Sparse Graphical Models for Cancer Signalling

by

Steven Mark Hill

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Centre for Complexity Science and Department of Statistics

May 2012

Contents

Acknowledgements

First and foremost I would like to thank my supervisor Sach Mukherjee. His unwavering support, dedication and guidance over the past few years has been invaluable. He always had time for discussions about my work and these were invariably stimulating and insightful, and conducted with much enthusiasm and patience.

I am also very grateful to all the biologists who played a part in generating experimental data that made the applications presented in this thesis possible. In particular, I would like to thank Gordon Mills, Paul Spellman, Joe Gray, Yiling Lu, Bryan Hennessy, Richard Neve, Jennifer Molina and Laura Heiser.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), through the University of Warwick Complexity Science Doctoral Training Centre (DTC). I am grateful to the Complexity DTC and EPSRC for the opportunity to conduct this research. I would like to thank all staff and students for creating an enjoyable and stimulating environment within which to do research. In particular, I thank Paul Chleboun, with whom I shared an office for over three years; along with the many research-based discussions, I shall miss our bizarre conversations and light-hearted banter. I also wish to thank the Complexity administrators over the past few years; Phil Richardson, Monica de Lucena and Jen Bowskill.

Several people have helped out in proof-reading this thesis. In this regard I wish to thank Linda Hill, Chris Oates, Anas Rana and Paul Chleboun.

Finally, I would like to thank my family for their ongoing support and encouragement. In particular, my most important thanks goes to my wife, Linda. The writing of this thesis impacted her the most; I am very grateful for her constant support, affection and patience, especially over the past few months.

Declarations

The work presented here is my own, except where stated otherwise. This thesis has been composed by myself and has not been submitted for any other degree or professional qualification.

- All experimental data used within this thesis was obtained by others, as indicated in each case. The experimental protocols in Appendix A were written by those that obtained the data.
- Some of the background information in Section 2.3.2.3 is adapted from material written by myself and contributed to a book chapter [Mukherjee *et al.*, 2010].
- The work in Chapter 3 has been published in *BMC Bioinformatics* (with some revisions) [Hill *et al.*, 2012].
- The work in Chapter 4 will shortly be submitted for publication.
- The work in Chapter 5 extends and improves upon work published by my supervisor and myself [Mukherjee and Hill, 2011], and will shortly be submitted for publication.

Abstract

Protein signalling networks play a key role in cellular function, and their dysregulation is central to many diseases, including cancer. Recent advances in biochemical technology have begun to allow high-throughput, data-driven studies of signalling. In this thesis, we investigate multivariate statistical methods, rooted in sparse graphical models, aimed at probing questions in cancer signalling.

First, we propose a Bayesian variable selection method for identifying subsets of proteins that jointly influence an output of interest, such as drug response. Ancillary biological information is incorporated into inference using informative prior distributions. Prior information is selected and weighted in an automated manner using an empirical Bayes formulation. We present examples of informative pathway- and network-based priors, and illustrate the proposed method on both synthetic and drug response data.

Second, we use dynamic Bayesian networks to perform structure learning of context-specific signalling network topology from proteomic time-course data. We exploit a connection between variable selection and network structure learning to efficiently carry out exact inference. Existing biology is incorporated using informative network priors, weighted automatically by an empirical Bayes approach. The overall approach is computationally efficient and essentially free of user-set parameters. We show results from an empirical investigation, comparing the approach to several existing methods, and from an application to breast cancer cell line data. Hypotheses are generated regarding novel signalling links, some of which are validated by independent experiments.

Third, we describe a network-based clustering approach for the discovery of cancer subtypes that differ in terms of subtype-specific signalling network structure. Model-based clustering is combined with penalised likelihood estimation of undirected graphical models to allow simultaneous learning of cluster assignments and cluster-specific network structure. Results are shown from an empirical investigation comparing several penalisation regimes, and an application to breast cancer proteomic data.

Chapter 1

Introduction

In recent years significant advances have been made in biochemical technology and techniques, resulting in an increasing availability of high-throughput data in molecular and cell biology. Examples of such technologies include DNA microarrays and high-throughput sequencing for transcriptomic and genomic data, and protein microarrays, mass spectrometry and flow cytometry for proteomic data. The availability of such data has motivated and driven a shift towards a 'systems' approach to biology instead of a more traditional reductionist approach. The reductionist approach focusses on the identification of individual molecular components and the study of their functions. However, cell behaviour arises from and is regulated by components such as genes and proteins acting in concert. Obtaining an understanding of this complex interplay and resulting functionality characterises the systems approach [Ideker *et al.*, 2001; Kitano, 2002; Ideker and Lauffenburger, 2003]. The approach is multi-disciplinary, combining biology with other fields such as mathematics, statistics, physics, engineering and computer science.

Biological networks are widely used to represent and visualise interplay between molecular components and are integral to the systems approach. Networks (or graphs; we use both terms interchangeably throughout this thesis) consist of a set of nodes, representing molecular components such as genes or proteins, and a set of (directed or undirected) edges which represent relationships or interplay between the components. Examples include gene regulatory networks [Hecker *et al.*, 2009] and protein signalling networks [Yarden and Sliwkowski, 2001; Sachs *et al.*, 2005]; the latter is the focus in this thesis. Discovery of the structure of these networks (i.e. the location of the edges) and the nature of the interactions represented by the edges (i.e. mechanisms and dynamics), in specific contexts, is an important goal in molecular biology. For example, investigating biological networks in disease states, such as cancer, can help in the discovery of processes that cause and sustain the disease, and can guide therapeutics [Pe'er and Hacohen, 2011].

Proteins play a key role in most cellular functions. Signalling is an important cellular process, which ultimately leads to cellular responses such as cell proliferation and apoptosis (programmed cell death). The nodes in protein signalling networks represent signalling proteins and edges represent interactions between proteins that result in the signal being transduced through the cell. Protein signalling pathways and networks have complex structures with combinatorial nonlinear interactions, cross-talk between pathways and feedback mechanisms [Citri and Yarden, 2006; Rubbi *et al.*, 2011]. Thus, in order to obtain a proper understanding of context-specific signalling processes, it is necessary to take a systems network approach rather than studying components in isolation. The development of high-throughput proteomics means that data is now available that enables many components to be investigated simultaneously [see e.g. Sachs *et al.*, 2005]. Further details of protein synthesis, protein signalling and signalling networks are provided in Section 2.1 and details of several protein assays are provided in Section 2.2.

Cancer is a prevalent disease; it was the cause of 28% of all deaths in the UK in 2009 [Cancer Research UK, 2012]. It is a genetic disease, with multiple DNA mutations resulting in dysregulation of cellular processes such as proliferation and apoptosis, ultimately leading to the transformation of a normal cell into a cancerous cell. Mutated, cancer-causing genes often code for signalling proteins, and so it is aberrant signalling that often causes the dysregulation in cellular processes. Indeed, aberrant signalling is heavily implicated in the six functional capabilities that are acquired during tumour development and are regarded as 'hallmarks' of cancer [Hanahan and Weinberg, 2000]; these are, self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Recently, two further potential hallmarks were added to the original six [Hanahan and Weinberg, 2011]; deregulating cellular energetics and avoiding immune destruction. Further details regarding protein signalling and cancer are provided in Section 2.1.3.

The key role that protein signalling plays in cancer formation and progression has implications for cancer therapy. Chemotherapy and radiotherapy are traditional cancer therapies that work by causing damage to cancer cells that inhibits cell division and induces apoptosis. However, the side-effects of these treatments can be severe due to healthy cells also being affected (for example, chemotherapy drugs target any rapidly dividing cells, not just tumour cells). Targeted cancer therapies are drugs that inhibit tumour growth by interfering with specific molecules that are involved in cell processes that are dysregulated in cancer, such as proliferation, apoptosis and angiogenesis. Signalling proteins are a common target for these therapies (specific examples are provided in Section 2.1.4) and so knowledge of signalling networks and mechanisms is important for identifying novel therapeutic targets. Since these therapies focus on specific molecules that play a role in carcinogenesis, they may be more effective than the more traditional approaches and may result in fewer side effects.

Classification of cancers is traditionally based on factors such as histopathological type, tumour grade and tumour stage. As the molecular biology of cancer becomes better understood, classifications based on molecular characteristics of tumours are being developed [Perou *et al.*, 2000; Sørlie *et al.*, 2001; TCGA-Network, 2011]. For example, Perou *et al.* [2000] used gene expression data to classify breast cancer tumours into four subtypes (basal, luminal, ErbB2 positive and normal), which were later refined into five subtypes that can be used as prognostic markers for overall and relapse-free survival [Sørlie *et al.*, 2001]. Cancer subtypes reflect the remarkable levels of molecular heterogeneity in cancer. This heterogeneity results in diverse responses to therapies across cancer patients; a drug that works for one patient may not work for another. This has driven the surge in interest in personalised therapies that tailor treatment regimes to the molecular characteristics of a patient's tumour [Majewski and Bernards, 2011]. The identification of molecular biomarkers or signatures that are predictive of response to a targeted therapy is a key goal for personalised medicine. Signalling proteins are potential sources of such predictive biomarkers due to their important role in cellular processes and the fact that many therapies target them. An example of such a biomarker is the HER2 receptor protein in breast cancer; see Section 2.1.4 for further details.

Along with the developments in experimental technology discussed above, continual improvements in computing power have enabled computationally-intensive simulations and statistical inference using increasingly complex mathematical and statistical models. A model is a simplification of the underlying system that captures features and relationships that are of interest and explain the observed data. Models vary in their levels of abstractness. Ordinary differential equations have been used to directly model biochemical reaction kinetics in signalling pathways [Schoeberl *et al.*, 2002; Chen *et al.*, 2009; Wang *et al.*, 2009a]. Such models provide a reasonably realistic representation of the underlying mechanisms and can be used for simulation and predictions, leading to novel insights and generation of hypotheses for further consideration. For example, perturbing a component in the model could aid in prediction of response to a certain drug treatment. However, these models are complex with many parameters and are generally not solvable analytically, resulting in their restriction to relatively small systems for both computational and statistical reasons. Also, the network structures are usually assumed to be known, with modelling focusing on the actual mechanisms involved in interactions. In contrast, more abstract, high level models, for example Boolean models or continuous, linear models, are often analytically tractable with fewer parameters, allowing larger systems to be modelled and network structures themselves to be inferred. While these models may not be as realistic or as predictively accurate, hypotheses can still be formed regarding network structure; this is the approach we take in this thesis, using continuous linear models. Once the structure is known, more detailed modelling can be used to investigate the underlying mechanisms between certain components.

Statistical models take variability into account via probability distributions. Variation in the data can either be systematic, due to an underlying signal in the system, or can be stochastic, due to measurement noise, variability within the biochemical cellular processes themselves [Ray, 2010], or unmodelled mechanisms and components in the system. Unlike deterministic models such as ordinary differential equations, statistical models take account of stochastic variation in the data, which can help to elucidate the true underlying interactions and mechanisms. Statistical inference can be used to select appropriate models, often employing a trade-off between model fit and model complexity that helps to avoid overfitting of the model to the data. This is particularly important for the majority of high-throughput data, which is of high dimensionality, but small sample size (the 'large p, small n' paradigm). The growing interest in biological networks has resulted in much interest in multivariate statistical methods, and in particular, graphical models.

Graphical models [Pearl, 1988; Lauritzen, 1996] are a class of statistical models in which a graph is used to represent probabilistic relationships between multiple interacting components. The nodes correspond to random variables for entities of interest (e.g. protein activation levels) and the edge structure describes statistical dependencies between the variables. Thus, the graph structure provides information about the joint probability distribution over all variables. Given data for the variables under study, statistical inference can be performed to determine the graph structure that best explains dependencies contained in the data. This is often a challenging problem for several reasons, including paucity of data, noisy data, combinatorial and nonlinear interactions, vast numbers of possible graph structures, and unobserved variables that impact interpretation of results. Graphical models will be introduced fully in Section 2.3.4 and background information on structure learning of graphical models will be given in Sections 2.3.5 and 2.3.6. Graph structure learning is also known as network inference and we use both terms interchangeably throughout the thesis.

This thesis concerns the learning of sparse graphical model structure from high-throughput proteomic data, with a focus on protein signalling networks in cancer. Graphical model structure learning has received much attention in the past decade, but most applications have focussed on gene regulatory networks, due primarily to the abundance of DNA microarray data. Since protein signalling plays a major regulatory role in the functionality of cells, with aberrant signalling implicated in cancer, and since signalling proteins are potential cancer therapeutic targets, it is also important to study molecular networks at the signalling level. We focus on structure learning of networks that are sparse; that is, networks with a small number of edges. This emphasis on sparse networks is important for several reasons. First, sparse network models have fewer parameters and so help to avoid statistical overfitting. Second, sparse networks are easier to interpret and thereby facilitate hypothesis generation. Third, the true underlying network structures are often sparse. This is thought to be the case for protein signalling networks due to the highly specific enzymatic reactions involved in signalling interactions [Beard and Qian, 2008].

Graphical model structure learning is used in this thesis to probe important questions in cancer signalling that have potential implications for cancer therapy. In Chapter 3 we consider the problem of identifying subsets of signalling proteins (i.e. protein signatures) that are predictive of drug response. This is a variable selection problem, but can also be framed as a graphical model structure learning problem (see Section 2.3.5.2). Chapters 4 and 5 both consider structure learning of protein signalling networks in cancer; Chapter 4 considers structure learning of signalling networks for individual cancer cell lines and Chapter 5 considers cancer subtype discovery and simultaneous structure learning of subtype-specific networks. The multivariate statistical and computational approaches we employ to address these problems incorporate a range of methods and models (in addition to directed and undirected graphical models). These include: Bayesian inference, frequentist (penalised) maximum likelihood inference, empirical Bayes approaches, linear regression models, Bayesian model selection and model averaging, variable selection, expectation-maximisation, and clustering. Background information for all these methods and models, and background on graphical model structure learning and structure learning in general will be given in Section 2.3.

Chapter 3 describes a Bayesian variable selection approach for identifying subsets of signalling proteins that jointly influence a biological response of interest.

Bayesian variable selection methods have been widely used for such inference problems [Lee *et al.*, 2003; Jensen *et al.*, 2007; Mukherjee *et al.*, 2009; Ai-Jun and Xin-Yuan, 2010; Li and Zhang, 2010; Yeung *et al.*, 2011]. An ever increasing amount of ancillary biological information is available, such as signalling pathway and network structures, that could be incorporated into inference to improve results. Bayesian approaches allow for such incorporation via biologically informative prior distributions, yet it is not always clear how information should be selected or weighted relative to primary data. In the proposed approach, empirical Bayes is used to automatically select and weight prior information in an objective manner. We develop informative pathway- and network-based priors and demonstrate on synthetic response data that the approach can aid prior elicitation and guard against misspecification of priors. In addition, the continuous linear model employed, together with sparsity constraints, results in a fast, exact procedure with very few user-set parameters, yet capable of capturing interplay between molecular players. An application of the approach is made to cancer drug response data.

Chapter 4 describes a graph structure learning approach, using directed graphical models known as dynamic Bayesian networks (DBNs), and applies the approach to learn a signalling network for an individual breast cancer cell line from proteomic time series data. DBNs have previously been used to infer gene regulatory networks from time series data [Husmeier, 2003; Perrin *et al.*, 2003; Kim *et al.*, 2003; Zou and Conzen, 2005; Grzegorczyk *et al.*, 2008; Grzegorczyk and Husmeier, 2011a; Rau *et al.*, 2010; Robinson and Hartemink, 2010; Li *et al.*, 2011]. As discussed above, the investigation of signalling networks in specific contexts, such as cancer, is an important problem. Yet, multivariate data-driven characterisations of context-specific signalling networks remains a challenging and open problem. Indeed, such multivariate data has only recently been made available as a result of advances in experimental proteomics [see e.g. Sachs *et al.*, 2005; Sheehan *et al.*, 2005]. The approach carries out inference within an exact framework and incorporates existing biology using an informative network prior, weighted automatically and objectively, relative to primary data, by empirical Bayes. This again results in a computationally efficient, exact method, with very few user-set parameters. Results on simulated data place the approach favourably relative to other existing structure learning approaches, and the network inferred from breast cancer proteomic data is used to generate hypotheses regarding novel context-specific signalling links, which are validated in independent experiments.

Chapter 5 describes an approach that combines network structure learning with clustering, allowing for the discovery of cancer subtypes that differ in terms of subtype-specific signalling network structure. Clustering approaches are often unable to, or often make assumptions that preclude, the modelling of cluster-specific network structure. Hence, if differences exist between clusters at the network level, these approaches may not be able to recover the correct clustering (or network structures). The proposed network-based clustering approach exploits recent results in penalised likelihood estimation of undirected graphical models [Friedman *et al.*, 2008] and combines this with model-based clustering [McLachlan and Basford, 1987; Fraley and Raftery, 1998; McLachlan and Peel, 2000; Fraley and Raftery, 2002] to permit simultaneous estimation of cluster assignments and cluster-specific networks. We perform an empirical investigation comparing several specific penalisation regimes, presenting results on both simulated data and high-throughput breast cancer protein signalling data. Our findings allow for some general recommendations to be made regarding penalisation regime and also demonstrate that network-based clustering can provide improved performance relative to clustering methods that disregard cluster-specific network structures.

The novel contributions of the thesis are as follows:

- Chapter 3:
  - We describe an empirical Bayes approach to automatically select and weight biologically informative priors in Bayesian variable selection, and develop examples of such priors based on pathway and network structure information.
  - We present an empirical investigation of variable selection with informative priors, selected and weighted by empirical Bayes. Results are shown on both synthetic and drug response data, and comparisons are made to alternative methods.
- Chapter 4:
  - We perform exact inference of DBN structure by exploiting a connection between DBN structure learning and variable selection. In particular, posterior edge scores are calculated using exact Bayesian model averaging. Empirical Bayes weighting of prior information is also carried out within the exact framework, along with relevant diagnostics.
  - We present an empirical investigation of the described structure learning approach. Results are shown on both simulated data and data from a synthetically constructed network in yeast [Cantone *et al.*, 2009]. The

    utility of prior information is assessed and comparisons are made to several other existing structure learning approaches for time series data.

  - We apply the approach to proteomic time series data from a breast cancer cell line. This contributes to the small number of recent studies in the literature concerning structure learning of cancer protein signalling networks [Guha *et al.*, 2008; Mukherjee and Speed, 2008; Ciaccio *et al.*, 2010; Bender *et al.*, 2010]. We generate testable hypotheses regarding novel signalling links, which are then subsequently validated in independent experiments (carried out by Gordon Mills' lab at MD Anderson Cancer Center, Houston). To the best of our knowledge, this is the first application of DBN structure learning to protein signalling time series data, and the first application of structure learning to protein signalling networks in cancer that results in independently validated hypotheses.

- Chapter 5:

  - We describe a network-based clustering approach that permits simultaneous estimation of cluster assignments and cluster-specific networks. The approach is similar to the one proposed by Zhou *et al.* [2009], but while they focus on clustering in combination with variable selection, our focus is on simultaneous clustering and network structure learning. In addition, we propose a more general form for the penalisation term than that in Zhou *et al.* [2009].
  - We present an empirical investigation comparing several specific penalisation regimes, using both simulated data and high-throughput breast cancer protein signalling data. General recommendations regarding penalisation regime are proposed based on the results.
  - The application to breast cancer data leads to the biologically interesting finding that heterogeneity between cancer subtypes at a transcriptional level is also reflected at the level of signalling network structure.

The thesis is organised as follows:
In Chapter 2 background information relevant to the thesis is provided. This covers biology and experimental approaches relevant to the applications, and methodological background. Chapter 3 presents a Bayesian variable selection method incorporating pathway- and network-based prior information, selected and weighted in an automatic, objective manner using an empirical Bayes formulation. The method is applied to discover subsets of signalling proteins that influence drug response.

In Chapter 4 context-specific structure learning of a signalling network in a breast cancer cell line is performed using a dynamic Bayesian network approach. The results lead to the generation of hypotheses regarding novel signalling links and their subsequent independent validation. Chapter 5 presents a study where clustering and network structure learning is combined, allowing for the simultaneous discovery of cancer subtypes and subtype-specific signalling network structure. Chapters 3-5 each end with a discussion of methods, results and directions for future work that are specific to the individual Chapter. Chapter 6 presents a general discussion containing points relevant to the thesis as a whole.

Chapter 2

Background

In this Chapter, we describe the background material relevant to this thesis. The focus of this thesis is the inference of sparse graphical model structure, where the graph structure represents protein signalling networks in cancer cells, or dependencies between signalling components and a response of interest. Section 2.1 outlines the biological process of protein signalling and its importance in cancer biology. Section 2.2 describes the various experimental techniques and technologies available for assaying signalling proteins, and Section 2.3 provides background on the statistical models and methods employed in the subsequent Chapters.

2.1 Biological background

2.1.1 Protein synthesis

Cells are the basic building blocks of living organisms. Prokaryotes, such as bacteria, are organisms that do not have cell nuclei or other complex cell structures and are mostly unicellular. In contrast, eukaryotes contain cell nuclei and other membrane enclosed entities such as mitochondria. All multicellular animals and plants, and some unicellular organisms such as yeast are eukaryotes. Proteins are integral to the majority of processes that occur within cells, and thus have a key role in all living organisms. Examples of protein function in cellular processes include, acting as enzymes for metabolic reactions, playing a part in forming and maintaining cell structure, aiding cell motility, and communicating inter- and intra-cellular signals.

Proteins are compounds consisting of linear polymer chains, which are built from a sequence of amino acids. These sequences define the structure and function of each type of protein in the cell, and are determined by genes. Genes are functional segments of DNA that code for proteins. Deoxyribonucleic acid (DNA) is

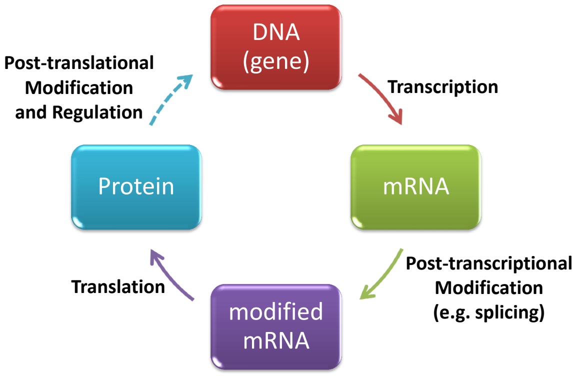

DNA (gene)
Transcription
mRNA
Post-transcriptional Modification (e.g. splicing)
modified mRNA
Translation
Protein
Post-translational Modification and Regulation

Figure 2.1: **Protein synthesis and genetic regulation.** Proteins are synthesised from DNA via production of mRNA. Synthesised proteins can directly or indirectly regulate gene expression. See text for details.

contained in the cell nucleus (in eukaryotes) and is a double-helix consisting of two long polymers composed of nucleotides adenine (A), guanine (G), cytosine (C) and thymine (T). The sequence of nucleotides are genetic instructions for development and functioning of the organism, carried out through production of proteins. All cells contain the same DNA, but the amount of each protein (protein expression) varies. This enables formation of cells of differing types and control of the processes within these cells.

The synthesis of proteins from DNA occurs via the production of ribonucleic acid (RNA), as stated by the central dogma of molecular biology. The main steps involved are transcription, splicing and translation as shown in Figure 2.1. In transcription, a protein called RNA polymerase copies the information contained in a gene into messenger RNA (mRNA), using the DNA sequence as a blueprint. Once transcribed the mRNA is usually modified before it can be translated into protein. An example of such a modification is RNA splicing, which removes sections of the RNA. The removed sections are known as introns and the remaining sections, known as exons, are spliced together. In translation, the spliced mRNA is synthesised into a protein by components of the cell known as ribosomes (usually found in the cytoplasm). Ribosomes read the mRNA sequence in triplets of nucleotides called codons, which specify amino acids to add to a growing peptide chain. Once translation is complete the protein folds into a complex three-dimensional structure.

The expression of genes (i.e. production of functional gene products) is highly regulated to control the amounts of protein produced, with regulation occurring in all steps of the process described above. One of the key ways gene expression is regulated is at the transcriptional level by transcription factors. Transcription factors are proteins that attach to the promoter region of a gene and activate or inhibit expression by promoting or blocking the recruitment of RNA polymerase. This gives rise to the notion of gene regulatory networks (GRNs). If a protein produced by an activated gene is a transcription factor, it can enter the nucleus and activate or inhibit expression of other genes. Thus, there is a regulatory relationship between the gene that codes for the transcription factor and the genes that the transcription factor binds to. This can be described by a network where nodes are genes and edges represent the regulatory relationships between them. However, regulatory relationships are often not direct; for example, a transcription factor may need to undergo a post-translational modification before being able to bind to a gene, or a gene may code for a protein that indirectly influences gene expression via its role in a signalling pathway or via it forming a complex with a transcription factor. GRNs are used to represent these indirect relationships also.

There is a substantial body of literature concerning structure learning of GRNs from gene expression data [see e.g. Bansal *et al.*, 2007; Hecker *et al.*, 2009, and references therein]. Due to widespread availability of high-throughput technology capable of performing genome-wide measurements, such as DNA microarrays and, more recently, RNA-sequencing, these network inference methods typically use data at the mRNA level, rather than the protein level. While mRNA levels may be correlated with total protein levels (although this has been shown to not always be the case [Greenbaum *et al.*, 2003]), post-translational modifications (PTMs) of proteins cannot be detected at the genomic level. PTMs are chemical modifications of the translated protein that can change the function of a protein or regulate its activity. Indeed, the number of distinct proteins in a cell is substantially more numerous than the number of genes. This complexity of the proteome is partly contributed to PTMs, with RNA splicing being another key process by which distinct proteins are synthesised from a single gene. PTMs are known to play a key role in cellular function and are also implicated in development and persistence of disease, as we shall discuss below. Thus, directly measuring and studying proteins and their PTMs can provide important insights into cellular processes and their dysregulation in disease, that would be hidden in genomic level studies. Examples of PTMs include those that attach certain chemical groups to specific amino acid residues, such as phosphates or acetates, or those that make structural changes to the protein, for example cleavage by a protease. Below we consider phosphorylation (the addition of phosphate groups), one of the most common PTMs [Khoury *et al.*, 2011], and the role it plays in intracellular signal transduction, in particular in the context of cancer.

2.1.2 Protein signalling

Protein signalling (also referred to as signal transduction) is the mechanism by which proteins on the cell membrane receive external signals and transduce them into the cell, where proteins communicate to process the signal and often transmit it to the cell nucleus to ultimately achieve a cellular response such as cell proliferation. In this way, cells are able to communicate with each other and adapt to their external environment. We now outline in more detail the mechanisms involved, using the epidermal growth factor (EGF) signalling pathway (also known as the ErbB pathway) as an illustration. The EGF signalling pathway is known to often be dysregulated in cancer cells; we discuss how dysregulation of signalling is central to carcinogenesis below and here focus on signalling mechanisms in a normal cell. A schematic of the EGF pathway is shown in Figure 2.2.

Central to protein signalling pathways are protein kinases. Protein kinases are enzymes that remove a high-energy phosphate group from ATP and transfer it to specific amino acids in other proteins, a process called phosphorylation (see Figure 2.3). The phosphate group is usually attached to one of three amino acids; serine, threonine or tyrosine. Serine/threonine kinases act on both serine and threonine residues and are the most common type of kinase, while tyrosine kinases act on tyrosine residues. Phosphorylation of a protein leads to a conformational change in its structure, causing it to become 'activated'. It is a reversible event; proteins can be 'deactivated' through removal of phosphate groups by enzymes called phosphatases. This mechanism allows signals to be transmitted through the cell with high levels of specificity and precision.

Signal transduction begins with the activation of a protein on the cell membrane, known as a receptor. In the majority of cases, activation occurs through binding of an extracellular ligand to the receptor. The ErbB receptors are a family of four receptor tyrosine kinases (RTKs), a specific type of transmembrane receptor. ErbB1 (EGFR; epidermal growth factor receptor), the first of the four ErbB family RTKs to be discovered, can be activated by several ligands, including EGF (see Figure 2.2). Growth factors are substances that stimulate cellular proliferation. ErbB1 receptors consist of an extracellular ligand-binding domain, a transmembrane domain and an intracellular protein tyrosine kinase domain. Binding of a ligand to ErbB1 enables formation of a homodimer with other bound ErbB1 receptors. This brings pairs of receptors into close proximity with one another, allowing the intracellular kinase domain of each receptor to phosphorylate the other receptor. This is known as autophosphorylation or transphosphorylation and leads to a conformational change that results in 'activation' of the receptors, i.e. the signal has now been

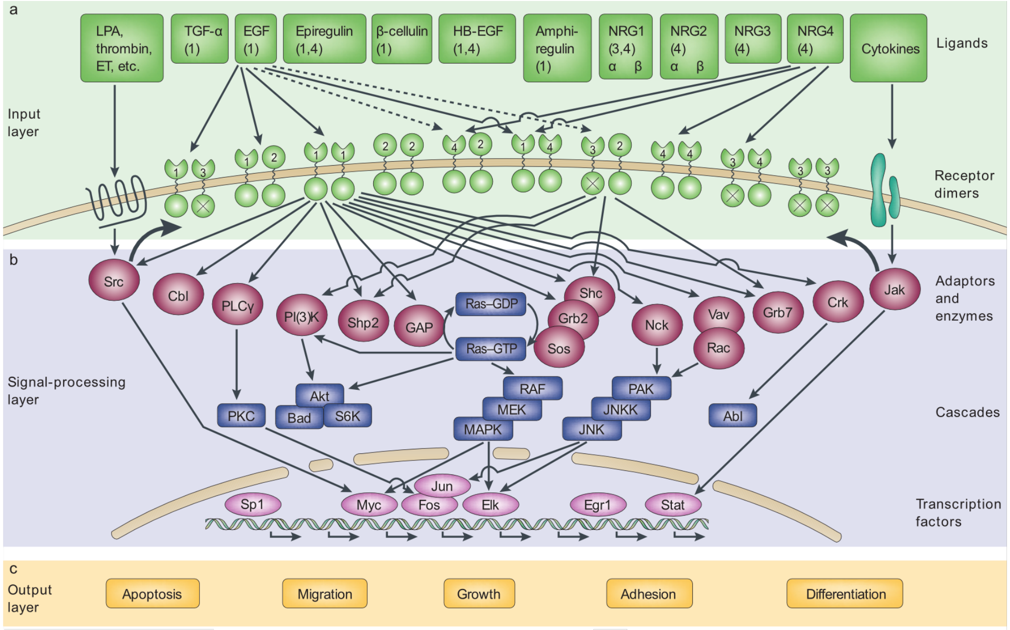


Figure 2.2: **The EGF (ErbB) signalling pathway.** (a) Ligands bind to cell surface receptors, leading to their activation. (b) Activated receptors transmit the signal to adaptor proteins which form a bridge between the receptor and various signalling pathways. These pathways consist of kinases that transmit the signal down to the cell nucleus via cascades of phosphorylations, resulting in activation of a transcription factor and therefore regulation of gene expression. (c) This results in changes in various cellular functions, such as growth and differentiation. Reprinted by permission from Macmillan Publishers Ltd: Molecular Cell Biology [Yarden and Sliwkowski, 2001], copyright (2001).

transduced to within the cell. As well as forming homodimers with other ErbB1 receptors, heterodimers with other members of the ErbB receptor family can also be formed. Indeed, ErbB2 (also known as HER2) does not have an extracellular ligand binding domain and is activated by forming heterodimers with other family members and homodimers with other ErbB2 receptors. ErbB3 has no intracellular kinase domain and so must also form heterodimers. Finally, ErbB4 is more similar to ErbB1 having both extracellular ligand binding and intracellular kinase domains.

