 37 random variables and 46 edges in the ALARM network. Each variable has a multinomial distribution, and so we use the natural multinomial-Dirichlet formulation (Heckerman *et al.*, 1995).

We drew 10 independent samples from the ALARM network, with sample sizes

$n = 100, 500, 1000, 2500, 5000$ respectively. To ensure a fair comparison, we fixed compute time, running each of the MCMC samplers for 30 minutes (on a single core of a cluster computer), and performed 10 independent runs starting from different initial graphs. In this time, MC^3 drew 800,000 samples; REV drew 7,500 samples; and our Gibbs sampler drew 20,000 samples. In all three cases, we discard the first quarter of the samples as burn-in.

5.3.2 Accuracy

We first assess the accuracy of the methods using ROC curves. Figure 5.1A is a plot of ROC curves at each sample size for the Gibbs sampler, MC^3, REV sampler, Xie-Geng's constraint-based method and the PC-algorithm. We see that some of the methods of inference have different properties at different sample sizes. The MC^3 sampler is particularly sensitive to the sample size. For smaller sample sizes ($n = 100, 250$) the MC^3 results are close to the most consistent of the methods with the true ALARM network, but for large sample sizes (e.g. $n = 2500, 5000$), the performance of MC^3 is extremely poor. The area under ROC curves decreases as the sample size increases (Figure 5.2), from 0.91 ($n = 100$) to 0.42 ($n = 5000$). This decrease is clearly unsatisfactory because increasing the sample size should improve the quality of the estimates. Indeed an area under the ROC curve less than 0.5 corresponds to a success rate that is worse than random.

The relationship between sample size and performance of the REV sampler is less clear. As shown in Figure 5.2, the area under the ROC curve for the REV sampler varies considerably between sample sizes, but no clear pattern emerges in the range of sample sizes considered here. The area under the ROC curve for the Gibbs sampler shows a slight pattern of increase with sample size, but is essentially stable at around 0.96.

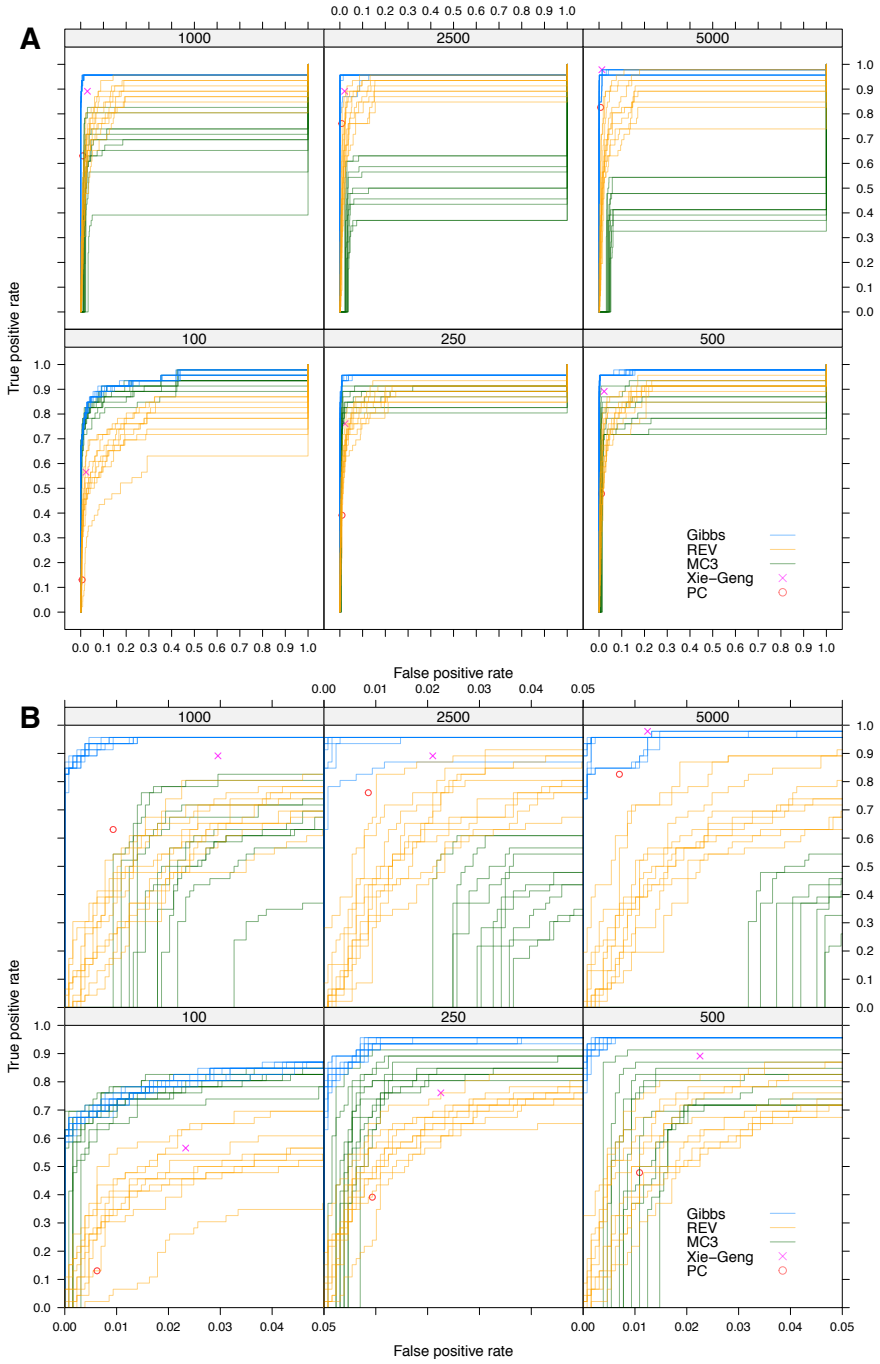


Figure 5.1: ROC curves given by estimated posterior distributions from 10 replications of our Gibbs sampler, MC^3, and the REV sampler for the synthetic data from the ALARM network. Point estimates from Xie-Geng's constraint-based method and the PC-algorithm are also shown. These plot the true positive rate (y-axis) against the false positive rate (x-axis) for a range of values of τ. In (A) the entire ROC curves are shown; in (B) a reduced range of false positives is shown.

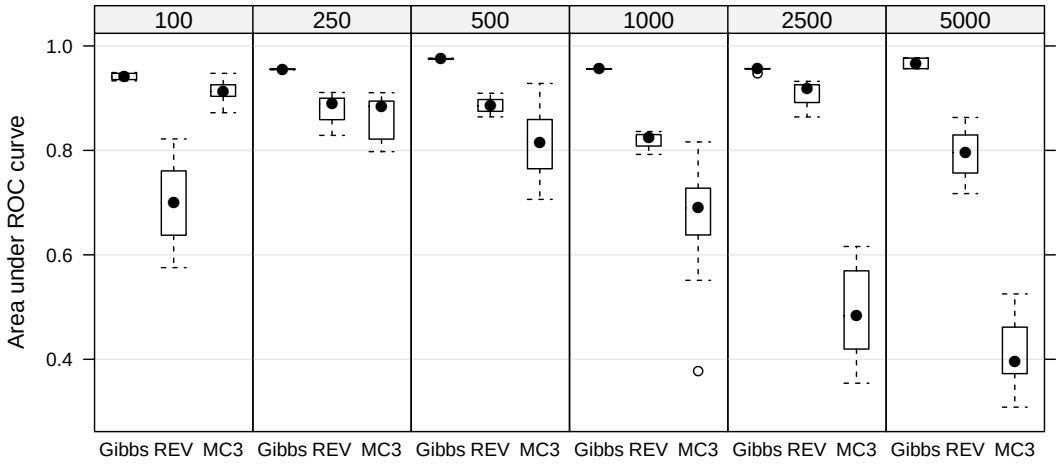


Figure 5.2: The distribution of the areas under the ROC curves, for $n = 100, \ldots, 5000$ for the synthetic data from the ALARM network.

The constraint-based methods (PC-algorithm and the Xie-Geng method) are also sensitive to the sample size. These methods give a point estimate, which is indicated on the ROC plot by a circle (PC-algorithm) and a cross (Xie-Geng). We see that these methods perform poorly when $n = 100$, but, as anticipated by the consistency of the PC-algorithm (Kalisch and Bühlmann, 2007), work well for large sample sizes. For $n = 5000$, the Xie-Geng method performs particularly well. It predicts 45 true positives, and 16 false positives.

Figure 5.1B shows that there is wide variation in the performance of the methods at low false positive rates. For example, for $n = 100$ for a false positive rate of 0 (corresponding to no incorrect edges), the Gibbs predicts 28.1 ± 0.74 (mean $\pm$ standard deviation) true edges; REV sampler 1.2 ± 1.8; and MC^3 14.4 ± 15.25. For $n = 5000$, for a false positive rate of 0, the Gibbs sampler finds 38.3 ± 4.00 true edges; the REV sampler 1.0 ± 2.82; and MC^3 never predicts any true edges. Since the true graph has 46 edges, these differences correspond to very important differences in the practical usefulness of the results given by the different methods.

We can also compare the accuracy of the methods by considering the SHD between the true graph and the graphs given by each of the estimators, at each sample size.

Table 5.1: Structural Hamming distances (SHDs) between the graph given by each method (MAP graph for MCMC method) and the true graph. The standard deviation of the SHDs is shown for the Bayesian Monte Carlo methods.

Method	$n = 100$	250	500	1000	2500	5000
Gibbs	20.1 ± 8.0	20.1 ± 8.0	30.3 ± 3.4	30.6 ± 11.2	19.91 ± 6.0	20.1 ± 8.0
REV	48.2 ± 10.7	48.2 ± 10.7	56.2 ± 11.9	54.3 ± 11.3	52.23 ± 16.9	48.2 ± 10.7
MC^3	90.7 ± 12.3	90.7 ± 12.3	55.8 ± 10.7	62.6 ± 13.4	76.92 ± 11.2	90.7 ± 12.3
Xie-Geng	50.0	40.0	34.0	43.0	32.0	17.0
PC	47.0	39.0	38.0	28.0	22.0	14.0

Table 5.1 shows the means and standard deviations of the SHDs. We see that the MAP estimator given by the Gibbs sampler is consistently the closest graph to the true graph, except in two cases in which the PC-algorithm is closer. The mean SHD across all sample sizes for the Gibbs sampler is 23.5, whereas for the REV sampler it is 42.8 and for MC^3 it is 51.0. We also see again that the constraint-based methods perform well with large sample sizes.

The number of edges not detected (as defined in Section 5.2.1) by each method also varies by sample size. For $n = 100$, the Gibbs sampler does not detect 1.5 ± 0.53 edges, the REV sampler 9.8 ± 3.46 edges and MC^3 2.8 ± 1.23 edges. For $n = 5000$, the Gibbs sampler does not detect 1.5 ± 0.53 edges, the REV sampler 4.9 ± 3.38 edges and MC^3 25.9 ± 3.31 edges. Again, by this metric it is clear that the results from the REV sampler and MC^3 is significantly less useful in practice. At large sample sizes, the constraint-based methods detect almost all of the edges. For the mid-sized samples, the Xie-Geng method detects most of the edges, but for the same number of true positives the Gibbs sampler gives far fewer false positives. For example, for $n = 1000$, the Xie-Geng method has 29 true positives and 12 false positives. To reach 29 true positives, no false positives are given by any of the 10 runs of the Gibbs sampler.

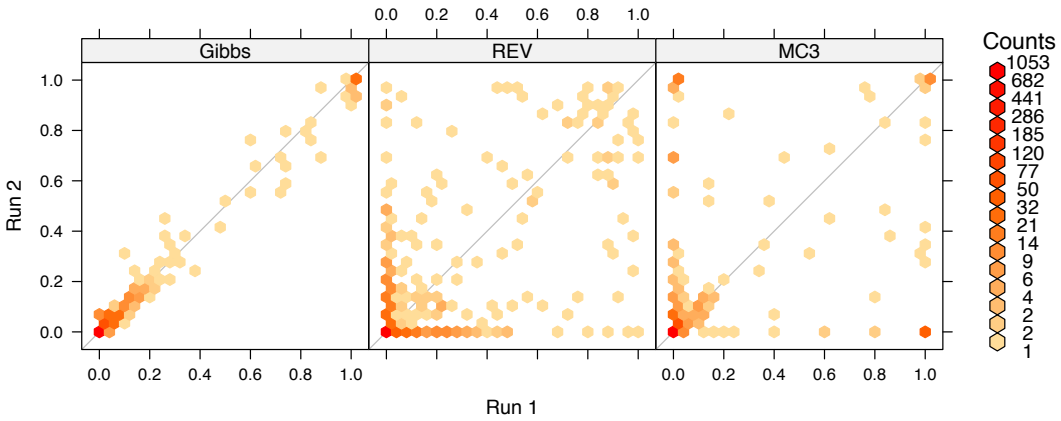


Figure 5.3: Convergence diagnostics for the MCMC samplers, for the ALARM data with $n = 1000$. The posterior edge probabilities given by two independent runs are plotted against each other. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. To avoid over-plotting, the points are binned into hexagonal areas (Carr *et al.*, 1987). When using the REV sampler or MC^3, for many edges there are extreme discrepancies between the two runs, in the sense that there are many edges have high probability in one run and low in the other. This pair of runs was typical of all the pairs of runs and sample sizes.

5.3.3 Monte Carlo stability

A good MCMC sampler gives consistent results across independent runs. The inconsistency of the REV and MC^3 samplers across independent runs is shown in Figure 5.3, which compares the posterior edge probabilities of two independent runs of the sampler for $n = 1000$. We see that there are many edges that have zero posterior probability in one run, but far greater than 0 posterior probability in the other. In contrast, there is close agreement between the two independent Gibbs runs. The run shown is typical of all runs, as shown in Figure 5.4.

The inconsistency of MC^3 and the REV sampler is highlighted by Figure 5.5. The figure shows that in almost all cases, there are no major discrepancies (as defined in Section 5.2.1) with the Gibbs sampler. In contrast, on average 5 major discrepancies are given by the REV sampler, at all sample sizes. The number of major

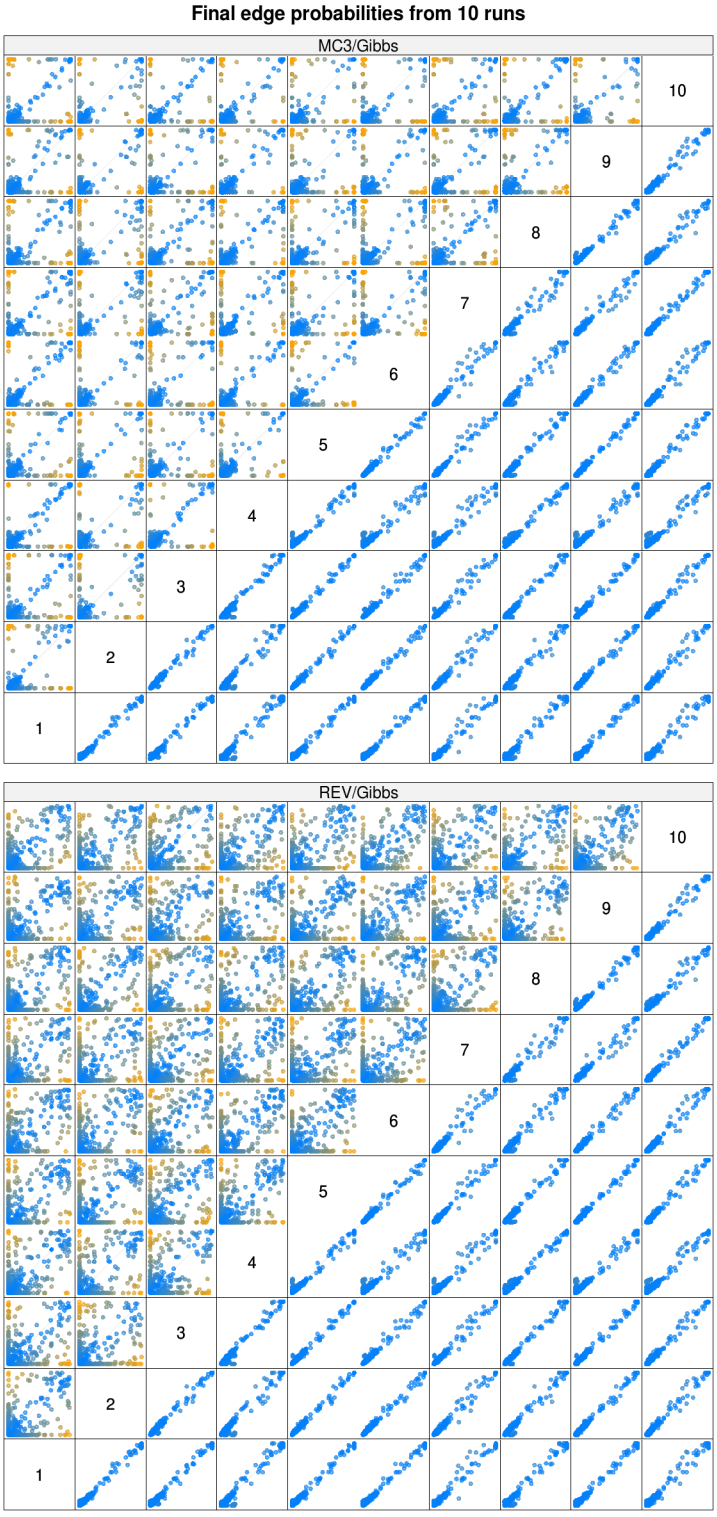


Figure 5.4: Convergence diagnostics for all 10 runs of each MCMC sampler for the ALARM data, with $n = 1000$. In each cell, the posterior edge probabilities given by two independent runs are plotted against each other. Each point represents a single edge. The lower half of both panels compares runs of the Gibbs sampler; the upper half compares runs of the MC^3 and the REV sampler respectively. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. The blue to orange colour scale represents the distance from this line, with orange points the furthest away.