In some cases, activation of a receptor by a ligand has a direct causal effect on the behaviour of the cell, or an indirect effect via a simple signal transduction mechanism. For example, activation of the Notch receptor causes its intracellular domain to be released from the cell membrane. The Notch fragment can then migrate to the cell nucleus where it acts as a TF, causing a change in gene expression and thus a change in cell behaviour due to different proteins being produced. However, in many cases, the signal transduction mechanism is much more complex, involving protein-protein interactions (physical binding of two or more proteins) and cascades of protein phosphorylations; this is the case for EGF signalling. Phosphorylation of RTKs creates binding sites for proteins with a SH2 [1] domain. An example of such a protein is Grb2, which also has a SH3 domain. Grb2 then binds, via its SH3 domain, to a protein called Sos, bringing Sos to the cell membrane where protein Ras is located. Proteins such as Grb2, which function as bridges between two other proteins (here, the RTK and Sos), are called adaptor proteins. Sos can then interact with and induce activation of Ras, a key intracellular signal transducer that transmits the signal to several pathways, including the MAPK/Erk and Akt pathways. These pathways involve signalling cascades of kinases. For example, in the MAPK/Erk pathway (see Figure 2.2), activation of Ras brings the serine/threonine kinase Raf to the cell membrane and binds to it, leading to its activation. The activated kinase Raf then phosphorylates the kinase MEK, which in turn phosphorylates the kinase MAPK (also known as Erk). In these kinase cascades, phosphorylation of a kinase leads to conformational changes and its functional activation, enabling it to phosphorylate and activate the next kinase in the pathway. Phosphorylated MAPK can then translocate to the nucleus, where it phosphorylates TFs, regulating their activity. In this way, the signal is passed down through the cell and into the cell nucleus, resulting in changes in gene expression and therefore regulation of cellular processes such as cell growth, differentiation and apoptosis.

It is important to note that, in addition to regulating gene expression, protein

[1] Protein (and protein domain) names are usually referred to using abbreviations. For example, HER2 stands for Human Epidermal growth factor Receptor 2, and SH2 stands for Src Homology 2. We use the abbreviated forms in this thesis.

signalling can also affect cell behaviour in other, more direct ways. For example, the intracellular serine/threonine kinase Akt phosphorylates the protein Bad, which plays a role in promoting apoptosis; phosphorylation inhibits its pro-apoptotic effect.

Further details of protein signalling in general can be found in Alberts *et al.* [2008, Chapter 15] and Weinberg [2006, Chapters 5 & 6], and of signalling in the ErbB network in Yarden and Sliwkowski [2001].

2.1.3 Protein signalling and cancer

Cancer is a genetic disease; many carcinogens (cancer-causing agents) cause alterations to DNA sequences. These alterations range from small single point mutations to larger aberrations such as deletions, insertions or translocations. Most mutations occur in somatic (body) cells and so, unlike germline mutations, are not passed on to offspring which can lead to a hereditary predisposition to developing cancer. However, somatic mutations can be passed on to new cells in the process of cell division. Cells have defense mechanisms such as DNA repair to correct mutations before cell division occurs, or if the DNA is unable to be repaired, apoptosis destroys the cell to prevent persistence of mutations in new cells. Evasion of these defense mechanisms and alterations in cell proliferation and differentiation processes can lead to carcinogenesis.

There are two main types of genes in which mutations can lead to development of cancer; oncogenes and tumour suppressor genes. Oncogenes are mutated genes that cause over-expression or an increase in activity of the genes protein product, which leads to enhanced cell proliferation and cell survival. The normal, unmutated version of an oncogene is referred to as a proto-oncogene. Tumour suppressor genes are inactivated by mutation and code for proteins that inhibit cell proliferation or promote apoptosis. Thus, inactivation of these genes again results in uncontrolled cell proliferation.

As described above, protein signalling pathways play a crucial role in most cellular functions, including cell proliferation, differentiation and apoptosis. Dysregulation of signalling pathways is implicated in most, if not all, of the alterations in cell physiology that lead to carcinogenesis, described by Hanahan and Weinberg [2000, 2011] and outlined in the Introduction. Oncogenes often code for proteins involved in mitogenic signalling pathways such as the EGF pathway; that is, signalling pathways that result in cell proliferation.

An example of how a genetic mutation results in aberrant signalling and therefore development of cancer is given by the ErbB2 oncogene, and its protein product, the ErbB2 (HER2) receptor in the EGF pathway in breast carcinomas.

The ErbB2 gene was found to be amplified (i.e. replication of the section of DNA containing the gene results in extra copies of the gene) in approximately a third of breast tumours [Slamon *et al.*, 1987]. This amplification leads to overexpression of the RTK HER2. As described above, HER2 does not require binding of a ligand to allow dimerisation; it is in a constitutively active conformation and so is free to bind with other HER2 receptors or other RTKs in the EGFR family. Heterodimers containing HER2 are formed preferentially and generate stronger intracellular signals than other combinations [Yarden and Sliwkowski, 2001]. Therefore, overexpression of HER2 leads to constitutive stimulation of Akt and MAPK pathways, conferring a growth advantage to the tumour cells (due to enhanced cell survival and cell proliferation). Breast cancers with an amplified ErbB2 gene are associated with short survival times and therefore poor prognosis [Slamon *et al.*, 1987; Sørlie *et al.*, 2001]. However, prognosis has been improved due to targeted therapies, which we discuss further below. Further details of the mechanisms involved in HER2 overexpression in breast cancer can be found in Emde *et al.* [2011].

In addition to mutations that dysregulate a signalling pathway through overexpression of a receptor, mutations can also affect other intracellular signalling transducers. An example of this type of mutation is given by the Abl oncogene in CML (chronic myelogenous leukemia). Abl codes for a nuclear tyrosine kinase that plays a role in apoptosis. The mutated version of Abl arises due to chromosomal translocation of the Abl gene to within the Bcr gene, resulting in a new fusion gene, Bcr-Abl. This fusion gene codes for a fusion protein, located in the cytoplasm rather than the nucleus, with constitutive activation of the Abl tyrosine kinase. The result is aberrant kinase signalling that affects the cell cycle, leading to uncontrolled proliferation of white blood cells (i.e. leukemia). Targeted therapies have also been developed for this mutation.

Further details regarding oncogenic signalling pathways can be found in Weinberg [2006, Chapters 5 & 6].

2.1.4 Targeted cancer therapy

We introduced targeted cancer therapy and the move towards stratified and personalised medicine in the Introduction. Recall that targeted therapies are drugs that target specific molecules involved in tumour formation and progression, and so interfering with these molecules can block the growth of cancer cells. Many of these targets are signalling proteins, which we focus on here and give some specific examples.

The first molecular target for cancer therapy was the estrogen receptor (ER), an intracellular receptor that activates when bound by the hormone estrogen. Approximately 75% of breast cancers are dependent on ER for proliferation. Activated ER can usually be found in the cell nucleus, where it acts as a transcription factor, regulating genes that control cell proliferation. In the 1970s a drug called Tamoxifen was approved for use against breast cancer. Tamoxifen is a competitive antagonist; that is, it competes with estrogen for binding of ER and binding does not lead to activation of ER. Therefore, it can prevent estrogen from binding and activating ER and thereby blocks cancer cell growth. However, it was not until the 1990s that analysis of clinical trial results demonstrated that Tamoxifen only provides benefits in ER positive early breast cancer tumours [EBCTCG (Early Breast Cancer Trialists' Collaborative Group), 1998]. Tamoxifen is an effective targeted breast cancer adjuvant therapy and has also been used as a preventive treatment [Jordan, 2006].

Trastuzumab (market name Herceptin) is a monoclonal antibody that binds selectively to the HER2 receptor and is used to treat breast cancer. Approximately a third of breast cancers are HER2 positive; that is, they overexpress the HER2 receptor (see above for details of aberrant HER2 signalling). The mechanism of action of Trastuzumab is not completely understood. Possibilities include the prevention of HER2 activation and therefore a reduction in the downstream signalling that leads to uncontrolled cell proliferation, antibody-dependent cell cytotoxicity (ADCC; binding of Trastuzumab to HER2 induces an immune response that attacks the cell), suppression of tumour angiogenesis (growth of new blood vessels), or removal of HER2 from the cell surface. Laboratory tests are performed on tumour samples to determine if it is HER2 positive, and will therefore respond to treatment with Trastuzumab. It was approved in 1998 and was one of the first treatments to be applied selectively based on molecular characteristics of tumours. When combined with chemotherapy, treatment with Trastuzumab results in improved survival and response rates. However, it can have severe side effects such as heart disease and resistence to the drug is usually developed within one year of commencement of treatment. Further details regarding trastuzumab can be found in Ménard *et al.* [2003]; Emde *et al.* [2011] and references therein.

An important class of drugs are those that inhibit the kinase activity of intracellular signalling proteins. An example of such a drug is the small molecule inhibitor Imatinib (market name Glivec or Gleevec) used to treat CML by binding to the Bcr-Abl fusion protein that is constitutively active in CML (as described above). Kinase inhibitors have high specificity for certain kinases and bind to the catalytic kinase domain, which blocks the binding of ATP or the substrate (or both), thereby preventing phosphorylation of the kinases downstream target (see Figure 2.3). For

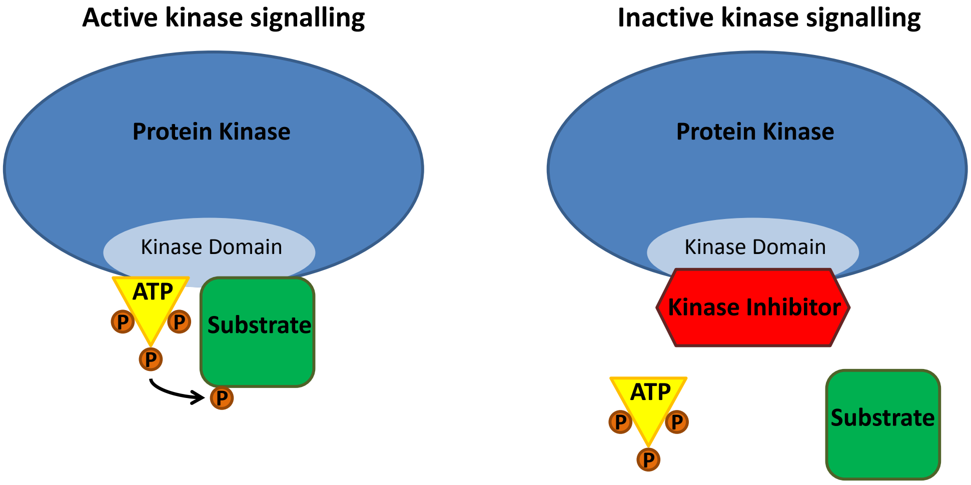


Figure 2.3: **Protein kinases and mechanism of action of kinase inhibitors.** Left: Protein kinases have a catalytic kinase domain that transfers a phosphate group from ATP to specific amino acids on protein substrates; a process called phosphorylation. This leads to 'activation' of the substrate and, in this way, a signal is transduced through the cell. Right: Kinase inhibitors block kinase signalling by binding to the kinase domain, which prevents binding of ATP or the substrate (or both) and therefore blocks phosphorylation of the substrate.

example, Imatinib binds to the ATP binding site of Abl. Imatinib has transformed CML treatment, improving five year survival rates from 30% to approximately 90%.

Kinase inhibitors are used in Chapters 3 and 4. In Chapter 3 an Akt inhibitor is used in the application of the variable selection approach to drug response data. In Chapter 4 Akt and Mek inhibitors are used to validate hypotheses generated from inference of signalling networks. In order to investigate whether a kinase inhibitor has the desired effect, it is necessary to monitor the targets of the kinase. For example, in the Raf-Mek-Erk cascade, a Mek inhibitor does not reduce the phosphorylation level of Mek, but the reduction in Mek's enzymatic activity should lead to a reduction in the phosphorylation level of Erk.

Many emerging targeted therapies are currently in clinical trials; for example, an overview for breast cancer is given by Alvarez *et al.* [2010].

2.1.5 Protein signalling networks

We have described above how signalling plays a key role in development and progression of cancer, and as a result, successful targeted therapies that interfere with specific signalling proteins have been developed. Therefore, it is important to study cells, both normal and cancerous, at the signalling network level. In particular, investigating phosphorylation of proteins can aid the discovery of new drug targets and cancer biomarkers [Yu *et al.*, 2007].

As described in the Introduction, signalling networks have complex structures and mechanisms. Therefore, multivariate approaches capable of modelling multiple interacting components are required in order to obtain a more complete understanding of normal signalling processes and their dysregulation in cancer. For example, consider structure learning of signalling networks. Interactions between signalling proteins were traditionally discovered by focussing on a single protein and performing multiple experiments to determine its upstream regulators and downstream targets (an example of such a discovery is the well-studied interaction between kinases Akt and GSK3; Akt phosphorylates GSK3 [Brazil and Hemmings, 2001]). Advances in high-throughput proteomics and the resulting multivariate data, together with advances in computing, have rendered feasible structure learning with multivariate models [Sachs *et al.*, 2005; Mukherjee and Speed, 2008; Bender *et al.*, 2010], and in specific contexts such as cancer, where signalling networks are thought to be 'rewired' [Pawson and Warner, 2007; Yuan and Cantley, 2008].

The protein signalling networks described above, and illustrated in Figure 2.2, contain several types of mechanisms, including protein-protein interactions and phosphorylation. In this thesis, we consider only phosphorylation signalling. This means that edges in the phosphorylation signalling network may not represent direct causal mechanisms. For example, an edge from a receptor to an upstream kinase (e.g. Raf) would be an indirect interaction via unobserved adaptor proteins (e.g. Grb2). Moreover, the signalling networks we consider do not contain any spatial information which may play an important part in signalling; for example, migration of an activated protein from the cytoplasm into the nucleus, or internalisation of a receptor. However, this simplified representation can still be useful for many analyses as we demonstrate in Chapters 3-5 where such signalling networks are used as prior information and as the object of inference.

It is worth emphasising that the phosphorylation signalling networks considered here are not the same as protein-protein interaction networks (PPI networks), which also receive significant attention in the literature. Protein-protein interaction networks mostly focus on protein interactions where two or more proteins physically bind to form a complex. Phosphorylation and PPI networks are, however, related. For example, phosphorylation of a protein may be required before a protein-protein interaction can occur [Shaywitz *et al.*, 2002]; this is the case for interactions between RTKs and adaptor proteins as described above.

2.2 Experimental background

In this section, we give a brief overview of several protein assays that can be used to probe protein phosphorylation. Such analyses are challenging due to the low abundance of phosphoproteins, low stoichiometries, and highly dynamic phosphorylation process. In recent years, several high-throughput approaches have been developed, although we note that these technologies are not yet on the same scale as DNA microarray and next generation sequencing technologies. We focus mainly on reverse phase protein arrays, which provide the data for our applications in Chapters 4 and 5.

2.2.1 Western blot

A Western blot [Burnette, 1981], also called a protein immunoblot, is used to detect the amount of specific proteins (at phosphoform and isoform level, if desired) in a sample. The procedure is as follows: Cells are lysed to release the proteins. The proteins are then separated according to their molecular weight by gel electrophoresis, usually SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis). This process uses an electric field to obtain separation. Proteins are then transferred from the gel to a special membrane, while retaining the same pattern of separation as on the gel; this is the 'blotting' part of the procedure. The membrane is probed for a protein of interest using antibodies. A primary antibody binds to the protein (ideally with high specificity) and then a secondary antibody is applied which binds to the primary antibody. The secondary antibody is labelled with a reporter enzyme, resulting in production of luminescence in proportion to the amount of protein bound by the primary antibody. This luminescence can then be 'photographed'.

Western blots are widely-used to assess phosphorylation states of proteins. They have reasonably good sensitivity and specificity, but require a relatively large amount of sample and significant human labour effort. Hence, they are not particularly suitable for large-scale proteomic analyses. Western blots are used in Chapter 4 to validate hypotheses generated from our statistical data analysis (see Figure 4.9(c)).

2.2.2 ELISA

Enzyme-linked immunosorbent assay (ELISA; [Engvall and Perlmann, 1971]) can also be used to determine the amount of (phospho)protein present in a sample. Here, we describe the 'Sandwich' ELISA method. A microplate containing a number of wells (often 96) is coated with a capture antibody specific to the protein of interest (but not specific to the phosphoform of interest). The sample (e.g. cell lysate) containing the protein of interest is added to the plate, resulting in binding of the protein to the capture antibody. A detection antibody specific to the protein phosphoform of interest is added, and then a secondary antibody with reporter enzyme is used, as in the Western blot assay, to determine the amount of phosphoprotein in the sample.

The results of ELISA are more easily quantifiable than for Western blot and the approach has higher specificity and sensitivity (due to the use of two antibodies). However, it is still not a high-throughput method. It is widely used in clinical settings, for example to perform HIV screening tests.

2.2.3 Mass spectrometry

Advancements in mass spectrometry (MS) technology and techniques over the last decade have enabled its use for high-throughput phosphoproteomic analyses. There are numerous alternative MS procedures and it is a very active field of research. Since MS data is not used in this thesis, we give only a very brief and general overview. Further details of mass spectrometry techniques and their application to (oncogenic) protein signalling can be found in Nita-Lazar *et al.* [2008]; Harsha and Pandey [2010]; Choudhary and Mann [2010].

The general MS procedure is as follows: Proteins are isolated from cell lysate and may also be separated by gel electrophoresis. They are then degraded into peptides; MS can be performed on whole proteins, but is then less sensitive. The peptides enter the mass spectrometer and are separated, vapourised and then ionised. The ions are then separated according to their mass-to-charge ratio, detected and quantified, resulting in a mass spectrum showing intensities against ratios. Proteins can then be identified using database matching algorithms. Quantitative MS techniques such as SILAC (stable isotope labelling by amino acids in cell culture) can be used to obtain relative abundances of proteins in different samples.

In the past, MS approaches have focussed on unbiased discovery of protein phosphorylation sites and their quantification. More recently, targeted MS approaches have been used to, for example, quantify temporal dynamics in specific signalling pathways [Wolf-Yadlin *et al.*, 2007], and has helped to improve reproducibility of MS results.

Probing phosphoproteins by MS is challenging due to their often low abundance relative to non-phosphorylated proteins. Enrichment methods are usually employed to help identify the phosphoproteins. The technology is capable of detecting thousands of phosphorylation sites in a single experiment.

2.2.4 Flow cytometry

Flow cytometry [Herzenberg *et al.*, 2002; Perez and Nolan, 2002] can measure the abundance of phosphoproteins on a single-cell level, in a high-throughput manner. A laser beam is aimed at a hydrodynamically focussed stream of liquid. Cells are passed through this stream, causing a scattering of the light beam. Fluorescent chemicals in the cell, attached to antibodies specific to certain phosphoproteins, are excited by the laser beam and emit light. The scattered and fluorescent light is measured by detectors allowing information to be obtained about the cell, such as abundance of phosphoproteins.

Flow cytometry can obtain thousands of samples (measurements from thousands of cells) in seconds and is also capable of sorting a heterogeneous mixture of cells in a technique called FACS (fluorescence-activated cell sorting). In contrast to experiments using lysed cells, cells are not destroyed during analysis by flow cytometry. Another advantage of flow cytometry is that variability on a single-cell level can be observed. Non-single-cell approaches assume that behaviour of cells can be captured from the population average (lysed cells). However, stochasticity at the single-cell level may be important in determining cell behaviour.

The number of phosphoproteins that can be measured simultaneously is, however, limited. This is due to spectral overlap between fluorescent markers, meaning that in practice, only up to ten proteins can be measured. Background noise (fluorescence) can also cause problems.

In Sachs *et al.* [2005] flow cytometry was used to measure 11 phosphoproteins and phospholipids in thousands of individual immune system cells, and the data was used to infer a signalling network. We use this data in simulations in Chapter 3. Flow cytometry is also used in clinical settings to diagnose blood cancers, for example.

2.2.5 Reverse phase protein arrays

Reverse phase protein arrays (RPPAs), first introduced in 2001 [Paweletz *et al.*, 2001], are a high-throughput technology capable of quantitative measurement of (phosphorylated) protein levels in thousands of biological samples simultaneously. The experimental procedure uses antibodies to detect proteins, and in this respect it is similar to Western blots and ELISA. The essence of the procedure is illustrated in Figure 2.4 and outlined below.

Cells are lysed and the solution is robotically spotted onto a nitrocellulose coated slide. The microarray slide contains many spots in a grid allowing for many samples to be immobilised and tested simultaneously, usually in replicate and using

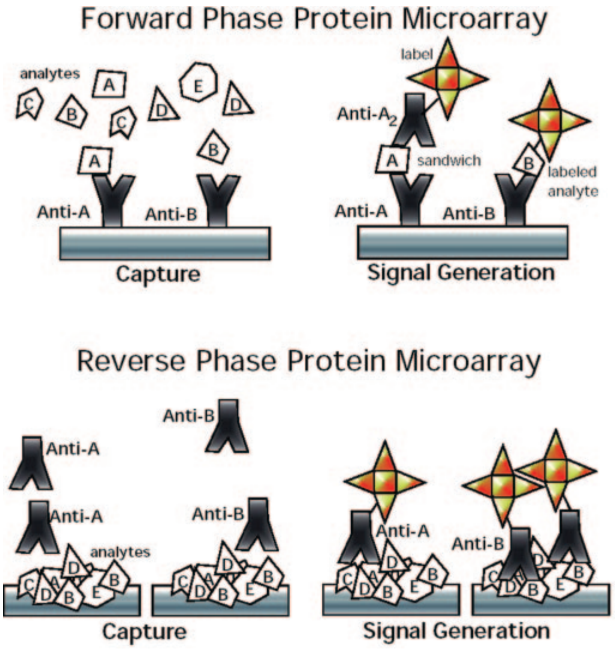


Figure 2.4: **Protein microarrays.** (a) Forward phase protein array. Protein antibodies are immobilised onto the slide which capture specific (phospho)proteins of interest (analytes) from cell lysate (a single sample). The analytes are detected using a sandwich antibody (as in ELISA), together with a labelled secondary antibody, or the analyte can be labelled directly to allow detection. (b) Reverse phase protein array. Cell lysates (multiple samples) are spotted onto the microarray slide, which is then probed with a single (phospho)protein specific primary antibody. A labelled secondary antibody is used to detect the primary antibody. Multiple slides can be used, each probed with different antibodies, to detect different proteins of interest. Figure reproduced from Sheehan *et al.* [2005].

dilution series (we explain dilution series further below). The array is then incubated with a primary antibody that binds specifically to the phosphoprotein of interest. This antibody is then detected using a labelled secondary antibody (as in ELISA and Western blot) and signal amplification. The emitted signal is quantified using a software package. Full details of RPPA protocol can be found in Tibes *et al.* [2006] and Hennessy *et al.* [2010]. If multiple microarray slides are spotted simultaneously with the same samples, each slide can be probed with a different antibody, thereby providing readouts for multiple samples and multiple proteins [see e.g. Sheehan *et al.*, 2005].

The samples are spotted onto the array in dilution series. Protein and phosphoprotein concentrations can vary greatly, so accurate measurements over a wide dynamic range are required. The dynamic range of measurements is extended by diluting each sample several times and spotting onto the array at each dilution step. Hence, if the protein concentration in the original undiluted sample is near saturation, it can still be detected in the diluted samples. Dilution series also aid the accurate quantification of protein concentrations. Quantification is usually carried out using response curves, that relate the observed signal intensities to the (phospho)protein concentrations. The fact that a single antibody is used for the whole slide motivates the use of a single response curve for all samples on the slide. For the RPPA data used in this thesis, a logistic model was used for the response curve (*R* package '*SuperCurve*' developed by the Department of Bioinfomatics and Computational Biology in MD Anderson Cancer Center [Hu *et al.*, 2007]).

There are two main types of protein microarrays, of which RPPAs are one. 'Reverse phase' refers to the fact that the cell lysate is immobilised on the slide and the array is probed with an antibody, which is a reversal of the procedure for antibody arrays, the other type of protein microarray. Antibody arrays are also referred to as forward phase protein arrays. For the antibody array, antibodies are immobilised onto the slide which capture proteins of interest from cell lysate. Each array spot contains a single type of antibody and the array is incubated with one sample only, providing readouts for multiple proteins from one sample (see Figure 2.4).

RPPAs are an emerging technology that enable measurements for a single (phospho)protein of interest to be obtained for hundreds to thousands of samples in a fast, automated, quantitative and economical manner. Multiple arrays can be used to probe for multiple (phospho)proteins; tens of proteins are often measured in the same experiment, providing an advantage over flow cytometry. The technique is also highly sensitive, requiring very small amounts of sample to enable detection of analytes; only 10^3 cells are required for an RPPA experiment, compared with 10^8 for mass spectrometry and 10^5 for Western blotting. Therefore, while mass spectrometry is a promising approach, RPPA is currently more sensitive and cost-effective. An advantage of RPPAs over antibody arrays is that either fewer antibodies are required or samples do not need to be directly labelled to allow detection of the analyte of interest. This can improve robustness and reproducibility of results. Also, RPPAs can use denatured lysates (proteins in the lysate have lost their three-dimensional conformation) which can allow antibodies to bind that previously would not have been able to do so, providing an advantage over tissue microarrays (another reverse-phase assay).

The main limitation of RPPAs is specificity of primary and secondary antibodies. The signal from a microarray spot could be due to cross-reactivity from unspecific binding and it is not possible to determine if this is the case from the RPPA results themselves. Therefore antibodies have to be carefully validated by Western blotting prior to their use in RPPA assays. Hennessy *et al.* [2010] is an example of such a validation study. The number of available validated antibodies is continuously growing.

RPPAs have been used in many studies to investigate cancer cell signalling, both in cancer cell lines [Tibes *et al.*, 2006] and in primary tumour samples [Sheehan *et al.*, 2005]. These studies include the profiling and comparison of active signalling pathways in different contexts; for example, between primary and metastatic tumours [Sheehan *et al.*, 2005] or between cancer subtypes [Boyd *et al.*, 2008], the identification of signalling biomarkers that are predictive of response to certain anti-cancer agents [Boyd *et al.*, 2008], the identification of optimal drug combinations [Iadevaia *et al.*, 2010] and structure learning of signalling networks [Bender *et al.*, 2010]. For further studies see, for example, Spurrier *et al.* [2008]; Hu *et al.* [2007] and references therein. RPPAs have promising utility in the development of personalised therapies; using RPPAs to investigate and compare signalling profiles in patient tumour cells and normal cells, and to monitor changes in phosphorylation through time, both pre- and post-treatment, could provide information that guides the discovery and application of targeted therapies. Indeed, RPPAs are currently involved in several clinical trials [Mueller *et al.*, 2010].

In Chapter 4, RPPA data for 20 phosphoproteins from a breast cancer cell line are used for structure learning of a signalling network, and in Chapter 5 RPPA data for 39 phosphoproteins across 43 breast cancer cell lines covering two breast cancer subtypes are used for clustering and network structure learning.

2.3 Methodological background

This section begins by reviewing the linear regression model, which is a core component of all the approaches proposed in the following Chapters. Both frequentist maximum likelihood and Bayesian approaches to inference for the linear model are outlined and compared. Inference of graphical model structure is often cast as a model selection problem and this is the route we take in the work here. A general overview of model selection and model averaging, including standard methods, is given in Section 2.3.2. We describe the specific case of variable selection in regression models in Section 2.3.3. In Section 2.3.4 we define and describe some technical details regarding directed and undirected graphical models. Methods for structure learning of graphical models are detailed in Sections 2.3.5 and 2.3.6. Several alternative methods for structure learning of molecular networks, in addition to those based on graphical models, have been proposed in the literature. We give a brief overview of some of these approaches in Section 2.3.7. Section 2.3.8 provides background information on empirical Bayes methods and clustering techniques are considered in Section 2.3.9.

2.3.1 Linear regression model

Due to its simplicity, interpretability and applicability, the linear regression model is arguably the most widely used statistical model in data analysis. Interactions between signalling components are complex and non-linear in nature, and so a linear model is not likely to be the most accurate representation of the underlying process. However, as we shall see below, using a Gaussian linear model can result in closed form expressions for high dimensional integrals and thereby provides significant computation gains.

The linear regression model assumes that a response variable Y is a linear combination of p predictor variables $X_1, \dots X_p$,

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \tag{2.1}$$

where $\{\beta_j\}$ are unknown regression coefficients and ϵ is a zero-mean noise term. The predictor variables X_j are often taken to be observed (experimentally measured) values of variables under study. For example, Y could represent drug response and X_j phosphorylation levels of cellular proteins. Alternatively, the predictor variables could also be transformations of measured entities (e.g. log-transform) or basis expansions. For example, if we initially have one predictor variable X, we could take

$X_j = X^j$ for $j = 1, \ldots, p$, resulting in p predictor variables with the dependence of Y on X given by a polynomial of degree p. In what follows, unless stated otherwise, predictor variables correspond directly to observed values of the variables under study.

Given a dataset of n samples, let variables Y_i and X_{ij} denote the response and j'th predictor in sample i respectively. Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$, $\mathbf{X}$ is the $n \times (p+1)$ design matrix with row i given by $(1, X_{i1}, \ldots, X_{ip})$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}$. We make the standard assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ where $\mathcal{N}$ denotes a Normal distribution, $\mathbf{I}_n$ is the $n \times n$ identity matrix and σ^2 is unknown variance. The linear regression inference problem is to determine the regression coefficients $\boldsymbol{\beta}$ from the data $(\mathbf{X}, \mathbf{Y})$. The inferred coefficients can then be used for prediction on new data and to assess the relative importance of individual predictors in influencing the response. Below we describe frequentist maximum likelihood and Bayesian approaches to inference for the linear model.

2.3.1.1 Frequentist maximum likelihood inference approach

In a frequentist formulation parameters $\boldsymbol{\Theta} = \left(\boldsymbol{\beta}, \sigma^2\right)$ are treated as fixed and unknown, while the observed response data $\mathbf{Y}$ is regarded as a single realisation of a repeatable process (with parameters $\boldsymbol{\Theta}$). We note that, in regression, predictor data $\mathbf{X}$ is regarded as fixed and known. Probability statements are interpreted as the limiting relative frequency as the number of repeats goes to infinity. Parameter estimation is based on the likelihood function,

$$\begin{aligned} L(\boldsymbol{\Theta} \mid \mathbf{Y}, \mathbf{X}) &= p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) & (2.3) \\ &= \mathcal{N}\left(\mathbf{Y} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}\right) & (2.4) \end{aligned}$$

The likelihood is the probability of observing the data given parameters $\boldsymbol{\Theta}$ (and predictor values $\mathbf{X}$), and is a function of $\boldsymbol{\Theta}$, but not a probability density function with respect to $\boldsymbol{\Theta}$. A standard approach to obtain parameter estimates is by maximising the likelihood; that is, by selecting the parameters that give the observed data the highest probability. This estimate is known as the *maximum* *likelihood estimate* (MLE) and we denote it by $\hat{\boldsymbol{\Theta}}_{\text{MLE}}$,

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \max_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta} \mid \mathbf{Y}, \mathbf{X}). \tag{2.5}$$

For the Gaussian linear regression model we have

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathsf{T}} Y \qquad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \left\| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{MLE}} \right\|_2^2. \tag{2.6}$$

where $\|\cdot\|_2$ is the Euclidean norm. We note that $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the same as the least squares estimate for $\boldsymbol{\beta}$ and that $\hat{\sigma}^2_{\text{MLE}}$ is a biased estimator. The unbiased estimator $s^2 = \frac{n}{n-p-1}\hat{\sigma}^2_{\text{MLE}}$ is often used. Since these estimates are based on a single realisation of the data-generating process, estimation uncertainty is based on hypothetical data that could have been observed. In practice this is achieved using the sampling distribution of the MLE to perform hypothesis tests and obtain confidence intervals; in regression the sampling distributions are $\hat{\boldsymbol{\beta}}_{\text{MLE}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\right)$ and $(n-p-1)s^2 \sim \chi^2_{n-p-1}$.