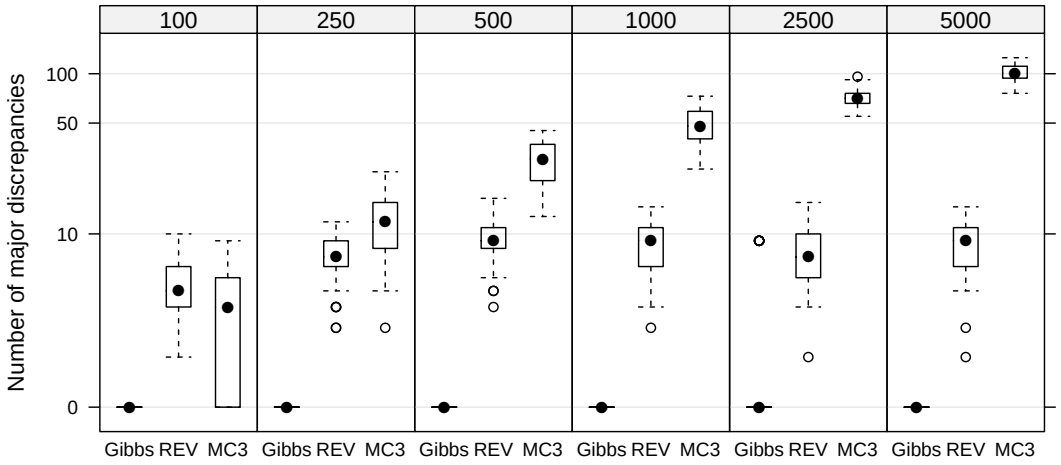


Figure 5.5: Major discrepancies between pairs of the 10 independent runs, for each MCMC sampler. For each pair of independent runs, the number of major discrepancies is the number of edges that have estimated posterior edge probability above 0.9 in one run and estimated posterior edge probability below 0.1 in the other run. The boxplot shows the range of discrepancies between runs. Each panel corresponds to one of the sample sizes $n = 100, 250, 500, 1000, 2500, 5000$.

discrepancies for MC^3 increases rapidly with sample size, from an average of 8 major discrepancies when $n = 100$ to an average of 90 when $n = 5000$. Results as variable as these are almost unusable because, with this instability, results that imply an edge has even very high probability are likely to be simply artefacts of the initial conditions of the sampler.

The area under the ROC curve (Figure 5.2) gives another indicator of the stability of the methods. It is clear that the area under the ROC curve for both the REV sampler and MC^3 varies considerably between runs. In contrast, the Gibbs sampler is very consistent between runs.

5.3.4 Marginal likelihood trace plot

The Gibbs sampler reaches a plateau of high posterior probability far more rapidly than the MC^3 or REV samplers. With $n = 1000$ it takes around 5,000 samples for the Gibbs sampler to reach a plateau on which it settles. The Gibbs sampler finds graphs

with higher log score than the REV sampler and MC^3, and does so consistently across independent runs. The maximum log score found by the Gibbs sampler across runs is -10499.34 ± 1.08, whereas for the REV sampler it is -10581.50 ± 82.12 and for MC^3 it is -11311 ± 341.69. The full trace plot of the marginal likelihoods, with $n = 1000$, is shown for all 10 independent runs in Figure B.1 on page 157.

5.4 Behavioral Risk Factor Surveillance System Survey data

The second data set we consider is the publicly available Behavioral Risk Factor Surveillance System Survey (BRFSS) (Centers for Disease Control and Prevention, 2008). This is a household-level random-digit telephone survey, collected by the U.S. Government's National Center for Chronic Disease Prevention and Health, that has been conducted throughout the United States since 1984. We consider the responses from New York in the 2008 survey. In this section, we describe the details of the simulations, and then assess the performance of the methods. We consider Monte Carlo stability, bootstrap stability, and finally the trace plots of the MCMC runs. The network given by thresholding the posterior edge probabilities from the Gibbs sampler at 0.5 (for the reasons described in Section 5.2) is shown in Figure B.6 on page 162.

5.4.1 Data and setup

We analysed the responses to 24 questions, which spanned most of the topics covered in BRFSS. All respondents who refused or were unsure of their response, or whose response is missing, to any of the 24 questions were removed from the analysis. The resulting sample size is 4,197.

We ran each MCMC sampler for 30 minutes. In this time, the Gibbs sampler drew

54,000 samples, the REV sampler 50,000 samples, and MC^3 1.8 million samples. In each case, the first quarter of the samples were discarded as burn-in. In addition, we ran the PC-algorithm and the Xie-Geng method on the BRFSS data.

5.4.2 Monte Carlo stability

The convergence of the three MCMC samplers is considered in Figures 5.6 and 5.7 by examining the agreement in edge probabilities between the runs. In Figure 5.6 the edge probabilities of two runs are compared. We see that there is considerable agreement between the edge probabilities given by the two Gibbs runs, but there is considerable disparity in the results from the REV sampler and MC^3. While both MC^3 and the REV sampler have 38 edges for which there is a disparity of 0.1 in posterior edge probability between the two runs, there are only 4 such edges for the Gibbs sampler. There are no edges with a disparity of 0.3 in posterior edges probability for Gibbs, but there are 33 for REV and 32 for MC^3. This pair of runs is typical of all pairs of runs, as shown in Figure 5.7. It is clear that in all pairs of runs, the REV and MC^3 runs have many edges in which there is a strong disagreement about edge probabilities. In contrast, there is good agreement between all of the runs of the Gibbs sampler. Indeed, there are no major discrepancies (Section 5.2.1) between any of the pairs of runs of the Gibbs sampler, whereas there are on average 11 majors discrepancies for REV sampler and 17 for the MC^3 sampler (Figure B.2 on page 158).

5.4.3 Marginal likelihood trace plot

The maximum log marginal likelihoods (log scores) found by each of the three MCMC samplers varies considerably. The maximum log score (mean $\pm$ standard deviation) encountered in each run reached by the samplers is -82121.81 ± 0 (Gibbs),

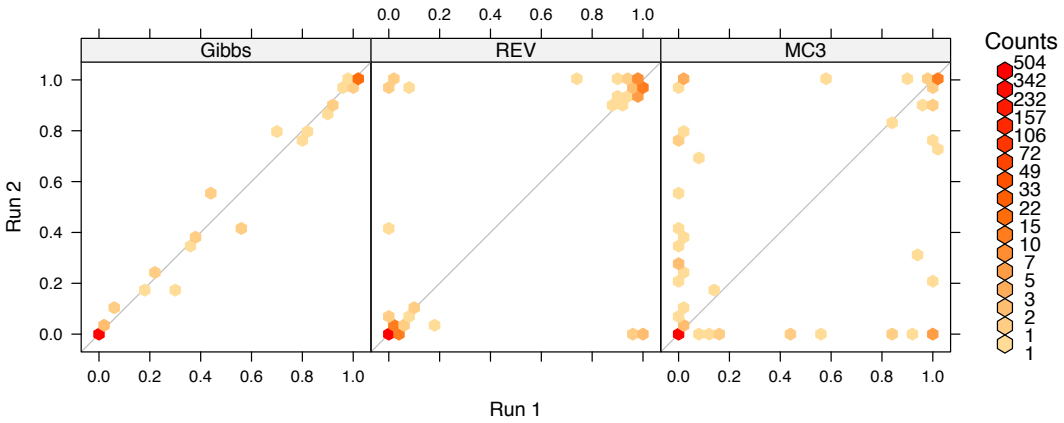


Figure 5.6: Convergence diagnostics for the MCMC samplers for the BRFSS data. The posterior edge probabilities given by two independent runs are plotted against each other. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. To avoid over-plotting, the points are binned into hexagonal areas (Carr *et al.*, 1987). We observe that the two Gibbs runs gives comparable posterior edge probabilities, but the MC^3 and REV sampler runs do not. This pair of runs was typical of all pairs.

-82172.7 ± 27.6 (REV) and -82198.43 ± 77.5 (MC^3). The highest scoring graph found by any of the runs of the REV sampler has log score -82139, which is 17 below the highest scoring graph (which was obtained consistently in all runs of the Gibbs sampler). This difference corresponds to a large difference in posterior probability: if the posterior distribution contained only (with a uniform graph prior) the modal graph from the Gibbs runs with log score -82121.81 and the modal graph from the REV runs with log score -82139, the Gibbs mode would have probability of unity (to 8 decimal places).

The number of samples until each sampler reaches a plateau also varies considerably. The Gibbs sampler reaches in all runs a plateau after around 500 samples, although in one run it is not reached until 10,000 samples have been drawn. The REV sampler takes longer to settle on a plateau, but even after doing so it does not reach a region with log score comparable to the plateau reached by the Gibbs sampler. Despite drawing an order of magnitude more samples, the MC^3 sampler becomes stuck in a region with yet lower log score.

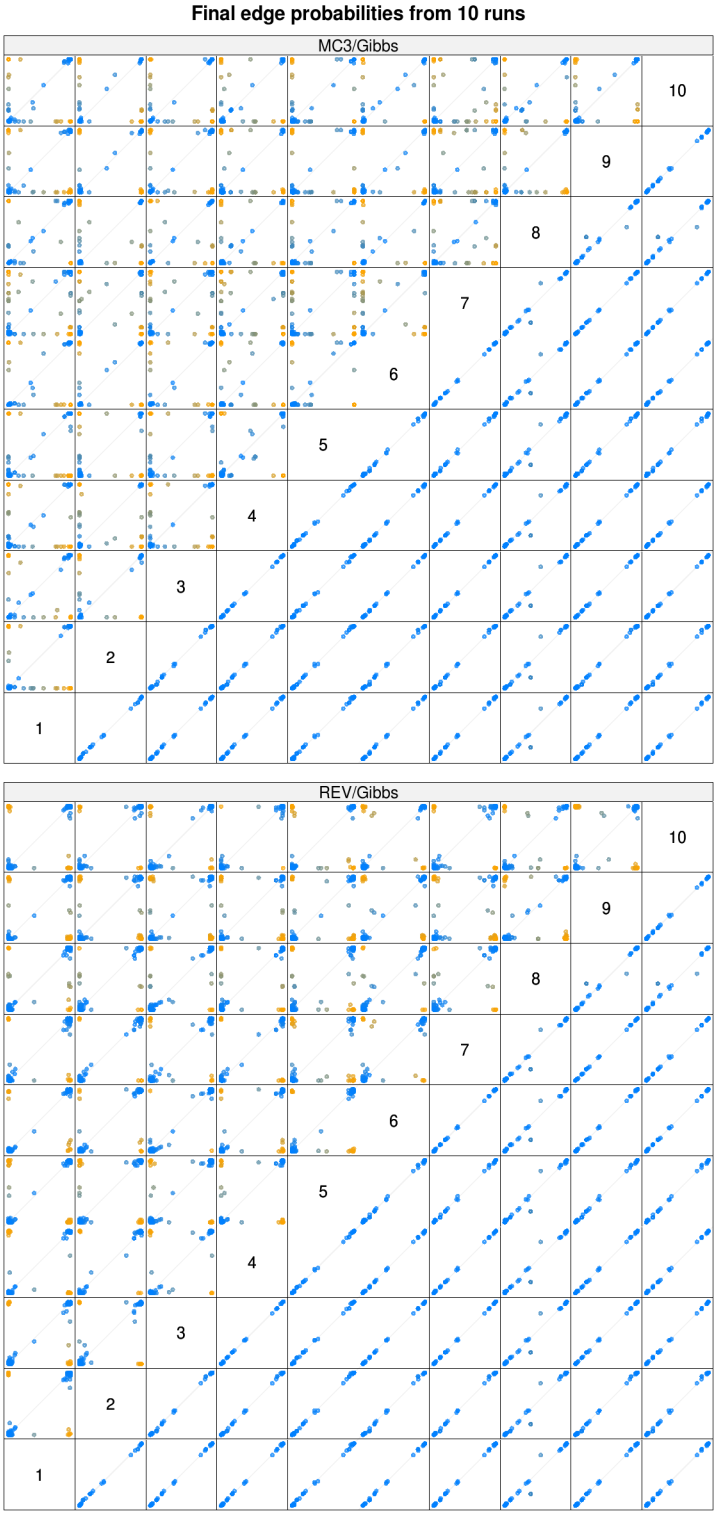


Figure 5.7: Convergence diagnostics for all 10 runs of each MCMC sampler for the BRFSS data. In each cell, the posterior edge probabilities given by two independent runs are plotted against each other. Each point represents a single edge. The lower half of both panels compares runs of the Gibbs sampler; the upper half compares runs of the MC^3 and the REV sampler respectively. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. The blue to orange colour scale represents the distance from this line, with orange points the furthest away.

A complete trace of the log score of each graph drawn by each sampler in each of 10 independent runs, initialised at disparate starting graphs, is shown in Figure B.3 on page 159.

5.4.4 Bootstrap stability

We drew 10 bootstrap replicates of the dataset, and for each estimated the Bayesian network using each inference method. We first threshold the edge probabilities of the MCMC methods such that the resulting Bayesian network has the same number of edges as the Bayesian network given by the PC-algorithm. For each of the MCMC methods and for the PC-algorithm the SHD between pairs of bootstrap replicates is shown in Figure 5.8. The Gibbs sampler has a mean SHD of 22 between pairs of replicates. This is the lowest mean among any of the methods. The next lowest mean of 30 is given by the PC algorithm. For a network with 24 nodes, this is a considerable increase in the number of edges that differ between bootstrap replications from a dataset with over four thousand samples. The results from thresholding the edge probabilities at 0.5 (for the reasons described in Section 5.2) and thresholding to match the number of edges in the graph given by the Xie-Geng procedure are shown in Figure B.5 (on page 161 in Appendix B).

5.5 Flow cytometry data

Single-cell data in molecular biology are often obtained using a technology called flow cytometry. Using this technique the number of variables that can be interrogated is severely limited for technical reasons (spectral overlap in relevant fluorophores). Recently, technology has been developed that applies atomic mass spectrometry to single-cell analysis, thereby allowing interrogation of larger numbers of variables than previously possible (Bendall *et al.*, 2011). We used single-cell data from Bendall

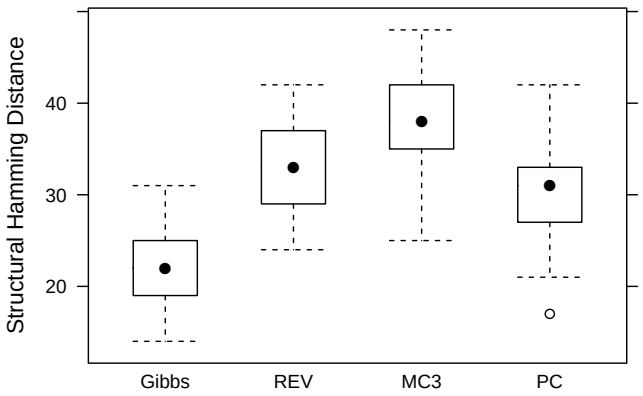


Figure 5.8: Stability of estimators of the BRFSS data across bootstrapping, as measured by SHDs, with the graph density made to match the graph given by the PC-algorithm. Using data from BRFSS, 10 bootstrap samples were drawn. Each estimator was run on each bootstrap sample. The structural Hamming distance (SHD) between the graphs from each bootstrap sample is shown for each estimator. Smaller SHD means that the graphs are structurally more similar and therefore the estimator is more stable.

et al. (2011) to infer Bayesian network structures. In this section, we describe the details of the data and simulations, and then assess the performance of the methods. We again consider Monte Carlo stability, bootstrap stability and finally the trace plots of the MCMC runs. We also give the resulting Bayesian network.

5.5.1 Data and setup

The single-cell nature of the data provided a large multi-variate sample over $n = 21,691$ cells that can reasonably be regarded as a random sample. We consider $p = 34$ measured cellular variables. We treat the data as independent replications, and model using a normal model with g-prior, as described in Section 2.5.3, with $g = n^{-1}$. A full specification of the data used is detailed in Appendix A.

We ran the Gibbs sampler until all 10 runs had essentially converged, which took 8.5 hours. In this time, the Gibbs sampler drew 480,000 samples. We ran MC^3 and

the REV sampler for the same amount of time, in which time the samplers drew 16 million and 4.5 million samples respectively. To reduce the computational demand of handling so many samples, we thin the samples drawn from both MC^3 and REV sampler so that only every 100^{th} sample is retained.

5.5.2 Monte Carlo stability

We examine the convergence properties of the samplers by examining the agreement in edge probabilities between the runs. For the flow cytometry data, there is again considerable agreement between the edge probabilities given by runs of the Gibbs sampler, but there is considerable disparity in the results from the independent runs of the REV sampler and MC^3, as shown for one pair of runs in Figure 5.9. Figure B.8 (on page 163) shows this comparison for each pair of runs of each MCMC sampler. This figure shows that there is little agreement between pairs of independent runs of both the MC^3 and REV samplers. In contrast all runs of the Gibbs samplers are consistent with each other, except for run 3. In fact, the figure shows that run 3 and the other runs of the Gibbs sampler are more in agreement that any pair of runs of either the MC^3 and REV samplers. Figure 5.9 shows the number of edges for which there is a considerable disagreement in two of the runs. On average, across the 45 pairs of runs, there are 12.9 edges for which there is a disparity of at least 0.1 in posterior edge probability between runs of the Gibbs sampler. For the REV sampler, the average is 50.4 edges, and for MC^3, the average is 151.4 edges. Comparing disparities of least 0.2 in posterior edge probability, an average of 6.1 edges differ between Gibbs runs, whereas the number of differences are 42.6 edges for the REV sampler and 138.3 for MC^3. The Gibbs sampler has no major discrepancies (as defined Section 5.2.1) between pairs of runs, while there are on average 25 majors discrepancies for REV sampler and 111 for the MC^3 sampler (Figure B.7 on page 162).

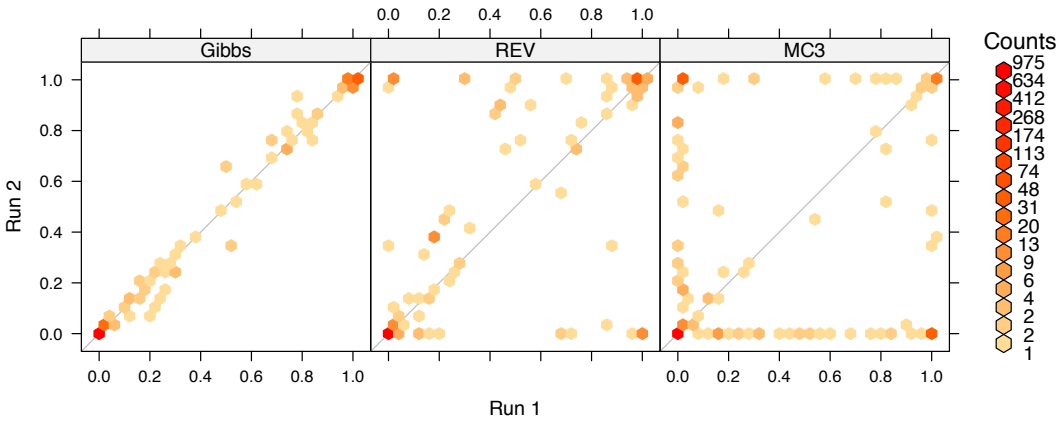


Figure 5.9: Convergence diagnostics for the MCMC samplers, for the flow cytometry data. The posterior edge probabilities given by two independent runs are plotted against each other. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. To avoid over-plotting, the points are binned into hexagonal areas (Carr *et al.*, 1987). When using the REV sampler or MC^3, for many edges there are extreme discrepancies between the two runs, in the sense that there are many edges have high probability in one run and low in the other. This pair of runs was typical of all pairs of runs and sample sizes.

5.5.3 Marginal likelihood trace plot

The Gibbs sampler for the flow cytometry data, as for the synthetic and BRFSS data, consistently finds a region of higher log marginal likelihood than the MC^3 and REV samplers. The maximum log marginal likelihood reached by the Gibbs sampler is $-3,641,282$, and this was reached in all 10 independent runs. In contrast, mean (standard deviation) of the maximum reached across the 10 runs of the MC^3 sampler is $-3,658,408 \pm 8091$, and for the REV sampler is $-3,642,879 \pm 4725$. The maximum log marginal likelihood reached by any of the 10 runs of the REV sampler is $-3,641,294$, which is 11 lower than the maximum log marginal likelihood that was reached in all 10 runs of the Gibbs sampler. The Gibbs sampler reaches a region of high log marginal likelihood after around $5,000$ samples in contrast to the REV sampler which, when it does reach such a region, only does so near the end of its 16 million samples. The full trace of the log marginal likelihood is shown in Figure B.4 on page 160.

5.5.4 Bootstrap stability

We finally studied stability of the methods under bootstrap replications for the flow cytometry data. We drew 10 bootstrap replicates of the dataset, and estimated the Bayesian network for each using each of the inference methods, thresholding the edge probabilities of the MCMC methods such that the resulting Bayesian network has the same number of edges as the Bayesian network given by the PC-algorithm. Figure 5.10 shows the SHD between pairs of bootstrap replicates for each of the MCMC methods and for the PC-algorithm. The Gibbs sampler has the lowest mean SHD (120) among any of the methods between pairs of replicates. The next lowest mean of 142 is given by the REV sampler. For a network with 34 nodes, this is a considerable increase in the number of edges that differ between bootstrap replications from a dataset with over four thousand samples. Note that the SHD are particularly high here because the PC-algorithm prefers a network with a high density. The results from thresholding the edge probabilities at 0.5 (for the reasons described in Section 5.2) and thresholding to match the number of edges in the graph given by the Xie-Geng procedure are shown in Figure B.10 (on page 165 in Appendix B).

5.6 Discussion

We have introduced a Gibbs sampler for structural inference of Bayesian networks. The sampler uses the idea of blocking to improve its rate of convergence, and we demonstrated empirically its utility on data from a large social science survey, and from molecular biology, as well as for simulated data, across a wide range of sample sizes. At low sample sizes, the MC^3 sampler performs reasonably. At large sample sizes the constraint-based methods perform well, and the computation required is quick. However, the existing methods are particularly unstable across Monte Carlo

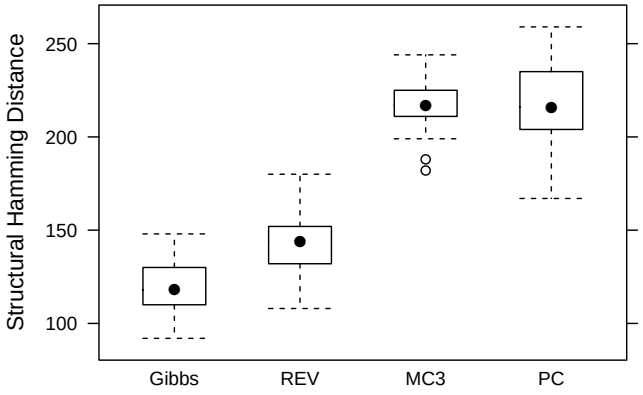


Figure 5.10: Stability of estimators for the flow cytometry data across bootstrapping, as measured by SHDs, with the graph density made to match the graph given by the PC-algorithm 10 bootstrap samples were drawn. Each estimator was run on each bootstrap sample. The structural Hamming distance (SHD) between the graphs from each bootstrap sample is shown, for each estimator. Smaller SHD means that the graphs are structurally more similar, and so the estimator is more stable.

replications or across bootstrap resamples. The instability of the PC-algorithm has been discussed before by Spirtes *et al.* (2000). In contrast, the Gibbs sampler consistently performs well and gives more stable results across the whole range of examples considered here.

In the Gibbs sampler introduced here, we used the exact posterior distribution from Bayesian variable selection to compute the required conditional probabilities. When evaluating the exact posterior, it has been noted previously that for some local models, it is advantageous to evaluate the marginal likelihoods in Gray code ordering (George and McCulloch, 1997). Nonetheless, for the exact posterior to be computationally tractable requires a maximum in-degree constraint to be enforced, as used by many other authors (e.g. Friedman and Koller, 2003; Koivisto and Sood, 2004). This requirement is not a significant drawback because in general models with a large in-degree are rarely useful in applications, and this is particularly true in a Bayesian setting in which the result accounts for model uncertainty. In this

setting, the effect of the constraint on the result is reduced because the Bayesian averaging over models reduces the rigidity of the constraint. For example, suppose an in-degree restriction of 3 is enforced, but the true in-degree of a particular node is 4. In this case, even though no model including all 4 parents can be considered, the posterior edge probability of all of the 4 nodes is likely to be high, unless there are particularly difficult non-linearities. Thus any form of model averaging will take heed of the influence of all 4 parents.

The Gibbs sampler builds on the simple heuristic that the parents of a node in a Bayesian network are similar to the independent variables chosen in Bayesian variable selection, with the node as the dependent variable, but adjusts this heuristic exactly to ensure acyclicity. By exactly adjusting for acyclicity, there is no need for heuristic choice of candidate parents in the manner of the Sparse Candidate algorithm (Friedman *et al.*, 1999).

However, the Gibbs sampler does have some similarities with the edge reversal proposal that the REV sampler (Grzegorczyk and Husmeier, 2008), a Metropolis-Hastings sampler, mixes with MC^3 proposals. The edge reversal proposal in the REV sampler reverses the direction of a particular edge $i \rightarrow j$, which is drawn uniformly at random from the set of edges in G. Then a new graph $G^{\odot}$ is created from G by removing all edges $\{(a, b) : a \in V, b \in \{i, j\}\}$, so that in $G^{\odot}$ neither i nor j have any parents. A proposal graph G' is constructed from $G^{\odot}$ by sampling new parents for both nodes i and j in the following manner. First, a parent set for node i—that is required to include node j—is sampled from the appropriate conditional distribution. Then, conditional on the choice of parents from i, a new parent set is sampled for node j, from the appropriate conditional distribution. The proposal G' is accepted according to the appropriate acceptance probability, as detailed in Grzegorczyk and Husmeier (2008).

The use of conditional distributions makes the REV sampler similar in one respect

to the Gibbs sampler proposed here. However, as we showed in this chapter, across a range of empirical examples the Gibbs sampler substantially outperforms the REV sampler. There are various possible explanations for this. The REV sampler does not use the natural conditional distribution and so requires an accept-reject step. An accept-reject step may be wasteful in settings in which evaluating the proposal distribution is relatively computationally expensive. Additionally, the proposal made by the REV sampler requires an existing edge whose direction is reversible in the posterior distribution. This requirement makes the REV sampler less flexible than the Gibbs sampler, which uses the full joint conditional distribution. The REV sampler is also unable to make moves considering more than two nodes simultaneously and in situations in which there are three highly correlated random variables, proposals that consider the three parent sets simultaneously are crucial in realising fast convergence of the MCMC sampler. Finally at least some MC^3 proposals must be used when using the REV sampler because the REV proposal is not irreducible by itself (Grzegorczyk and Husmeier, 2008), and these simple MC^3 forms are not tailored to the local shape of the posterior distribution. Grzegorczyk and Husmeier (2008) make REV proposals with probability $^1/_{15}$, and so the majority of steps are based on simple MC^3 proposals. Using such small proposals is likely to make the sampler less efficient.

An appealing aspect of this approach is that it harnesses the connection between Bayesian variable selection and structural inference of Bayesian networks. This connection has been widely studied and exploited for undirected graphs (e.g. Meinshausen and Bühlmann, 2006), but for directed graphs the connection is complicated by the acyclicity requirement. The Gibbs sampler accounts for this, enabling it to exploit the relationship with variable selection. The Gibbs sampler thus may ease the adaption of theoretical results about Bayesian variable selection (e.g. Scott and Berger, 2010) to the case of Bayesian networks.

Chapter 6

Exploratory network analysis of large social science questionnaires

There are now many large surveys of individuals that include questions covering a wide range of behaviours. Such surveys contain a vast amount of information. Surprisingly, not many studies have taken advantage of the availability of such rich data to investigate the possibility of unexpected and complex relationships in the data. In this chapter, we describe how structural inference for (dynamic) Bayesian networks can be used to explore relationships between variables in such data and present this information in an interpretable format for subject-matter practitioners.

The remainder of this chapter is organised as follows. We first introduce the aim of the study. We shall focus the study particularly on adolescent depression. We then introduce the Add Health dataset that we use. We finally present and discuss the resulting Bayesian network, focusing on depression, and provide estimates of how different variables affect the probability of depression via the overall probabilistic

structure given by the Bayesian network.

6.1 Introduction

6.1.1 Aims and background

Hypotheses that correspond to complex, multifactorial causes of symptoms and outcomes play an important role in the social sciences and in public health. The usual approach to exploring such hypotheses is through regression-based approaches. Considerable insight can be gained through such approaches, but it is sometimes overly constraining to fix a particular quantity as the dependent variable, especially if the goal is to explore the possibility of unexpected relationships between the data. Instead, we can consider a number of variables on an equal footing, and study the possibility of unexpected relationships in the data.

Consideration of unexpected relationships between factors requires datasets that incorporate a wide range of topics. Here, we investigate longitudinal data from the Add Health survey of adolescents in the US. However, such data are now widely available for representative samples of populations in many countries, and for many sub-groups of interest. Many of these datasets are derived from surveys that are general in scope, and are not collected to study any one particular question. For example, in the US, the health of the whole population is representatively sampled annually for the Behavioral Risk Factor Surveillance System (BRFSS) survey, and the Add Health study, which we use here, followed a cohort of young people from 1994 until 2008. Data from both of these have been used in scores of studies, but these commonly focus on one specific aspect, often using the data to evaluate existing hypotheses. Given the wide scope inherent in the design of these studies and the large samples available in many cases, we can broaden the scope of the analysis by considering richer structures. In this chapter, we discuss the potential that such a

more explorative approach yields. We do not seek to make conclusive causal claims, but instead suggest that a broader approach may uncover important aspects that have been neglected.

6.1.2 Adolescent depression

Our focus will be on depression among adolescents in the US, drawing on data from the National Longitudinal Study of Adolescent Health (Add Health). It is estimated that around 1–6% of adolescents each year are affected by depression (Costello *et al.*, 2003, 2006). The effects of depression in this age-group are wide-ranging (Thapar *et al.*, 2010), and include the stigma associated with poor mental health more generally (Patel *et al.*, 2007). There is considerable evidence that there is a wide range of causal factors for depression amongst adolescents, spanning biological, psychological and social domains. Understanding these causal factors and separating them from the consequences of depression has been recognised as an important aim (Barnett and Gotlib, 1988). Some of the relevant causal factors may interact and the approach taken here accounts for this.