2.3.1.2 Bayesian inference approach

A Bayesian analysis of the linear model was first presented by Lindley and Smith [1972]. In the Bayesian approach, the parameters $\boldsymbol{\Theta}$ are unknown, but are regarded as random variables. Therefore, unlike in the frequentist approach, probability statements can be made regarding the parameters themselves. In particular, the distribution over $\boldsymbol{\Theta}$ given the observed data is of interest, $p(\boldsymbol{\Theta} \mid \mathbf{Y}, \mathbf{X})$. By Bayes' theorem we have,

$$\begin{aligned} p(\boldsymbol{\Theta} \mid \mathbf{Y}, \mathbf{X}) &= \frac{p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta})}{p(\mathbf{Y} \mid \mathbf{X})} && (2.7) \\ &\propto p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) && (2.8) \end{aligned}$$

where $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta})$ is the likelihood (2.3) and $p(\boldsymbol{\Theta})$ is a prior distribution over parameters, assigning probabilities to $\boldsymbol{\Theta}$ before observing any response data [2]. In the Bayesian framework, probability is interpreted as 'degree of belief' and the prior distribution allows subjective prior beliefs to enter the analysis. Thus, Bayes' theorem combines, in a coherent manner, information about parameters from prior beliefs (the prior distribution) with information from the observed data (the likelihood),

[2] We note that technically we should write the prior as $p(\boldsymbol{\Theta} \mid \mathbf{X})$, and some prior formulations do indeed depend on the predictor data. However, for notational simplicity we suppress this possible dependence.

to form the posterior distribution $p(\mathbf{\Theta} \mid \mathbf{Y}, \mathbf{X})$. Since the posterior is a density, it provides more information about $\mathbf{\Theta}$ than a single point estimate (such as the MLE) and allows uncertainty to be intuitively quantified (e.g. with credible intervals), based on the actual observed data. This is in contrast to the frequentist approach, where confidence intervals are based on hypothetical datasets that could have been observed.

One of the challenges of the Bayesian approach is specifying the prior distribution. While at large sample sizes a (reasonably defined) prior will have a small influence on the posterior, this is not the case at small sample sizes. One school of thought in the Bayesian community is that priors should be objective or 'non-informative'. That is, they should, in some sense, play a minimal role in the posterior. Such priors can be hard to define, but one approach is to use a so-called Jeffreys priors, named after Harold Jeffreys [Jeffreys, 1961], which are invariant to reparameterisation of the parameter space (Bayesian inferences are, in general, not invariant to such transformations, unlike the MLE). For example, if we assume that σ^2 is known in the linear regression model (2.2), and we take the prior for $\boldsymbol{\beta}$ to be flat (i.e uniform; this is a Jeffreys prior), $p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2) \propto 1$, then by (2.8) the posterior is proportional to the likelihood,

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \sigma^2) \propto p(\mathbf{Y} \mid \mathbf{X}, \mathbf{\Theta}). \tag{2.9}$$

Hence it is clear that, under this non-informative flat prior formulation, the *maximum a posteriori* (MAP) estimate for $\boldsymbol{\beta}$ (the mode of the posterior) coincides with the MLE. Moreover, it can be shown that the posterior has distribution $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\right)$ [see e.g. Gelman *et al.*, 2003]. So the variance of the posterior is the same as the variance of the MLE. This demonstrates that the frequentist MLE and Bayesian approaches result in essentially the same inferences in this case; the main difference is in the interpretation of interval estimates. We note that the flat prior is improper (the integral over parameter space is infinite). However, in many cases (as here) an improper prior does not lead to an improper posterior.

Bayes' theorem enables the posterior to be easily found up to proportionality (2.8). However, the normalising constant $p(\mathbf{Y} \mid \mathbf{X})$ in (2.7) is required to calculate actual posterior probabilities and make inferences (other than finding the MAP estimate). Except for special cases, this normalising constant, which can be expressed as an integral,

$$p(\mathbf{Y} \mid \mathbf{X}) = \int p(\mathbf{Y} \mid \mathbf{X}, \mathbf{\Theta}) p(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta} \tag{2.10}$$

is not obtainable in closed form. These special cases are where the prior is *conjugate*;

that is, the prior has the same parametric form as the posterior [see e.g. Bernardo and Smith, 1994]. When the prior is not conjugate and the integral above cannot be evaluated numerically, computational methods are needed to sample from the posterior and calculate posterior probabilities of interest. Conjugate priors therefore offer substantial computational gains. We refer the interested reader to Bolstad [2010] for details of methods in computational Bayesian statistics. We briefly mention Markov chain Monte Carlo methods in Section 2.3.3 below.

A conjugate prior choice for the Gaussian linear model is a normal inverse-gamma (NIG) distribution; a Gaussian prior for $\boldsymbol{\beta}|\sigma^2$ and an inverse-gamma prior for σ^2,

$$\begin{aligned} p(\boldsymbol{\beta},\sigma^2) &= p(\boldsymbol{\beta}\,|\,\sigma^2)p(\sigma^2) \\ &= \mathcal{N}\left(\boldsymbol{\beta}\,|\,\mathbf{m},\sigma^2\mathbf{V}\right) IG(\sigma^2\,|\,a,b) \end{aligned} \tag{2.11}$$

The inverse-gamma distribution $IG(\sigma^2\,|\,a,b)$ is parameterised as

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)}\left(\sigma^2\right)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right). \tag{2.12}$$

where $a,b > 0$ and $\Gamma(a)$ is the Gamma function. We denote the NIG distribution by $NIG(\mathbf{m},\mathbf{V},a,b)$. The parameters of a prior distribution are referred to as *hyper-parameters*. Since the prior is conjugate, applying Bayes' theorem (2.7) results in a NIG posterior [see e.g. Denison *et al.*, 2002],

$$p(\boldsymbol{\beta},\sigma^2\,|\,\mathbf{Y},\mathbf{X}) = NIG(\mathbf{m}^*,\mathbf{V}^*,a^*,b^*) \tag{2.13}$$

where

$$\mathbf{m}^* = \left(\mathbf{V}^{-1}+\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m}+\mathbf{X}^\mathsf{T}\mathbf{Y}\right) \tag{2.14}$$

$$\mathbf{V}^* = \left(\mathbf{V}^{-1}+\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \tag{2.15}$$

$$a^* = a+\frac{n}{2} \tag{2.16}$$

$$b^* = b+\frac{1}{2}\left(\mathbf{m}^\mathsf{T}\mathbf{V}^{-1}\mathbf{m}+\mathbf{Y}^\mathsf{T}\mathbf{Y}-(\mathbf{m}^*)^\mathsf{T}\left(\mathbf{V}^*\right)^{-1}\mathbf{m}^*\right). \tag{2.17}$$

The posterior distribution over parameters can be used to obtain a posterior predictive distribution for new data $(Y',\mathbf{X}')$,

$$p(Y'\,|\,\mathbf{X}',\mathbf{Y},\mathbf{X}) = \int p(Y'\,|\,\mathbf{X}',\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta},\sigma^2\,|\,\mathbf{Y},\mathbf{X})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\sigma^2 \tag{2.18}$$

where the first term in the integrand is distributed as $\mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and the second is the posterior. Since the posterior is a density, the parameters can be marginalised (integrated) out. This has advantages over simply plugging a point estimate for the parameters (e.g. the MLE) into $p(\mathbf{Y}' \,|\, \mathbf{X}', \boldsymbol{\beta}, \sigma^2)$, because no information is lost and it takes parameter uncertainty into account. For the NIG prior, the posterior predictive distribution (2.18) has a Student distribution with mean $\mathbf{X}'\mathbf{m}^*$ [see Denison *et al.*, 2002].

2.3.2 Model selection and averaging

Selection of an appropriate model is a trade-off between fit to (training) data and predictive capability on independent (test) data. A model with higher complexity (more parameters) will have an improved fit to data, but will not necessarily provide better predictions because the model may be overfitting the data. We illustrate this point with an example in linear regression. Suppose the true underlying relationship between the response Y and a predictor variable X is a cubic polynomial. However, the model we consider is a polynomial with degree k. That is, (2.1) with $X_j = X^j$ for $j = 1, \ldots, k$. Here, k can be interpreted as model complexity, with larger k producing a model with higher complexity (more parameters). This model is fitted to training data, using least squares (i.e. the MLE), for various values of k, and predictive capability is then assessed on independent test data. Figure 2.5 shows average training data error and average test data error, as a function of k. Training data and test data error are given by the sum of squared error loss (SSE), where

$$\mathrm{SSE}(\hat{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{Y}) = \left\| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2^2 \tag{2.19}$$

and $\hat{\boldsymbol{\beta}}$ is estimated on the training data (other loss functions could also be used). As k (model complexity) increases, the model fits the training data increasingly well. However, if k gets too large the model is unable to generalise to new data, resulting in large test data error. Equally, if k is too small, the model does not have enough flexibility to give a reasonable approximation of the underlying function, resulting in large train data and test data error.

In order to select a model that has a good fit to data, but does not overfit, the principle of Occam's razor can be applied. This says that if two models fit the data equally well, the least complex model should be favoured. Parsimonious models can also be preferable as they enable easier interpretation of results.

The space of models depends on context. Variable selection, or feature selection, is a model selection problem in regression, in which a model corresponds

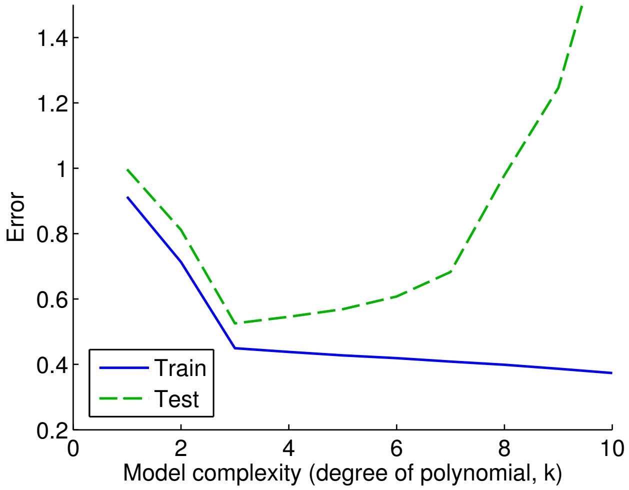


Figure 2.5: **Model fit, model complexity and predictive capability.** Polynomials of degree k, for $k = 1, \ldots, 10$, were fitted to training data (generated using a cubic polynomial). Fitted models were then used to make predictions on independent test data. Average training data error and average test data error are shown as a function of k (model complexity). (Average values taken over 100 train/test dataset pairs, each with sample size $n = 50$).

to a subset of predictors and we wish to determine which subset of predictors best explains the response. We discuss variable selection further in Section 2.3.3 below and apply it in Chapter 3 to discover proteins that influence drug response. In Chapter 4, the space of models consists of protein signalling network structures. Network structure learning can be performed by selecting the model (network) that is best supported by the data (see Section 2.3.5).

In order to select a model, a scoring function is required to assign each model a score (that takes fit to data and model complexity into account), and a method is needed to search over the model space (if the space is too large for exhaustive enumeration). We consider the latter in the context of variable selection in Section 2.3.3 and consider methods of scoring models below. Non-Bayesian methods are briefly outlined, before focussing on Bayesian model selection, which we use in Chapters 3 and 4.

2.3.2.1 Non-Bayesian scoring methods

The Akaike Information Criterion (AIC) [Akaike, 1974] is one means of selecting a model from a finite set of models $\mathcal{M} = \{M\}$. It is defined as follows,

$$AIC(M) = -2l(\hat{\boldsymbol{\Theta}}_{\text{MLE}}(M)) + 2d(M) \tag{2.20}$$

where $l(\hat{\mathbf{\Theta}}_{\text{MLE}}(M))$ is the log-likelihood evaluated at the MLE for model M and $d(M)$ is the number of parameters in model M. The likelihood term assesses fit to data, while the second term is a complexity penalty that helps to prevent overfitting. The chosen model is the one that minimises (2.20). AIC has its routes in information theory; asymptotically, it is a measure of the information lost, as quantified by Kullback-Leibler divergence, when using a given model instead of the true model. A corrected version of AIC, called AICc, takes sample size into account, and for least squares regression, minimising AIC is equivalent to minimising Mallows' C_p statistic [Mallows, 1973]. See Claeskens and Hjort [2008] for further details.

Another widely-used method is multifold cross-validation (CV), which we outline in the context of least squares linear regression. If the dataset $(\mathbf{X}, \mathbf{Y})$ under study contains enough data, it can be split into a training set and testing set to enable assessment of model performance. However, this is often not the case, motivating the use of multifold cross-validation. The data is partitioned into S subsets of (roughly) equal size. We denote these subsets by $(\mathbf{X}^{(s)}, \mathbf{Y}^{(s)})$ for $s = 1, \ldots, S$. The algorithm consists of S iterations. In iteration s, the model is trained (using least squares/MLE) on all data save that in subset $(\mathbf{X}^{(s)}, \mathbf{Y}^{(s)})$ (training data). We denote the resulting parameter estimates by $\hat{\mathbf{\Theta}}^{(-s)}$. The predictive performance of the estimate is then assessed on the held-out data subset $(\mathbf{X}^{(s)}, \mathbf{Y}^{(s)})$ (test data) using a loss function. This is repeated S times, allowing each subset to play the role of test data. For linear regression, using the sum of squared error loss (2.19) results in the following CV score,

$$\text{CV} = \sum_{s=1}^{S} \text{SSE}\left(\hat{\boldsymbol{\beta}}^{(-s)}, \mathbf{X}^{(s)}, \mathbf{Y}^{(s)}\right). \tag{2.21}$$

To perform model selection, CV scores are calculated for each model M and the model with lowest CV score is chosen. The case $S = n$ is known as leave-one-out-cross-validation (LOOCV); in iteration s, estimation is performed using all data samples except sample s. Multifold CV can be computationally expensive as estimation is performed S times for each model. We shall use multifold CV in Chapters 3-5 to assess predictive capability of models and to select tuning parameters.

We note that multifold CV can also be used within a Bayesian framework. Instead of using point estimates for parameters (such as the MLE) to make predictions on test data, the expected value of the posterior predictive distribution (2.18) can be used. This takes parameter uncertainty into account and, if conjugate priors are used, can be calculated in closed-form.

Penalised likelihood-based approaches can also be used to simultaneously perform parameter estimation and model selection. We discuss these further in Section 2.3.3.

2.3.2.2 Bayesian model selection

Bayesian model selection is based on the posterior distribution over models $P(M|\mathbf{Y})$, where $\mathbf{Y}$ denotes data (for regression all distributions are also conditional on $\mathbf{X}$). In addition to this being a natural approach, the use of posterior probabilities can be formally motivated in a decision theory framework. If it is assumed that the model space $\mathcal{M}$ contains the true model, then for a 0-1 loss function (with zero loss obtained only on selection of the true model), selecting the model with highest posterior probability is equivalent to minimising the expected posterior loss [Bernardo and Smith, 1994]. The assumption that the model space contains the true model is almost certainly incorrect. However, it is hoped that at least one of the models provides a reasonable approximation to the truth, and the posterior probabilities can still be interpreted as the relative evidence in favour of a model [Wasserman, 2000].

By Bayes' theorem, the posterior probability of model M is given by

$$P(M \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid M)P(M)}{p(\mathbf{Y})} \tag{2.22}$$

where $p(\mathbf{Y} \mid M)$ is the marginal likelihood and $p(M)$ is a prior distribution over the model space ('model prior'). We consider the marginal likelihood further below. The prior $p(M)$ can either be chosen to be objective, by using a flat prior for example, or can be used to incorporate prior beliefs, based on existing domain knowledge. We take the latter approach in Chapters 3 and 4.

The marginal likelihood is the probability of observing the data under a given model, and can be obtained by integrating out parameters $\boldsymbol{\Theta}$,

$$p(\mathbf{Y} \mid M) = \int p(\mathbf{Y} \mid M, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \mid M) \mathrm{d}\boldsymbol{\Theta} \tag{2.23}$$

where the integrand is a product of the likelihood and prior over parameters. This marginalisation enables inference to take parameter uncertainty into account, which is in contrast to the model selection approaches described above, where the MLE is used.

The AIC model selection approach (2.20) uses a complexity penalty to avoid overfitting. In the Bayesian framework, complexity is automatically taken into account through the marginal likelihood. This penalisation occurs because a more

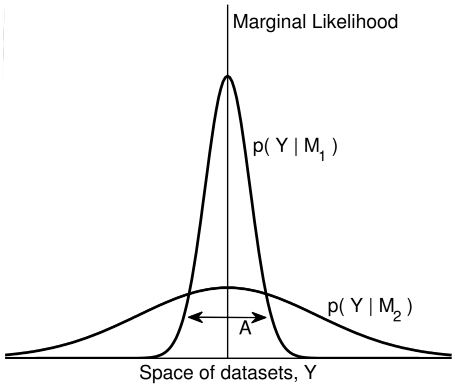


Figure 2.6: **Illustration of Occam's razor effect of marginal likelihood.** The horizontal axis represents the space of possible datasets and the vertical axis is the marginal likelihood. The marginal likelihood distribution $p(\mathbf{Y} \mid M)$ is shown for two models M_1 and M_2, with M_2 more complex than M_1 (i.e. M_2 has more parameters). The simpler model M_1 can make a limited range of predictions, while the more complex model M_2 can predict a greater variety of datasets, resulting in a more diffuse marginal likelihood $p(\mathbf{Y} \mid M_2)$. As a result, if the observed data lies in the region denoted by A, the simpler model will have a larger marginal likelihood. (Figure adapted from Mackay [1995] and Denison *et al.* [2002].)

complex model has a larger (higher dimensional) parameter space and so is capable of predicting a greater variety of datasets. This is illustrated in Figure 2.6, where two models M_1 and M_2 are considered, with M_2 more complex than M_1. In the scenario depicted, the simpler model M_1 can only predict a subset of the datasets that M_2 can predict. Thus, M_2 has a more diffuse marginal likelihood $p(\mathbf{Y} \mid M_2)$ than M_1 (since both must integrate to one over the space of datasets). This means that if the observed dataset is supported by both M_1 and M_2, the simpler model will have a larger marginal likelihood. A more detailed explanation of this automatic Occam's razor effect can be found in Mackay [1995].

If conjugate parameter priors $p(\mathbf{\Theta} \mid M)$ are used, the marginal likelihood integral (2.23) can be found exactly in closed-form. This has computational advantages over asymptotic approximate methods for calculating marginal likelihoods (for example, Laplace's method), especially when many models are under consideration. Details of approximate methods can be found in Kass and Raftery [1995] and references therein.

Competing models can be compared using posterior odds ratios. The posterior odds in favour of model M_i and against model M_j are given by

$$\frac{P(M_i \mid \mathbf{Y})}{P(M_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid M_i)}{p(\mathbf{Y} \mid M_j)} \cdot \frac{P(M_i)}{P(M_j)}. \tag{2.24}$$

That is, a product of the marginal likelihood ratio and prior odds ratio. The marginal likelihood ratio is known as the Bayes factor in favour of M_i and against M_j. The Bayes factor can be interpreted as the ratio between posterior odds and prior odds and is a measure of how much the prior odds in favour of M_i have been increased (or decreased) after observing data. If the prior over model space $P(M)$ is flat, then using the Bayes factor to select a model is equivalent to finding the model with highest posterior probability. Interpretations of Bayes factor values have been proposed by Jeffreys [1961] and modified by Kass and Raftery [1995]. We note that Bayes factors can be ill-defined when using improper or diffuse parameter priors. Further information on inference with Bayes factors can be found in Kass and Raftery [1995].

The Bayesian information criterion (BIC) or Schwarz criterion [Schwarz, 1978] is a model scoring method similar in spirit to AIC. It is defined as

$$BIC(M) = -2l(\hat{\mathbf{\Theta}}_{\text{MLE}}(M)) + \log(n)d(M) \tag{2.25}$$

where $l(\hat{\mathbf{\Theta}}_{\text{MLE}}(M))$ is the log-likelihood evaluated at the MLE for model M and $d(M)$ is the number of parameters in model M. BIC provides an approximation to the (log) marginal likelihood $p(\mathbf{Y} \mid M)$, with equivalence holding asymptotically as $n \to \infty$. Hence, when the integral in (2.23) is intractable, BIC provides an efficient method for calculating an approximation, and, under a flat prior on model space, it can be used to find the model with largest posterior probability. The complexity penalty in BIC is larger than that in AIC, and so leads to sparser models. While it is asymptotically consistent for model selection (unlike AIC), it may not perform well at small sample sizes, and it makes implicit assumptions regarding the parameter prior. BIC also has an information-theoretic interpretation in terms of minimum description length [see e.g. Hastie *et al.*, 2003].

2.3.2.3 Bayesian model averaging

As described above, Bayesian model selection aims to find the 'best' model M^* from a finite set of models $\mathcal{M}$, where 'best' is defined as the model which maximises posterior probability; that is, $M^* = \text{argmax}_{M \in \mathcal{M}} P(M \mid \mathbf{Y})$. M^* is often referred to as the MAP (maximum a posteriori) model. At large sample sizes, the posterior

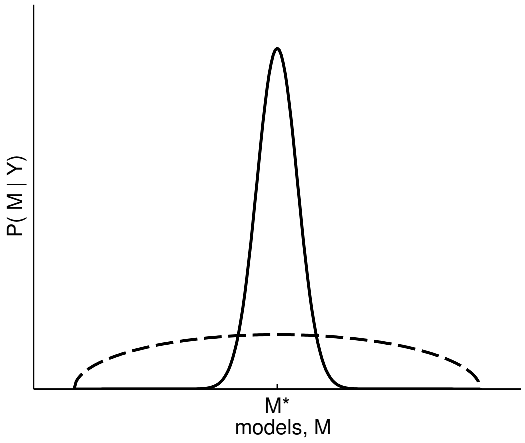


Figure 2.7: **Model uncertainty.** The horizontal axis represents the space of models $\mathcal{M}$ and the vertical axis shows posterior probability $P(M|\mathbf{Y})$ of model M given data $\mathbf{Y}$. Two posterior distributions are shown. The solid line is the posterior arising from a dataset with a large number of samples n; the distribution has a single sharp peak around the best model M^*. The dashed line is the posterior from a dataset with small n; the distribution is diffuse with many models having high scores. In this setting there is more uncertainty regarding the best structure.

distribution $P(M \mid \mathbf{Y})$ is likely to have a single well-defined peak at M^*, and M^* will be the model that best describes the data. However, at the small sample sizes that are typical of molecular data, the posterior is likely to be diffuse (Figure 2.7 illustrates this scenario). In this case, the single MAP model M^* may not be a good representation of the information contained in the entire posterior; there may be other models with posterior scores close to that of the MAP model, yet these models may display different features to M^*. Moreover, M^* will not necessarily be the model that best describes the data.

Bayesian model averaging is an intuitive way to address the issue described above. It allows inferences to be made using information from across the entire posterior distribution (not just the MAP model) and thereby takes model uncertainty into account. In particular, posterior probabilities for features of interest can be calculated by averaging over the space of models, weighting each model by its posterior probability. Then, a feature that appears in M^* but not in any other high scoring model will receive a low posterior score. A toy example in linear regression illustrates this point. Suppose our model space consists of two models M_1 and M_2, where M_1 says that response $\mathbf{Y}$ is independent of all predictor variables, and model

M_2 has $\mathbf{Y}$ dependent on predictor $\mathbf{X}_1 = (X_{11}, \ldots, X_{n1})$. So we have

$$M_1 : Y_i = \beta_0 + \epsilon_i \tag{2.26}$$

$$M_2 : Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \tag{2.27}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Suppose the feature of interest is whether the response is dependent on predictor $\mathbf{X}_1$. If M_2 has a higher posterior score than M_1, one could conclude that this dependence does indeed exist, but if we have $P(M_2 \,|\, \mathbf{Y}, \mathbf{X}) = 0.51$ and $P(M_1 \,|\, \mathbf{Y}, \mathbf{X}) = 0.49$ then M_1 is almost as probable as M_2. Averaging over the two models tells us that the posterior probability of the dependence existing is 0.51, which reflects the uncertainty associated with this feature.

In general, the posterior probability of a feature of interest ζ can be calculated as follows. For example, as seen in the regression example above, it could be the existence of a dependence between response and a certain predictor. Let $\mathbb{1}_M(\zeta)$ be an indicator function, evaluating to unity if and only if model M contains feature ζ. Then the posterior probability of ζ can be written as the posterior expectation of $\mathbb{1}_M(\zeta)$,

$$\begin{aligned} P(\zeta \,|\, \mathbf{Y}) &= \mathbb{E}\left[\mathbb{1}_M(\zeta)\right]_{P(M \,|\, \mathbf{Y})} && (2.28) \\ &= \sum_M \mathbb{1}_M(\zeta) P(M \,|\, Y). && (2.29) \end{aligned}$$

That is, the sum of the posterior probabilities of those models that contain feature ζ. It can be shown that averaging over all models in this way provides better predictive capability over using any single model [Madigan and Raftery, 1994].

One of the main challenges in implementing Bayesian model averaging is due to computational constraints. The space of models $\mathcal{M}$ can be very large, precluding explicit enumeration of the sum over graph space in (2.29) and also calculation of the normalising constant $p(\mathbf{Y})$ in (2.22), which can also be expressed as a sum over the graph space,

$$p(\mathbf{Y}) = \sum_M P(\mathbf{Y} \,|\, M) P(M). \tag{2.30}$$

Therefore, while it is easy to calculate the posterior score of a model up to proportionality (if conjugate parameter priors are used to give a closed-form marginal likelihood), it is difficult to calculate the absolute probabilities when the model space is large. Methods to deal with this issue include placing restrictions on the model space and summing over a subset of models, or using Markov chain Monte Carlo to obtain samples from the posterior distribution which can be used to approximate posterior probabilities of interest (2.29). We discuss this approach further in Section 2.3.3.

Bayesian model averaging takes both model uncertainty and parameter uncertainty into account (the latter through the marginal likelihood), while also allowing prior information to be incorporated into inference in a coherent manner. Other non-Bayesian model averaging approaches are available. For example, AIC scores can be used to assign weights to each model and obtain a weighted average parameter estimate [see e.g. Claeskens and Hjort, 2008]. Bootstrapping [Efron and Tibshirani, 1993] and, in particular, bagging (bootstrap aggregating) [Breiman, 1996] methods also have a model averaging flavour. Unlike the Bayesian approach, these methods calculate a point parameter estimate (e.g. MLE) under each model, rather than averaging over parameter space, and they are not as easily interpretable because the model weights do not have an intuitive probabilistic interpretation.

Further details on model selection and model averaging can be found in Claeskens and Hjort [2008], and reviews focussing on Bayesian methods are provided by Hoeting *et al.* [1999] and Wasserman [2000].

2.3.3 Variable selection in the linear model

Variable selection, also known as feature selection, is a specific form of model selection most frequently used with supervised learning problems. Variable selection in the linear regression model with p predictor variables (see (2.2)) aims to select a subset of the predictor variables that best explains variation in the response variable. Here, the space of models consists of all possible subsets of the predictor variables.

There are several reasons why performing variable selection may be useful. First, it can help to ameliorate the effects of overfitting and therefore improve prediction accuracy over using all p predictors. In settings where $p > n$, which is often the case for the types of molecular data considered here, the MLE (2.6) can not be found using all p predictors because $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is not invertible. This means some form of dimensionality reduction or regularisation is required; variable selection is one way to achieve this. Second, it can increase interpretability of results by obtaining a smaller set of predictors, each of which play a significant role in explaining the response. This can help to generate hypotheses for testing in follow-up experiments. The assumption that many of the predictors do not have a substantial influence on the response can be a reasonable one in many settings. Third, performing variable selection in a Bayesian framework, with model averaging, can improve robustness of results due to model and parameter uncertainty being taken into account.

We assume from here on that the predictors and response have been stan-

p	5	10	25	50	100	500	1000
$\lvert\Gamma\rvert$	32	1024	$\approx 10^7$	$\approx 10^{15}$	$\approx 10^{30}$	$\approx 10^{150}$	$\approx 10^{301}$

Table 2.1: **Number of models $|\Gamma|$ for varying number of predictors p.**

dardised to have zero mean, unit variance across the n samples. This means that we can drop the intercept term β_0 from the model. Thus the design matrix $\mathbf{X}$ for all p predictors is now of size $n \times p$.

An inclusion indicator vector $\gamma = (\gamma_1, \ldots, \gamma_p)^\mathsf{T} \in \{0,1\}^p$ specifies which predictors are contained in the model. That is, predictor j is included in the model if and only if $\gamma_j = 1$. We use γ to denote both the inclusion indicator vector and the model it specifies; γ takes the place of M in Section 2.3.2 above, and we let Γ denote the model space. We let $|\gamma| = \sum_j \gamma_j$ be the number of non-zeros in γ (i.e. the number of predictors in the model) and $\mathbf{X}_\gamma$ be the $n \times |\gamma|$ matrix obtained by removing from $\mathbf{X}$ those columns j for which $\gamma_j = 0$. Similarly, for regression coefficient vector $\boldsymbol{\beta}$, $\boldsymbol{\beta}_\gamma$ is obtained from $\boldsymbol{\beta}$ by removing components β_j for which $\gamma_j = 0$.

Given model γ we have the reduced linear model

$$\mathbf{Y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}. \tag{2.31}$$

All equations in Sections 2.3.1 and 2.3.2 still apply here, with $\mathbf{X}$, $\boldsymbol{\beta}$ and M replaced by $\mathbf{X}_\gamma$, $\boldsymbol{\beta}_\gamma$ and γ respectively. In particular all the model scoring methods described above can be applied in the variable selection setting.

2.3.3.1 Greedy search methods

For p potential predictors there are 2^p possible models. Hence the number of models grows exponentially with p, precluding an exhaustive search of the model space Γ for all but small p; Table 2.1 shows the size of the model space for various values of p. When an exhaustive search is not possible, search algorithms can be used to find high scoring models.

One of the most common search methods is forward stepwise selection. This starts with the empty model containing no predictors, and then adds the single predictor that leads to the most improvement in the chosen scoring criterion. This process continues, adding one predictor at a time, until a stopping rule terminates the algorithm (e.g. no improvement in score can be achieved by adding any single predictor). Backwards stepwise selection is a similar approach which starts with the full model containing all predictors and removes one predictor at a time. These greedy algorithms are relatively fast but are not guaranteed to find the highest scoring model since the additions/deletions at each step are only locally optimal. Results are also dependent on the scoring criteria and stopping rule.

These heuristic algorithms are generally used with non-Bayesian scoring methods. We discuss Bayesian variable selection below, giving some details further to those provided on Bayesian model selection and averaging in Section 2.3.2.