6.1.3 Graphical models

As throughout this thesis, we use graphical models as the statistical framework within which the relationship between variables is studied, and focus on the structure of the model, as given by the graph. The use of graphs helps to make the interpretation of the model simpler. The structure of the model suggests how the different components of the system interact, which may be helpful in understanding the system as a whole.

Surveys often have a large sample-size. This clearly increases the precision of inference. However, it may mean that the posterior distribution over Bayesian networks

(or graphs) is concentrated on disparate graphs. In such situations, the standard MC^3 sampler converges very slowly to the posterior distribution. Instead, we use the Gibbs sampler introduced in Chapter 4, which moves more freely through graph space. Whilst the PC-algorithm (Spirtes *et al.*, 2000; Korb and Nicholson, 2011), as described in Section 2.7.2, has properties that often make it attractive in such contexts, we found that the results in this situation were not robust (see Section 6.4).

6.2 Data and methods

6.2.1 Add Health

The data that we use are drawn from the National Longitudinal Study of Adolescent Health (Add Health) that explores health-related behaviour of adolescents (Harris *et al.*, 2009) in the US. The questionnaire contains over 2000 questions that cover many aspects of adolescent behaviours and attitudes. We consider the representative sample of adolescents from Waves I and II of the in-home section, and the parental questionnaire from Wave I of the study. The analysis is not feasible when the data is not complete (see Section 7.2), and so individuals with missing data were removed from the study. Removing incomplete samples leaves 5975 individuals in the study.

Our measure of depression is a self-assessed scale based upon the Centre for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977). Two questions from the 20-item scale are omitted from Add Health, and two are modified, and so we scale the score given by the available questions (Goodman, 1999). A Receiver Operating Characteristic (ROC) analysis showed that thresholds of 24 for females and 22 for males provided the best agreement with clinical assessments of depression (Roberts *et al.*, 1991). We use this threshold to create a binary indicator of depression status.

Many of the remainder of the variables that we consider (Table 6.1) are drawn from

the risk factors described in the depression literature, and the mental health literature more generally. A recent review (Patel *et al.*, 2007) described a wide range of factors that are associated with poor mental health in young people, including gender, poverty, violence and the absence of social networks in the local neighbourhood. The quality of relationships with parents is also thought to be important, especially with the mother (Holt *et al.*, 2008), as are parental alcohol problems (Obot and Anthony, 2004) and parental discord (Holt *et al.*, 2008). The individual's use of alcohol, drugs, smoking and HIV/AIDS are all also associated with depression (Brown *et al.*, 1996; Battles and Wiener, 2002). Physical exercise has been proposed in some studies as a useful intervention for the management of depression, but many of these studies have been deemed to be poor quality (Larun *et al.*, 2006).

Table 6.1: The labels used in the plots below, the number of categories (r), and the exact wording of the question. The ID(s) of the relevant variables in the Add Health dataset are in parentheses. See www.cpc.unc.edu/projects/addhealth for full details of all of these questions.

Label	**r**	**Question**
Female	2	Interviewer, please confirm that R's sex is (male) female. (BIO_SEX)
Hispanic/Latino	2	Are you of Hispanic or Latino origin? (H1GI4)
White	2	What is your race? [White] You may give more than one answer (H1GI6A)
Black/African American	2	What is your race? [Black or African American] You may give more than one answer (H1GI6B)
American Indian/ Native American	2	What is your race? [American Indian or Native American] You may give more than one answer (H1GI6C)
Asian/Pacific Islander	2	What is your race? [Asian or Pacific Islander] You may give more than one answer (H1GI6D)
Other race	2	What is your race? [Other] You may give more than one answer (H1GI6E)

Skips school	4	[If SCHOOL YEAR:] During this school year [If SUMMER:] During the 1994–1995 school year how many times HAVE YOU SKIPPED/DID YOU SKIP school for a full day without an excuse? (H1ED2; H2ED2)
Experiences prejudice	3	[If SCHOOL YEAR:] Students at your school are prejudiced [If SUMMER:] Last year, the students at your school were prejudiced. (H1ED21; H2ED17)
In physical fights	4	In the past 12 months, how often did you get into a serious physical fight? (H1DS5; H2FV16)
Didn't present to doctor	2	Has there been any time over the past year when you thought you should get medical care, but you did not? (H1GH26; H2GH28)
Severely injured	3	Which of these best describes your worst injury during the past year? (H1GH54; H2GH47)
Have HIV/AIDS	2	Have you ever been told by a doctor or a nurse that you had... HIV/AIDS (H1CO16D; H2CO19D)
Seen shooting	3	During the past 12 months, how often did each of the following things happen? You saw someone shoot or stab another person. (H1FV1; H2FV1)
Mother warm/loving	4	Most of the time, your mother is warm and loving toward you. (H1PF1; H2PF1)
Been suspended	2	Have you ever received an out-of-school suspension from school? (H1ED7; H2ED3)
Been expelled	2	Have you ever been expelled from school? (H1ED9; H2ED5)
Good health	3	In general, how is your health? Would you say... (H1GH1; H2GH1)
Talks to neighbours	2	In the past month, you have stopped on the street to talk with someone who lives in your neighborhood? (H1NB2; H2NB2)
Age	5	Age at interview, computed from date of birth, and date of interview (Constructed from IYEAR, IMONTH, IDAY, H1GI1Y, H1GI1M)
Live with mother	2	Indicator variable (Constructed from H1HR3A-T; H2HR4A-Q)
Live with father	2	Indicator variable (Constructed from H1HR3A-T; H2HR4A-Q)
Smoker	4	Frequency of smoking (Constructed from H1TO1/2/5; H2TO1/5)

Drinks alcohol	4	Frequency and amount of drinking alcohol (Constructed from H1TO12/15/18; H2TO15/19/22)
Exercises	3	Amount of exercise (Constructed from H1DA4/5/6; H2DA4-6)
Depressed	2	Rescaled CES-D, following (Goodman, 1999) (Constructed from H1FS1-18; H2FS1-18)
Victim of violence	2	Indicator variable (Constructed from H1FV2-6; (H2FV2-5)
Family bereavement	3	Number of bereavements (Constructed from H1NM2/F2, H1FP24A1-5; H2NM4/F4, H2FP28A1-3)
Strong academically	4	Quartiles (Constructed from H1ED11-4; H2ED7-10)
Drug user	2	Indicator variable (Constructed from H1TO30/34/37/41; H2TO44/50/54/58)
Family poor	5	Census Bureau measure of poverty (Constructed from H1HR2/3/7/8, PA55)
Parents unhappy together	4	(Parent asked.) Do you and your partner argue/talk of separating? (Constructed from PB19/20)
Parent drinks	4	(Parent asked.) Number/frequency of drinks (Constructed from PA61/2)
Householder smokes	3	(Parent asked.) Either parent or others in household smokes (Constructed from PA63/4)
Has learning disability	2	(Parent asked.) Does (he/ she) have a specific learning disability, such as difficulties with attention, dyslexia, or some other reading, spelling, writing, or math disability? (PC38)
Parents aid decisions	5	(Parent asked.) How often would it be true for you to make each of the following statements about {child's name}? {Child's name} and you make decisions about (his/ her) life together. (PC34B)

6.2.2 Methods

We will use structural inference of Bayesian networks to explore the relationships between variables in the Add Health study. We use the usual multinomial-Dirichlet formulation (Section 2.5.2). We choose a graph prior $\pi(G) \propto 1$ that is flat across the space of graphs, and we will approximate the posterior distribution using MCMC.

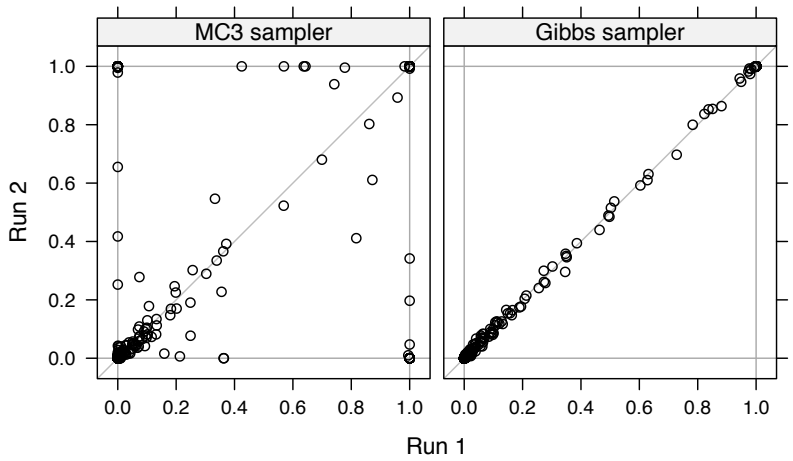


Figure 6.1: Convergence diagnostics for MC^3 (left) and the Gibbs sampler (right) for the Add Health data. The posterior edge probabilities given by two independent runs are plotted against each other. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. The two Gibbs runs give similar posterior edge probabilities, but the MC^3 runs do not. (5 runs of 750,000 samples (MC^3) or 100,000 samples (Gibbs) of each sampler were performed; the first half of the samples were discarded as burn-in; mean Pearson correlation between runs was 0.9999 ± 0.0002 (standard deviation) for Gibbs and 0.6322 ± 0.0477 for MC^3.)

6.3 Results

The variables that we consider are detailed in Table 6.1. As is common when using graphical models (Cox and Wermuth, 1996), all of these variables were grouped, initially into 'Background', 'Wave I' and 'Wave II', and then refined into whether the question asked about the long- or short-term, as shown in Table 6.2. These groups define constraints on the Bayesian networks that are considered. Specifically, no edges can be directed backwards through the groups. Edges, however, are allowed within groups. For example, no edge is allowed to be directed into 'Gender', and no edge can pass backwards in time, for example, from Depression at Wave II to Depression at Wave I. Additionally, no edge can pass from a short-term variable to a long-term variable in the same wave, for example, from Depressed at Wave I to Have HIV/AIDS at Wave I.

Table 6.2: The groupings of the variables that were used to determine constraints on the Bayesian networks. Each variable in the analysis is either a Background variable, or from Wave I or Wave II of the Add Health study. Within each wave of the study, variables were further classified into whether they asked about the short- or long-term.

Background	Wave I Long-term	Wave I Short-term	Wave II Long-term	Wave II Short-term
Female	Skips school	Householder smokes	Seen shooting	Smoker
Age	Experiences prejudice	Smoker	Alcohol	Live with mother
Hispanic/Latino	In physical fights	Live with mother	Drug user	Live with father
White	Didn't present to doctor	Live with father	Mother warm/loving	Talks neighbours
Black/African American	Severely injured	Parent drinks	Have HIV/AIDS	Exercises
American Indian/Native American	Have HIV/AIDS	Talks neighbours	Family bereavement	Depressed
Asian/Pacific Islander	Seen shooting	Exercises	Experiences prejudice	
Other race	Mother warm/loving	Depressed	Been expelled	
Has learning disability	Been suspended		Been suspended	
	Been expelled		Victim of violence	
	Good health		In physical fights	
	Alcohol		Strong academically	
	Victim of violence		Didn't present to doctor	
	Family bereavement		Skips school	
	Strong academically		Severely injured	
	Drug user		Good health	
	Family poor			
	Parents unhappy together			
	Parents aid decisions			

We precomputed the local scores, and then drew 100,000 samples (the first half of which were discarded as burn-in) using the Gibbs sampler (Chapter 4), which took 30 minutes (on a single core of a cluster computer). The graph space was constrained such that no node had more than 3 parents, to ensure Equation 1 could be evaluated.

We ran 5 independent samplers, with disparate initial states. This enables a simple test of convergence to be performed that compares the posterior edge probabilities obtained from each of the independent runs (Robert and Casella, 2004). The agreement between runs can be examined graphically by plotting the edge probabilities against each other (Figure 6.1). Mean Pearson correlation coefficients between edge probabilities from pairs of runs were 0.9999 ± 0.0002 (standard deviation) for the Gibbs sampler and 0.6322 ± 0.0477 for MC^3. The agreement between the independent runs of the Gibbs sampler gave us confidence in our results, in contrast to the large disagreements between MC^3 runs. In addition, cumulative edge probability plots for each edge showed regular excursions around the mean (Yu and Mykland, 1998), and a numerical diagnostic (Gelman and Rubin, 1992) monitoring the number edges in the sampled graph also clearly suggested that sufficient samples had been drawn ($\hat{R} \approx 1.0$).

The samples drawn using MCMC allow the posterior distribution of Bayesian networks to be approximated. In particular, the samples can be used to estimate the posterior edge probability $P(e \mid \mathbf{X})$ with $e \in E$. Figure 6.2 displays all edges with posterior probability of at least 0.5.

Our focus is on depression, the parents of which in Figure 6.2 we observe are "Didn't present to doctor" and "Gender". It is important, however, to note that the model does not imply that these are the only factors that are important. For example, "Drug user" at Wave I is related to depression through "Didn't present to doctor"

Figure 6.2: Summary network for the Add Health variables considered. The edge colours are given by the Kendall correlation coefficients between the two variables, with green edges corresponding to positive correlation, and red edges to negative correlation. The strength of the correlation is indicated by the transparency of the line, with greater transparency indicating weaker correlation. The variables 'Depressed (1)', 'Depressed (2)' and their parents are shown in bold.

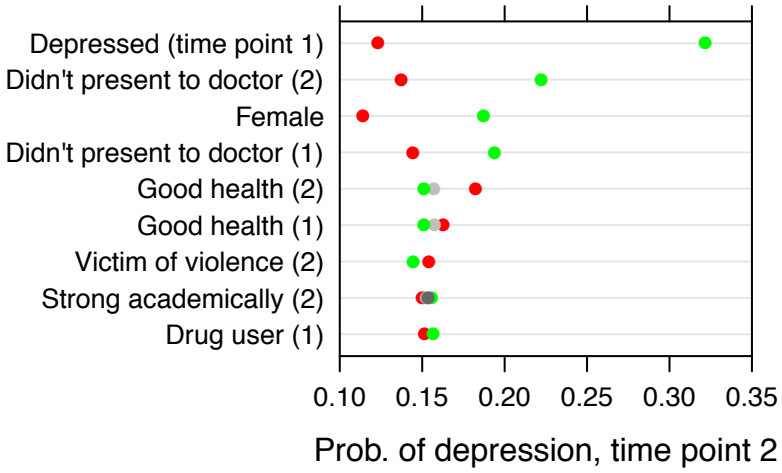


Figure 6.3: Conditional probability of depression. The conditional probability of being depressed at Wave II given the variable indicated is changed to the level indicated by the colours, conditional on the DAG shown in Figure 6.2. For binary variables, ● is true, and ● is false; shades of grey indicate intermediate levels. Wave number (time point) is indicated in parentheses. Only variables for which the conditional probability differed between levels by at least 0.005 are displayed.

at Wave I and II (Figure 6.2).

These effects are shown in Figure 6.3, which gives the conditional probability of being depressed at Wave 2 when a particular variable is set to a specific value. We see that general health, violence, academic performance and drug use all affect the conditional probability of depression at Wave II. To compute this probability, links from the parents of the variable in which we 'intervene' are removed; this is equivalent to the 'do-operator' in the terminology of Pearl (Pearl, 2009).

The analysis reveals the interaction between the many aspects of life that have an impact on depression. The connection between the depression and its two parents in Figure 6.2 have been previously discussed in the literature. The importance of gender in depression is particularly extensively documented in the literature (Patel *et al.*, 2007). The connection to a failure in seeking medical care even when the individual thinks they should has also been discussed in the literature, often in

terms of poor accessibility of health care services for young people (Rickwood *et al.*, 2007; Patel *et al.*, 2007). Several decades of research have revealed the complex causation of depression in young people, as suggested by this study (Patel *et al.*, 2007).

6.4 Discussion

There is a large amount of information held in large social science questionnaires. In this chapter we have examined a graphical model approach to inferring structure amongst the variables in such questionnaires. In contrast to the standard regression-based approaches, a graphical model approach forgoes the need to specify a particular variable as the response. Instead, a more comprehensive estimate of the entire structure of the underlying system can be obtained. Regression approaches posit a particular conditional-independence structure, while graphical approaches allow consideration of more general structures.

The limitations of this study include those of all similar studies using observational data that are collected for multiple audiences. These forms of data, including the longitudinal data used here, do not permit strong causal conclusions to be drawn. In particular there may be important variables that we have not included in the analysis. However, the results are consistent with studies that have used other research approaches including experimental designs. The connection between an individual not seeking medical care when they think they should and depression supports current practice guidance in the UK (National Institute for Health and Clinical Excellence, 2005) where there is an emphasis on providing access to health care through the school system rather than expecting young people to seek health care themselves. Not seeking medical care despite believing it should be sought is a complex factor because it captures both barriers to receiving medical care within

the individual, such as lacking motivation to seek care, and barriers within the individual's environment, such as poor access to care. This complexity may mean that the variable encapsulates various different characteristics related to depression, and thus may form a 'marker' for depression. However, the use of a form of the question "Has there been any time over the past year when you thought you should get medical care, but you did not?" as a screening question in different contexts needs further consideration.

This method of analysis clarifies the complexity of depression and suggests why when using traditional methods of analysis it can be difficult to clarify whether or not factors such as experiences in the family, in the wider community and at school impact on the experience of depression for young people. It may also suggest why interventions for prevention of depression have not yet been demonstrated to be cost effective (Merry, 2007).

We performed structural inference for the Bayesian network using a Gibbs sampler (introduced in Chapter 4), because MC^3 did not mix in a reasonable time. We have also found the Gibbs sampler to be superior to the REV sampler (Grzegorczyk and Husmeier, 2008), and it has the advantage of avoiding the need to consider an order prior as required by order MCMC methods (Ellis and Wong, 2008; Friedman and Koller, 2003), which induces a bias that can only be corrected exactly by NP-hard computation of a correction factor.

An alternative to the MCMC method used here is the PC-algorithm (Spirtes *et al.*, 2000; Korb and Nicholson, 2011), described in Section 2.7.2. This method is computationally efficient and is asymptotically consistent. However, to test whether the sample size available here is sufficient to reach the asymptotic regime, we applied the PC-algorithm (without constraints) to 10 different subsamples, each containing 90% of the data. We found that these results differed significantly, with a mean 84 in structural Hamming distance between the pairs of completed partially directed

acyclic graphs (CPDAGs) given for the subsamples.

Chapter 7

Discussion

In this thesis we have applied and developed statistical methods that manage uncertainty about the structure or form of models. We have demonstrated the application of structural inference of Bayesian networks and related models for large social science datasets, and have developed a new method for approximating the relevant posterior distribution. In Chapter 3 we compared models for flexible discrete models of risk taking. Then in Chapters 4 and 5, we developed and tested a novel MCMC sampler for structural inference of Bayesian networks. In Chapter 6 we used this to investigate depression in adolescents.

There are various extensions and further areas in application, modelling, and estimation that would be interesting to investigate in future work. We first discuss extensions that are particularly relevant to Chapter 3 and then to Chapter 6. We then consider extensions of the modelling framework that we use throughout the thesis. Finally we consider developments of the MCMC methodology, and other alternative approaches.

7.1 Well-being and risky behaviour

In Chapter 3, we gave evidence in support of a relationship existing between well-being and risk taking. There is further work that could be done to extend the results. Some of the evidence in the chapter is not definitive because happiness cannot be randomly assigned by an experimenter. Even assuming the direct relationship between the factors exists, without randomly assigning happiness, we can not be sure whether subjective well-being affects risk taking behaviour, or vice-versa, or whether there are effects in both directions.

One approach to investigating this further is through a simple experiment, that assesses the risk taking characteristics of individuals when they are induced to be more and less satisfied. The selection of the method by which satisfaction is controlled exogenously (i.e. by the experimenter) would clearly be key. In particular, the difference between short-term and long-term well-being would need consideration. It is likely that only short-term satisfaction could be controlled, but the effects of this may be unlike the effects of long-term satisfaction.

7.2 Depression in adolescents

In Chapter 6 we highlighted the importance of an adolescent's feeling they should have seen a doctor, but did not. This effect has been mentioned before in the literature (as described in Section 6.4) but has not been highlighted before.

Our finding came from an exploratory analysis of observational data, from which it is not usually possible to draw strong causal conclusions. Two areas of particular concern are the removal of samples with missing data and the possibility that important variables that have been omitted, so are latent. It is possible to handle missing data formally, for example by using structural EM (Friedman, 1998), and

similarly consider latent variables (e.g. shared genetics driving both child and parent behaviour). However, doing so whilst robustly exploring large model spaces remains an open challenge. Tackling these computational and inferential issues is a key area for future research.

Another area for future research is to consider the complementary predictive model. The model that we consider in Chapter 6 includes within-time-slice edges, for example from "Didn't present to doctor (wave 2)" to "Depressed (wave 2)". The model is thus explanatory, rather than predictive. A predictive model can be constructed by including only the depression indicator at Wave 2. This model may be useful when the aim is early identification of adolescents at risk of future depression, as might be the case in clinical practice.

7.3 Model enhancements

In the following section, we consider generalisations and issues with the likelihood, priors and posterior summaries that we use in this thesis.

7.3.1 Errors-in-variables models

In regression, the predictor variables are typically assumed to be observed without error. However, in a Bayesian network model variables act as both predictors and outcomes. In using a regression model for the local likelihood, we are thus assuming that the variables are observed without error when they act as predictors, but are observed with a form of error or randomness when they act as outcomes. Errors-in-variables models (e.g. Dellaportas and Stephens, 1995), which acknowledge errors in the observation of predictors, may thus provide an improved method for modelling in this setting.

7.3.2 Parameter priors

In the applications we used a multinomial-Dirichlet model for the local conditional distributions, which yields a closed-form marginal likelihood. This specification has the advantage of being a very flexible model; it is non-parametric in the sense that no constraints are placed on the distribution, allowing its form to be guided by the data. However, the number of parameters in the local distributions for this model increases exponentially with the number of parents, which may mean that overly-sparse models are preferred. This increase in the number of parameters is particularly problematic when the sample size of the available data is small, because models with many parameters cannot be assessed adequately without a large dataset. The large sample size of the datasets used here minimises this issue, but it would nonetheless be worthwhile to consider more compact parameterisations. However, estimating such models (Friedman and Goldszmidt, 1996) significantly increases the complexity of the model space, which makes such an approach computationally challenging in this setting.

There are also unsatisfactory aspects to the g-prior when it is used for model selection. These relate to when the improper prior with $g \to \infty$ is used, and the bounded nature of the associated Bayes factors when overwhelming evidence implies one particular model. These are discussed in Liang *et al.* (2008) and Berger and Pericchi (2001). Various alternatives to the g-prior that ameliorate some of its unsatisfactory aspects have been advanced, including using a mixture of g-priors (Liang *et al.*, 2008), an approach that has been generalised by Deltell (2011).

7.3.3 Model priors

In this thesis, we used a prior that is flat over the space of DAGs. An unsatisfactory aspect of this prior is that it places higher prior probability on models that include

more variables. Consider regression with p covariates. The problem arises because there are more models with $p_\gamma+1$ variables than with p_γ variables, when $p_\gamma+1 < p/2$. This has been noted by various authors, including Scott and Berger (2010). Taking a uniform prior on the number of variables in the model gives the following prior.

$$\pi(M_\gamma) = \frac{1}{p+1}\binom{p}{p_\gamma}^{-1}$$

Alternatively, a beta prior can be used for the inclusion probability of each possible predictor (Ley and Steel, 2009). Where there is particular knowledge of interactions, the prior developed by Chipman (1996) that assigns differing prior weight to particular interactions may be useful.

7.4 Posterior approximation

In this section, we consider improvements to the MCMC methodology used in this thesis, and discuss the merits of other approaches.

7.4.1 Convergence diagnostics

Assessing convergence of Markov Chains on a space as large as the space of Bayesian networks is not straightforward. In the thesis, we focused on comparisons of posterior edge probabilities across runs.

One of the most satisfactory methods in general for assessing convergence is to examine regeneration times. Regeneration in Markov Chain occurs at times $\{\tau_t : t = 1, \ldots, T\}$ when, conditioned on τ_t for some $t \in \{1, \ldots, T\}$, the sample paths of the Markov Chain before and after τ are independent. While the usual Central Limit Theorem does not apply to Equation 2.5, if we observe the Markov chain until a fixed number of regenerations have occurred, the usual sample mean based upon the

samples drawn between the first and last regeneration is a consistent estimator for the mean. This idea was proposed as a convergence assessment method by Mykland *et al.* (1995); see also Robert (1995).

On discrete spaces, regenerations are easy to define: incursion into any subset of the state space can be defined to constitute a regeneration. We would like regenerations to occur frequently, and so it is sensible to choose the regeneration set to be of significant posterior mass. The posterior mode is a good choice, if a good estimate of it is available.

7.4.2 Order approaches

An alternative MCMC sampler for structural inference in Bayesian networks is MCMC in order-space (Friedman and Koller, 2003; Ellis and Wong, 2008; Eaton and Murphy, 2007). This approach samples total orders rather than Bayesian networks directly. Often this improves the mixing of the sampler. However, to use this sampler, a prior over the space of order must be constructed. Unfortunately, the number of total orders with which a Bayesian network is consistent is not constant, and so only an approximation to standard graph priors can be used in this approach (Ellis and Wong, 2008; Eaton and Murphy, 2007). In contrast, no prior over order space is required for the REV sampler, or for the Gibbs sampler used here. Given this, and results in Grzegorczyk and Husmeier (2008) that suggest that the REV sampler matches the performance of order MCMC, we view the REV sampler and Gibbs sampler as more satisfactory.

These approaches have more recently led to exact methods (Koivisto and Sood, 2004; Parviainen and Koivisto, 2009; Tamada *et al.*, 2011) using dynamic programming. However, the exact methods are extremely computationally demanding, and the same form of graph priors is required for these methods as for order space MCMC.

7.4.3 Improvements to the Gibbs sampler

Highlighting the connection to Bayesian variable selection, as the Gibbs sampler does, suggests further possibilities. Evaluation of the exact posterior of the associated variable selection problem is required for exact sampling from the conditional distribution $P(F_W \mid G_{-W}, \mathbf{X})$. When this is computationally prohibitive, alternatives are possible. In particular, we can substitute a Metropolis step in place of the Gibbs step when the required conditional distribution is not available. This form of sampler is known as Metropolis-within-Gibbs (Müller, 1991). When $|W| = 1$, the conditional distribution $P(F_W \mid G_{-W}, \mathbf{X})$ is identical to the posterior distribution of the corresponding Bayesian variable selection. This correspondence means that the Metropolis-within-Gibbs move can exploit algorithms designed for variable selection. The most straightforward move is a component-wise Gibbs move of the form used by Smith and Kohn (1996). However, using such a 'small' move negates the advantages of the Gibbs sampler introduced here. Instead, a blocked Gibbs move for the variable selection, as discussed by George and McCulloch (1997) and Kohn *et al.* (2001), is more appropriate. Other alternatives include the version of the Swendsen-Wang algorithm proposed by Nott and Green (2004). When $|W| > 1$ there are more complications.

A drawback of all of these variations is that the exact form of the conditional distribution is no longer used. Instead, a single draw from a random-walk type Metropolis proposal is made. As is often the case with random-walk Metropolis proposals, it is difficult to make large moves without the acceptance probability becoming small. One approach that may be useful in this context is Multiple-try Metropolis (Liu *et al.*, 2000), in which a set of proposals is drawn, and then a final proposal is sampled from the set of proposals. While such a sampler will usually make proposals with larger acceptance probabilities, the extra computation required to draw the set of proposals may mean that, adjusting for computation time, it is

not an improvement.

Throughout Chapter 5, we used $|W| = 3$ and found this yielded a sampler with attractive properties. In some settings it may be advantageous to use $|W|$ as a tuning parameter for the algorithm. There is clearly a trade-off: increasing $|W|$ increases the time taken to evaluate $P(F_W \mid G_{-W}, \mathbf{X})$, but also increases move size, which should improve the convergence rate of the sampler. Another possibility is to choose $|W|$ at each step according to some distribution, so that a mixture of different block sizes is used.

Another marginal improvement in the properties of the sampler may be possible by converting the Gibbs sampler into a Metropolised Gibbs sampler, in which the graph sampled at each step is always different from the current graph. As noted by Liu (1996), the asymptotic variance of an estimator based upon such a sampler will be lower than a Gibbs sampler. However, sometimes the Gibbs sampler converges faster (Frigessi *et al.*, 1993).

The Gibbs sampler does not work in all situations. In Chapter 5 we gave an example of the Gibbs sampler using data from the New York part of the BRFSS study. The Gibbs sampler converges well for this example, but does not converge rapidly for the corresponding variables for the full BRFSS study. The difficulty stems from the large sample size, which makes the posterior distribution 'peaky' to such an extent that even the Gibbs sampler does not mix well. There are number of possible approaches that help. Tempering, as in Barker *et al.* (2010), may be another approach that would help in this situation, because it would reduce the 'peakiness' of the distribution being explored.

7.4.4 Generalising the approach

A key part of the thesis is an improved MCMC sampler for structural inference of Bayesian networks. The Gibbs sampler works by considerably increasing the variance of the proposal distribution. In a Metropolis-Hastings framework, this would usually be undesirable because of the concomitant decrease in the acceptance rate. Instead, we consider the appropriate conditional distribution of large blocks of random variables, thus constructing a Gibbs sampler. Approximating discrete distributions with enormous sample spaces is a common problems in many areas of statistics. Using a Gibbs sampler on discrete spaces with large blocks may also be useful in these contexts.

Appendix A

Data used in Chapter 5

The following variables from the BRFSS data were used in the analysis.

SEX, _AGE_G, _RACEGR2, MARITAL, _CHLDCNT, _INCOMG, USEEQUIP, _HCVU65, MEDCOST, _SMOKER3, _ASTHMST, _RFDRHV3, _RFBING4, QLREST2, _RFSEAT3, _TOTINDA, _BMI4CAT, DIABETE2, EMTSUPRT, LSATISFY, _EXTETH2, _AIDTST2, _DENVST1, IMONTH

The following quantities were included in our analysis of the flow cytometry data, including the binding of antibodies, viability, DNA content. Full details can be found in Bendall *et al.* (2011).