2.3.3.2 Bayesian variable selection

The posterior distribution over models is given by Bayes' theorem in (2.22), which we reproduce here (up to proportionality) using our variable selection notation,

$$P(\gamma \mid \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma) P(\gamma). \tag{2.32}$$

Placing a conjugate $NIG(\mathbf{m}, \mathbf{V}, a, b)$ prior (2.11) on parameters $\Theta_\gamma = (\boldsymbol{\beta}_\gamma, \sigma^2)$ yields a closed-form expression for the marginal likelihood integral (2.23) [see e.g. Denison *et al.*, 2002],

$$p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma) = \frac{|\mathbf{V}^*|^{\frac{1}{2}}\, b^a \Gamma(a^*)}{|\mathbf{V}|^{\frac{1}{2}}\, \pi^{\frac{n}{2}} \Gamma(a)} (b^*)^{-a^*} \tag{2.33}$$

where $\mathbf{V}^*, a^*$ and b^* are given in (2.14)-(2.17) (with $\mathbf{X}$ replaced by $\mathbf{X}_\gamma$).

There is a significant body of research investigating parameter prior specification for Bayesian variable selection in the linear model, and in particular, the choice of hyperparameters in the NIG prior. We first discuss the Gaussian prior for $\boldsymbol{\beta}_\gamma$. There are two popular choices for $\mathbf{V}$, the prior covariance of $\boldsymbol{\beta}_\gamma$; $\mathbf{V} = c\mathbf{I}_{|\gamma|}$ [George and McCulloch, 1997; Raftery *et al.*, 1997; Li and Zhang, 2010] and $\mathbf{V} = c\left(\mathbf{X}_\gamma^\mathsf{T} \mathbf{X}_\gamma\right)^{-1}$ [George and McCulloch, 1997; Smith and Kohn, 1996; Raftery *et al.*, 1997; Lee *et al.*, 2003; Nott and Green, 2004], where $c > 0$. The former renders the components of $\boldsymbol{\beta}$ conditionally independent given γ, while the latter results in Zellner's g-prior [Zellner, 1986], which uses the predictor data to introduce prior dependence between components in an intuitive manner (it is proportional to the variance of the MLE for $\boldsymbol{\beta}_\gamma$). We use the latter formulation in Chapters 3 and 4 and provide further details in Section 3.2.1. The most common choice for $\mathbf{m}$, the prior mean of $\boldsymbol{\beta}_\gamma$, is $\mathbf{m} = \mathbf{0}$. This is a neutral choice reflecting indifference between positive and negative values of regression coefficients; we use $\mathbf{m} = \mathbf{0}$ in our analyses. Using the MLE for $\boldsymbol{\beta}_\gamma$ has also been proposed [Kohn *et al.*, 2001], resulting in a prior that depends on the response data $\mathbf{Y}$. Priors with dependence on the data can no longer be considered as Bayesian in the strictest sense and are referred to as *empirical Bayes* priors (see Section 2.3.8). We now turn our attention to the hyperparameters a, b

of the inverse-gamma prior for σ^2. Several heuristic methods have been proposed to set a and b based on data considerations [George and McCulloch, 1997; Raftery *et al.*, 1997; Chipman *et al.*, 2001]. However, the choice can be avoided by using $p(\sigma^2 \mid \gamma) \propto \sigma^{-2}$, the limit of the inverse-gamma prior as $a, b \to 0$ [Smith and Kohn, 1996; Nott and Green, 2004]. The resulting prior is non-informative and improper, but still yields a proper posterior. We use this limiting case in Chapters 3 and 4.

Calculating the posterior distribution over models (2.32) requires specifying the model prior $P(\gamma)$. A common choice of prior assumes that the *a priori* inclusion probabilities $P(\gamma_j)$ are independent and Bernoulli distributed with success parameter π_j.

$$P(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j}. \tag{2.34}$$

The prior is also often simplified to have $\pi_j = \pi$, resulting in a single hyperparameter. This hyperparameter may be a user-defined constant or may itself have a Beta prior [Nott and Green, 2004]. In the former case, taking $\pi = \frac{1}{2}$ results in a flat prior over Γ, a popular default non-informative choice [George and McCulloch, 1997; Smith and Kohn, 1996], or small values are often chosen to promote parsimonious models [George and McCulloch, 1997; Lee *et al.*, 2003] (it could be argued that this is not necessary due to the marginal likelihood already penalising complex models). Another option is to use an empirical Bayes approach to set π in an objective data-driven manner, as described in George and Foster [2000]. Empirical Bayes approaches are discussed in Section 2.3.8. We do not use this standard prior formulation in Chapter 3, but exploit domain knowledge in the form of protein signalling networks to specify a biologically informative model prior.

While greedy search methods can be used to find models with high posterior probabilities (easily calculated up to proportionality using (2.32)), a preferable approach is to take model uncertainty into account by model averaging. In particular, averaging over the entire space of models enables calculation of posterior inclusion probabilities for each individual predictor,

$$P(\gamma_j = 1 \mid \mathbf{Y}, \mathbf{X}) = \sum_{\gamma : \gamma_j = 1} P(\gamma \mid \mathbf{Y}, \mathbf{X}). \tag{2.35}$$

This is a special case of (2.29) with ζ as the feature '$\gamma_j = 1$'. These inclusion probabilities are a measure of the importance of each individual predictor in determining the response.

When p is too large to enumerate the entire posterior over models, Markov chain Monte Carlo (MCMC) [Robert and Casella, 2004] can be used to sample from the posterior. The essence of this approach is to construct and simulate from a Markov chain with state space Γ and stationary distribution $P(\gamma|\mathbf{Y},\mathbf{X})$. Simulating the chain for a sufficiently large number of iterations allows samples to be drawn from the posterior distribution. These samples, which we denote by $\gamma^{(1)},\ldots,\gamma^{(T)}$, can then be used to calculate asymptotically valid estimates for the posterior expectation $\mathbb{E}\left[\phi(\gamma)\right]_{P(\gamma\,|\,\mathbf{Y},\mathbf{X})}$ of any function $\phi(\gamma)$,

$$\hat{\mathbb{E}}\left[\phi(\gamma)\right]_{P(\gamma\,|\,\mathbf{Y},\mathbf{X})} = \frac{1}{T}\sum_{t=1}^{T}\phi(\gamma^{(t)}). \tag{2.36}$$

In particular, from (2.28), we see that taking $\phi(\gamma)$ to be the indicator function $\mathbb{1}_{\gamma}(\gamma_j = 1)$ yields an estimate for the inclusion probabilities,

$$\hat{P}(\gamma_j = 1\,|\,\mathbf{Y},\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\gamma^{(t)}}(\gamma_j^{(t)} = 1). \tag{2.37}$$

One approach to construct and simulate from such a Markov chain is the popular ‘Markov chain Monte Carlo model composition’ (MC^3) method proposed by Madigan *et al.* [1995] for model selection with graphical models (see Section 2.3.5), and subsequently adapted for variable selection in linear regression models by Raftery *et al.* [1997]. MC^3 uses a Metropolis-Hastings sampler; in each iteration a model is drawn from the ‘neighbourhood’ of the current model according to a proposal distribution, and then accepted or rejected according to an ‘acceptance probability’, which is defined in such a way as to ensure convergence of the chain to the posterior. In the variable selection case the neighbourhood of a given model consists of any model that can be obtained from the given model by adding or removing a predictor. See Raftery *et al.* [1997] for full details. Assessing convergence of MCMC algorithms is well-known to be a non-trivial problem. Many diagnostics verify necessary, but not sufficient conditions for convergence.

Alternatives to the above MC^3 approach include the Occam’s window method proposed by Madigan and Raftery [1994] which averages over a subset of models by eliminating those with low posterior probabilities and those that are unnecessarily complex, and the stochastic search variable selection (SSVS) method of George and McCulloch [1993]. SSVS is similar to MC^3, but does not completely remove predictors from the model. Instead, predictors are assigned coefficients with very small values and MCMC is used to sample from the joint space of models and parameters. In Chapter 3 we calculate exact inclusion probabilities by restricting the size of the model space; an approach similar in spirit to Occam’s window.

The posterior predictive distribution $p(Y'|\mathbf{X}',\mathbf{Y},\mathbf{X},\gamma)$ for new data $(Y',\mathbf{X}')$

given in (2.18) can be combined with model averaging to calculate the expected value of Y', taking model uncertainty into account,

$$\mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{Y}, \mathbf{X}\right] = \sum_{\gamma} \mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{Y}, \mathbf{X}, \gamma\right] P(\gamma \mid \mathbf{Y}, \mathbf{X}) \tag{2.38}$$

where $\mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{Y}, \mathbf{X}, \gamma\right] = \mathbf{X}'\mathbf{m}^*$ when a conjugate NIG parameter prior is used.

An overview of Bayesian variable selection, including prior specification and details of computational aspects can be found in Chipman *et al.* [2001] and Clyde and George [2004]. Bayesian variable selection methods have been applied to address various questions in molecular biology (see Chapter 3 for references).

2.3.3.3 Shrinkage methods

The variable selection approaches outlined above all work by either including or excluding predictors in a model, and searching (or averaging) over model space. An alternative is shrinkage methods, which regularise the problem by shrinking regression coefficients $\boldsymbol{\beta}$ towards zero. This discourages complex models because models that overfit tend to have larger coefficients. Shrinkage is achieved through minimisation of a penalised negative log-likelihood (penalised sum of squares), with a penalty on the magnitude of the coefficients,

$$l(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda p(\boldsymbol{\beta}) \tag{2.39}$$

where $p(\boldsymbol{\beta})$ is the penalty term and $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage. Minimising (2.39) is equivalent to minimising $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ subject to the constraint $p(\boldsymbol{\beta}) \leq t$ (where t depends on λ). The tuning parameter is often set using cross-validation or BIC.

Ridge regression is a popular approach that uses an L_2 penalty, $p(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^p \beta_j^2$, and shrinks each coefficient in proportion to its magnitude. However, it does not shrink coefficients to be exactly zero, so while it regularises the problem and allows estimation in small sample size settings, it does not perform variable selection as such. To obtain a subset of predictors, absolute coefficients can be thresholded.

Lasso (least absolute shrinkage and selection operator) regression [Tibshirani, 1996] is similar in spirit to ridge regression but simultaneously performs shrinkage and variable selection. It uses an L_1 penalty instead of the L_2 penalty used in ridge regression, $p(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. This has the effect of shrinking some coefficients to exactly zero, with the sparsity of the inferred model controlled by λ (or t). Figure 2.8 illustrates, for $p = 2$, why the lasso has this effect while ridge

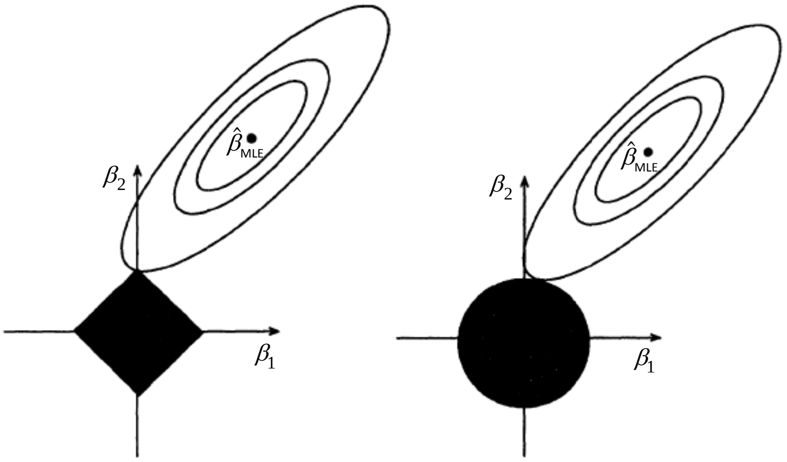


Figure 2.8: **Estimation picture for the lasso (left) and ridge regression (right).** For two predictors ($p = 2$), contours are shown for the residual sum of squares $\sum_{i=1}^{n}(\mathbf{Y} - \mathbf{X}_i\boldsymbol{\beta})^2$ and the shaded areas are constraint regions $p(\boldsymbol{\beta}) \leq t$ where $p(\boldsymbol{\beta}) = |\beta_1| + |\beta_2|$ for lasso and $p(\boldsymbol{\beta}) = \beta_1^2 + \beta_2^2$ for ridge regression. (Figure adapted from Tibshirani [1996]).

regression does not. Contours of the residual sum of squares $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ are elliptical and centred at the (unpenalised) MLE $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, and the constraint region $p(\boldsymbol{\beta}) \leq t$ is a diamond for lasso and a disk for ridge regression. The solution to (2.39) is the first place where the contours hit the constraint region. For lasso, this can happen at a corner, resulting in a zero coefficient.

As noted by Tibshirani [1996], the lasso estimate for $\boldsymbol{\beta}$ can also be interpreted as the Bayesian MAP estimate (that is, the mode of the posterior distribution $p(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})$; see (2.7)) under independent Laplace priors on the individual coefficients β_j. Indeed, fully Bayesian analyses for the lasso have been proposed in the literature [Park and Casella, 2008]. While Bayesian approaches can provide some advantages (e.g. intuitive estimates of uncertainty and taking uncertainty into account), they are generally computationally expensive. This is in contrast to the lasso regression problem, which can be solved very efficiently using a slightly modified version of the least angle regression (LARS) algorithm [Efron *et al.*, 2004].

LARS is similar to forward stepwise selection (see Section 2.3.3.1), but does not completely add predictors into the model. It begins by adding to the model the predictor most correlated with the response. Coefficients of included predictors are moved towards the least squares estimate, until a predictor not in the model has higher correlation with the residual. This predictor is then added to the model.

An alternative efficient method for solving the lasso regression problem is with a coordinate descent approach [Friedman *et al.*, 2007, 2010]. This approach minimises the penalised negative log-likelihood (2.39) one parameter (i.e. one component of $\boldsymbol{\beta}$) at a time, whilst keeping all other parameters fixed.

Lasso-based variable selection has also been used to probe questions in molecular cancer biology. For example, Li and Li [2008] use lasso regression to identify genes that are related to survival time from glioblastoma (the most common primary malignant brain tumour).

2.3.4 Graphical models

In this Section we give a brief overview of graphical models. Full technical details can be found in Pearl [1988]; Lauritzen [1996]; Jordan [2004] and Koller and Friedman [2009].

As described in the Introduction, graphical models are a class of statistical models that use a graph-based representation to intuitively and compactly describe probabilistic relationships between multiple interacting components. They consist of a graph $G = (V, E)$, in which each node (or vertex) in V corresponds to a random variable of interest (describing phosphorylation level of a protein, for example) and the edges E represent dependencies between these variables. Formally, the graph structure encodes conditional independence statements regarding the variables. There are two main types of graphical models, Bayesian networks (BNs) and Markov random fields (MRFs) (also called Markov networks), and we use both types in this thesis. Bayesian networks are directed graphical models, whereas Markov random fields have undirected edges. Below we describe each type, and in particular, the conditional independencies they encode, and explain the relationship between them.

2.3.4.1 Bayesian networks

BNs are directed acyclic graphs (DAGs). That is, the edges between nodes are directed and they do not form directed cycles. Figure 2.9 shows an example of a BN with five nodes, $X_1, \ldots, X_5$, and five edges. Note that the same notation is used to represent the node and the associated random variable. Before proceeding we define some graph terminology. If a directed edge exists from node X_i to node X_j, then X_i is called a *parent* of X_j, and X_j is called a *child* of X_i (e.g. in Figure 2.9, X_1 is a parent of X_2). Similarly, if a directed path exists between node X_i and node X_j, then X_i is called an *ancestor* of X_j, and X_j is called a *descendant* of X_i (e.g. in Figure 2.9, X_4 is a descendant of X_1).

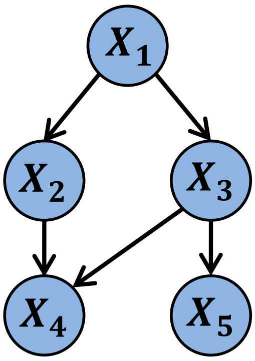


Figure 2.9: **Example of a 5-node Bayesian network (BN).** The edge structure of the BN enables the joint distribution over all variables $X_1, \ldots, X_5$ to be factorised into a product of local conditional distributions (see (2.40) and (2.41)).

The graph structure implies the following 'local Markov property': Each variable is conditionally independent of its non-descendants given its parents. Importantly, this enables us to factorise the global joint distribution over all variables $X_1, \ldots, X_p$ (the likelihood) into a product of local conditional distributions, with each variable dependent only on its parents,

$$p(X_1, \ldots, X_p \mid G, \Theta) = \prod_{j=1}^{p} p(X_j \mid X_{\pi_G(j)}, \theta_j) \tag{2.40}$$

where $\pi_G(j) \subseteq \{1, \ldots, p\}$ is an index set for the parents of node X_j, $X_{\pi_G(j)} = \{X_k | k \in \pi_G(j)\}$ is a corresponding data set including only those variables in $\pi_G(j)$, and $\theta_j \subseteq \Theta$ are parameters for the local conditional distribution for X_j. For example, for our toy BN in Figure 2.9 we obtain the following factorisation (parameters have been suppressed here),

$$p(X_1, \ldots, X_5 \mid G) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1)p(X_4 \mid X_3, X_2)p(X_5 \mid X_3). \tag{2.41}$$

We note that the local Markov property conditional independence statements are not the only independencies implied by the graph structure. The notion of graphical separation between (sets of) nodes in the graph can be used to determine global conditional independences. In BNs this is known as *d-separation* (directed separation); if A, B, C are three sets of nodes in G and A and B are d-separated given C, then A and B are conditionally independent given C [see e.g. Koller and Friedman, 2009, for full details].

To fully specify the BN and how exactly the variables depend on each other, it is necessary to specify the functional form of the local conditional distributions and associated parameters, which are then sufficient to fully determine the global joint distribution via (2.40). The factorisation enables the joint distribution to be modelled with far fewer parameters than would be required if modelled directly, and so helps to avoid overfitting. Common choices for the local conditional distributions are multinomial for a discrete variable and Gaussian for a continuous variable. In the discrete case, there are only a finite number of possibilities for the values of each variable. This means a local conditional distribution can be fully represented by a table that specifies probabilities for values of a variable given each possible combination of values for its parent variables. In the continuous case there is no such tabular representation; a variable depends on its parent variables through the mean of the Gaussian distribution. For example, for our toy BN in Figure 2.9 a conditional could be $p(X_4|X_2, X_3, \theta_4) = \mathcal{N}(X_4|\beta_2 X_2 + \beta_3 X_3, \sigma_4^2)$ for some parameters $\beta_2, \beta_3, \sigma_4^2$. We focus on continuous variables and Gaussian distributions throughout this thesis.

The correspondence between graph structures and conditional independence statements implied by the graph is not one-to-one; different graph structures can represent the same set of conditional independence assertions. Such graphs are said to be equivalent. Verma and Pearl [1990] derived the following useful characterisation for equivalent BNs: two BNs are equivalent if and only if they have the same skeleton and the same v-structures. The *skeleton* of a graph is the undirected graph obtained by converting all directed edges to undirected edges, while a *v-structure* is a set of three nodes X_1, X_2, X_3 such that X_1 and X_2 are parents of X_3 and there is no edge between X_1 and X_2. For example, Figure 2.10 shows four 3-node BNs, all with the same skeleton structure. The first three graphs (a,b and c) do not contain a v-structure, while the fourth graph (d) does. Hence, it follows from the above characterisation that the first three graphs are equivalent and the fourth graph is not equivalent to any of the others. This can also be seen directly by factorising the joint distribution into a product of local distributions using (2.40); these factorisations are shown in Figure 2.10. Following an application of Bayes' theorem we see that the first three graphs result in an identical factorisation, and this factorisation differs from that for the fourth graph. The space of graph structures can be partitioned into equivalence classes; disjoint subsets where all graphs in the same subset are equivalent [Chickering, 1995]. An equivalence class can be represented by a *completed partially directed acyclic graph* (CPDAG), an acyclic graph with directed and undirected edges, where a directed edge from X_j to X_k means all BN structures in the equivalence class have that directed edge, while an undirected between X_j and

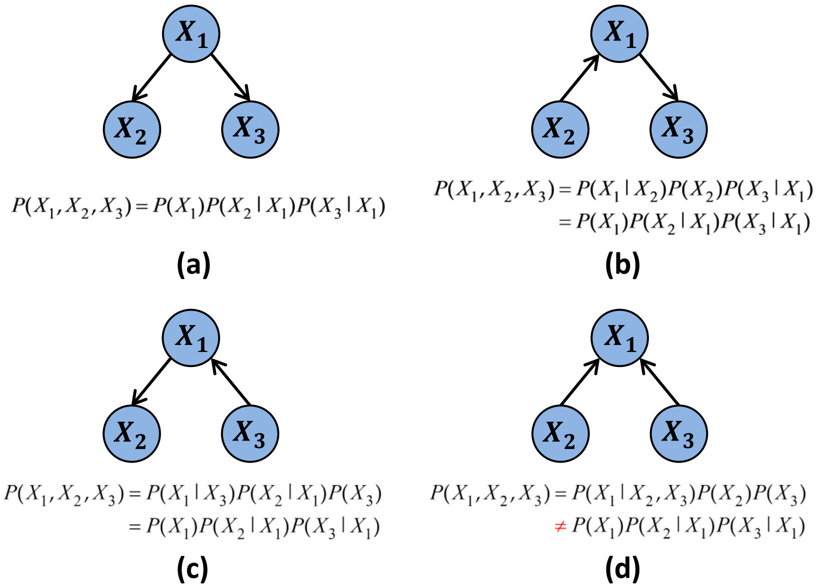

X_1
X_2
X_3
$P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)$
(a)
X_1
X_2
X_3
$P(X_1, X_2, X_3) = P(X_1 \mid X_2)P(X_2)P(X_3 \mid X_1)$
$= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)$
(b)
X_1
X_2
X_3
$P(X_1, X_2, X_3) = P(X_1 \mid X_3)P(X_2 \mid X_1)P(X_3)$
$= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)$
(c)
X_1
X_2
X_3
$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3)P(X_2)P(X_3)$
$\neq P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)$
(d)

Figure 2.10: **Equivalence of BNs.** Four 3-node BNs are shown, each with the same skeleton structure. Below each graph the factorisation of the joint distribution implied by the graph structure is given (see (2.40)). Applying Bayes' theorem to the factorisation in (b) and (c), we see that BNs (a), (b) and (c) are equivalent (describe the same independencies), whereas BN (d) describes a different set of independencies.

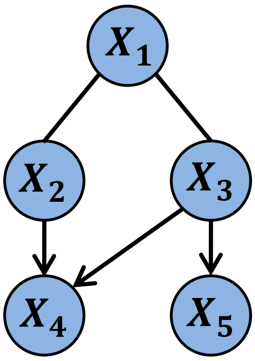


Figure 2.11: **CPDAG for the equivalence class containing the BN in Figure 2.9.**

X_k means some BN structures in the equivalence class have a directed edge from X_j to X_k and others have it in the opposite direction from X_k to X_j. Figure 2.11 shows the CPDAG representation for the equivalence class containing the example BN in Figure 2.9. This equivalence between BNs has implications for the inference of graph structure from data and for the interpretation of edges in the graph as representing causal relationships; we discuss this further below.

2.3.4.2 Markov random fields and Gaussian graphical models

MRFs are undirected graphical models that, like BNs, describe conditional independencies between variables. Figure 2.12 shows an example of a MRF with five nodes and six edges. To describe the independencies we'll need the following terminology. Two nodes that are connected by an edge in G are said to be *adjacent* or *neighbours* in G. If A, B, C are three sets of nodes in G, then C *separates* A and B if any path between a node in A and a node in B contains a node in C (this is analogous to the notion of d-separation for BNs). The graph structure encodes the following three Markov properties, which can be shown to be equivalent (under mild conditions) [see e.g. Koller and Friedman, 2009]. Examples of each, relating to Figure 2.12, are given in parentheses, with $\perp\!\!\!\perp$ denoting independence.

- *Pairwise Markov property*: Any two non-adjacent variables are conditionally independent given all other variables. ($X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3, X_5$.)

- *Local Markov property*: A variable is conditionally independent of all other variables given its neighbours. ($X_4 \perp\!\!\!\perp X_1, X_5 \mid X_2, X_3$.)

- *Global Markov property*: Any two sets of variables A, B are conditionally independent given a set of variables that separate A and B. ($A = \{X_1\}, B = \{X_4, X_5\}, C = \{X_2, X_3\}$: $A \perp\!\!\!\perp B \mid C$.)

In the case where the joint distribution over all variables $(X_1, \ldots, X_p)$ takes a multivariate Gaussian distribution, MRFs are referred to as Gaussian MRFs or Gaussian graphical models (GGMs) [Dempster, 1972; Speed and Kiiveri, 1986; Rue and Held, 2005]. We use the latter label throughout this thesis, but note that the former is perhaps a more precise description. There is a direct relationship between the structure of a GGM and the covariance matrix of the Gaussian distribution, which we denote by $\mathbf{\Sigma}$. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the inverse covariance matrix, also known as the precision matrix, and let ρ_{jk} be the partial correlation coefficient between variables X_j and X_k given all other variables. This is a measure of association between X_j and X_k, after having removed the effects of all other variables. For example, X_j and X_k could be highly correlated (i.e. have a Pearson correlation coefficient of close to one), but the association may be explained away by both of them depending on a third variable X_l. In this case, the partial correlation ρ_{jk} would be less than the standard correlation as the effect of all other variables (including X_l) has been taken into account. For variables that are jointly Gaussian distributed, zero partial correlation $\rho_{jk} = 0$ is equivalent to conditional independence of X_j and X_k given all other variables. Hence, from the pairwise Markov property above we have $\rho_{jk} = 0$ if and only if no edge exists in G between X_j and X_k (i.e. $(j,k) \notin E$). Moreover, since ρ_{jk} has the following relationship with the precision matrix $\mathbf{\Omega} = (\omega_{jk})$,

$$\rho_{jk} = -\frac{\omega_{jk}}{\sqrt{\omega_{jj}\omega_{kk}}} \tag{2.42}$$

we find that non-zero (off-diagonal) entries in the precision matrix correspond to edges in a GGM, that is $\omega_{jk} \neq 0$ if and only if $(j,k) \in E$.

2.3.4.3 Relationship between Bayesian networks and Markov random fields

BNs and MRFs do not necessarily describe the same independencies. Indeed, each can encode conditional independence statements that the other cannot. However, they are related and, since we use both BNs and MRFs in this thesis, we shall outline some details of the relationship between them.

Given a BN graph structure G, a corresponding MRF structure $\tilde{G}$ can be obtained by finding the *moral graph* of G. This is done by taking the skeleton of G and then adding an undirected edge between nodes X_j and X_k if they are both parents of the same node in G. The example MRF in Figure 2.12 is the

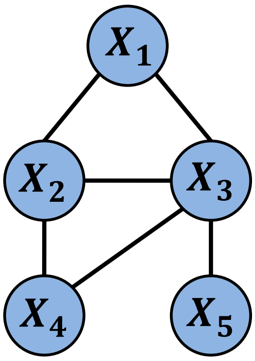


Figure 2.12: **Example of a 5-node Markov random field (MRF).** Any pair of adjacent nodes are conditionally independent given all other variables. This graph structure is the moral graph of the BN in Figure 2.9.

moral graph of the example BN in Figure 2.9. All the independencies represented by the moral graph $\tilde{G}$ are also represented by G, and $\tilde{G}$ is minimal in the sense that this is no longer true if any edge from $\tilde{G}$ is removed. To illustrate this point, suppose the edge between X_2 and X_3 in Figure 2.12 is removed. Then we would have $X_2 \perp\!\!\!\perp X_3 \,|\, X_1, X_4$, but this does not hold for the BN in Figure 2.9. However, the addition of edges in the process of obtaining the moral graph, such as the edge between X_2 and X_3 in Figure 2.12, means independence information is lost. While, as just noted, the addition of such an edge is necessary so that $\tilde{G}$ is minimal, it means there are independencies described by the BN G that are not described by the MRF $\tilde{G}$. For example, in Figure 2.9 we have $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$, but this does not hold in Figure 2.12. In general, the MRF $\tilde{G}$ (the moral graph of G) encodes the same conditional independencies as G if $\tilde{G}$ has the same skeleton as G (i.e. no edges are added).

A BN structure G can also be obtained from a MRF structure $\tilde{G}$, although the process is not as straight-forward. In brief, an ordering is placed on the nodes so that directed edges can only go from X_j to X_k if X_j precedes X_k in the ordering. The edges in $\tilde{G}$ can then be converted to directed edges according to this ordering. However, some additional edges may need to be added to avoid G having independencies not described by $\tilde{G}$. Again, this addition of edges leads to a loss of independence information. A BN exists that encodes the same independencies as the MRF if and only if any cycles of greater than 3 nodes in the MRF contain a ‘shortcut’ (an edge between two non-adjacent nodes in the cycle).

2.3.4.4 Inference and learning

There are three main tasks that can be performed using graphical models; inference, parameter learning and structure learning [see e.g Koller and Friedman, 2009]. We outline each of these in turn.
Inference: The graph structure and parameters are known, providing a complete model for the relationships between variables. This can then be used to answer queries. For example, finding the posterior distribution of some variables when other variables are observed.
Parameter learning: The graph structure is known, describing the existence of dependencies between variables, but the parameters are unknown and need to be learned from data. For example, for Bayesian networks, the functional form of the local conditionals are usually assumed to belong to some family of distributions, but the parameters that fully specify the conditionals, and hence the nature of relationships between variables, are learned from data. Both maximum likelihood methods and Bayesian inference methods are often used for this purpose.
Structure learning: Also referred to as network inference, structure learning is the process of learning the graph structure from data; that is, finding a graph structure that describes the conditional independencies present in the data. This is the task we perform in this thesis. Parameters of the underlying distributions are often also considered as part of this process. Different approaches exist for BNs and MRFs; we discuss some of these approaches in Sections 2.3.5 and 2.3.6.

2.3.4.5 Applications

Graphical models have been employed as a modelling tool in a wide range of applications. For example, BNs have been applied to molecular networks, and in particular the inference of network structure [Friedman *et al.*, 2000; Sachs *et al.*, 2005] (we apply BNs for this purpose in Chapter 4), information retrieval, risk management, clinical decision support and forensic science [Pourret *et al.*, 2008]. MRFs have been applied to image processing and computer vision [Li, 2009], and to applications in statistical physics using the Ising model (one of the earliest types of MRFs).

When systems under study contain variables that interact with a natural directionality, for example a flow of information between variables, then BNs may provide a better model than MRFs. Similarly, MRFs may be a better choice when interactions are more symmetrical in nature. However, wider modelling choices and inference or learning methods can also influence the choice of graphical model. Indeed, in Chapter 5 we use undirected GGMs to model signalling networks (which have a clear natural directionality), because they naturally fit into the mixture model-based framework we employ there and estimation of GGM structure can be very computationally efficient.

2.3.5 Graph structure learning: Bayesian networks

Methods for learning the graph structure of Bayesian networks fall under two main categories; constraint-based methods and score-based methods. We describe both below, but focus mainly on score-based methods, which we shall use in Chapter 4. The reader is also referred to Heckerman [1998] and Needham *et al.* [2007] for tutorials on learning with BNs.

2.3.5.1 Constraint-based methods

Constraint-based methods use statistical tests to find conditional independencies in the data. A BN structure (or, more precisely, an equivalence class of BN structures) can then be constructed that reflects the inferred independencies. The algorithms are based on the inductive causation (IC) algorithm by Pearl and Verma [1991] and consist of three main steps. First, the skeleton (undirected graph) of the BN structure is determined using independence tests of the form $X_i \perp\!\!\!\perp X_j \mid A$ where A is a subset of other variables. If no A is found such that independence holds, then an undirected edge is added between X_i and X_j. Restrictions are usually placed on A to reduce the number of independence tests required, which would otherwise be prohibitively large. For example, a constraint may be placed on the cardinality of A. Second, triplets of nodes in the skeleton that could potentially form a v-structure are considered. The independence tests carried out in the previous step can be used to determine their existence, and edges are assigned directions accordingly. Third, some of the remaining undirected edges in the graph can be assigned a direction in cases where directionality is compelled due to the acyclicity constraint and the fact that no further v-structures can be created (as this would contradict the second step). This results in a CPDAG representation for the equivalence class.