191-DNA, 193-DNA, 103-Viability, 115-CD45, 139-CD45RA, 141-pPLCgamma2, 142-CD19, 144-CD11b, 145-CD4, 146-CD8, 148-CD34, 150-pSTAT5, 147-CD20, 152-Ki67, 154-pSHP2, 151-pERK1/2, 153-pMAPKAPK2, 156-pZAP70/Syk, 158-CD33, 160-CD123, 159-pSTAT3, 164-pSLP-76, 165-pNFkB, 166-IkBalpha, 167-CD38, 168-pH3, 170-CD90, 169-pP38, 171-pBtk/Itk, 172-pS6, 174-pSrcFK, 176-pCREB, 175-pCrkL, 110_114-CD3

Appendix B

Additional figures for Chapter 5

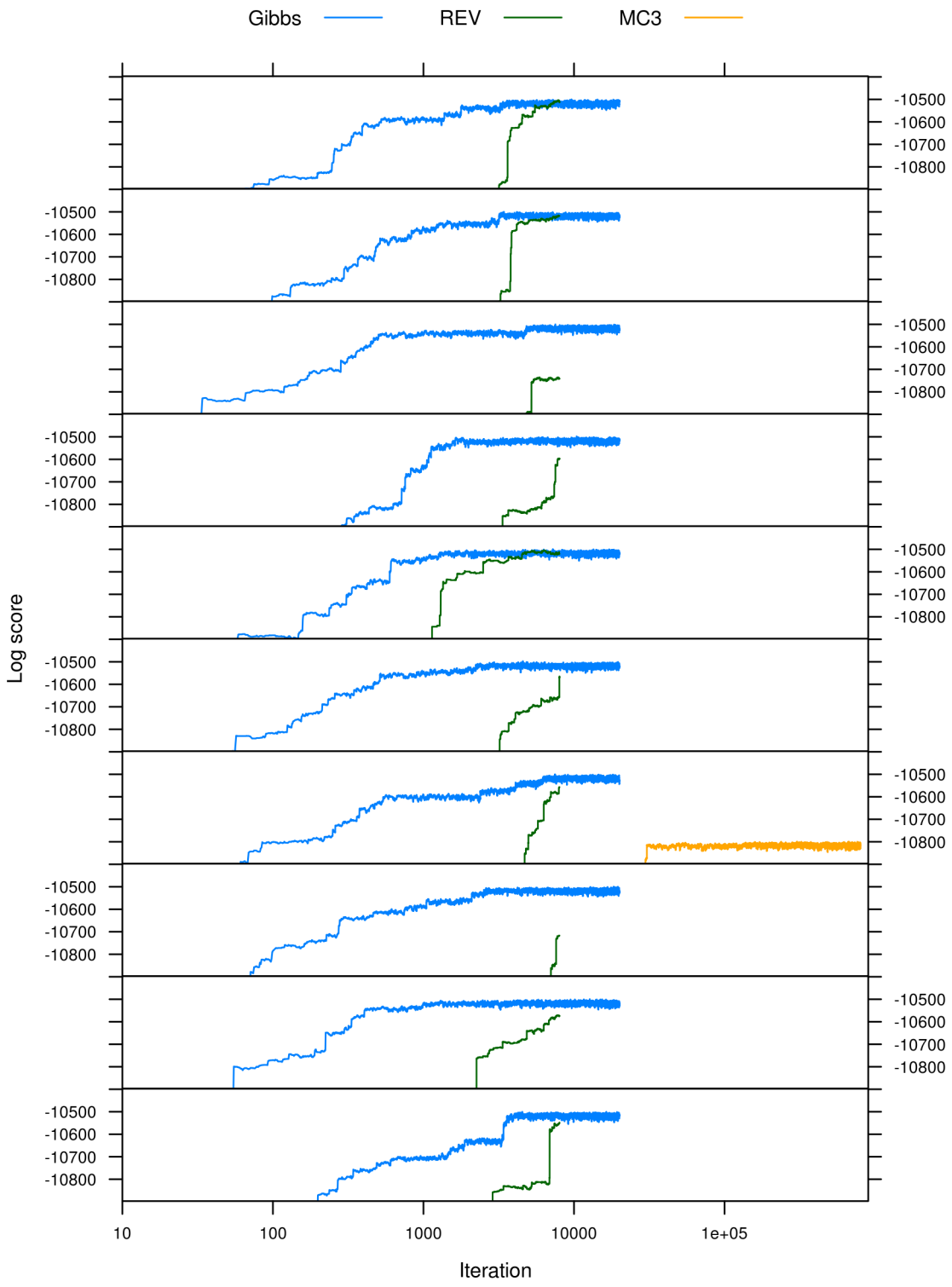


Figure B.1: Log scores of the graphs visited by the three MCMC samplers in 10 independent runs on the ALARM data, with $n = 1000$, initialised at disparate initial conditions. Iteration number is displayed on a $\log_{10}$ scale. Each sampler was run for 30 minutes. In this time, MC^3 drew the most samples. However, neither MC^3 nor the REV sampler routinely reach the plateau reached by the Gibbs sampler.

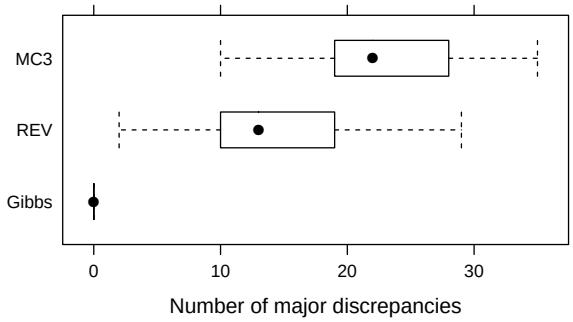


Figure B.2: Major discrepancies between pairs of the 10 independent runs, for each MCMC sampler on the BRFSS data. For each pair of independent runs, the number of major discrepancies is the number of edges that have estimated posterior edge probability above 0.9 in one run and estimated posterior edge probability below 0.1 in the other run. The boxplot shows the range of discrepancies between runs.

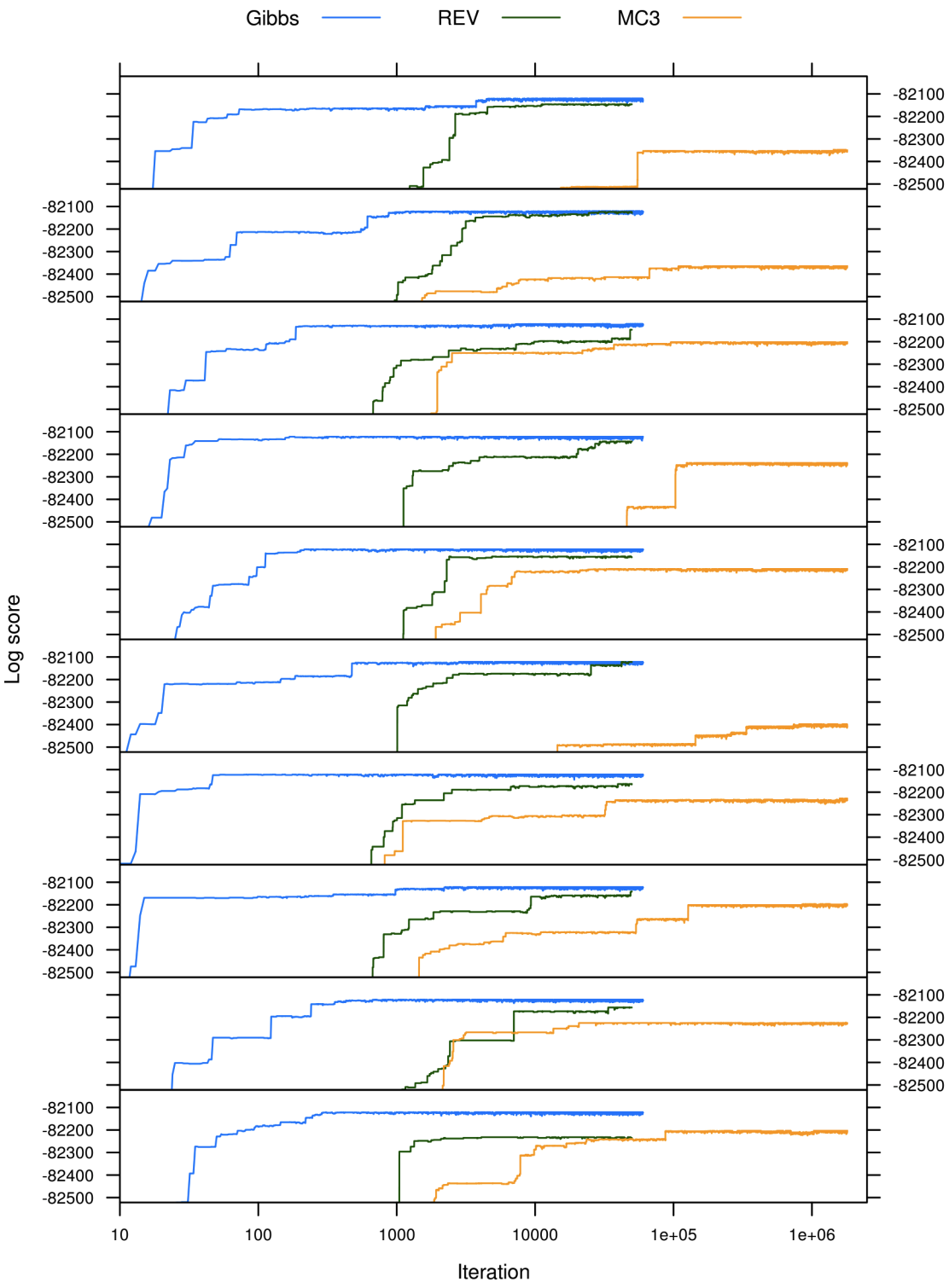


Figure B.3: Log scores of the graphs visited by the three MCMC samplers in 10 independent runs on the BRFSS data, initialised at disparate initial conditions. Iteration number is displayed on a $\log_{10}$ scale. Each sampler was run for 30 minutes. In this time, MC^3 drew the most samples. However, neither MC^3 nor the REV sampler reach the plateau reached by the Gibbs sampler.

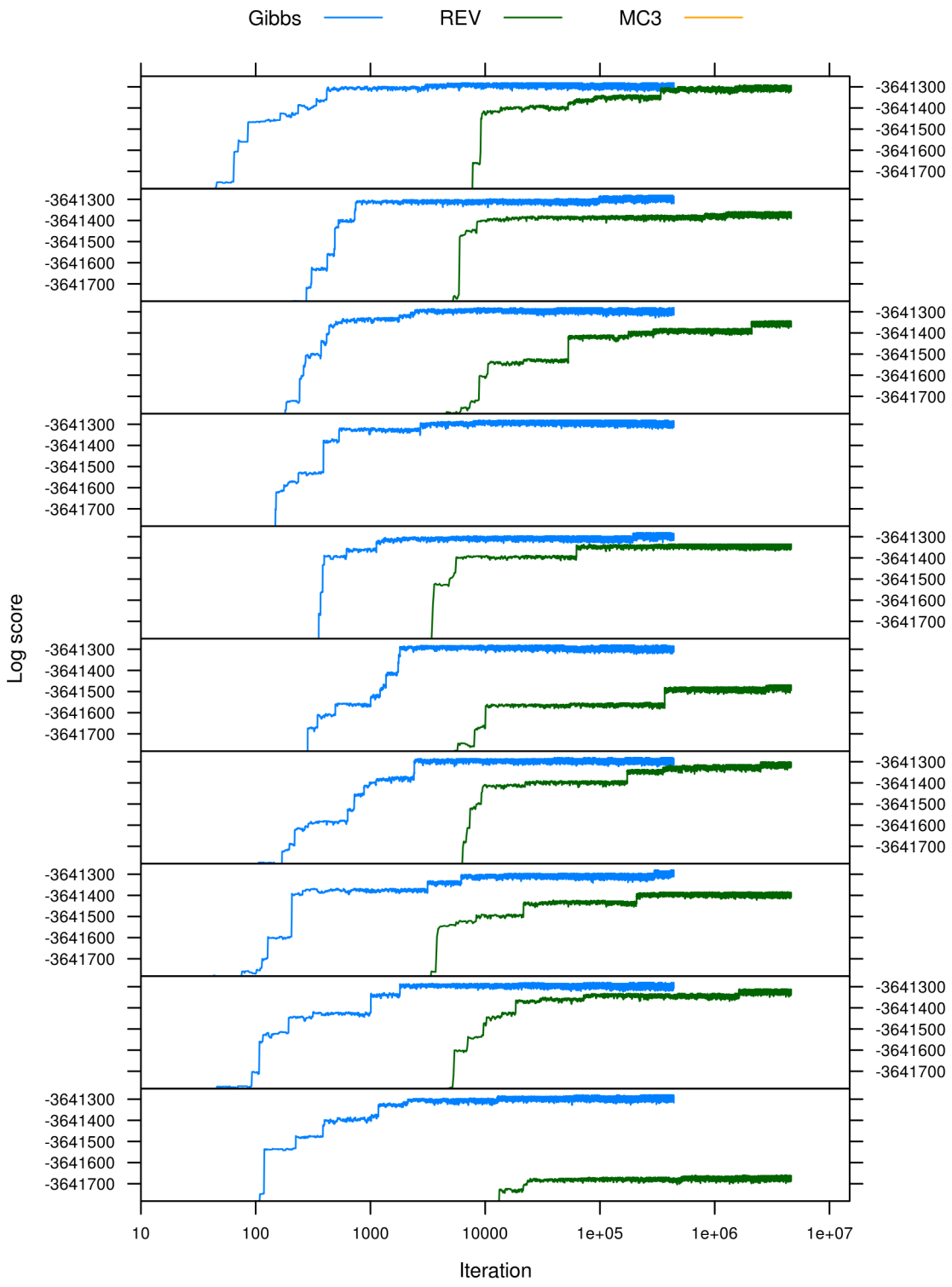


Figure B.4: Log scores of the graphs visited by the three MCMC samplers in 10 independent runs on the flow cytometry data, initialised at disparate initial conditions. Iteration number is displayed on a $\log_{10}$ scale. Each sampler was run for 8.5 hours. In this time, MC^3 drew the most samples, but never breached -3646099 in log score, and so is not shown. The REV sampler also does not reach the plateau reached by the Gibbs samplers.

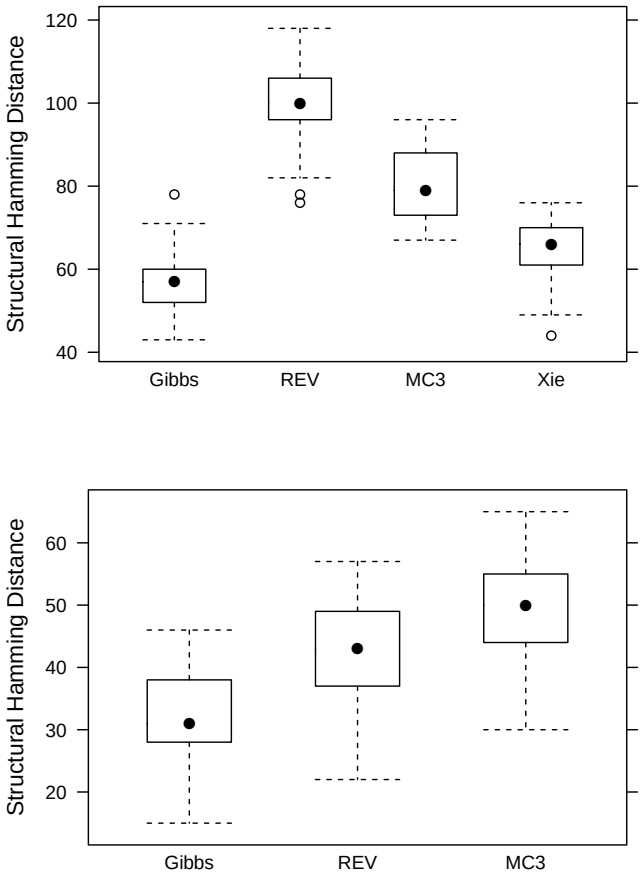


Figure B.5: Stability of estimators across bootstrapping. 10 bootstrap samples were drawn from the BRFSS data. The structural Hamming distance (SHD) between the graphs given by each estimator on each bootstrap sample is shown. Smaller SHD means that the graphs are structurally more similar, and so the estimator is more stable. In (A) the edge probabilities were thresholded so that the resulting graphs had the same edges as the point estimate graph given by the Xie-Geng method. In (B) the graphs are given by thresholding at 0.5 the edge probabilities from the 3 MCMC samplers.

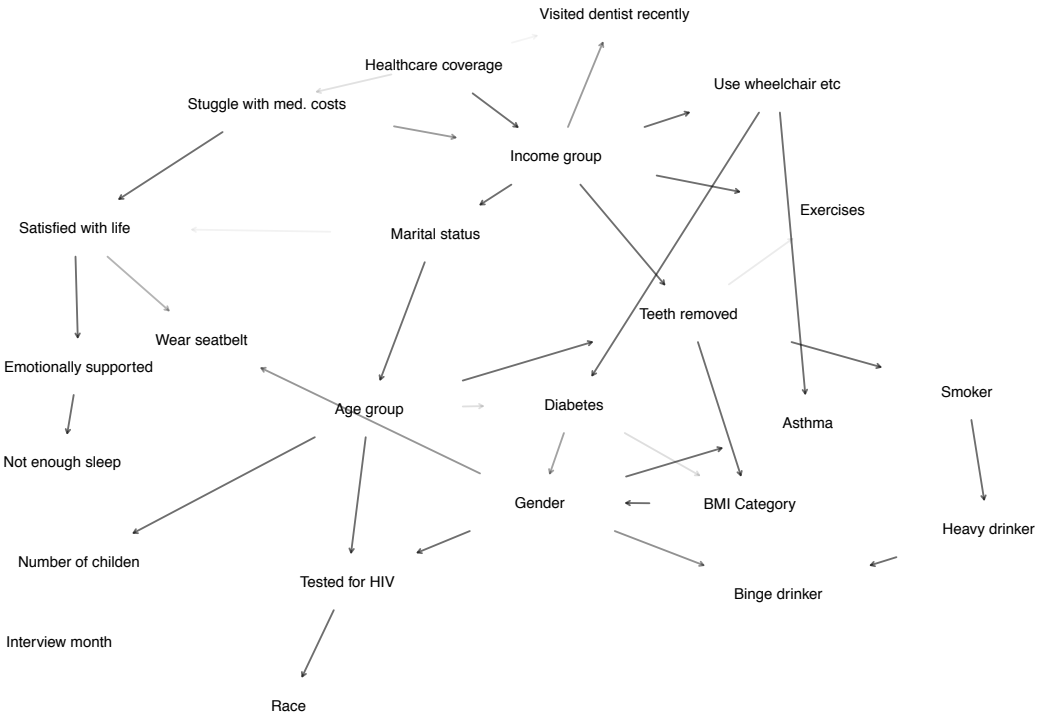


Figure B.6: The edges with posterior edge probability greater than 0.5, as given by the Gibbs sampler for the BRFSS data. The gray-to-black scale gives an indication of the posterior edge probability. Note that no hard constraints were specified to ensure, for example, an indegree of 0 for 'Age Group'; such constraints were omitted to keep the implementations of the various methods simple.