There are several choices for the independence tests in the first step. Common choices are Pearson's χ^2 test and mutual information tests (equivalent to a log-likelihood ratio test) for discrete data, and Student's t test, Fisher's Z test and mutual information tests (all three are based on partial correlation coefficients) for continuous data [see e.g. Lehmann and Romano, 2008, for details of statistical independence tests].

Constraint-based approaches are intuitive because they closely follow the conditional independence interpretation of a Bayesian network. Moreover, they can be computationally efficient and some enjoy asymptotic guarantees [Kalisch and Bühlmann, 2007]. However, in the challenging small sample size and noisy data setting, it is by no means guaranteed that constraint-based methods perform well. Since BNs are constructed from results of many independence tests (i.e. a multiple hypothesis testing problem), the overall results can be sensitive to failures in these tests.

Specific examples of constraint-based algorithms for BN structure learning include the path consistency (PC) algorithm [Spirtes *et al.*, 1993], grow-shrink (GS) algorithm [Margaritis and Thrun, 2000] and incremental association Markov blanket (IAMB) algorithm [Tsamardinos *et al.*, 2003]. Such constraint-based methods have been used to infer molecular networks, and in particular, gene regulatory networks, with recent examples including Li *et al.* [2011] and Zhang *et al.* [2012].

2.3.5.2 Score-based methods

Score-based methods for BN structure learning use model selection criteria (see Section 2.3.2) to score candidate graph structures given the observed data and searching over the space of graphs to find those with high scores. We denote graph structures by G and the graph space by $\mathcal{G}$ (instead of M and $\mathcal{M}$ respectively in Section 2.3.2), and let $\mathbf{X}$ denote a $n \times p$ data matrix, where p is the number of variables (nodes in the BN) and n is the sample size.

Scoring functions: Any of the scoring methods described in Section 2.3.2, taking fit to data and model complexity into account, can be applied to score BN structures. The methods described were AIC, BIC, CV and Bayesian scoring. The likelihood function for a BN $p(\mathbf{X} \mid G, \mathbf{\Theta})$ is given in (2.40) and factorises into a product of local conditional distributions for each variable. This allows maximum likelihood estimates for parameters (for a given graph G), required for AIC and BIC, to be obtained by independently maximising each local distribution. We note that the 'Bayesian' in 'Bayesian network' does not refer to the methodology of parameter or structure learning for BNs; non-Bayesian approaches are equally applicable, although the Bayesian scoring approach we now describe is widely-used.

The Bayesian score for a BN with graph structure G is given by its posterior probability, given by Bayes' theorem in (2.22) and reproduced here (up to proportionality) using the graph notation,

$$P(G \mid \mathbf{X}) \propto p(\mathbf{X} \mid G) P(G). \tag{2.43}$$

Similarly, the marginal likelihood integral in (2.23) becomes

$$p(\mathbf{X} \mid G) = \int p(\mathbf{X} \mid G, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \mid G) \mathrm{d}\boldsymbol{\Theta}. \tag{2.44}$$

For continuous data, $p(\mathbf{X} \mid G)$ can be calculated with the widely-used 'BGe' score, proposed by Geiger and Heckerman [1994]. We give a brief outline of this scoring metric here, and refer the interested reader to the reference for full details. Each of the local conditionals in the factorised likelihood (2.40) is assumed to be a linear Gaussian distribution, with mean of X_j dependent only on the values of its parents,

$$p(X_j \mid X_{\pi_G(j)}, \theta_j) = \mathcal{N}\left(X_j \mid m_j + \sum_{k=1}^{j-1} \beta_{jk}(X_k - m_k), \sigma_j^2 \right) \tag{2.45}$$

where m_j is the unconditional mean of X_j, σ_j^2 is conditional variance, and β_{jk} are coefficients reflecting the strength of the dependence between X_j and X_k. Hence we have $\beta_{jk} \neq 0$ if and only if $k \in \pi_G(j)$. It is also assumed, without loss of generality, that the variables are labelled so that $\pi_G(j) \subseteq \{1, \ldots, j-1\}$. Each of these local conditionals can be regarded as a linear model.

Linear Gaussian local conditional distributions of the form (2.45) result in a joint multivariate Gaussian likelihood $p(X_1, \ldots, X_p \mid G, \boldsymbol{\Theta})$ with mean $\mathbf{m} = (m_1, \ldots, m_p)$ and covariance $\boldsymbol{\Sigma}$. The precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ can be found in terms of the local conditional parameters β_{jk} and σ_j^2 using the following recursive formula,

$$\boldsymbol{\Omega}(j) = \begin{pmatrix} \boldsymbol{\Omega}(j-1) + \sigma_j^2 \boldsymbol{\beta}_j \boldsymbol{\beta}_j^\mathsf{T} & -\sigma_j^2 \boldsymbol{\beta}_j \\ -\sigma_j^2 \boldsymbol{\beta}_j^\mathsf{T} & \sigma_j^2 \end{pmatrix} \tag{2.46}$$

for $j = 2, \ldots, p$, where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j,j-1})^\mathsf{T}$ and $\boldsymbol{\Omega}(1) = \sigma_1^2$. $\boldsymbol{\Omega}(p)$ gives the precision matrix for the multivariate Gaussian. For example, if we take the example BN given in Figure 2.9 with linear Gaussian local conditionals of the form (2.45), then the joint distribution $p(X_1, \ldots, X_5 \mid G, \boldsymbol{\Theta})$ is a multivariate Gaussian with precision matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2 + \beta_{21}^2\sigma_2^2 + \beta_{31}^2\sigma_3^2 & -\beta_{21}\sigma_2^2 & -\beta_{31}\sigma_3^2 & 0 & 0 \\ -\beta_{21}\sigma_2^2 & \sigma_2^2 + \beta_{42}^2\sigma_4^2 & \beta_{42}\beta_{43}\sigma_4^2 & -\beta_{42}\sigma_4^2 & 0 \\ -\beta_{31}\sigma_3^2 & \beta_{42}\beta_{43}\sigma_4^2 & \sigma_3^2 + \beta_{43}^2\sigma_4^2 + \beta_{53}^2\sigma_5^2 & -\beta_{43}\sigma_4^2 & -\beta_{53}\sigma_5^2 \\ 0 & -\beta_{42}\sigma_4^2 & -\beta_{43}\sigma_4^2 & \sigma_4^2 & 0 \\ 0 & 0 & -\beta_{53}\sigma_5^2 & 0 & \sigma_5^2 \end{pmatrix}.$$

Recall that the non-zero off-diagonal entries of a Gaussian precision matrix correspond to edges in a GGM (see (2.42)). We see that the precision matrix above represents the example GGM structure in Figure 2.12, which is the moral graph for the example BN.

Geiger and Heckerman [1994] first derive a marginal likelihood score $p(\mathbf{X}|G_c)$ for a 'complete' graph structure G_c consisting of all possible edges ($\beta_{jk} \neq 0$ for all $k < j$). The parameter prior $p(\mathbf{\Theta} \mid G_c)$ where $\mathbf{\Theta} = (\mathbf{m}, \mathbf{\Omega})$ is taken to be a normal-Wishart distribution. That is, $\mathbf{m} \mid \mathbf{\Omega} \sim \mathcal{N}(\boldsymbol{\mu}_0, \nu\mathbf{\Omega}^{-1})$ with $\nu > 0$, and $\mathbf{\Omega} \sim \mathcal{W}(\alpha, \mathbf{T}_0)$, a Wishart distribution with $\alpha > p - 1$ degrees of freedom and precision matrix $\mathbf{T}_0$. Geiger and Heckerman [1994] suggest setting the hyperparameters $\boldsymbol{\mu}_0, \nu, \alpha, \mathbf{T}_0$ using a prior Gaussian BN specified by the user. This choice of prior is conjugate for the multivariate Gaussian likelihood $p(\mathbf{X} \mid G_c, \mathbf{\Theta})$, resulting in a closed form marginal likelihood (2.44). We note that this closed form is only obtained when the data is complete (no missing values).

Assumptions of prior parameter independence and parameter modularity are made. Parameter independence means that the parameters θ_j of the local conditional distributions are a priori independent ($p(\mathbf{\Theta} \mid G) = \prod_{j=1}^{p} p(\theta_j \mid G)$), and parameter modularity means that the prior over local parameters θ_j depends only on parent variables (if for two graphs G_1, G_2 we have $\pi_{G_1}(j) = \pi_{G_2}(j)$, then $p(\theta_j \,|\, G_1) = p(\theta_j \,|\, G_2)$). Under these assumptions, the BGe score for any BN structure G can be calculated in closed form using scores for complete graphs,

$$p(\mathbf{X} \mid G) = \prod_{j=1}^{p} \frac{p(\mathbf{X}_{\{j\} \cup \pi_G(j)} \mid G_c)}{p(\mathbf{X}_{\pi_G(j)} \mid G_c)} \tag{2.47}$$

where $\mathbf{X}_A$ for $A \subseteq \{1, \ldots p\}$ is the data $\mathbf{X}$ restricted to only those variables in A. In Chapter 4 we shall use a scoring function that differs from, but is related to, the BGe formulation described here.

A corresponding scoring metric has also been developed for discrete data, using multinomial likelihoods and Dirichlet priors. It was initially proposed by Buntine [1991] and Cooper and Herskovits [1992], then further developed by Heckerman *et al.* [1995]. The resulting scoring function is called the 'BDe' metric.

The Bayesian score (2.43) requires specification of a prior over graph space $P(G)$ (the 'graph prior' or 'network prior'). A common choice is to simply use a flat prior, so that all graphs are *a priori* equally plausible. More informative choices have also been proposed; see Chapter 4, in which we shall use a biologically informative prior, for further discussion. The network prior is usually defined to be modular, which means that it factorises into a product of local priors for the parent

p	2	3	4	5	8	10
$\lvert\mathcal{G}\rvert$	3	25	543	29281	$\approx 7.8 \times 10^{11}$	$\approx 4.2 \times 10^{18}$

Table 2.2: **Number of BN graph structures $|\mathcal{G}|$ for varying number of variables (nodes) p.** Calculated using a recursive formula by Robinson [1973].

set of each variable,

$$p(G) = \prod_{j=1}^{p} P(\pi_G(j)). \tag{2.48}$$

This property, together with the factorisation of the likelihood and parameter modularity, results in the posterior score being modular (also referred to as a decomposable score),

$$P(G \,|\, \mathbf{X}) \propto \prod_{j=1}^{p} p(\mathbf{X}_j \,|\, \pi_G(j)) P(\pi_G(j)) \tag{2.49}$$

where $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^\mathsf{T}$ denotes data for variable j in the BN (column j of $\mathbf{X}$) and $p(\mathbf{X}_j \,|\, \pi_G(j))$ is the contribution of variable X_j and its parents to the marginal likelihood $p(\mathbf{X}|G)$. The modular Bayesian score provides computational gains when searching for the best scoring graph, which we discuss below.

Search methods: As was the case with variable selection, the size of the model space precludes an exhaustive search for the highest scoring BN graph structure. Indeed, the number of graphs grows super-exponentially with the number of variables (see Table 2.2), and this graph space search problem is known to be NP-hard [Chickering *et al.*, 1995]. Hence, heuristic search approaches are required.

One popular heuristic search algorithm is the greedy hill-climbing method, which is a local search procedure [see e.g Heckerman *et al.*, 1995]. At each step in the search it considers the neighbourhood of the current graph G, which contains all graphs that can be obtained from G by removing, adding or reversing the direction of a single edge (and satisfy the acyclicity constraint). The graph in the neighbourhood of G that results in the biggest improvement in the score is selected. The procedure is iterated until no improvement in score is found. Since only single edge changes are considered, if the score is modular, then the score of a proposed graph can be efficiently calculated from the score of the current graph by recalculating only the components of the score that are affected by the edge change. This method is a generalised version of the K2 algorithm proposed by Cooper and Herskovits [1992] in which an ordering over the nodes is assumed and edges are only added. These algorithms are only guaranteed to find local maxima, but methods that have been proposed to help overcome this include random restarts, tabu search and simulated annealing. See Koller and Friedman [2009] for further details and Chickering *et al.* [1995] for a comparison of some of these methods. Friedman *et al.* [2000] combine the greedy hill-climbing method with the *sparse candidate algorithm* [Friedman *et al.*, 1999]. This facilitates efficient learning by placing restrictions on the possible parents for each variable and thereby reducing the size of the graph space.

When using the Bayesian posterior score (2.43), instead of searching for the highest scoring (MAP) graph structure, model uncertainty can be taken into account by performing Bayesian model averaging (see Section 2.29). This proceeds analogously to the Bayesian variable selection approach described in Section 2.3.3. In particular, posterior edge probabilities can be calculated by averaging over the entire posterior distribution,

$$P(e = (j,k) \mid \mathbf{X}) = \sum_{G:e\in G} P(G \mid \mathbf{X}) \tag{2.50}$$

where $e = (j,k)$ denotes an edge from X_j to X_k and $e \in G$ means edge e is contained in graph G.

As described for Bayesian variable selection in Section 2.3.3.2, MCMC can be used to sample from the posterior $P(G \mid \mathbf{X})$ and provide asymptotically valid estimates for the posterior edge probabilities. The popular MC^3 structure MCMC approach [Madigan *et al.*, 1995] also proceeds as described above, except the neighbourhood of a graph G, as defined by Madigan *et al.* [1995], contains those acyclic graphs that can be obtained from G by adding or removing an edge. Giudici and Castelo [2003] extended the definition of the neighbourhood to include edge reversals. These are the same local moves as used in the greedy search algorithms and so the computational gains resulting from the modularity of the Bayesian score are also experienced here. The main computational bottleneck in the MCMC algorithm is the large number of acyclicity checks required to find the neighbourhoods.

An alternative MCMC approach, proposed by Friedman and Koller [2003], searches over the space of orders rather than the space of BN graph structures. Given a node ordering $\prec$, a graph G must satisfy the following statement: if $k \in \pi_G(j)$, then $X_k \prec X_j$. In words, for any given variable X_j, its parents must precede X_j in the ordering. Therefore, any graph that is consistent with the order $\prec$ is guaranteed to be acyclic. This enables efficient calculation of summations over the whole graph space when the posterior score is modular (see Friedman and Koller [2003] for further details, and Chapter 4, where this efficiency is also exploited, but derived from a different perspective). Since an ordering is not usually known *a priori*, Friedman and Koller [2003] perform an MCMC search over the $p!$ possible orders, resulting in samples from the posterior $P(\prec \mid \mathbf{X})$, which can then be used to obtain samples from $P(G \mid \mathbf{X})$. This MCMC method is shown to have better convergence and mixing properties than MC^3, but the joint prior distribution over graphs and orders $P(G, \prec)$ introduces bias into results since graphs consistent with more orders are favoured. Several methods have been proposed to ameliorate this bias: Ellis and Wong [2008] have proposed an improvement to order space MCMC, based on importance sampling; Eaton and Murphy [2007a] have combined the exact order-space dynamic programming method of Koivisto [2006] with MCMC over BN graph structures; and Grzegorczyk and Husmeier [2008] propose a new edge reversal move for MCMC over graph structures that results in improved MCMC convergence and mixing, but without the bias of order-space MCMC.

Causality: While it may seem natural to interpret the directed edges in an inferred BN structure as causal interactions between variables (for example, protein X_j directly phosphorylates protein X_k), it is necessary to proceed with caution. Causal interpretations of Bayesian networks, known as *causal networks*, have been proposed in the literature [see e.g. Pearl and Verma, 1991]. In a causal network, parents of a variable are regarded as its immediate causes and a causal Markov assumption is assumed: given a variables immediate causes (parents), it is independent of its earlier causes. Under this assumption, a causal network also satisfies the Markov independencies of the corresponding Bayesian network. However, the converse does not necessarily hold. As we saw in Section 2.3.4, graph structures can be equivalent (represent the same independencies), and so if the scoring function assigns identical scores to equivalent structures (a property known as score equivalence), it is not possible to distinguish between equivalent structures based on the data alone (the Bayesian BGe and BDe scores satisfy this property). As such, the best that can be done is to determine the equivalence class of the true underlying graph structure (i.e. the CPDAG representation). The only edges that can be given a causal interpretation are those that are directed in the CPDAG (as these edges have the same direction in all equivalent graph structures). Due to the importance of equivalence classes, structure learning methods have been proposed that search over the space of equivalence classes [Chickering, 2002].

Interventional data has been shown to be useful in elucidating directionality of edges in Bayesian networks [Cooper and Yoo, 1999; Markowetz *et al.*, 2005; Eaton and Murphy, 2007b]. In an intervention, values of some variables are set by an influence from outside the system (for example, by gene knockouts or protein kinase inhibition), whereas observational data consists of passive measurements of variables. An intervention can break the symmetry within an equivalence class, resulting in different scores being assigned to equivalent graph structures and thereby allowing directionality to be determined.

However, in order to be able to make any causal deductions from the data, it is necessary to assume that there are no hidden or latent variables that explain away the dependence between observed variables. This assumption is very unlikely to hold in the molecular biology domain due to the vast number of molecular players within a cell and the limited number of variables that can be measured in an experiment. Hence, for any inferred edge from X_j to X_k, the possibility that this dependence is indirect and occurs via an unobserved variable X_h cannot in general be ruled out. We note that this effect of hidden variables on the interpretability of results is not unique to Bayesian networks.

Benefits and limitations of BNs: Due to several attractive properties, BNs have been widely-used to infer the structure of molecular networks such as gene regulatory networks [Friedman *et al.*, 2000] and protein signalling networks [Sachs *et al.*, 2005] (see Chapter 4 for further references). Some of the benefits of BNs are: they offer an intuitive graphical representation; the probabilistic framework allows them to handle noisy data; they can model combinatorial relationships between variables; they can describe both direct (causal) interactions and indirect relationships that involve unobserved variables; if a dependence exists between a pair of variables, an edge is not inferred if this dependence is explained by other edges in the graph (for example, in Figure 2.9, X_1 and X_4 may be highly correlated, but no edge exists between them because the dependence is explained by an indirect relationship via X_2); and they are also capable of handling missing data and hidden variables in a principled manner, although the marginal likelihood can no longer be found in closed-form (methods for structure learning in this setting include the structural EM algorithm proposed by Friedman [1998]).

However, BNs also suffer from a number of limitations, of which we consider three here. First, the acyclicity constraint precludes the modeling of feedback loops, which are known to play an important role in regulatory mechanisms within the cell. For example, the graph structure in Figure 2.13(a) contains a feedback loop, but is not a valid BN. Second, BNs are 'static' models that assume samples obtained in an experiment are independent. Thus, they are not particularly suitable for analysing time-resolved molecular datasets, that are now relatively common-place, especially for gene expression data. Third, equivalence of BNs reduces the ability to determine directionality of edges, as described above. These three limitations can be overcome

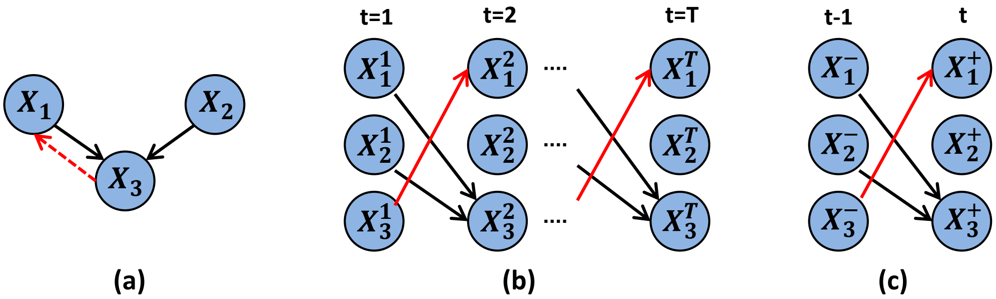


Figure 2.13: **Dynamic Bayesian networks.** (a) A (static) graph structure containing a feedback loop. This graph is not a Bayesian network (BN) because it does not satisfy the acyclicity condition. (b) The graph structure in (a) can be 'unrolled' through time to give a dynamic Bayesian network (DBN) structure in which each component is represented at multiple time points. DBNs are able to model feedback loops. (c) Due to the homogeneity of the graph structure through time, the 'unrolled' DBN in (b) can be 'collapsed' into two time slices representing adjacent time points.

using *dynamic Bayesian networks* (DBNs) which we now describe. From here on, BNs refers to static Bayesian networks.

Dynamic Bayesian networks (DBNs): DBNs extend BNs by incorporating an explicit time element. They can be regarded as BNs "unrolled" through time, with each variable now represented at multiple time points, as shown in Figure 2.13(b). The feedback loop in Figure 2.13(a) can be modelled by the DBN since it no longer creates a cycle. Further, when edges are restricted to be forwards in time only, the network structure of a DBN is fully identifiable because there are no longer any equivalent structures.

Let p denote number of variables under study and T denote number of time points sampled. DBNs associate a random variable with each of the p components at each time point. Let these pT variables be denoted by X_i^t and let $\mathbf{X}^t = \left(X_1^t, \ldots, X_p^t\right)$ be the corresponding random vector at time t. Thus, the full, "unrolled" graph, with each X_i^t explicitly represented as a vertex (Figure 2.13(b)) contains T time slices, each with p nodes ($p = 3$ in Figure 2.13). To facilitate inference over large spaces of candidate graph structures several simplifying assumptions are usually made when DBNs are employed for structure learning of molecular networks [Murphy and Mian, 1999; Husmeier, 2003; Kim *et al.*, 2003]. In particular, first-order Markov and stationarity assumptions are made: each variable at a given time is conditioned only on variables at the previous time point, $p(\mathbf{X}^t \,|\, \mathbf{X}^1, \ldots, \mathbf{X}^{t-1}) = p(\mathbf{X}^t \,|\, \mathbf{X}^{t-1})$, with the conditional probability distribution being time independent. Moreover, this first-order dependence may be sparse, with each component at time t depending on only a subset of components at time $t-1$. The sparsity pattern is described by the edge structure of the network and the above assumptions result in this edge structure being homogeneous through time. This gives a model in which a directed acyclic graph G, with two vertices for each protein, representing adjacent time points, is sufficient to describe the pattern of dependence (the "collapsed DBN"; see Figure 2.13(c)). In DBN structure learning, the edge set of the graph G is the object of inference. The sparse, time-invariant dependence leads to a model with far fewer parameters than a model with a full, time-varying dependence structure. We note that it is a common assumption to assume that edges are only permitted forwards in time [see e.g. Husmeier, 2003; Rau *et al.*, 2010]. This guarantees acyclicity of the DBN, removing the need for computationally expensive acyclicity checks during structure learning. It also facilitates the exact inference approach used in Chapter 4.

The factorisation of the joint distribution over all data is similar to that for a BN (2.40), except each variable now depends on its parents at the previous time step,

$$p(\mathbf{X} \mid G, \mathbf{\Theta}) = \prod_{j=1}^{p} p(X_j^1 \mid \psi_j) \prod_{t=2}^{T} p(X_j^t \mid X_{\pi_G(j)}^{t-1}, \theta_j) \tag{2.51}$$

where $\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^T)$ denotes the complete data, $X_{\pi_G(j)}^t = \left\{X_k^t \mid k \in \pi_G(j)\right\}$ is time t data for the parents of variable j, and $\{\theta_j\}$ and $\{\psi_j\}$ are parameters that fully define the conditional distributions for X_j^t $(t \geq 2)$ and X_j^1 respectively.

Since the marginal $p(X_j^1)$ does not depend on graph G, it will be omitted from here on. In the interests of notational simplicity, we introduce the vector $\mathbf{X}_j^+ = \left(X_j^2, \ldots, X_j^T\right)^{\mathsf{T}}$ to denote all data for variable j in the "current" (second) time slice of the "collapsed DBN" (Figure 2.13(c)) and $\mathbf{X}_j^- = \left(X_j^1, \ldots, X_j^{T-1}\right)^{\mathsf{T}}$ to denote corresponding data in the "previous" (first) time slice. This allows us to remove the product over time above and express the likelihood in the following simple form:

$$p(\mathbf{X} \mid G, \mathbf{\Theta}) = \prod_{j=1}^{p} p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-, \theta_j) \tag{2.52}$$

(up to a multiplicative constant that does not depend on graph G). Structure learning proceeds as for BNs except the likelihood (2.40) is replaced by (2.52).

DBNs were first proposed for structure learning of gene regulatory networks by Friedman *et al.* [1998] and Murphy and Mian [1999], and have been employed

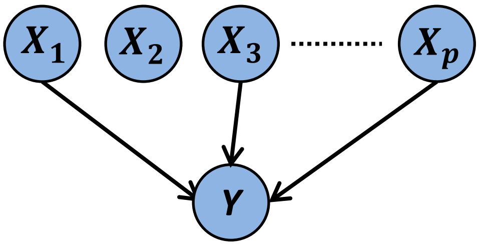


Figure 2.14: **BN structure learning and variable selection.** The variable selection problem of selecting a subset of p predictors $X_1, \ldots, X_p$ that best explains the response Y can be regarded as a BN structure learning problem. A subset of predictors γ can be represented by a BN where $\gamma_j = 1$ if and only if X_j is a parent of Y.

for this purpose many times since (see Chapter 4 for references). Further technical details of DBNs, including their relationship to state space models and hidden Markov models can be found in Murphy [2002].

Bayesian networks and variable selection: The variable selection problem described in Section 2.3.3 can be cast as a BN structure learning problem. The p predictor variables $X_1, \ldots, X_p$ and response variable Y are represented as $p + 1$ nodes in a graph G, as shown in Figure 2.14. The variable selection problem is to find a subset of predictors γ that best explains the response. A model γ can be represented as a BN G where any edge must go from a predictor variable to the response variable, and $\gamma_j = 1$ if and only if X_j is a parent of Y in G. Then, finding the best model γ in the variable selection setting is equivalent to finding the best BN graph structure G. We exploit this connection in Chapter 4.

2.3.6 Graph structure learning: Gaussian graphical models

In this section we consider structure learning methods for GGMs. For structure learning methods for general MRFs, the interested reader is referred to Koller and Friedman [2009].

Learning the structure of a GGM is equivalent to identifying the location of non-zero elements in the precision matrix (for the joint multivariate Gaussian distribution over all variables); see Section 2.3.4. There is a rich literature on sparse precision matrix estimation in the context of Gaussian graphical models, with the seminal 'covariance selection' paper by Dempster [1972] first proposing sparse estimation by setting entries in the precision matrix to zero. Edwards [2000] provides a review of standard approaches, such as greedy stepwise backward selection, for identifying zero entries in the precision matrix. The general procedure for such methods is as follows: First, the covariance matrix is estimated, usually using the unbiased sample covariance matrix. This is then inverted and estimates for the partial correlations ρ_{jk} are calculated using (2.42). Hypothesis tests are then applied to each coefficient to identify those that are significantly different from zero, and thereby find the structure of the GGM. The stepwise procedure has a couple of drawbacks; multiple comparisons are not taken into account and it is computationally intensive. Drton and Perlman [2004] addressed these issues by proposing a method that uses a conservative confidence interval to select the model in a single step.

These standard approaches are only valid when the sample size n is larger than the number of variables p. When $n < p$, as is typically the case for molecular data, the sample covariance matrix is not positive definite and hence cannot be inverted to find estimates for the partial correlations. In addition, the hypothesis tests are based on asymptotics and so may be unreliable at small sample sizes. Several methods have been proposed that enable stable estimates of the precision matrix to be obtained in challenging 'large p, small n' settings. We outline a few approaches below.

Schäfer and Strimmer [2005a] propose three related methods for estimating the precision matrix, based on the Moore-Penrose pseudoinverse and bagging. In particular, they find the best approach at small sample sizes is to use bagging to obtain variance-reduced estimates of the covariance matrix, and then invert this estimate using the pseudoinverse. In a subsequent study, Schäfer and Strimmer [2005b] propose a shrinkage approach to obtain a regularised estimate of the covariance matrix and demonstrated its superior performance in comparison to the bagging/pseudoinverse approach. The shrinkage approach combines the unconstrained unbiased sample covariance matrix estimator S, which will have a high variance due to the large number of parameters that need estimating from small sample size data, with a constrained estimator T that has lower variance, but is a more biased estimate. They are combined in a weighted average to obtain a new, improved estimator S^*,

$$S^* = \lambda T + (1 - \lambda)S \tag{2.53}$$

where $\lambda \in [0, 1]$ is the shrinkage intensity, controlling the relative contribution of S and the shrinkage target T, which they take to be a diagonal matrix with unequal variances. This choice of target ensures that S^* is positive-definite, allowing it to be inverted to obtain partial correlations. An optimal shrinkage intensity, minimising mean squared error, is obtained analytically using the Ledoit-Wolf Lemma [Ledoit and Wolf, 2003]. Once estimation of the partial correlations is complete, model selection (identification of edges in the GGM) is carried out using an empirical Bayes approach (see Section 2.3.8) combined with large-scale multiple testing.

Other recent approaches have focussed on using ℓ_1 penalisation to regularise estimation of GGM structure. Meinshausen and Bühlmann [2006] use lasso regression (see Section 2.3.3) to perform neighbourhood selection for each node in the graph. This approach sets a subset of regression coefficients to zero, and so automatically performs model selection, with no additional hypothesis testing required. A sparse precision matrix can subsequently be obtained via constrained maximum likelihood estimation using the inferred sparse graph structure. Several authors, including Friedman *et al.* [2008], have proposed maximum penalised likelihood estimators with an ℓ_1 penalty applied to the precision matrix. This approach simultaneously performs model selection and parameter estimation, and is the approach we employ in Chapter 5, where further details can be found.

Bayesian approaches have also been proposed [Dobra *et al.*, 2004; Jones *et al.*, 2005], which allow the posterior distribution over graph structures to be explored, but are more computationally intensive than the shrinkage and penalised likelihood approaches.

GGMs have many of the same attractive properties as BNs, including the ability to distinguish direct interactions between observed variables from indirect interactions. Since they are undirected models, they are, in contrast with BNs, able to model feedback loops. However, the directionality of BNs can aid interpretation of results, and BNs can model both discrete and continuous data.

GGMs have been applied for structure learning of molecular networks, although they are not as widely used as BNs. Early work applied standard methods to a small number of genes [Waddell and Kishino, 2000] or a small number of clusters of genes [Toh and Horimoto, 2002] so that $p < n$ and the sample covariance matrix is invertible. More recent work has applied GGMs to gene expression data with $p > n$, with many of the studies referenced above including such applications [Schäfer and Strimmer, 2005a,b; Dobra *et al.*, 2004; Jones *et al.*, 2005; Banerjee *et al.*, 2008]. While gene regulatory networks are the main focus for applications, GGMs have also been applied to protein signalling networks [Friedman *et al.*, 2008] and metabolic networks [Krumsiek *et al.*, 2011].

2.3.7 Other structure learning methods

In this thesis we use graphical models (both DBNs and GGMs) to infer the structure of protein signalling networks. However, many other structure learning methods have been proposed that are not based on graphical models. For completeness, we briefly outline below some of the methods that are more prominent in the literature. Reviews and comparisons of these (and other) methods can be found in Werhli *et al.* [2006]; Bansal *et al.* [2007]; Markowetz and Spang [2007]; Cantone *et al.* [2009]; Hecker *et al.* [2009] and Altay and Emmert-Streib [2010].

2.3.7.1 Clustering

Clustering methods can be used to group together molecular components that display similar characteristics. They are a very popular approach for analysing and visualising gene expression data [e.g. Eisen *et al.*, 1998], with genes being grouped together based on similarity in gene expression profiles, the idea being that genes in the same cluster are likely to be functionally related. This notion is often referred to as the *guilt-by-association heuristic.* It can loosely be regarded as a structure learning approach by placing an (undirected) edge between each pair of genes in the same cluster, but since the resulting networks are fully connected, there is no distinction between direct and indirect influences, and no indication of which associations are strongest. We discuss clustering methods in more detail in Section 2.3.9 below.

2.3.7.2 Relevance networks

The relevance network approach to structure learning [Butte and Kohane, 2000] is again based on the guilt-by-association heuristic. An edge is placed between a pair of variables if they show a high level of association according to a similarity metric, such as Pearson's correlation coefficient or mutual information (MI). These similarity measures are symmetric, resulting in undirected graph structures. Unlike Pearson correlation, MI is applied to discrete data and, since it does not assume a linear dependence between variables, can pick out nonlinear dependencies. It is a measure of the degree of independence between two variables, with a value of zero implying independence, and larger scores indicating a higher degree of dependence. Methods of calculating MI include a histogram technique to discretise continuous data into bins and thereby calculate probabilities, and Gaussian kernel density estimation. Further information about MI can be found in Steuer *et al.* [2002].

Similarity scores are calculated for all pairs of variables, and those that exceed a threshold have an undirected edge placed between them. However, since pairs of components are considered in isolation from all other components, the approach is unable to distinguish between direct and indirect influences. To remedy this, a pruning step is usually employed, removing edges where the association can be better explained by an indirect influence.

Several methods, based on the MI relevance network approach, have been proposed in the literature. They differ on the methods to calculate MI, obtain the initial graph structure from the pairwise scores, and prune the graph. Two such approaches are ARACNe [algorithm for reconstruction of accurate cellular networks; Basso *et al.*, 2005] and CLR [context likelihood of relatedness; Faith *et al.*, 2007]. Both of these methods, and the original relevance network method [Butte and Kohane, 2000] were demonstrated on gene expression data. A comparison of these MI-based methods can be found in Altay and Emmert-Streib [2010].

Unsurprisingly, relevance networks have an improved performance for structure learning over clustering [Bansal *et al.*, 2007]. However, due to their ability to directly model combinatorial relationships and conditional independencies in the data, and therefore distinguish between direct and indirect interactions, graphical models can perform better than relevance networks, as demonstrated by Werhli *et al.* [2006]. Hartemink [2005] also show superior performance of DBNs over ARACNe when applied to time series data. This is likely due to the fact that the MI-based methods assume that samples are independent, which is not the case for time series data.

2.3.7.3 Ordinary differential equations

Structure learning methods based on ordinary differential equations (ODEs) model changes in a variable (e.g. expression level of a gene or phosphorylation level of a protein) as a function of all other variables and of external perturbations that affect some of the variables (e.g. gene knockouts or drug treatments). To help avoid overfitting and to facilitate inference of parameters, a linear ODE is often used. That is, each variable X_j, has an ODE of the following form,

$$\dot{X}_j(t) = \sum_{k=1}^{p} a_{jk} X_k + b_j U_j \tag{2.54}$$

where $\dot{X}_j(t)$ is the rate of change of variable X_j at time t, and a_{jk} and b_j represent the strength of influence on $\dot{X}_j(t)$ of variable X_k and perturbation U_j respectively. The parameters a_{jk} also encode the network structure; if a_{jk} is non-zero a directed edge exists in the graph from X_j to X_k. Hence the structure learning problem is to identify the non-zero parameters a_{jk}.

Gardner *et al.* [2003] proposed the ODE-based method NIR (network identification by multiple regression) for steady-state data. At steady-state we have

$\dot{X}_j(t) = 0$ which means (2.54) becomes a standard system of linear equations, with unknown parameters a_{jk}. The effect of the perturbations on each variable (i.e. $b_j U_j$) is assumed to be known. NIR proceeds by making a network sparsity assumption, which assumes that $K < p$ variables influence any given variable; that is, there are K non-zero a_{jk} for each j. An exhaustive search is performed over all possible subsets of K variables, and the subset resulting in least square parameter estimates that best fit the data is chosen. Good performance of NIR has been demonstrated when the majority of variables are affected by a perturbation [Gardner *et al.*, 2003; Bansal *et al.*, 2007]. However, one limitation is that it requires knowledge of the targets of perturbations in the first place, which may not always be known. Indeed, once parameters have been learnt, the model can be used to predict targets of a perturbation.

TSNI (time series network identification) is a method for time series data proposed by Bansal *et al.* [2006]. No steady-state assumptions are made here, and so the derivatives $\dot{X}_j(t)$ have to be estimated from the data. The targets of the perturbations are assumed to be unknown and are inferred along with the graph structure (parameters a_{jk}). Interpolation and dimensionality reduction methods (PCA) are used to regularise the problem and obtain parameter estimates. Although it can be used to infer a whole graph structure, TSNI is mainly focussed on inferring the perturbation target along with a local network structure around the target.

In contrast to the probabilistic approach provided by graphical models, these ODE-based methods are deterministic. This means that they are arguably less well placed to deal with the noise that is ubiquitous in molecular data. As discussed in the Introduction, ODEs offer a rich modelling framework, with, for example, non-linear ODEs based on biochemical reaction kinetics providing a reasonably realistic mechanistic model for signal transduction pathways [Schoeberl *et al.*, 2002; Chen *et al.*, 2009; Wang *et al.*, 2009a]. However, such models require a large number of parameters, leading to overfitting if estimated from small sample data, and moreover, even if a large amount of data were available, parameter estimation and ODE simulation is likely to be mathematically and computationally challenging. For these reasons, in spite of their reduced ability to model realistic dynamic behaviour, linear models are utilised. Linear ODE models have many similarities with linear Gaussian BNs. Indeed, it is not difficult to incorporate linear ODEs within a DBN structure learning framework [Li *et al.*, 2011].

2.3.7.4 Gaussian processes

One approach that does combine ODEs with non-linear models in a tractable manner is the Gaussian processes approach by Äijö and Lähdesmäki [2009]. Here, a non-parametric model is used, instead of a linear model, to relate the change in a variable to all other variables. This avoids making strong assumptions regarding the form of the regulatory function, which in this approach, is learned from the data using Gaussian processes. Gaussian processes are stochastic processes that consist of a collection of random variables and every finite subset of these random variables has a joint Gaussian distribution. Here, they provide a non-parametric prior distribution over regulatory functions. Unlike the linear ODE-based approaches outlined above, noise is explicitly taken into account within the model and inference is carried out within a Bayesian framework. Therefore, the procedure offers all the benefits of Bayesian model selection (see Section 2.3.2). In particular, it takes model complexity into account automatically via the marginal likelihood (which can be found in closed form), it allows prior knowledge to be integrated into inference via the network prior, and Bayesian model averaging takes model uncertainty into account. Äijö and Lähdesmäki [2009] demonstrate their method on gene expression data with favourable results.

2.3.8 Empirical Bayes

We consider again the Bayesian inference approach where we are interested in the posterior distribution over parameters $\mathbf{\Theta}$ given data $\mathbf{Y}$, which by Bayes' Theorem is proportional to the product of the likelihood and parameter prior,

$$p(\mathbf{\Theta} \mid \mathbf{Y}, \lambda) \propto p(\mathbf{Y} \mid \mathbf{\Theta}) p(\mathbf{\Theta} \mid \lambda). \tag{2.55}$$

Here, the prior distribution $p(\mathbf{\Theta} \mid \lambda)$ has its own parameters (hyperparameters), which we denote by λ. For example, the NIG prior for the linear model in (2.11) has hyperparameters $\mathbf{m}, \mathbf{V}, a, b$.

If λ is unknown, a fully Bayesian approach takes uncertainty in the hyperparameters into account by placing a prior $p(\lambda)$ over λ and marginalising to give,

$$p(\mathbf{\Theta} \mid \mathbf{Y}) = \frac{\int p(\mathbf{Y} \mid \mathbf{\Theta}) p(\mathbf{\Theta} \mid \lambda) p(\lambda) \mathrm{d}\lambda}{p(\mathbf{Y})} \tag{2.56}$$

$$= \int p(\mathbf{\Theta} \mid \mathbf{Y}, \lambda) p(\lambda \mid \mathbf{Y}) \mathrm{d}\lambda. \tag{2.57}$$

The prior distribution over λ could in turn have its own unknown parameters, over which another prior could be placed. Specifying a model over several levels in this way is known as *hierarchical modelling.* We assume here that the parameters for $p(\lambda)$ are known.

The empirical Bayes (EB) approach uses the data to estimate hyperparameters. In particular, a point estimate for λ is obtained using the marginal distribution of all the data,

$$p(\mathbf{Y} \mid \lambda) = \int p(\mathbf{Y} \mid \mathbf{\Theta}) p(\mathbf{\Theta} \mid \lambda) \mathrm{d}\mathbf{\Theta}. \tag{2.58}$$

This point estimate, denoted $\hat{\lambda}$, is usually obtained by maximum likelihood or method of moments estimation. Here we consider the maximum marginal likelihood approach,

$$\hat{\lambda} = \max_{\lambda} p(\mathbf{Y} \mid \lambda). \tag{2.59}$$

This estimate for λ is then used in the posterior distribution $p(\mathbf{\Theta}|\mathbf{Y}, \hat{\lambda})$ and Bayesian inference proceeds as if λ were known. Therefore the EB approach replaces the integral in (2.56) with a maximisation procedure, which can have computational advantages. Since $p(\lambda \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \lambda) p(\lambda)$, the EB maximum marginal likelihood estimate can be regarded as a MAP estimate of the posterior $p(\lambda \mid \mathbf{Y})$ under a flat prior $p(\lambda)$. Moreover, if the posterior $p(\lambda \mid \mathbf{Y})$ is sharply peaked around $\hat{\lambda}$, then the integral (2.57) is approximately equal to $p(\mathbf{\Theta} \mid \mathbf{Y}, \hat{\lambda})$, and so the fully Bayesian and EB approaches give similar results in this case. The EB approach is also related to James-Stein estimation [Stein, 1955]; see Carlin and Louis [2008] for details.

We illustrate the EB approach on a simple example, taken from Carlin and Louis [2008]. Let $p(Y_j \mid \mu_j) = \mathcal{N}(Y_j \mid \mu_j, \sigma^2)$ for $j = 1, \ldots, p$ where σ^2 is known and let μ_j have prior $p(\mu_j \mid \lambda) = \mathcal{N}(\lambda, \tau^2)$ where τ^2 is known. The (Y_j, μ_j) pairs are independent of each other. It can be shown that the marginal distributions $p(Y_j \mid \lambda)$ are independent and distributed as $\mathcal{N}(Y_j|\lambda, \sigma^2+\tau^2)$. The marginal likelihood $p(\mathbf{Y}|\lambda)$, where $\mathbf{Y} = (Y_1, \ldots, Y_p)$, can then be maximised with respect to λ to give the EB estimate $\hat{\lambda} = \bar{Y} = \frac{1}{p} \sum_{j=1}^{p} Y_j$ (i.e. the sample mean). This results in the posterior,

$$p(\mu_j \mid Y_j, \hat{\lambda}) = \mathcal{N}(\mu_j \mid \omega \bar{Y} + (1 - \omega) Y_j, (1 - \omega) \sigma^2) \tag{2.60}$$

where $\omega = \sigma^2 / (\sigma^2 + \tau^2)$. Therefore, the posterior mean for μ_j is $\omega \bar{Y} + (1 - \omega) Y_j$, which depends on all the data and not just Y_j; information is 'borrowed' from all components to estimate μ_j.

The EB approach described here is in fact a parametric EB approach [Morris, 1983] since the parameter prior has a parametric form. Non-parametric EB approaches [Robbins, 1955], which we do not discuss here, also exist. Further details about EB approaches can be found in Carlin and Louis [2008].

EB approaches have been used in several areas of statistics and bioinformatics. Perhaps the most well-known application of EB in bioinformatics is for assessing differential gene expression [Smyth, 2004], where as seen in the example above, information is borrowed across genes to help obtain more stable inferences for each individual gene at small sample sizes. Other examples where EB has been used to set hyperparameters include: Bayesian variable selection in the linear model [George and Foster, 2000], the Bayesian Lasso [Park and Casella, 2008], structure learning with Gaussian processes [Äijö and Lähdesmäki, 2009], and structure learning with Gaussian graphical models [Schäfer and Strimmer, 2005a].

In Chapters 3 and 4 we use an EB approach to automatically set the model prior hyperparameters that select and/or weight the prior information.

2.3.9 Clustering

Clustering is an unsupervised learning method, often used to analyse and visualise high-dimensional data. The aim is to group data objects together into subsets (clusters) so that those within the same cluster have a higher similarity with each other than with those in other clusters. Importantly, it differs from supervised classification, where class labels are known and the aim is to learn a model from the labelled data that can then be used to classify new unlabelled data. In clustering, there are no observed class labels.

Clustering is a popular approach for analysing gene expression data. In molecular biology applications, clustering can be used to either group variables together [e.g. to find sets of co-regulated genes; Eisen *et al.*, 1998; Toh and Horimoto, 2002], group samples together [e.g. to discover disease subtypes characterised by similar gene expression patterns; Golub *et al.*, 1999; Alizadeh *et al.*, 2000], or to simultaneously group both variables together and samples together [e.g. to find subsets of genes with similar expression patterns in a subset of experimental conditions; ‘bi-clustering’ methods; Alon *et al.*, 1999; Madeira and Oliveira, 2004].

Below we outline some of the most widely-used clustering approaches, and we do so in the context of clustering samples together. All the methods can also be applied to cluster variables together. The reader is referred to Datta and Datta [2003]; Thalamuthu *et al.* [2006]; Kerr *et al.* [2008] and de Souto *et al.* [2008] for reviews and comparisons of various clustering methods, including those outlined below.

2.3.9.1 K-means

K-means uses squared Euclidean distance as a measure of similarity,

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2^2 \tag{2.61}$$

where $\mathbf{x}_i$ is a p-dimensional data sample.

The number of clusters K is fixed in advance and the algorithm is initialised by selecting K of the data samples to be the initial cluster means, denoted by $\mathbf{m}_1, \ldots, \mathbf{m}_K$. Let $C(i) \in \{1, \ldots, K\}$ denote the cluster assignment for sample i. The algorithm then proceeds in an iterative fashion as follows,

1. **Assign:** Given current cluster means $\{\mathbf{m}_1, \ldots, \mathbf{m}_K\}$, assign each sample to the closest cluster mean,

$$C(i) = \underset{k=1,\ldots,K}{\operatorname{argmin}}\ d(\mathbf{x}_i, \mathbf{m}_k). \tag{2.62}$$

2. **Calculate means:** Given current cluster assignment function C, recalculate cluster means,

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{i:C(i)=k} \mathbf{x}_i \tag{2.63}$$

   where N_k is the number of samples currently assigned to cluster k.

3. **Repeat:** Iterate steps 1 and 2 until assignments remain constant.

The algorithm is trying to minimise the within-cluster sum-of-squares,

$$\underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{i:C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) \tag{2.64}$$

where $\mathbf{m}_k$ is the cluster mean, as defined in Step 2 of the algorithm. However, it is only guaranteed to find a local minimum, so performing multiple runs with random initialisations is advisable. It can also be sensitive to noise and outliers in the data, and requires the user to select the number of clusters.

Figure 2.15(a) shows an application of K-means to simulated data consisting of two clusters, each with 50 samples. The method works well when clusters are of a spherical shape and are clearly separated. When this is not the case, as in Figure 2.15(b), K-means can fail.

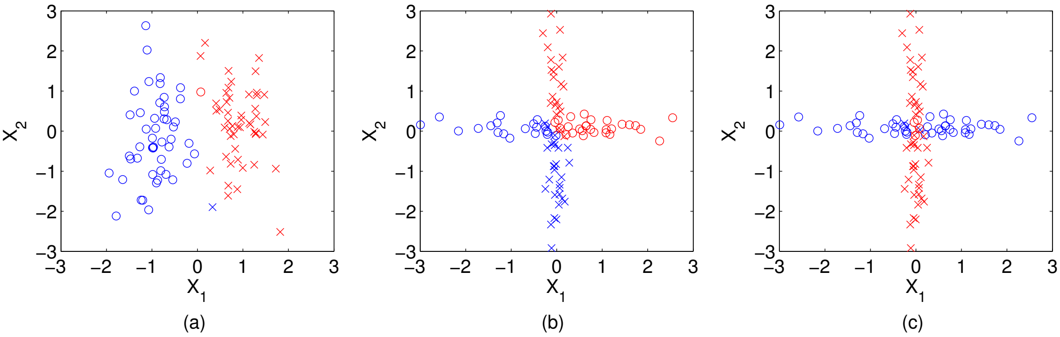


Figure 2.15: ***K*-means and model-based clustering.** (a) Results of K-means clustering applied to simulated data consisting of two clusters, each with 50 two-dimensional samples. Circles and crosses denote true cluster labels; red and blue denote cluster assignments returned by K-means. Here the clusters have a clear separation and have a relatively spherical shape. (b) K-means clustering applied to simulated data with clusters that overlap and have an elliptical shape. (c) Model-based clustering applied to the same data as in (b). Hard assignments are obtained by assigning samples to the cluster with highest responsibility for that sample.

Applications of K-means in molecular biology include clustering of genes into transcriptional regulatory subnetworks [Tavazoie, 1999] and discovery of subtypes of gliomas based on gene expression signatures [Shai *et al.*, 2003].

2.3.9.2 Hierarchical clustering

As the name suggests, hierarchical clustering forms a hierarchy, where at each level in the hierarchy clusters are created by merging clusters at the previous level. This hierarchy can be represented graphically in the form of a tree, known as a dendogram.

In agglomerative hierarchical clustering, the algorithm begins with each data sample in a cluster by itself. Then, the closest pair of clusters is merged into a single cluster. This is repeated until only one cluster containing all samples exists. A less popular method is divisive clustering, which starts with all samples in one cluster, and then at each iteration splits a cluster into two. Closeness between two clusters is often assessed using the average intercluster distance (i.e. the average of all pairwise distances between samples, where the samples are in different clusters). This is known as average-linkage clustering. The distance metric is chosen by the user; results are not invariant to this choice.

The dendogram allows visualisation of global patterns in the data. Clusters can be obtained by cutting the dendogram at a certain level, which is decided upon by the user. Higher levels in the dendogram represent the merging of clusters with a larger intercluster dissimilarity (higher average intercluster distance). The idea is to select a level where samples within clusters have a greater similarity with each other than to samples in other clusters.

Like K-means, hierarchical clustering can be sensitive to noise and outliers in the data. However, it does not require pre-specification of the number of clusters.

There have been many applications of hierarchical clustering, including the clustering of genes into functional modules [Eisen *et al.*, 1998], identification of cancer types [Nielsen *et al.*, 2002], and discovery of new cancer subtypes [Alizadeh *et al.*, 2000].

2.3.9.3 Model-based clustering

Model-based clustering [McLachlan and Basford, 1987; Fraley and Raftery, 1998; McLachlan and Peel, 2000; Fraley and Raftery, 2002] with finite Gaussian mixture models is a popular approach to clustering that, unlike the approaches above, is rooted in an explicit statistical model. Each mixture component corresponds to a cluster. Given data, the aim is to estimate the parameters of the mixture model and calculate probabilities of samples belonging to each cluster (known as *responsibilities*). Therefore, this method performs 'soft' clustering assignments; it does not assign each sample to a single specific cluster as K-means and hierarchical clustering do. However, 'hard' assignments can be obtained by assigning each sample to the cluster (mixture component) with largest responsibility for that sample.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a finite Gaussian mixture distribution,

$$f(\mathbf{x}_i; \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \tag{2.65}$$

where the mixing proportions π_k satisfy $0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1$, f_k is the p-dimensional multivariate Gaussian density with component-specific mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{\Sigma}_k$, and $\mathbf{\Theta} = \{(\pi_k, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) : k = 1, \ldots, K\}$ is the set of all unknown parameters. The log-likelihood for the sample is given by

$$l(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \right). \tag{2.66}$$

Maximizing the log-likelihood is difficult due to its non-convexity. The Expectation-Maximisation (EM) algorithm [Dempster *et al.*, 1977] can be used to obtain maximum likelihood estimates and calculate responsibilities. Let z_i be a latent variable satisfying $z_i = k$ if sample $\mathbf{x}_i$ belongs to cluster k. Then we have $P(z_i = k) = \pi_k$ and $p(\mathbf{x}_i \mid z_i = k) = f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The log-likelihood for the complete data $\{\mathbf{x}_i, z_i\}_{i=1}^n$ is

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log(\pi_{z_i}) + \log\left(f_{z_i}\left(\mathbf{x}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)\right). \tag{2.67}$$

In the E-step of the EM algorithm, given current estimates of the parameters $\boldsymbol{\Theta}^{(t)}$, we compute

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) &= \mathbb{E}\left[l_c(\boldsymbol{\Theta}) \mid \{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{\Theta}^{(t)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \left[\log(\pi_k) + \log\left(f_k\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)\right] \end{aligned} \tag{2.68}$$

where $\tau_{ik}^{(t)}$ is the posterior probability of sample $\mathbf{x}_i$ belonging to cluster k (the responsibility of cluster k for sample $\mathbf{x}_i$),

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t)} f_k\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} f_j\left(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}. \tag{2.69}$$

In practice, in order to perform the M-step below, it is only necessary to calculate the responsibilities. Therefore, the E step is carrying out a soft assignment based on the current estimates for the parameters.

In the M-step $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$ is maximised with respect to $\boldsymbol{\Theta}$ to give the following new estimates for the parameters $\boldsymbol{\Theta}^{(t+1)}$, based on the current responsibilities,

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{n} \tag{2.70}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}} \tag{2.71}$$

$$\mathbf{S}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^{\mathsf{T}}}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}. \tag{2.72}$$

The E and M steps are iterated until the change in the likelihood $l(\boldsymbol{\Theta}^{(t)})$ is below a threshold. It is only guaranteed to find a local maximum of $l(\boldsymbol{\Theta})$ so multiple random intialisations are usually performed.

Like K-means, the number of clusters K needs to be set in advance by the user. However, since this method is model-based, the problem of setting the number of clusters can be cast as a model selection problem, and so methods such as information criteria (e.g. BIC) and likelihood-based cross-validation can be used to set K.

Figure 2.15(c) shows an application of model-based clustering to the same data as in Figure 2.15(b). We see that model-based clustering is successful in recovering the true clusters where K-means was not, due to its ability to identify clusters with an elliptical shape.

McLachlan *et al.* [2002] applied model-based clustering to cluster cancer tissues using gene expression data. We apply model-based clustering in Chapter 5 to cluster breast cancer cell lines on the basis of signalling network connectivity.

Chapter 3

Integrating biological knowledge into variable selection: an empirical Bayes approach

3.1 Introduction

Ongoing advancements and cost reductions in biochemical technology are enabling acquisition of ever richer datasets. As discussed in Chapter 1, in many settings, in both basic biology and medical studies, it may be important to model the relationship between assayed molecular entities, such as genes, proteins or metabolites, and a biological response of interest.

Molecular players may act in concert to influence biological response: this has motivated a need for multivariate methods capable of modelling such joint activity. When sample sizes are small-to-moderate, as is often the case in molecular studies, robust modelling of joint influences becomes especially challenging. However, often it is likely that only a small number of players are critical in influencing the response of interest. Then, the challenge is to identify appropriate variable subsets.

Statistical variable selection methods have been widely used in the bioinformatics domain to discover subsets of influential molecular predictors. Both Bayesian [Lee *et al.*, 2003; Jensen *et al.*, 2007; Mukherjee *et al.*, 2009; Ai-Jun and Xin-Yuan, 2010; Li and Zhang, 2010; Yeung *et al.*, 2011] and penalised likelihood approaches [Li and Li, 2008; Wu *et al.*, 2009] have been used in a diverse range of applications. These include, the identification of sets of genes that can discriminate between cancer cells and normal cells, or between different (sub)types of cancer (based on gene expression data) [Lee *et al.*, 2003; Ai-Jun and Xin-Yuan, 2010]; inference of gene regulatory relationships from gene expression data [Jensen *et al.*, 2007; Yeung *et al.*, 2011]; identification of DNA motifs that are associated with gene expression levels [Li and Zhang, 2010]; identification of genes that are related to cancer survival rates [Li and Li, 2008] and discovery of sets of signalling proteins that are predictive of drug response [Mukherjee *et al.*, 2009].

Bayesian approaches can facilitate the integration of ancillary information regarding variables under study through prior probability distributions over predictor subsets. Ongoing development of online tools and databases have meant that such information is widely available, and depending on context, may include networks and pathway maps, public gene expression datasets, molecular interaction databases, ontologies and so on. However, while the idea of incorporating such information into variable selection has a clear appeal, it is not always obvious what information should be included nor how it should be weighted. Indeed, many existing Bayesian variable selection approaches do not attempt integrative analyses exploiting such information and instead employ standard priors that do not specify preferences for particular variables, but may, for example, encode a preference for sparse models [Brown *et al.*, 2002; Mukherjee *et al.*, 2009]. Several Bayesian variable selection studies have put forward simple approaches for incorporating prior knowledge by independently assigning each variable a prior probability of being included in the model [George and McCulloch, 1997; Chipman *et al.*, 2001; Lee *et al.*, 2003; Ai-Jun and Xin-Yuan, 2010] (Bernoulli-distributed prior; see (2.34)). However, subjectively setting such hyperparameters for each variable may be difficult. As a result, in practice, each variable is usually assigned the same prior probability. Furthermore, prior independence may be a questionable assumption, since molecular variables are unlikely to influence a response independently of one another.

We develop a variable selection procedure in which an empirical Bayes approach is used to objectively select between a choice of informative priors incorporating ancillary information ('biologically informative priors') and also to objectively weight the contribution of the prior to the overall analysis. Empirical Bayes approaches have previously been used with Bayesian variable selection, but only to set the success parameter π in the standard Bernoulli-distributed prior [George and Foster, 2000]. A related, yet often more computationally intensive approach is to perform a fully Bayesian analysis and marginalise out the hyperparameter. Such an approach has also been applied to Bayesian variable selection (with Bernoulli-distributed prior) [Nott and Green, 2004] and also in a structure learning setting with biologically informative priors [Werhli and Husmeier, 2007].

The work presented here is motivated by questions concerning the relationship between signalling proteins and drug response in human cancers. In the protein signalling setting (as also in gene regulation) there is now much information available, both in the literature and in diverse online resources, concerning relevant pathways and networks. We therefore develop pathway- and network-based informative priors for this setting, applying the empirical Bayes method proposed to automatically select and weight the prior and thence carry out variable selection.

The relationship between response and predictors is modelled using a continuous, linear model with interaction terms. In this way we avoid data discretisation, which can lose information, yet retain the ability to capture combinatorial interplay. We take advantage of biochemically-motivated sparsity constraints to permit exact inference, thereby avoiding the need for approximate approaches such as Markov chain Monte Carlo (MCMC). This enables the calculation of exact probability scores over which variables are likely to be influential. The overall procedure is computationally fast: empirical Bayes analysis and subsequent calculation of posterior (inclusion) probabilities for 52 predictors via full model averaging required only 10 minutes (in MATLAB R2010a on a standard single-core personal computer; code freely available at `http://go.warwick.ac.uk/stevenhill/IBKVS`). Moreover, the overall procedure we put forward is simple from the user perspective, requiring very few user-set parameters and no MCMC convergence diagnostics.

The remainder of this Chapter is organised as follows. In Section 3.2 we refer to the background information given in Chapter 2 on Bayesian linear models and Bayesian variable selection, and then describe the particular details of our Bayesian variable selection approach. We describe in turn, a linear model including interactions between predictors, exact computation of posterior inclusion probabilities, biologically informative pathway-based priors, and empirical Bayes analysis to objectively select and weight prior information. In Section 3.3 we illustrate our method on published single cell proteomic data [Sachs *et al.*, 2005] with synthetic response data, and on breast cancer proteomic and drug response data. In Section 3.4 we conclude with a discussion of the merits and shortcomings of our work, make further comparisons of our approach to those existing in the literature and highlight directions for further work.

3.2 Methods

The reader is referred to Sections 2.3.1 and 2.3.3 for background information on the Bayesian linear model and Bayesian variable selection respectively. The notation used in this Chapter also follows that used in the aforementioned sections (in particular, see Equations (2.2), (2.31) and surrounding text).

3.2.1 Bayesian linear model with interaction terms

We include interaction terms in the linear model, $\mathbf{Y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \epsilon$ (see also (2.31)), to enable combinatorial relationships between predictors included in model γ and response $\mathbf{Y}$ to be captured. Let $\mathbf{X}_{i\gamma}$ be a $1 \times |\gamma|$ vector denoting sample i data for predictors included in model γ. Given model γ, response Y_i depends in a non-linear fashion on the included predictors $\mathbf{X}_{i\gamma}$ while remaining linear in the regression parameters. In particular, the mean for Y_i is a linear combination of included predictors and all possible products of included predictors. For example, if $|\gamma| = 3$ with $\gamma_1 = \gamma_2 = \gamma_3 = 1$, then the mean for variable Y_i is a linear combination of the three included predictors X_{ij} (for $j = 1, 2, 3$), the three possible pairwise products of included predictors $X_{ij}X_{ik}$ and the product of all included predictors $X_{i1}X_{i2}X_{i3}$. We extend the $n \times |\gamma|$ predictor (design) matrix $\mathbf{X}_\gamma$ and regression coefficient vector $\boldsymbol{\beta}_\gamma$ to include the interaction terms and corresponding coefficients respectively, and we denote the extended versions by $\bar{\mathbf{X}}_\gamma$ and $\bar{\boldsymbol{\beta}}_\gamma$. The likelihood now takes the form

$$p(\mathbf{Y} \mid \gamma, \bar{\mathbf{X}}_\gamma, \bar{\boldsymbol{\beta}}_\gamma, \sigma^2) \sim \mathcal{N}\left(\bar{\mathbf{X}}_\gamma \bar{\boldsymbol{\beta}}_\gamma, \sigma^2 \mathbf{I}\right). \tag{3.1}$$

All columns in $\bar{\mathbf{X}}_\gamma$ are standardised to have zero mean and unit variance.

Recall from Section 2.3.3.2 that, in Bayesian variable selection, the object of interest is the posterior distribution over models $P(\gamma \mid \mathbf{Y}, \mathbf{X})$, which is given (up to proportionality) by Bayes' theorem in (2.32) and is also reproduced here,

$$P(\gamma \mid \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma) P(\gamma), \tag{3.2}$$

where $p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma)$ is the marginal likelihood, and $P(\gamma)$ is the model prior and is the main focus of this Chapter. General background regarding the marginal likelihood, including its ability to automatically control for model complexity, was given in Section 2.3.2.2.

As discussed in Section 2.3.3.2, for the prior on parameters $\Theta_\gamma = (\boldsymbol{\beta}_\gamma, \sigma^2)$ we take a limiting case of the conjugate $NIG(\mathbf{m}, \mathbf{V}, a, b)$ prior (2.11), following Zellner [1986]; Smith and Kohn [1996] and Nott and Green [2004]. Specifically, we take $\mathbf{m} = \mathbf{0}$, $\mathbf{V} = c\left(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma\right)^{-1}$ (Zellner's g-prior) and $a = b = 0$. Hence, the prior for $\bar{\boldsymbol{\beta}}_\gamma$ given γ and σ^2 is given by

$$p(\bar{\boldsymbol{\beta}}_\gamma \mid \gamma, \bar{\mathbf{X}}_\gamma, \sigma^2) \sim \mathcal{N}\left(\mathbf{0}, c\sigma^2 \left(\bar{\mathbf{X}}_\gamma^\top \bar{\mathbf{X}}_\gamma\right)^{-1}\right) \tag{3.3}$$

and the prior for σ^2 is $p\left(\sigma^2\right) \propto \sigma^{-2}$.