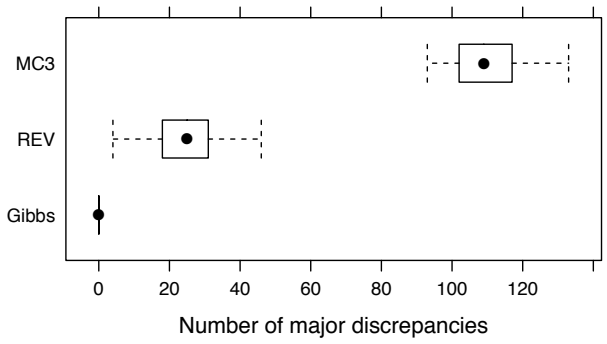


Figure B.7: Major discrepancies between pairs of the 10 independent runs, for each MCMC sampler on the flow cytometry data. For each pair of independent runs, the number of major discrepancies is the number of edges that have estimated posterior edge probability above 0.9 in one run and estimated posterior edge probability below 0.1 in the other run. The boxplot shows the range of discrepancies between runs.

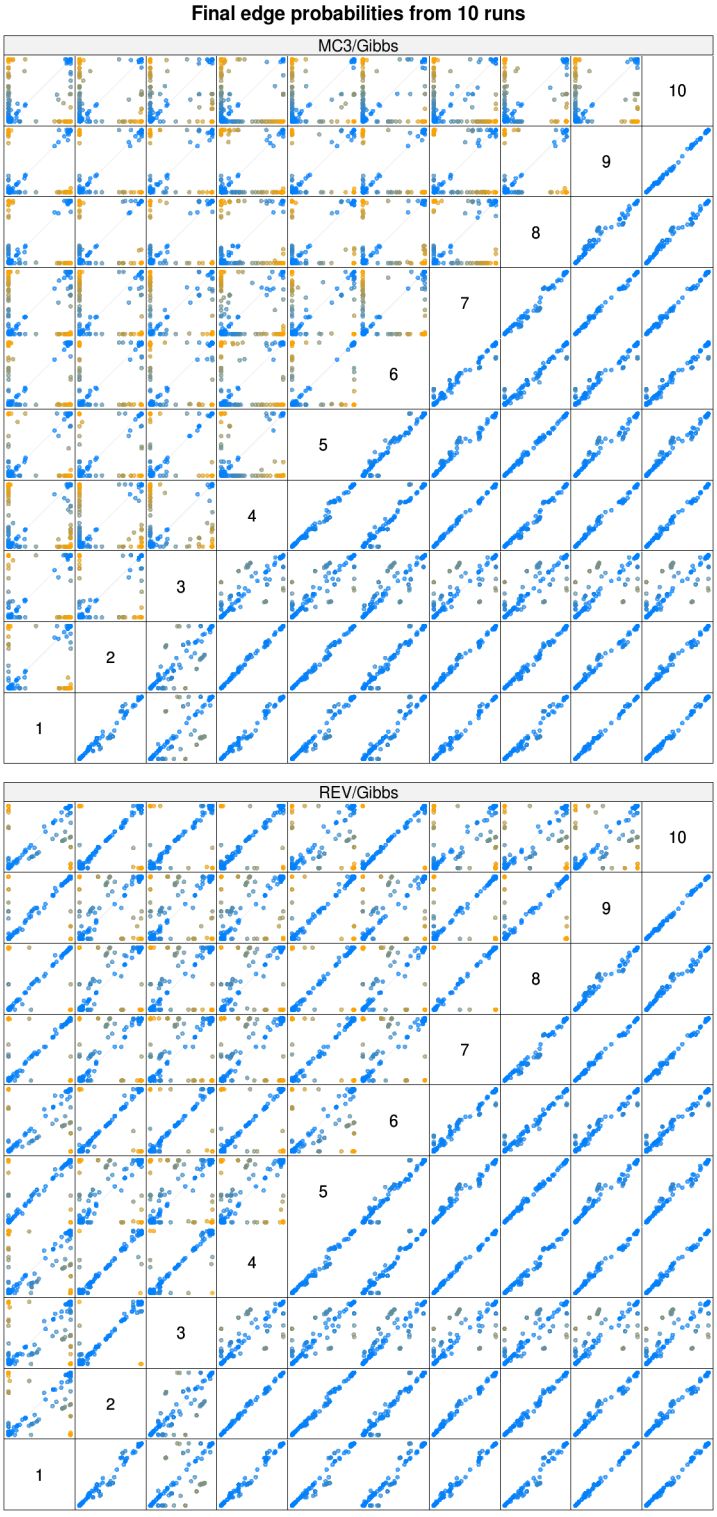


Figure B.8: Convergence diagnostics for all 10 runs of each MCMC sampler for the flow cytometry data. In each cell, the posterior edge probabilities given by two independent runs are plotted against each other. Each point represents a single edge. The lower half of both panels compares runs of the Gibbs sampler; the upper half compares runs of the MC^3 and the REV sampler respectively. When the two runs give the same estimates of the posterior edge probabilities, all of the points appear on the line $y = x$. The blue to orange colour scale represents the distance from this line, with orange points the furthest away.

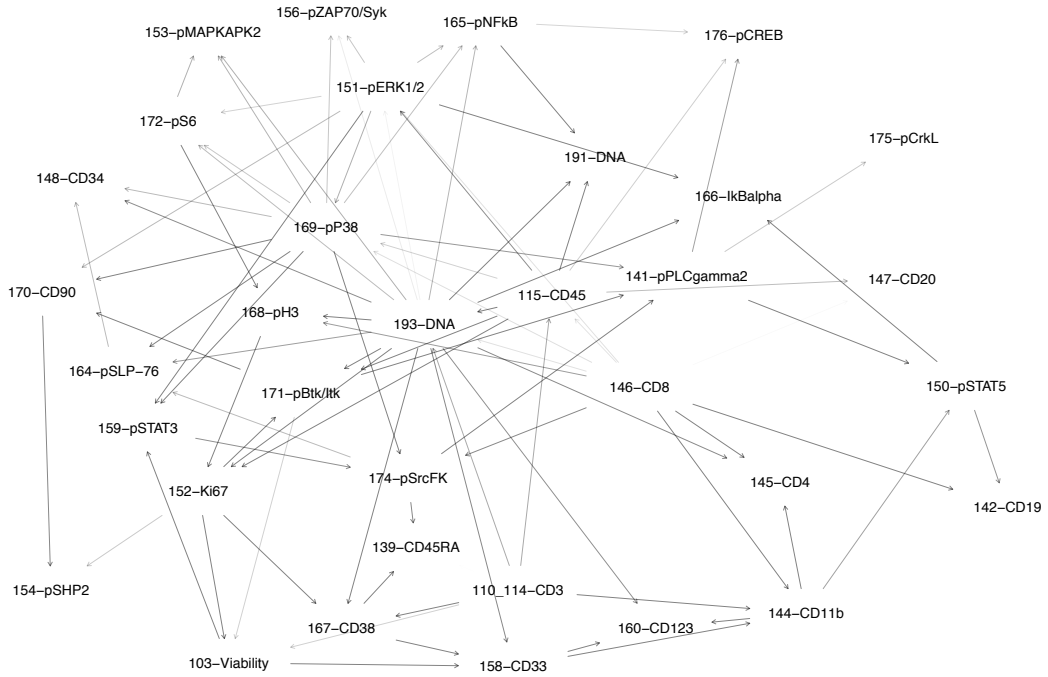


Figure B.9: The edges with posterior edge probability greater than 0.5, as given by the Gibbs sampler for the flow cytometry data. The gray-to-black scale gives an indication of the posterior edge probability.

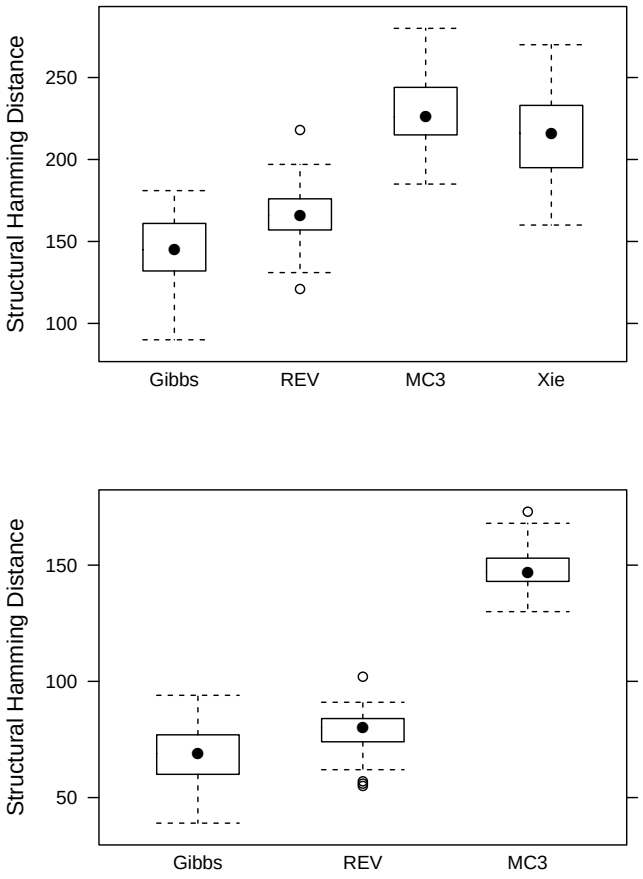


Figure B.10: Stability of estimators for the flow cytometry data across bootstrapping, as measured by SHDs. 10 bootstrap samples were drawn from the flow cytometry data. The structural Hamming distance (SHD) between the graphs given by each estimator on each bootstrap sample is shown. Smaller SHD means that the graphs are structurally more similar, and so the estimator is more stable. In (A) the edge probabilities were thresholded so that the resulting graphs had the same edges as the point estimate graph given by the Xie-Geng method. In (B) the graphs are given by thresholding at 0.5 the edge probabilities from the 3 MCMC samplers.

Appendix C

Software

The software developed for this thesis is `structmcmc`[1], a R-package (R Development Core Team, 2011) for performing Bayesian structural inference for Bayesian networks using Markov chain Monte Carlo (MCMC). The software implements both MC^3 (Section 2.6.4), and the Gibbs sampler (Chapter 4). Exact posterior distributions can also be computed for small networks ($p \leq 6$, or so), as described in Section 2.6.1.

The analyses using the REV sampler, PC-algorithm and the Xie-Geng method were conducted using versions of existing software. Marco Grzegorczyk and Dirk Husmeier provided their reference implementation of the REV sampler, which is implemented in MATLAB. We used a modified (faster) version of the implementation of the PC-algorithm (Section 2.7.2) contained in `pcalg` (Kalisch *et al.*, 2011). The implementation of Xie and Geng (2008) used is that which accompanies the original paper[2].

[1] Available at `http://go.warwick.ac.uk/rgoudie/structmcmc`
[2] Available at `http://www.mathworks.com/matlabcentral/fileexchange/20678`

Bibliography

Acid, S., de Campos, L. M., Fernández-Luna, J. M., Rodríguez, S., Rodríguez, J. M. and Salcedo, J. L. (2004) A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. *Artificial Intelligence in Medicine*, **30**, 215–232.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Anderson, L. R. and Mellor, J. M. (2008) Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, **27**, 1260–1274.

Angelopoulos, N. and Cussens, J. (2008) Bayesian learning of Bayesian networks with informative priors. *Annals of Mathematics and Artificial Intelligence*, **54**, 53–98.

Argyle, M. (2001) *The Psychology of Happiness*. Hove: Routledge.

Barbieri, M. M. and Berger, J. O. (2004) Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.

Barker, D. J., Hill, S. M. and Mukherjee, S. (2010) MC4: A Tempering Algorithm for Large-Sample Network Inference. *Lecture Notes in Computer Science*, **6282**, 431–442.

Barnett, P. A. and Gotlib, I. H. (1988) Psychosocial functioning and depression: Distinguishing among antecedents, concomitants, and consequences. *Psychological Bulletin*, **104**, 97–126.

Barsky, R. B., Juster, F. T., Kimball, M. S. and Shapiro, M. D. (1997) Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement study. *Quarterly Journal of Economics*, **112**, 537–579.

Battles, H. B. and Wiener, L. S. (2002) From adolescence through young adulthood: Psychosocial adjustment associated with long-term survival of HIV. *Journal of Adolescent Health*, **30**, 161–168.

Beinlich, I. A., Suermondt, H. J., Chavez, R. M. and Cooper, G. F. (1989) The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Second European Conference on Artificial Intelligence in Medicine*, pp. 247–256. Springer-Verlag, Berlin.

Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton University Press: Princeton, NJ.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D. and Nolan, G. P. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–696.

Benjamin, D. J., Heffetz, O., Kimball, M. S. and Rees-Jones, A. (2010) Do people seek to maximize happiness? Evidence from new surveys. NBER Working Paper 16489.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Berger, J. O. and Pericchi, L. R. (2001) Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (ed. P. Lahiri), pp. 135–207. Institute of Mathematical Statistics.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Bertrand, M. and Mullainathan, S. (2001) Do people mean what they say? Implications for subjective survey data. *American Economic Review*, **91**, 67–72.

Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**, 192–236.

Besag, J. E. (1994) Discussion of "Markov chains for exploring posterior distributions". *Annals of Statistics*, **22**, 1734–1741.

Blanchflower, D. G. and Oswald, A. J. (2004) Well-being over time in Britain and the USA. *Journal of Public Economics*, **88**, 1359–1386.

Blanchflower, D. G. and Oswald, A. J. (2008) Is well-being U-shaped over the life cycle? *Social Science and Medicine*, **66**, 1733–1749.

Bland, J. M. and Altman, D. G. (1995) Multiple significance tests: the Bonferroni method. *British Medical Journal*, **310**, 170.

Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Box, G. E. P. and Draper, N. R. (1987) *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L., eds. (2011) *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC.

Brooks, S. P. (1998) Quantitative convergence assessment for Markov chain Monte Carlo via cusums. *Statistics and Computing*, **8**, 267–274.

Brooks, S. P. and Roberts, G. O. (1998) Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, **8**, 319–335.

Brown, R. A., Lewinsohn, P. M., Seeley, J. R. and Wagner, E. F. (1996) Cigarette smoking, major depression, and other psychiatric disorders among adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, **35**, 1602–1610.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997) Model selection: an integral part of inference. *Biometrics*, **52**, 603–618.

Buhlmann, P., Kalisch, M. and Maathuis, M. H. (2010) Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika*, **97**, 261–278.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Buntine, W. (1991) Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pp. 52–60. San Mateo, CA: Morgan Kaufmann.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer, 2nd edition.

Carr, D. B., Littlefield, R. J. and Nicholson, W. L. (1987) Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, **32**, 424–436.

Centers for Disease Control and Prevention (2008) *Behavioral Risk Factor Surveil-*

lance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **158**, 419–466.

Chickering, D. M. (2002) Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, **2**, 445–498.

Chipman, H. (1996) Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, **24**, 17–36.

Claeskens, G. and Hjort, N. L. (2008) *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Clark, A. E. (2003) Unemployment as a social norm: psychological evidence from panel data. *Journal of Labor Economics*, **21**, 323–351.