Several approaches have been proposed in the literatue for setting the hyperparameter c in Zellner's prior. For example, Smith and Kohn [1996] take a 'large' value of $c = 100$ so that the prior contains little information relative to the likelihood; Liang *et al.* [2008] consider a fully-Bayes approach and place a prior distribution on c; George and Foster [2000] set c using an empirical Bayes approach; and Gustafson [2000]; Kohn *et al.* [2001]; Nott and Green [2004] set c equal to the sample size n. We take the latter approach here, which yields a 'unit information prior' [Kass and Wasserman, 1995] with the amount of information about the parameter equal to the amount of information in one observation.

With these choices of prior, the closed form marginal likelihood in (2.33), obtained by integrating out parameters, becomes

$$p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma) \propto (1+n)^{-\frac{2^{|\gamma|}-1}{2}} \left(\mathbf{Y}^\mathsf{T}\mathbf{Y} - \frac{n}{n+1} \mathbf{Y}^\mathsf{T} \bar{\mathbf{X}}_\gamma \left(\bar{\mathbf{X}}_\gamma^\mathsf{T} \bar{\mathbf{X}}_\gamma \right)^{-1} \bar{\mathbf{X}}_\gamma^\mathsf{T} \mathbf{Y} \right)^{-\frac{n}{2}}. \tag{3.4}$$

We note that, in contrast to the widely-used (non-limiting) NIG prior, this formulation has no free hyperparameters and enjoys attractive invariance properties under rescaling [Kohn *et al.*, 2001].

Calculation of the marginal likelihood (3.4) requires inversion of a $\left(2^{|\gamma|}-1\right) \times \left(2^{|\gamma|}-1\right)$ matrix $\bar{\mathbf{X}}_\gamma^\mathsf{T} \bar{\mathbf{X}}_\gamma$, which could result in problems with matrix singularity when sample size n is too small relative to $|\gamma|$. In this work, due to a restriction on $|\gamma|$ (discussed below), we do not encounter such matrix singularity issues.

3.2.2 Exact posterior inclusion probabilities

We are interested in calculating posterior inclusion probabilities for each predictor, as given in (2.35) and reproduced here,

$$P(\gamma_j = 1 \mid \mathbf{Y}, \mathbf{X}) = \sum_{\gamma:\gamma_j=1} P(\gamma \mid \mathbf{Y}, \mathbf{X}). \tag{3.5}$$

Inclusion probabilities are a measure of the importance of each individual predictor in determining the response and are calculated by model averaging. Background information on Bayesian model selection and model averaging can be found in Sections 2.3.2.2 and 2.3.2.3.

Instead of employing MCMC methods to sample from the posterior distribution over the vast space of models ($|\Gamma| = 2^p$) and obtain asymptotically valid estimates for inclusion probabilities (3.5), we calculate exact inclusion probabilities by enforcing a restriction on the number of predictors that are allowed to be included in the model. That is, we only allow γ with $|\gamma| \leq d_{max}$ for some $d_{max} \in \mathbb{N}$. This gives

$$|\Gamma| = \sum_{d=0}^{d_{max}} \binom{p}{d}. \tag{3.6}$$

Thus, instead of being exponential in p, the model space Γ has polynomial size of order $p^{d_{max}}$, thereby allowing explicit calculation of posterior inclusion probabilities via (3.5). We take $d_{max} = 4$, giving $|\Gamma| = 294,204$ for the $p = 52$ predictors in the cancer drug response application below; the original size of Γ was of order 10^{15}.

3.2.3 Biologically informative model priors

We now turn our attention to the model prior $P(\gamma)$. As discussed in Section 2.3.3.2, a common choice of prior assumes that the *a priori* inclusion probabilities are independent and have Bernoulli marginal distributions $P(\gamma_j)$ with success parameter π. These priors provide no information regarding specific predictors and do not utilise domain knowledge. Employing predictor dependent hyperparameters π_j enables incorporation of prior knowledge that some predictors are more important than others. However, utilising such a prior may be difficult in practice due to the many hyperparameters that must be subjectively specified. We note also in this formulation, prior inclusion probabilities are still independent.

In many molecular biology settings, there is much information available regarding predictors which may be used to construct biologically informative model priors. This could be network and pathway structures, providing information on relationships between predictors, or information from publicly available datasets. However, it may not be obvious precisely *how* such information should be used and it is usually possible to encode several different, apparently plausible priors. We are therefore interested in investigating the question of how to choose between such priors.

Suppose we have S priors to choose from, with each prior, indexed by $s \in \{1, \ldots, S\}$, encoded by a function $f_s : \{0,1\}^p \to \mathbb{R}$ which scores a proposed model γ according to the prior information. We take the overall prior to be of a form similar to that used in Mukherjee *et al.* [2009],

$$P(\gamma \mid s, \lambda) \propto \exp\{\lambda f_s(\gamma)\} \tag{3.7}$$

where s is a hyperparameter (the 'source parameter') that selects amongst priors and λ is a hyperparameter controlling the overall strength of the prior.

Here, we consider two simple pathway-based priors. Proteins can be organised into pathways within the overall signalling network structure. These pathways consist of signalling cascades in which proteins transduce signals from the cell membrane down to the cell nucleus, ultimately leading to a cellular response. Therefore, signalling pathways can be associated with particular cellular functions. For example, the AKT/PI3K pathway is known to play a role in cell proliferation and apoptosis. The priors we consider capture information regarding two pathway-based features, via functions f_1 and f_2 respectively: (i) the number of pathways represented by proteins included in a model and (ii) the intra-pathway distances between proteins included in a model. Below we proceed to give details for each, making use of the following notation. We let $\mathcal{E}_k \subseteq \{1, \ldots, p\}$ denote the set of proteins contained in pathway k, $k \in \{1, \ldots, K\}$, and we let $\mathcal{E}_k^\gamma = \gamma \cap \mathcal{E}_k$ be the set of proteins that are both in model γ and in pathway k. We note that a protein is both allowed to be a member of more than one pathway or to not be a member of any pathways. If there is no prior information available, the pathway-based priors reduce to a flat prior over model space. Figure 3.1 illustrates properties of the two priors.

3.2.3.1 Number of pathways (f_1).

The first pathway-based feature encodes the notion that predictors that are influential in determining response may belong to a small number of pathways or, in contrast, may be spread across many pathways. We encode this belief by a function $f_1(\gamma)$ which counts the number of pathways represented in a model γ. Specifically, $f_1(\gamma) = \max(0, K_\gamma - 1)$ where K_γ is the pathway count given by $\min_A |A|$ for $A \subseteq \{1, \ldots, K\}$ satisfying

$$\bigcup_{k \in A} \mathcal{E}_k^\gamma = \bigcup_{k=1}^{K} \mathcal{E}_k^\gamma. \tag{3.8}$$

This definition prevents the empty model being *a priori* most probable and avoids double counting (proteins that are members of multiple pathways are considered to be a member of only one pathway for the purpose of calculating $f_1(\gamma)$, and this single pathway is selected to minimise the pathway count; see Figure 3.1). If the strength parameter λ is negative, the prior increasingly penalises models as number of pathways increases, whereas a positive value results in a prior that prefers models representing many pathways.

3.2.3.2 Intra-pathway distance (f_2).

The second feature we consider is that variables which jointly influence the response may either be close to each other in a network sense, or may in fact be far apart in the

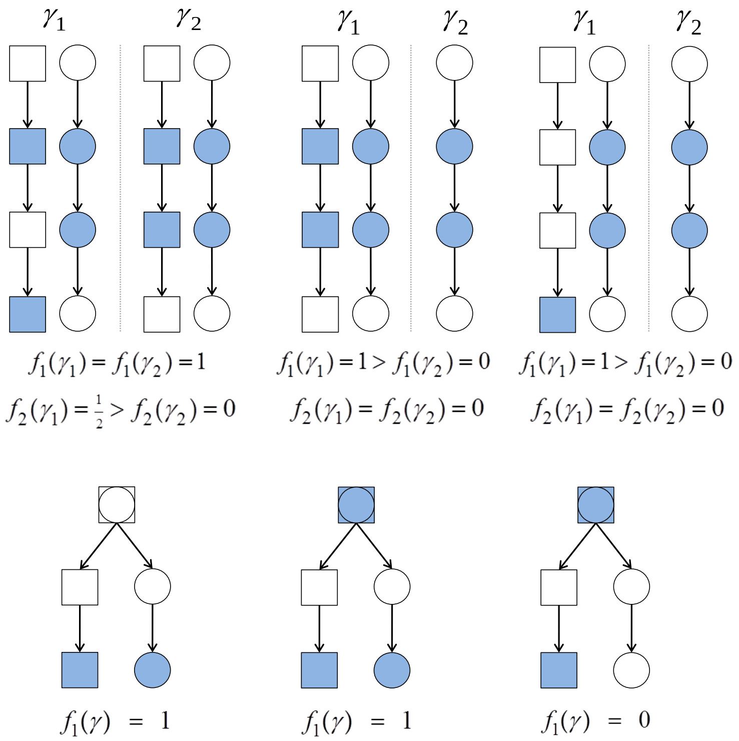

γ_1
γ_2
γ_1
γ_2
γ_1
γ_2
$f_1(\gamma_1) = f_1(\gamma_2) = 1$
$f_2(\gamma_1) = \frac{1}{2} > f_2(\gamma_2) = 0$
$f_1(\gamma_1) = 1 > f_1(\gamma_2) = 0$
$f_2(\gamma_1) = f_2(\gamma_2) = 0$
$f_1(\gamma_1) = 1 > f_1(\gamma_2) = 0$
$f_2(\gamma_1) = f_2(\gamma_2) = 0$
$f_1(\gamma) = 1$
$f_1(\gamma) = 1$
$f_1(\gamma) = 0$

Figure 3.1: **Properties of pathway-based priors.** Priors are encoded by functions $f_1(\gamma)$ (number of pathways) and $f_2(\gamma)$ (intra-pathway distance). Shaded components are contained in model γ and shapes represent different pathways. (Note that the pathway count and intra-pathway distance have a value of unity subtracted from them to obtain $f_1(\gamma)$ and $f_2(\gamma)$. See Sections 3.2.3.1 and 3.2.3.2 for full details.) Top row: Comparisons of the scoring functions. Top left - γ_1 has larger intra-pathway distance than γ_2; Top middle - distance is agnostic to number of pathways; Top right - addition of a singleton has no effect on distance. Bottom row: The root component in each network is in both pathways. However, $f_1(\gamma)$ is defined so as to avoid double counting. Bottom left/middle - both the circle and square pathways are required to obtain all components contained in γ, giving a pathway count of 2 and a score $f_1(\gamma) = 1$; Bottom right - taking the square pathway only is sufficient to obtain all components in γ, giving a pathway count of 1 and a score of $f_1(\gamma) = 0$.

network. This is done by a function $f_2(\gamma)$ which gives the average distance between pairs of proteins that are both in γ and in the same pathway. Specifically, the distance between two proteins j_1 and j_2, denoted $d(j_1, j_2)$, is the number of edges in the shortest (undirected) path between them. Then, we define $f_2(\gamma) = \max(0, D_\gamma - 1)$ where D_γ is the average of all $d(j_1, j_2)$ with $j_1, j_2 \in \mathcal{E}_k^\gamma$ for some k. In order for the distance to be defined for any two proteins in a pathway, we assume that the network topology for a pathway consists of a single connected component (in the undirected sense). We term a protein included in γ as a *singleton* if there are no other included proteins in the same pathway (i.e. protein j is a singleton if $\mathcal{E}_k^\gamma = \{j\}$ for some k). For models that only contain singletons or the empty model we set $D_\gamma = 0$. The function $f_2(\gamma)$ defined in this way satisfies a number of natural desiderata. It is agnostic to $|\gamma|$ and to the pathway count K_γ (see Figure 3.1). Also, it avoids double counting and is indifferent between models including only singletons and models with the smallest possible average distance of $D_\gamma = 1$. A negative strength parameter λ results in a prior that penalises larger intra-pathway distances, while a positive value encourages larger distances.

3.2.4 Empirical Bayes

We set the prior source parameter s and strength parameter λ in an objective manner using empirical Bayes (see Section 2.3.8 for an introduction to empirical Bayes methods). Specifically, we maximise the following marginal likelihood,

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{X}, s, \lambda) &= \mathbb{E}\left[p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma)\right]_{P(\gamma \mid s, \lambda)} \\ &= \sum_\gamma p(\mathbf{Y} \mid \gamma, \mathbf{X}_\gamma) P(\gamma \mid s, \lambda). \end{aligned} \tag{3.9}$$

For a given choice of hyperparameters, calculation of the marginal likelihood entails a summation over the model space. This can be calculated exactly by exploiting the model space restriction described above. The marginal likelihood score is calculated over a grid of hyperparameter values and those resulting in the largest score are used for variable selection.

3.2.5 Prediction

Given already observed data $(\mathbf{X}, \mathbf{Y})$, we can use the posterior predictive mean $\mathbb{E}[Y' \mid \mathbf{X}', \mathbf{X}, \mathbf{Y}]$ to predict the value of new response Y' from new predictor data

$\mathbf{X}'$. This entails averaging over models as shown in (2.38), which we reproduce here,

$$\mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{X}, \mathbf{Y}\right] = \sum_{\gamma} \mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{X}, \mathbf{Y}, \gamma\right] P(\gamma \mid \mathbf{Y}, \mathbf{X}) \tag{3.10}$$

where the choice of parameter prior described in Section 3.2.1 gives

$$\mathbb{E}\left[Y' \mid \mathbf{X}', \mathbf{X}, \mathbf{Y}, \gamma\right] = \frac{n}{n+1} \bar{\mathbf{X}}'_{\gamma} \left(\bar{\mathbf{X}}_{\gamma}^{\mathsf{T}} \bar{\mathbf{X}}_{\gamma}\right)^{-1} \bar{\mathbf{X}}_{\gamma}^{\mathsf{T}} \mathbf{Y} \tag{3.11}$$

and the model posterior $P(\gamma \mid \mathbf{Y}, \mathbf{X})$ is calculated via (3.2), (3.4) and (3.7). Equation (3.11) can be used to predict the value of Y' under a single model, instead of averaging over models; we use it below to make prediction with the MAP model.

3.3 Results

We first show an application of our proposed approach to synthetic response data generated from a published study of cell signalling, and then further illustrate the approach with an analysis of proteomic data and drug response from breast cancers.

3.3.1 Synthetic response data

In ongoing studies, such as that presented below, objective performance comparisons may be challenging, since we usually do not know which molecules are truly influential in driving biological response. At the same time, in fully synthetic data it can be difficult to mimic realistic correlations between variables within a pathway or across a network. For this reason, we empirically assessed the methods proposed using published single-cell, phosphoproteomic data [Sachs *et al.*, 2005] obtained by flow cytometry (see Section 2.2.4), with responses generated from that data. This preserved pathway-related correlation structure between predictors but permitted objective assessment. The dataset consists of $p = 11$ proteins and $n_{tot} = 853$ samples.[1]

Figure 3.2 shows a network and pathway structure for the 11 proteins for use with the biologically informative priors of the type described above. We used a network structure based on the one given in Sachs *et al.* [2005] (which reflects current knowledge of signalling interactions) and assigned the proteins into four pathways.

We considered two simulation models, γ_1^* and γ_2^*, each of which is a predictor subset consisting of three proteins: PIP3, ERK1/2, p38 for Simulation 1, and Raf,

[1] The complete dataset from Sachs *et al.* [2005] contains data obtained under nine different conditions, corresponding to different interventions. Here, we use the baseline dataset which contains 853 samples.

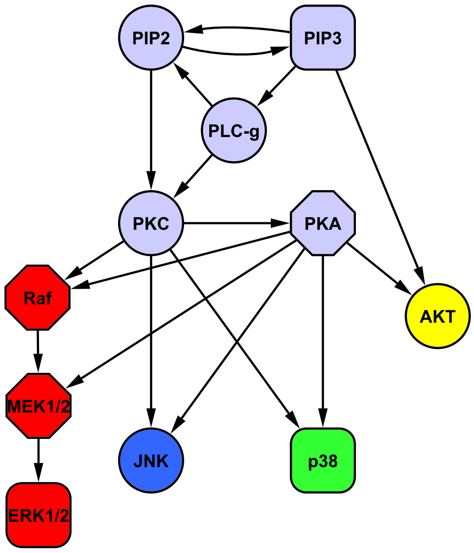


Figure 3.2: **Protein network and pathway structure for biologically informative priors in the synthetic response data study.** Responses were generated from published phosphoproteomic data [Sachs *et al.*, 2005] consisting of 11 proteins and 853 samples (baseline data only). Network structure shown here is based on that given in Sachs *et al.* [2005]. Proteins were divided into four pathways, denoted by node colours red, blue, green and yellow. The grey nodes are each members of all four pathways. The square and octagonal proteins influence the response in simulation models γ_1^* and γ_2^* respectively.

Mek1/2, PKA for Simulation 2. In each case, the three proteins were chosen to be favoured by a particular prior. γ_1^* is favoured by the intra-pathway distance prior ($s = 2$) with positive λ; the proteins included in γ_1^* had a large average intra-pathway distance (d(PIP3,ERK1/2)=4 and d(PIP3,p38)=3 (both undirected paths via AKT and PKA), giving average distance of 3.5 and score of $f_2(\gamma_1^*) = 2.5$) and incorporated a medium number of pathways (red and green pathways, giving score of $f_1(\gamma_1^*) = 1$). γ_2^* is favoured by either the number of pathways prior ($s = 1$) with negative λ or the intra-pathway distance prior ($s = 2$) with negative λ; the proteins included in γ_2^* had both a small intra-pathway distance (distance between each pair of proteins is unity, giving average distance of unity and score of $f_2(\gamma_2^*) = 0$) and incorporated a small number of pathways (red pathway only, giving score of $f_1(\gamma_2^*) = 0$). Since, by construction, each model is favoured by a particular prior, we can test the ability of the empirical Bayes approach to select appropriate hyperparameter values. Response data Y were generated using (3.1); $Y = A + BC + \epsilon$, where A, B, C are the three influential variables, and ϵ is independent Gaussian noise.

We are especially interested in the small-sample regime that is often of interest in molecular studies. We therefore subsampled (without replacement) $n = 35$ training data from the dataset (this matched the sample size of the drug response study reported below), and assessed predictive ability on the remaining, held-out data ($\tilde{n} = n_{tot} - n = 818$).

Subsampling was repeated to give 5,000 training/test pairs, over which results are reported below. At each iteration, only small-sample training data was used for inference. The empirical Bayes method was employed to set prior source and strength parameters (using training data only), with $\lambda \in [-5, 5]$ (this specification permits a flat prior if empirical Bayes analysis supports neither prior). Posterior inclusion probabilities were then calculated as described above.

We assessed performance by comparing the true underlying model γ^* to the model γ_τ obtained by thresholding posterior inclusion probabilities at level τ. γ^* can be compared to γ_τ using the number of true positives (TPs) and number of false positives (FPs). TPs are predictors included in γ_τ and also included in γ^* (i.e. $\gamma^* \cap \gamma_\tau$), while FPs are predictors included in γ_τ but not included in γ^* (i.e. $\gamma_\tau \backslash \gamma^*$). For results from each small-sample dataset, a receiver operating characteristic (ROC) curve was constructed by plotting number of TPs against number of FPs for varying thresholds τ. Figure 3.3 shows average ROC curves over the 5,000 iterations, together with area under the ROC curve (AUC). AUC is a summary of the curve and provides a measure of variable selection accuracy, with higher scores indicating better performance. The score is normalised to take a value between 0 and 1. The Bayesian variable selection (BVS) method with empirical Bayes and linear model with interaction terms ('EB') is compared with six other approaches:

(i) BVS with flat prior and linear model with interaction terms ('flat +int');

(ii) BVS with a prior that is incorrect with respect to the true, underlying model: intra-pathway distance prior (f_2) favouring small distances ($\lambda = -5$) for Simulation 1 and large distances ($\lambda = 5$) for Simulation 2 ('incorrect');

(iii) BVS with flat prior and linear model with no interaction terms ('flat -int');

(iv) penalised-likelihood lasso regression [Tibshirani, 1996] using a linear model with pair-wise interaction terms (see Section 2.3.3.3 and below for further details; 'Lasso');

(v) BVS with a Markov random field prior [Li and Zhang, 2010] and linear model with interaction terms (Simulation 2 only, see below for further details; 'MRF prior'); and

(vi) absolute correlation coefficients between each predictor and response ('corr').

Recall from Section 2.3.3.3 that lasso regression performs variable selection by placing an L_1 penalty on the regression coefficients. This has the effect of shrinking a subset of regression coefficients to exactly zero; the predictors with non-zero coefficients are taken as the inferred model. Sparsity of the inferred model is controlled by a tuning parameter, which we set by 5-fold cross-validation (see Section 2.3.2.1 for background information on cross-validation). This method results in a single inferred model (i.e. point estimate). However, a full ROC curve can still be obtained by thresholding absolute regression coefficients.

Markov random field priors have previously been used in Bayesian variable selection to take network structure of predictors into account [Li and Zhang, 2010; Stingo and Vannucci, 2011]. A Markov random field is an undirected graphical model $G = (V, E)$ in which vertices V represent variables (here, the predictors) and edges E represent probabilistic relationships between them. Background information regarding graphical models and Markov random fields can be found in Section 2.3.4. Let $A = (a_{i,j})$ be a binary symmetric matrix with $a_{i,j} = 1$ if and only if edge $(i, j) \in E$. Then, the Markov random field prior is given by

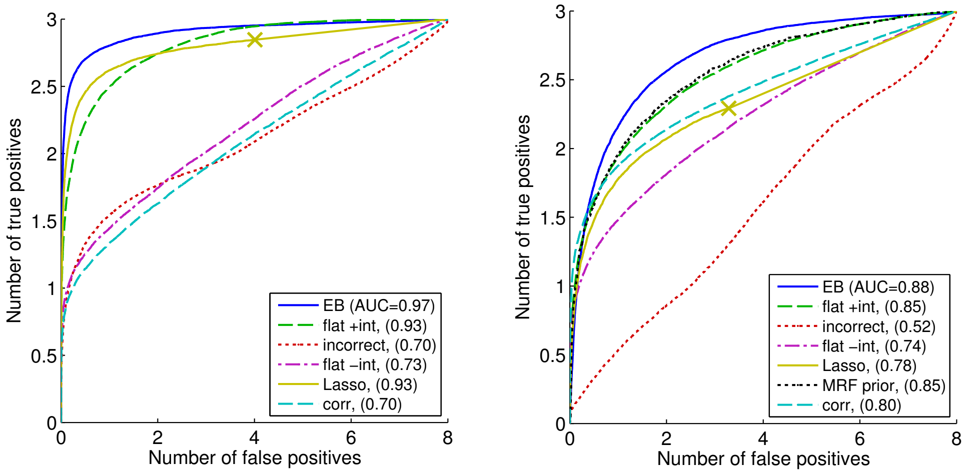
$$P(\gamma \mid \lambda) \propto \exp\left\{\lambda \gamma^{\mathsf{T}} A \gamma\right\}. \tag{3.12}$$

This prior encourages selection of predictors whose neighbours in G are also included in the model. The strength of this preference is controlled by a parameter $\lambda \geq 0$. We apply this prior to Simulation 2, where the underlying true model contains predictors that are neighbours in the network. We use a linear model with interaction terms and set λ with empirical Bayes. The graph structure G is obtained from the structure shown in Figure 3.2 by converting all directed edges to undirected edges.

We observe that, in both simulations, the automated empirical Bayes analysis, with pathway-based priors, improves performance over the flat prior and provides substantial gains over an incorrect prior. The empirical Bayes approach selected the correct prior in 90% of iterations for Simulation 1 and 95% of iterations for Simulation 2 (for Simulation 1 correct prior parameters were $s = 2$ with $\lambda > 0$, median value of λ selected was $\lambda = 3$; for Simulation 2 correct prior parameters were $s = 1$ or $s = 2$ with $\lambda < 0$, median value of λ selected was $\lambda = -5$ for both $s = 1$ and $s = 2$). Since the Lasso regression method does not incorporate prior information, it is unsurprising that it is also outperformed by the empirical Bayes approach. However, in Simulation 2 it does not perform at all well, with reduced performance compared to simply looking at correlations between predictors. Intriguingly, the strength parameter λ for the Markov random field prior was set to zero by empirical Bayes in 91% of iterations in Simulation 2. Thus, its performance is almost identical to that of a flat prior. We discuss this further in Section 3.4 below. Due to its inability to model combinatorial interplay, the linear model without interaction terms is outperformed by the linear model with interaction terms.

The failure of the incorrect prior illustrates the importance of prior elicitation. Moreover, our results demonstrate that the proposed empirical Bayes approach can select a suitable prior automatically, even under very small sample conditions (here $n = 35$). If the data is not in agreement with a proposed prior, then it is desirable that $\lambda = 0$ is selected by empirical Bayes, resulting in a flat prior. To test this, we used the model in Simulation 2 with a prior that favoured models with predictors from many pathways (i.e. number of pathways prior with λ restricted to be non-negative). This prior does not reflect the true, underlying model, which contains a small number of pathways. Empirical Bayes analysis successfully selected $\lambda = 0$ in 98% of iterations.

For each dataset, we used the posterior predictive distribution ((3.10); calculated via exact model averaging) to predict responses for held-out test data. Mean absolute predictive errors, obtained by averaging over all 5,000 train/test iterations, are shown in Table 3.1 ('MA'). The empirical Bayes approach with pathway-based priors shows some improvement in predictive accuracy over a flat prior (and Markov random field prior in Simulation 2), and substantial improvements over both the 'incorrect' prior and a baseline linear model (without interaction terms) including all



(a) Simulation 1

(b) Simulation 2

Figure 3.3: **Synthetic response data, average ROC curves.** Number of true positives plotted against number of false positives for Simulations 1 and 2. Proteomic data from Sachs *et al.* [2005] were used to create response data with true underlying model known to favour a particular prior: Simulation 1 - distance prior with positive λ; Simulation 2 - either distance prior or number of pathways prior with negative λ. Average ROC curves are obtained from 5000 iterations. Legend - 'EB': Bayesian variable selection (BVS) using empirical Bayes to select and weight prior automatically (and linear model with interaction terms); 'flat +int': BVS with a flat prior (and linear model with interaction terms); 'incorrect': BVS with a wrong prior with respect to true, underlying protein set (and linear model with interaction terms; see main text for details); 'flat -int': BVS with flat prior and linear model with no interaction terms; 'Lasso': Lasso linear regression with interaction terms (curve produced by thresholding absolute regression coefficients, while marker 'X' is single model obtained by taking only predictors with non-zero coefficients); 'MRF prior': BVS with a Markov random field prior [Li and Zhang, 2010] (and linear model with interaction terms); 'corr': absolute Pearson correlations between each protein and response. Area under the (average) ROC curve ("AUC") appears in parentheses.

11 predictors (i.e. no variable selection). Lasso regression offers the best predictive performance, with slight gains over Bayesian variable selection with empirical Bayes (we note that prediction used regression coefficients obtained by maximum penalised likelihood estimation; the alternative of using Equation (3.11) with the single model corresponding to non-zero coefficients gave very poor predictive accuracy, inferior to the baseline linear approach; data not shown). We also found that model averaging provided gains relative to prediction using the MAP model (and Equation (3.11)), with a 5% and 7% decrease in error on average for Simulation 1 and Simulation 2 respectively (see Table 3.1, 'MAP').

The only user-set parameters in the proposed method are d_{max} (the maximum number of predictors allowed in a model), and the range of values for the prior strength parameter λ to consider in the grid search optimisation in empirical Bayes. We sought to check the sensitivity of our results to these parameters. As described in 'Methods' above, we set $d_{max} = 4$ and considered $\lambda \in [-5, 5]$. We compared the posterior inclusion probabilities inferred from 50 iterations of Simulation 2 to those obtained using (i) an increased maximum number of included predictors of $d_{max} = 5$; (ii) Markov chain Monte Carlo-based (MCMC) inference with no restriction on number of included predictors, and (iii) an increased range for the prior strength $\lambda \in [-10, 10]$ (see Figure 3.4). We found very close agreement in all cases, indicating that results reported do not depend on the sparsity restriction or the chosen range for λ.

In simulation 2, the smallest value of $\lambda = -5$ was selected by empirical Bayes in a majority of iterations. The true, underlying model has the minimum possible number of pathways and intra-pathway distance. Hence, the strong (negative) prior strength is appropriate because it causes the prior to heavily penalise any model not satisfying these minima. Under the increased range for λ, the smallest value ($\lambda = -10$) was still selected in these iterations, but results were almost identical. This indicates that the prior was already having close to maximal influence at the lower value of $\lambda = -5$.

3.3.2 Cancer drug response data

The ability to be able to predict an individual patient's response to drug treatment, based on molecular characteristics of cancer cells, is an important goal for personalised cancer medicine. Previous studies have attempted to make such predictions from high-throughput molecular data. For example, Staunton *et al.* [2001] predict drug response from gene expression profiles and Boyd *et al.* [2008] attempt to find protein biomarkers that are predictive of drug response from phosphoproteomic

	EB prior	Flat prior	‘Incorrect’ prior	Lasso	MRF prior	Baseline linear
Simulation 1						
MA	0.789±0.003	0.813±0.003	0.845±0.003	0.781±0.003	-	1.00±0.002
MAP	0.816±0.004	0.871±0.005	0.883±0.003			
Simulation 2						
MA	0.886±0.004	0.891±0.003	0.947±0.003	0.875±0.003	0.907±0.004	1.00±0.002
MAP	0.937±0.005	0.963±0.005	1.019±0.005		0.974±0.005	

Table 3.1: **Synthetic response data, predictive errors from held-out test data.** Predictions using small-sample training data (n = 35) and held-out test data (n = 818; total of 5,000 train/test pairs) for Simulation 1 and Simulation 2 (see text for details). Results shown are mean absolute predictive errors ± SEM for the following methods: ‘EB prior’: Bayesian variable selection (BVS) with biologically informative pathway-based prior with source and strength parameters set by empirical Bayes; ‘Flat prior’: BVS with a flat prior; “Incorrect’ prior’: BVS with a wrong prior with respect to true, underlying protein set (see text for details); ‘Lasso’: Lasso linear regression with interaction terms; ‘MRF prior’: BVS with a Markov random field prior [Li and Zhang, 2010]; ‘Baseline linear’: linear regression with all 11 predictors. All these methods included interaction terms except ‘Baseline linear’. For BVS, predictions were made using the posterior predictive distribution with exact model averaging (‘MA’) and using the *maximum a posteriori* model (‘MAP’).