Clyde, M. and George, E. I. (2004) Model uncertainty. *Statistical Science*, **19**, 81–94.

Cooper, G. F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Coppersmith, D. and Winograd, S. (1990) Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, **9**, 251–280.

Costello, E. J., Erkanli, A. and Angold, A. (2006) Is there an epidemic of child or adolescent depression? *Journal of Child Psychology and Psychiatry*, **47**, 1263–1271.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G. and Angold, A. (2003) Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, **60**, 837–844.

Cox, D. R. (1990) Role of models in statistical analysis. *Statistical Science*, **5**, 169–174.

Cox, D. R. and Wermuth, N. (1996) *Multivariate Dependencies Models, Analysis and Interpretation*. London: Chapman & Hall.

Cox, D. R. and Wermuth, N. (2004) Causality: a statistical view. *International Statistical Review / Revue Internationale de Statistique*, **72**, 285–305.

Cussens, J. (2008) Bayesian network learning by compiling to weighted MAX-SAT. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pp. 105–112. Corvallis, OR: AUAI Press.

Cussens, J. (2011) Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pp. 153–160. Corvallis, OR: AUAI Press.

Dawid, A. P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**, 1–31.

Dellaportas, P. and Stephens, D. A. (1995) Bayesian analysis of errors-in-variables regression models. *Biometrics*, **51**, 1085–1095.

Deltell, A. F. (2011) *Objective Bayes criteria for variable selection*. Ph.D. thesis, Universitat de València.

Demetrescu, C., Eppstein, D., Galil, Z. and Italiano, G. F. (2010) Dynamic Graph Algorithms. In *Algorithms and Theory of Computation Handbook: General Concepts and Techniques* (eds. M. J. Atallah and M. Blanton), pp. 9.1–9.28. Boca Raton, FL: Chapman and Hall/CRC.

Demetrescu, C. and Italiano, G. F. (2005) Trade-offs for fully dynamic transitive closure on DAGs: breaking through the $O(n^2)$ barrier. *Journal of the ACM*, **52**, 147–156.

Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Di Tella, R. and MacCulloch, R. (2005) Partisan social happiness. *Review of Economic Studies*, **72**, 367–393.

Diener, E. (1984) Subjective well-being. *Psychological Bulletin*, **95**, 542–575.

Diener, E., Diener, M. and Diener, C. (1995) Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology*, **69**, 851–864.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. and Wagner, G. G. (2011) Individual risk attitudes: measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, **9**, 522–550.

Dolan, P. and Kahneman, D. (2008) Interpretations of utility and the implications for the valuation of health. *Economic Journal*, **118**, 215–234.

Dolan, P. and White, M. P. (2007) How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, **2**, 71–85.

Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 45–97.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Physics Letters B*, **195**, 216–222.

Dudoit, S. and van der Laan, M. J. (2008) *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Easterlin, R. A. (1974) Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz* (eds. P. A. David and M. W. Reder), pp. 89–125. New York: Academic Press.

Easterlin, R. A. (2003) Explaining happiness. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11176–11183.

Eaton, D. and Murphy, K. (2007) Bayesian structure learning using dynamic programming and MCMC. In *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pp. 101–108. Corvallis, OR: AUAI Press.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407–499.

Ellis, B. and Wong, W. H. (2008) Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association*, **103**, 778–789.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fernández, C., Ley, E. and Steel, M. F. J. (2001a) Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, **100**, 381–427.

Fernández, C., Ley, E. and Steel, M. F. J. (2001b) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, **16**, 563–576.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E. and Falk, A. (2007) Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, **318**, 1305–1308.

Fong, C. and McCabe, K. (1999) Are decisions under risk malleable? *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 10927–10932.

Fowler, J. H. and Christakis, N. A. (2008) Dynamic spread of happiness in a large

social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, **337**, a2338.

Freedman, D. A. (1983) A note on screening regression equations. *The American Statistician*, **37**, 152–155.

Frey, B. S. and Stuzer, A. (2002) What can economists learn from happiness research? *Journal of Economic Literature*, **40**, 402–435.

Friedman, N. (1998) The Bayesian structural EM algorithm. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pp. 129–138. Morgan Kaufmann Publishers Inc.

Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Friedman, N. and Goldszmidt, M. (1996) Learning Bayesian networks with local structure. In *Proceedings of the Twelfth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pp. 252–260. San Francisco, CA: Morgan Kaufmann.

Friedman, N. and Koller, D. (2003) Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, **50**, 95–125.

Friedman, N., Nachman, I. and Pe'er, D. (1999) Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 206–215. San Francisco, CA: Morgan Kaufmann.

Frigessi, A., Stefano, P., Hwang, C.-R. and Sheu, S.-J. (1993) Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating

dynamics. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**, 205–219.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Gelman, A. and Shirley, K. (2011) Inference from Simulations and Monitoring Convergence. In *Handbook of Markov Chain Monte Carlo* (eds. S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Shor, B., Bafumi, J. and Park, D. (2007) Rich state, poor state, red state, blue state: what's the matter with Connecticut? *Quarterly Journal of Political Science*, **2**, 345–367.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721–741.

George, E. I. and Foster, D. P. (2000) Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731–747.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483.

Geyer, C. J. (2011) Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo* (eds. S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng). Boca Raton, FL: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall/CRC.

Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73(2)**, 123–214.

Giudici, P. and Castelo, R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, **50**, 127–158.

Goodman, E. (1999) The role of socioeconomic status gradients in explaining differences in US adolescents' health. *American Journal of Public Health*, **89**, 1522–1528.

Goudie, R. J. B., Mukherjee, S. N., De Neve, J.-E., Oswald, A. J. and Wu, S. (2011) Happiness as a driver of risk-avoiding behavior. *CESifo Working Paper Series*.

Grzegorczyk, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.

Hamming, R. W. (1950) Error detecting and error correcting codes. *The Bell System Technical Journal*, **29**, 147–160.

Harris, K. M., Halpern, C. T., Whitsel, E. A., Hussey, J., Tabor, J., Entzel, P. and Udry, J. R. (2009) The National Longitudinal Study of Adolescent Health: Research Design.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Heckerman, D., Geiger, D. and Chickering, D. M. (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

Hobert, J. P., Robert, C. P. and Goutis, C. (1997) Connectedness conditions for the convergence of the Gibbs sampler. *Statistics & Probability Letters*, **33**, 235–240.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoeting, J., Madigan, D., Raftery, A. E. and Volinsky, C. (1999) Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–401.

Holt, S., Buckley, H. and Whelan, S. (2008) The impact of exposure to domestic violence on children and young people: A review of the literature. *Child Abuse & Neglect*, **32**, 797–810.

Hsiang, T. C. (1975) A Bayesian view on ridge regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **24**, 267–268.

Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271.

Jaakkola, T., Sontag, D., Globerson, A. and Meila, M. (2010) Learning Bayesian Network Structure using LP Relaxations. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, volume 9 of *Journal of Machine Learning Research Workshop and Conference Proceedings*, pp. 358–365.

Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica*, **47**, 263–291.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-Algorithm. *Journal of Machine Learning Research*, **8**, 613–636.

Kalisch, M., Maechler, M. and Colombo, D. (2011) *pcalg: Estimation of CPDAG/-PAG and causal inference using the IDA algorithm.* R package version 1.1-4.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Kimball, M. S. (1993) Standard risk aversion. *Econometrica*, **61**, 589–611.

Kimball, M. S. and Willis, R. (2006) Utility and happiness. Working paper, University of Michigan, Michigan, MI.

King, V. and Sagert, G. (2002) A fully dynamic algorithm for maintaining the transitive closure. *Journal of Computer and System Sciences*, **65**, 150–167.

Kirkcaldy, B. and Furnham, A. (2000) Positive affectivity, psychological well-being, accident and traffic deaths, and suicide: an international comparison. *Studia Psychologia*, **42**, 97–105.

Kohn, R., Smith, M. and Chan, D. (2001) Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, **11**, 313–322.

Koivisto, M. and Sood, K. (2004) Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, **5**, 549–573.

Korb, K. B. and Nicholson, A. E. (2011) *Bayesian Artificial Intelligence.* Boca Raton, FL: Chapman & Hall/CRC Press.

Larun, L., Nordheim, L. V., Ekeland, E., Hagen, K. B. and Heian, F. (2006) Exercise in prevention and treatment of anxiety and depression among children and young people. *The Cochrane Database of Systematic Reviews*.

Laud, P. W. and Ibrahim, J. G. (1995) Predictive model selection. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 247–262.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **50**, 157–224.

Layard, R. (2005) *Happiness: Lessons from a New Science*. London: Allen Lane.

Lehmann, E. L. (1990) Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science*, **5**, 160–168.

Ley, E. and Steel, M. F. J. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, **24**, 651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008) Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410–423.

Liu, J. S. (1996) Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, **83**, 681–682.

Liu, J. S., Liang, F. and Wong, W. H. (2000) The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, **95**, 121–134.

Luttmer, E. F. P. (2005) Neighbors as negatives: relative earnings and well-being. *Quarterly Journal of Economics*, **120**, 963–1002.

MacEachern, S. N. and Berliner, L. M. (1994) Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. C. (1995) Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, **63**, 215–232.

Meek, C. (1995) Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pp. 403–410. San Francisco, CA: Morgan Kaufmann.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–473.

Merry, S. N. (2007) Prevention and early intervention for depression in young people — a practical possibility? *Current Opinion in Psychiatry*, **20**, 325–329.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Mukherjee, S. and Speed, T. P. (2008) Network inference using informative priors. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 14313–14318.

Müller, P. (1991) A generic approach to posterior integration and Gibbs sampling. Technical report, Purdue University.

Munro, I. (1971) Efficient determination of the transitive closure of a directed graph. *Information Processing Letters*, **1**, 56–58.

Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90**, 233–241.

National Institute for Health and Clinical Excellence (2005) *Depression in Children and Young People*. London: NICE.

Needham, C. J., Bradford, J. R., Bulpitt, A. J. and Westhead, D. R. (2007) A primer on learning in Bayesian networks for computational biology. *PLoS Computational Biology*, **3**, e129.

Nott, D. J. and Green, P. J. (2004) Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, **13**, 141–157.

Obot, I. S. and Anthony, J. C. (2004) Mental health problems in adolescent children of alcohol dependent parents: Epidemiologic research with a nationally representative sample. *Journal of Child & Adolescent Substance Abuse*, **13**, 83–96.

Offer, A. (2006) *The Challenge of Affluence: Self-Control and Well-being in the United States and Britain Since 1950*. Oxford: Oxford University Press.

Offer, A., Pechel, R. and Ulijaszek, S. (2010) Obesity under affluence varies by welfare regimes: the effect of fast food, insecurity, and inequality. *Economics and Human Biology*, **8**, 297–308.

O'Hagan, A. and Forster, J. (2004) *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Chichester: John Wiley and Sons.

Oswald, A. J. (1997) Happiness and economic performance. *Economic Journal*, **107**, 1815–1831.

Oswald, A. J. and Wu, S. (2010) Objective confirmation of subjective measures of human well-being: evidence from the USA. *Science*, **327**, 576–579.

Parviainen, P. and Koivisto, M. (2009) Exact structure discovery in Bayesian networks with less space. In *Proceedings of the Twenty-Fifth Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pp. 436–443. Corvallis, OR: AUAI Press.

Patel, V., Flisher, A. J., Hetrick, S. and McGorry, P. (2007) Mental health of young people: a global public-health challenge. *The Lancet*, **369**, 1302–1313.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press, 2nd edition.

Pickrell, T. M. and Ye, T. J. (2008) Traffic safety facts: seat belt use in 2008 – overall results. Research Note DOT HS 811 036, National Highway Traffic Safety Administration, Washington, DC.

Pischke, J. S. (2011) Money and happiness: evidence from the industry wage structure. NBER Working Paper 17056.

Pittau, M. G., Zelli, R. and Gelman, A. (2009) Economic disparities and life satisfaction in European regions. *Social Indicators Research*, **96**, 339–361.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radcliff, B. (2001) Politics, markets, and life satisfaction: the political economy of human happiness. *American Political Science Review*, **95**, 939–952.

Radloff, L. (1977) The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, **1**, 385–401.

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory.* Cambridge, MA: MIT Press.

Rickwood, D. J., Deane, F. P. and Wilson, C. J. (2007) When and how do young people seek professional help for mental health problems? *The Medical Journal of Australia*, **187**, S35–S39.

Ripley, B. D. (1979) Algorithm AS 137: Simulating spatial patterns: Dependent samples from a multivariate density. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28**, 109–112.

Robert, C. P. (1995) Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, **10**, 231–253.

Robert, C. P. (2007) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* New York: Springer.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods.* New York: Springer.

Roberts, G. O. (1998) Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastics Reports*, **62**, 275–283.

Roberts, G. O. and Rosenthal, J. S. (1998) Markov-chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, **26**, 5–20.

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Methodological)*, **59**, 291–317.

Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

Roberts, R. E., Lewinsohn, P. M. and Seeley, J. R. (1991) Screening for adolescent depression – a comparison of depression scales. *The Journal of the American Academy of Child and Adolescent Psychiatry*, **30**, 58–66.

Robins, J. M. and Greenland, S. (1986) The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, **123**, 392–402.

Robins, J. M., Mark, S. D. and Newey, W. K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, **48**, 479–495.

Robinson, R. (1973) Counting Labeled Acyclic Digraphs. In *New Directions in Graph Theory* (ed. F. Harary), pp. 239–273. New York: Academic Press.

Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer-Verlag.

Rubin, D. B. (2005) Causal inference using potential outcomes. *Journal of the American Statistical Association*, **100**, 322–331.

Sapienza, P., Zingales, L. and Maestripieri, D. (2009) Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 15268–15273.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Scott, J. and Berger, J. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587–2619.

Senn, S., Graf, E. and Caputo, A. (2007) Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, **26**, 5529–5544.

Silander, T., Kontkanen, P. and Myllymäki, P. (2007) On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the Twenty-Third Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pp. 360–367. AUAI Press: Corvallis, OR.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Smith, J. Q. (2010) *Bayesian Decision Analysis*. Cambridge: Cambridge University Press.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–343.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990) Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Spirtes, P. and Glymour, C. (1991) An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, **9**, 62–72.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*. Cambridge, MA: The MIT Press.

Steptoe, A. and Wardle, J. (2005) Positive affect and biological function in everyday life. *Neurobiology of Aging*, **26S**, 108–112.

Stevenson, B. and Wolfers, J. (2008) Happiness inequality in the United States. *Journal of Legal Studies*, **37**, S33–S79.

Tamada, Y., Imoto, S. and Miyano, S. (2011) Parallel algorithm for learning optimal Bayesian network structure. *Journal of Machine Learning Research*, **12**, 2437–2459.

Thaler, R. H. and Johnson, E. J. (1990) Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice. *Management Science*, **36**, 643–660.

Thapar, A., Collishaw, S., Potter, R. and Thapar, A. K. (2010) Managing and preventing depression in adolescents. *British Medical Journal*, **340**, c209.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–1728.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, **65**, 31–78.

Udry, J. R. (1998) *The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994-1996*. Los Altos, CA.

Verma, T. and Pearl, J. (1992) An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-92)*, pp. 323–330. San Mateo, CA: Morgan Kaufmann.

Verma, T. S. and Pearl, J. (1990) Equivalence and synthesis of causal models. In *Proceedings of the Proceedings of the Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pp. 220–227. New York, NY: Elsevier Science.

Wasserman, L. (2000) Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Werhli, A. V. and Husmeier, D. (2007) Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, **6**.

Wild, B., Kenwright, J. and Rastogi, S. (1985) Effect of seat belts on injuries to front and rear seat passengers. *British Medical Journal (Clinical research ed.)*, **290**, 1621–1623.

Williamson, J. (2005) *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

Xie, X. and Geng, Z. (2008) A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research*, **9**, 459–483.

Yu, B. and Mykland, P. (1998) Looking at Markov samplers through cusum path plots: a simple diagnostic idea. *Statistics and Computing*, **8**, 275–286.

Zeckhauser, R. J. and Viscusi, W. K. (1990) Risk within reason. *Science*, **248**, 559–564.

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti* (eds. P. K. Goel and A. Zellner), pp. 233–243. Amsterdam: North-Holland.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.