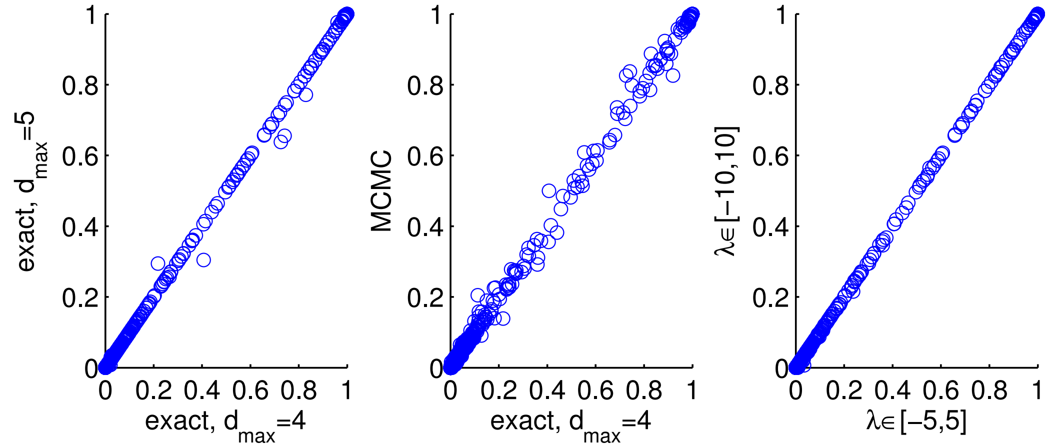


Figure 3.4: **Synthetic response data; effect of sparsity restriction and range of prior strength parameter.** Results reported in Figure 3.3, for the empirical Bayes approach, were obtained by exact model averaging with the number of predictors included in a model restricted to not exceed $d_{max} = 4$. Posterior inclusion probabilities for 50 simulated datasets from Simulation 2 were compared with results obtained by exact model averaging with an increased maximum number of included predictors of $d_{max} = 5$ (left) and using Markov chain Monte Carlo-based model averaging with no sparsity restriction (centre). Sensitivity to the range of prior strength parameter values considered by empirical Bayes was also assessed by comparing the posterior inclusion probabilities obtained with $\lambda \in [-5, 5]$ to those obtained with an increased range of $\lambda \in [-10, 10]$ (right).

data. Indeed, signalling proteins have potential to be predictive biomarkers since, as discussed in Section 2.1.3, aberrant signalling is heavily implicated in almost every aspect of cancer biology and signalling proteins are targets for many emerging cancer therapies. Here, we apply the methods proposed to probing phosphoproteomic influences on response to an anti-cancer agent, Triciribine.

Phosphoprotein abundance was assayed in a high-throughput manner using the KinetWorks$^{\text{TM}}$ system (Kinexus Inc, Vancouver, Canada; this is a high-throughput version of protein immunoblotting, which is outlined in Section 2.2.1), for 52 proteins related to epidermal growth factor (EGF) signalling, in each of 35 breast cancer cell lines (see Tables A.1 and A.2 for details). The EGFR signalling network plays a central role in breast cancer biology (see Section 2.1) and the cell lines used have previously been shown to retain much of the biological heterogeneity of primary tumours [Neve *et al.*, 2006]. GI50 (log transformed) was used to quantify response to Triciribine for each of the 35 cell lines [Heiser *et al.*, 2011]. GI50 is the concentration that causes 50% growth inhibition compared to a baseline. A network (with a total of five pathways) was constructed using canonical signalling pathway maps available at the online repository `cellsignal.com` (see Figure 3.5). In particular, signalling pathway maps for the MEK/MAPK, PI3K/AKT, mTOR and insulin receptor pathways were used. The network includes indirect edges via components not included in our study and edges between protein phosphoforms and isoforms. Edges marked as 'tentative' in the online repository were not included in our network. Proteins were assigned into one or more of the following pathways, denoted in Figure 3.5 by node colour: PI3K/AKT (red), MEK/MAPK (yellow), JNK/JUN (blue), SRC/JAK-STAT (purple) and HSP27 (green). Proteins with no network or pathway information available in the repository were left unconnected and not assigned to a pathway (white nodes in Figure 3.5).

Figure 3.6 shows marginal likelihood scores arising from empirical Bayes. This selects the intra-pathway distance prior ($s = 2$) with hyperparameter $\lambda = 5$ (i.e. a prior promoting larger distances). Due to the small sample size, we tested robustness of this choice by running empirical Bayes with each data sample removed. The same prior was selected in 86% of the iterations.

Figure 3.7 shows posterior inclusion probabilities obtained under three prior regimes: empirical Bayes (intra-pathway distance prior with $\lambda = 5$), flat prior and an "incorrect" prior that is not optimal according to the empirical Bayes analysis (number of pathways prior with $\lambda = -5$). Phospho-IR and phospho-RB(S259) stand out in the empirical Bayes analysis. Triciribine targets AKT, which inhibits apoptotic processes and is heavily implicated in cancer signalling [Yang *et al.*, 2004]. IR (in-

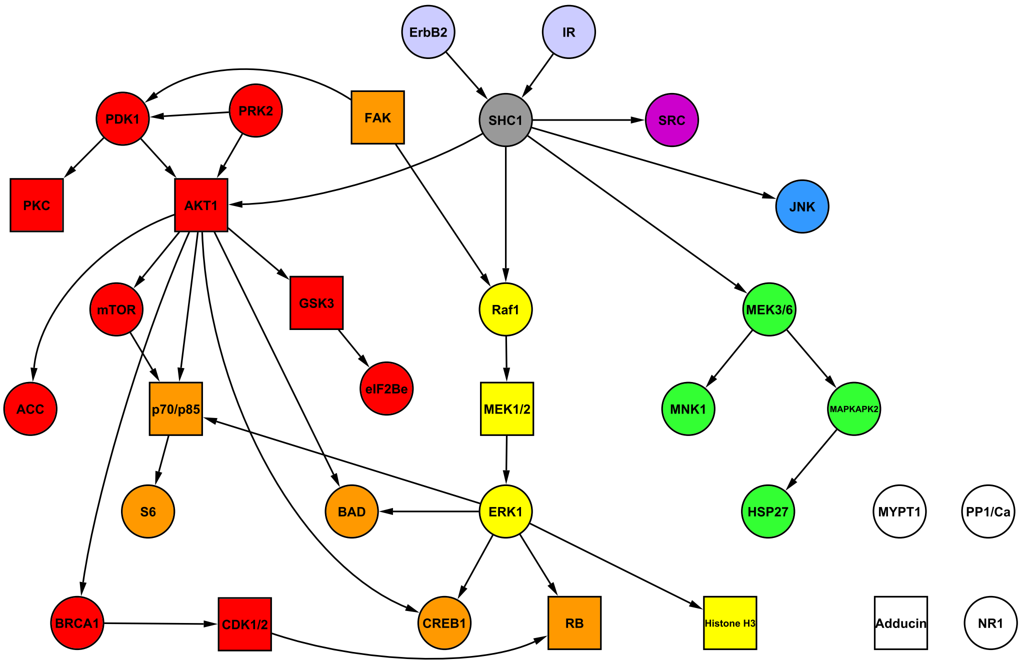


Figure 3.5: **Network and pathway structure for biologically informative priors in the cancer drug response data study.** Network constructed using information from `cellsignal.com`. Square nodes represent fully connected subnetworks consisting of iso-forms and phospho-forms of the named protein (see Table A.1). Node colouring represents pathway structure. Red, blue, yellow, green and purple nodes denote 5 pathways. Orange nodes are in both the red and yellow pathways. Light grey nodes are in all 5 pathways. Dark grey node is in all pathways except the purple pathway. White nodes are not assigned to a pathway.

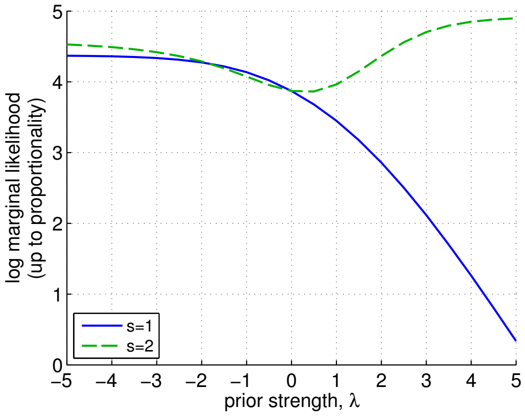
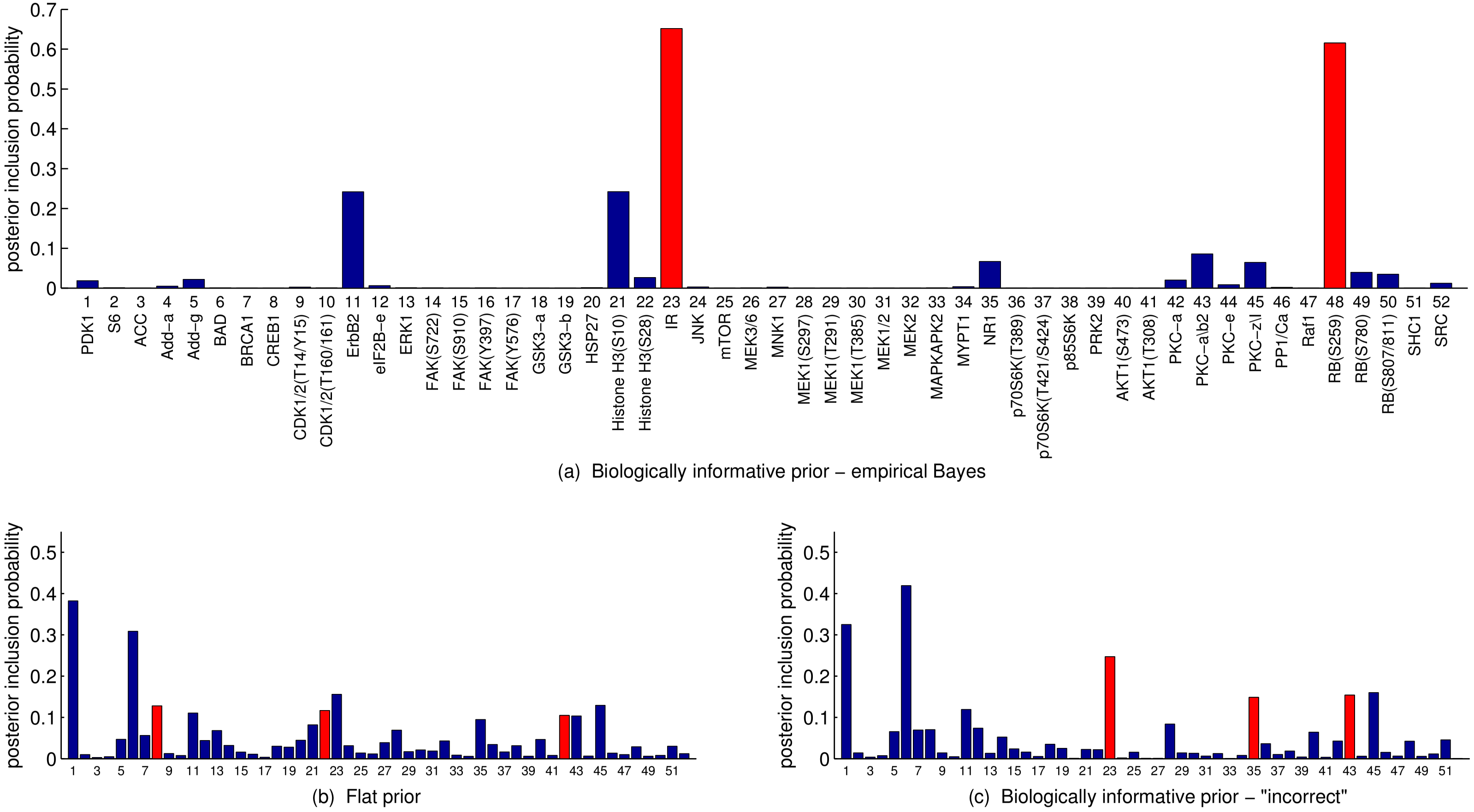


Figure 3.6: **Drug response data, empirical Bayes analysis.** Parameters controlling source of prior information and prior strength (s and λ respectively) were set objectively using the data. Log marginal likelihood (calculated exactly up to a constant) is plotted against λ for $s = 1$ (number of pathways prior) and $s = 2$ (intra-pathway distance prior). Parameters were set to the values with maximal marginal likelihood: $s = 2$ and $\lambda = 5$.

sulin receptor) is a tyrosine kinase receptor, known to stimulate the AKT pathway [Burgering and Coffer, 2002], and it has been suggested that the RB/E2F pathway, which is also known to play a role in cancer [Nevins, 2001], has an effect on AKT activity via transcriptional regulation [Chaussepied and Ginsberg, 2004]. Hence, the salience of IR and RB accords with known biology and drug mechanism. The MAP model for each prior regime is highlighted in red in Figure 3.7. We note that these models do not always contain the proteins with highest inclusion probabilities.

We performed Leave-One-Out-Cross-Validation (LOOCV), making predictions for the held-out test sample using both posterior model averaging (3.10) and the MAP model (3.11). The full variable selection approach, including selection of hyperparameters with empirical Bayes, was carried out at each cross-validation iteration. Table 3.2 shows mean absolute predictive errors, with comparisons made as in the synthetic response data study above. For the 'incorrect' prior, the prior source parameter not selected by empirical Bayes was used, along with the optimal strength parameter λ for that prior. Mirroring the synthetic data results, we observe that prior elicitation with empirical Bayes provides a small increase in mean predictive accuracy over a flat prior and an 'incorrect' prior, and Lasso regression has lowest mean predictive error. We note, however, that due to the very small sample size, differences in mean predictive error between these regimes are not conclusive. Yet, they all show a clear improvement over the baseline linear approach, and model averaging results in an average 36% decrease in predictive error over us-



(a) Biologically informative prior – empirical Bayes

(b) Flat prior

(c) Biologically informative prior – "incorrect"

Figure 3.7: **Drug response data, posterior inclusion probabilities.** Obtained via exact model averaging with (a) biologically informative pathway-based model prior with parameters set objectively using empirical Bayes ($s = 2$ (intra-pathway distance), $\lambda = 5$ - see Figure 3.6), (b) flat prior and (c) "incorrect" biologically informative prior that that contradicts the empirical Bayes analysis ($s = 1$ (number of pathways), $\lambda = -5$). Posterior inclusion probabilities provide a measure of how influential each protein is in determining drug response. Proteins contained in the single MAP model are shaded red.

	EB prior	Flat prior	'Incorrect' prior	Lasso	MRF prior	Baseline linear
MA	0.85±0.12	0.86±0.11	0.93±0.15	0.77±0.09	0.86±0.11	1.00±0.14
MAP	1.00±0.16	1.26±0.17	1.22±0.17		1.26±0.17	

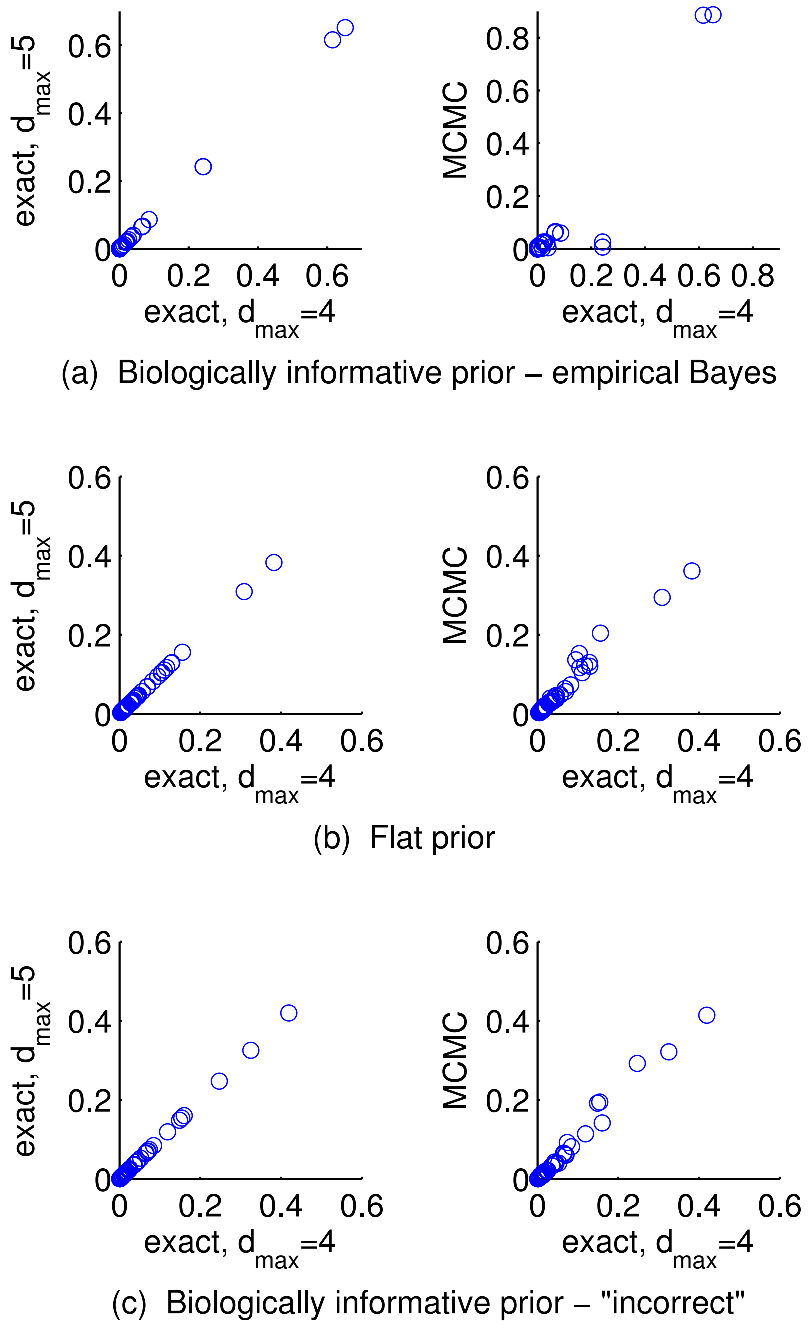
Table 3.2: **Drug response data, predictive errors from cross-validation.** Predictions using leave-one-out-cross-validation (see text for details). Results shown are mean absolute predictive errors ± SEM for the following methods: 'EB prior': Bayesian variable selection (BVS) with biologically informative pathway-based prior with source and strength parameters set by empirical Bayes; 'Flat prior': BVS with a flat prior; "Incorrect' prior': BVS with a wrong prior with respect to true, underlying protein set (see text for details); 'Lasso': Lasso linear regression with interaction terms; 'MRF prior': BVS with a Markov random field prior [Li and Zhang, 2010]; 'Baseline linear': linear regression with all 11 predictors. All these methods included interaction terms except 'Baseline linear'. For BVS, predictions were made using the posterior predictive distribution with exact model averaging ('MA') and using the *maximum a posteriori* model ('MAP').

ing MAP models. The prior strength parameter for the Markov random field prior was set to $\lambda = 0$ by empirical Bayes in every cross-validation iteration, resulting in identical performance to the flat prior.

We again checked sensitivity of results to the restriction on the number of predictors included in a model, $d_{max} = 4$. The results in Figure 3.7 were compared with those obtained using an increased maximum number of included predictors of $d_{max} = 5$ and using MCMC-based inference with no such restriction (see Figure 3.8). The strong agreement between $d_{max} = 4$ and $d_{max} = 5$ suggests that the minor differences observed between $d_{max} = 4$ and MCMC are a result of inherent Monte Carlo error. We also see a close agreement between results in Figure 3.7a (using $\lambda \in [-5, 5]$) and those obtained by optimising over the increased range of $\lambda \in [-10, 10]$ (see Figure 3.9). This shows that results reported do not depend on the sparsity restriction or the range of values considered for the prior strength parameter.

3.4 Discussion

Model priors incorporating biological information can play an important role in Bayesian variable selection, especially at the small sample sizes characteristic of molecular studies. In applications where there are multiple sources of prior information, or multiple possible prior specifications, the empirical Bayes approach we put forward permits objective selection and weighting. We demonstrated, on synthetic response data, that a biologically informative prior, with hyperparameters set by empirical Bayes, can have benefits over both a flat prior and a subjectively



(a) Biologically informative prior – empirical Bayes

(b) Flat prior

(c) Biologically informative prior – "incorrect"

Figure 3.8: **Drug response data, effect of sparsity restriction.** Posterior inclusion probabilities in Figure 3.7 were obtained by exact model averaging with the number of predictors included in a model restricted to not exceed d_{max} = 4. These results were compared with results obtained by exact model averaging with an increased maximum number of included predictors of d_{max} = 5 (left column) and using Markov chain Monte Carlo-based model averaging with no sparsity restriction (right column).

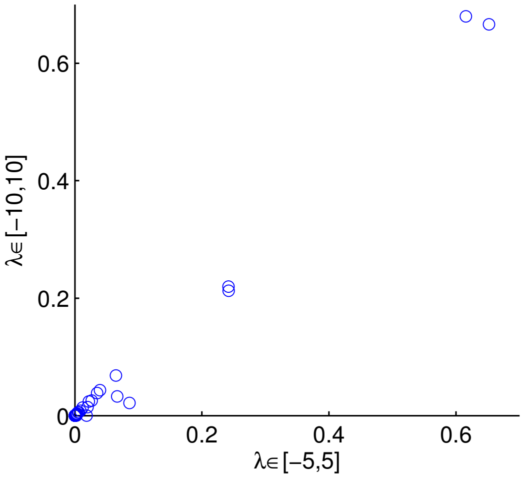


Figure 3.9: **Drug response data; sensitivity to range of prior strength parameter.** Posterior inclusion probabilities reported in Figure 3.7a were obtained using the proposed empirical Bayes approach, with prior strength parameter optimised over the interval $\lambda \in [-5, 5]$. These results are compared to those obtained with an increased range of $\lambda \in [-10, 10]$.

formed prior which is incorrect with respect to the underlying system. Thus the approach can aid prior elicitation and has potential to significantly improve results by guarding against the use of mis-specified priors. We also observed that, while lasso regression can offer some improvement in predictive performance over the Bayesian approaches, its accuracy in selecting the correct underlying model (i.e. variable selection) can be poor, thereby affecting interpretability of results. The proposed empirical Bayes approach offers clear gains in this respect. We have also shown an application on cancer drug response data and obtained biologically plausible results.

We developed informative priors in the context of protein signalling based on two high-level features derived from network information: the number of pathways a subset of predictors incorporates and the intra-pathway distance between proteins in a proposed model. This formulation used the entire network structure in an intuitive way, removing the need to specify individual prior probabilities for each variable and avoiding assumptions of prior independence between variables.

Our pathway-based priors form part of a growing literature on exploiting existing domain knowledge to aid inference, especially in the small sample setting. For example, recent variable selection studies also make use of graph structure within a Bayesian Markov random field prior [Wei and Li, 2008; Li and Zhang, 2010; Monni and Li, 2010] and within a non-Bayesian framework [Li and Li, 2008; Binder and Schumacher, 2009; Slawski *et al.*, 2010], essentially preferring models containing predictors that are neighbours in the graph. This is similar in spirit to the special case of our prior where the network consists of a single pathway and short intra-pathway distances are strongly preferred. We compared our pathway-based priors to the Markov random field prior, but found that empirical Bayes frequently set the prior strength parameter to zero, essentially preferring a flat prior. This is possibly due to the parameterisation of the Markov random field prior, which is not agnostic to the number of included predictors in the model $|\gamma|$; addition of a predictor to a model could lead to a substantial increase in the prior score. Indeed, it has previously been noted that Markov random field priors can be unstable with the occurance of phase transitions in $|\gamma|$ [Li and Zhang, 2010]. Hence, the prior prefers less sparse models, but these models do not agree well with the data, as more complex models are penalised by the marginal likelihood. In contrast, our distance prior is based on an average distance measure and so is somewhat indifferent to $|\gamma|$. We note that biologically informative priors have also been used for classification [Zhu *et al.*, 2009; Stingo and Vannucci, 2011; Guillemot *et al.*, 2011] and network structure learning [Bernard and Hartemink, 2005; Werhli and Husmeier, 2007; Mukherjee and Speed, 2008].

We used a continuous regression framework with interaction terms. While discrete models are naturally capable of capturing non-linear interplay between components, the discretisation process results in a loss of information. Continuous models avoid this loss, but the response is usually assumed to depend linearly on predictors. The product terms in our model provide the possibility of capturing influences on the response of interest by interplay between predictors, including higher-order interactions. Chipman [1996] and Jensen *et al.* [2007] have employed a related approach allowing pairwise interactions only. We note that, under our formulation, model complexity grows rapidly with number of included predictors. However, complex models are naturally penalised by the marginal likelihood formulation giving overall sparse, parsimonious models, yet allowing for complex interplay via product terms.

We carried out variable selection using exact model averaging. This was made possible by means of a sparsity restriction. Sparsity constraints have been employed in previous work in Bayesian variable selection [Jiang, 2007; Mukherjee *et al.*, 2009] and also in the related setting of inference of gene regulatory networks [Husmeier, 2003; Werhli and Husmeier, 2007]. The sparsity-constrained approach proposed is attractive as it yields exact posterior probabilities and facilitates exact empirical Bayes analysis. Sparsity is a reasonable assumption in settings where it is likely that only a few predictors play a key role in influencing a response. In such

	Linear model without interaction terms				Linear model with interaction terms			
	$d_{max} = 2$	$d_{max} = 3$	$d_{max} = 4$	$d_{max} = 5$	$d_{max} = 2$	$d_{max} = 3$	$d_{max} = 4$	$d_{max} = 5$
p=30	0.1	1.1	8.7	9.5	0.4	4.7	38.6	374.6
p=60	0.5	10.5	114.3	-	1.8	39.4	661.6	-
p=120	2.8	116.3	-	-	8.2	350.1	-	-
p=500	150.3	-	-	-	238.7	-	-	-

Table 3.3: **Illustrative computation times.** Computation times (in seconds) for proposed Bayesian variable selection procedure, using empirical Bayes to select between two priors ($S = 2$) and to set the prior strength parameter λ (optimisation performed over ten values of λ). Results shown for varying values of d_{max} (maximum number of predictors allowed in a model) and p (total number of predictors), for both a linear model without interaction terms and a linear model with interaction terms. (Data and model priors generated using random numbers and results are averages over three iterations. Computation performed on a standard single-core personal computer; 1.6GHz, 2GB RAM. '-' denotes a (p, d_{max}) regime where the procedure failed due to insufficient memory.)

settings, and where data is of small-to-moderate dimensionality, our exact approach is computationally efficient and deterministic with no requirement of MCMC convergence diagnostics. This, together with empirical Bayes and the choice of parameter priors, results in the overall approach having very few user-set parameters.

In applications of higher dimensionality, where the exact calculation is no longer feasible, empirical Bayes can still be performed using an approximate conditional marginal 'likelihood' approach as seen in George and Foster [2000] and Yuan and Lin [2005]. This involves optimisation over the model space instead of averaging. MCMC, with the selected hyperparameter values, can then be used to estimate inclusion probabilities. Alternatively, a fully Bayes MCMC approach could be taken, which places a prior on the hyperparameters and integrates them out [see e.g. Nott and Green, 2004].

Illustrative computational times for our approach are shown in Table 3.3, for four values of p (number of predictors) and four values of d_{max} (maximum number of predictors allowed in a model). We also considered linear models with and without interaction terms. Empirical Bayes was used to select between two priors ($S = 2$) and to set the prior strength parameter (optimisation performed over ten values of λ). The computation time scales as $d_{max}p^{d_{max}}$ for the model without interaction terms and $(2^{d_{max}}-1)p^{d_{max}}$ for the model with interaction terms. We see that the approach is fast on datasets of moderate dimensionality (~100 variables) with $d_{max} = 3$. We note that shortage of memory was the limiting factor on our machine. Computational time could also be easily improved by using multiple cores to calculate empirical Bayes marginal likelihood scores for multiple values of λ simultaneously.

We showed examples of automated selection between multiple sources of ancillary information, but, rather than selecting a single source, the methods proposed could be generalised to allow combinations of complementary information sources as seen in Jensen *et al.* [2007]. While our priors were based on pathway and network structure, the methods can also permit integration and weighting of publicly available data, which while plentiful, can be of uncertain relevance to a given study.

Chapter 4

Dynamic Bayesian networks reveal novel signalling links in a cancer cell line

4.1 Introduction

Protein signalling plays a central role in diverse cellular functions including proliferation, apoptosis and differentiation. An emerging literature suggests that signalling networks may be 'rewired' in specific contexts, including cancer [Pawson and Warner, 2007; Yuan and Cantley, 2008]. Indeed, signalling and aberrations thereof are central to almost every aspect of cancer biology [Hanahan and Weinberg, 2000, 2011]. An introduction to protein signalling and cancer can be found in Section 2.1. In Section 2.1.3 we described, using the ErbB2 oncogene as an example, how genetic mutations can lead to dysregulated signalling. However, in general, the manner in which genomic aberrations in specific cancers are manifested at the level of signalling networks is not currently well understood.

Signalling is a multi-dimensional, dynamic process in which post-translational protein modifications (e.g. phosphorylation) play a key role. Therefore, elucidating signalling networks in a data-driven manner, specific to a context of interest, such as a cell line, cell type, tissue, or disease state, requires the ability to probe post-translational modification states in multiple proteins through time and across samples. However, proteomic analyses on this scale remain challenging. Several high-throughput technologies are described in Section 2.2. Flow cytometry assays yield large sample size, single-cell datasets [Sachs *et al.*, 2005], but on account of spectral overlap in fluorophores do not scale well to systems-level studies. Mass spectrometry approaches are promising, but currently have limited sample throughput and are often not sensitive enough to detect low abundant phosphoproteins.

At the same time, the modelling of signalling connectivity poses statistical challenges. Noise, both intrinsic and experimental, is ubiquitous in this setting and network components may interact in a complex, non-linear manner [Lauffenburger and Linderman, 1993; Citri and Yarden, 2006; Rubbi *et al.*, 2011]. In addition, candidate networks may differ with respect to model dimension (i.e. network models with more edges have a larger number of parameters). Analyses that do not account for this run the risk of overfitting the model to the observed data, preferring networks that are over-complex, yet not predictive. This makes the trade-off between fit-to-data and model parsimony a crucial one in network modelling. An introductory discussion of model complexity and model selection is given in Section 2.3.2.

In this Chapter we present a data-driven approach to the characterisation of signalling networks (Figure 4.1). We exploit reverse-phase protein array technology (see Section 2.2.5) to interrogate dynamic signalling responses in a defined set of 20 phosphoproteins, including members of MAPK, STAT and AKT pathways, in a specific breast cancer cell line. The analysis is rooted in graphical models; in particular, we use directed graphical models known as dynamic Bayesian networks (DBNs). For background information on graphical models, including static Bayesian networks (BNs) and DBNs, see Sections 2.3.4 and 2.3.5.2. Previous studies have applied DBNs to gene expression time series data for structure learning of gene regulatory networks [Husmeier, 2003; Perrin *et al.*, 2003; Kim *et al.*, 2003; Zou and Conzen, 2005; Grzegorczyk *et al.*, 2008; Grzegorczyk and Husmeier, 2011a; Rau *et al.*, 2010; Robinson and Hartemink, 2010; Li *et al.*, 2011]. DBNs have so far not been used for structure learning of protein signalling networks. This is in contrast to static Bayesian networks, which have previously been employed to infer both gene regulatory networks [Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Friedman and Koller, 2003; Tamada *et al.*, 2003; Imoto *et al.*, 2003; Werhli and Husmeier, 2007] and protein signalling networks [Sachs *et al.*, 2005; Werhli and Husmeier, 2007; Ellis and Wong, 2008; Mukherjee and Speed, 2008; Guha *et al.*, 2008; Ciaccio *et al.*, 2010].

We carry out inference regarding network topology within a score-based Bayesian framework (see Section 2.3.5.2), with existing biology incorporated via an informative prior distribution on networks. Model averaging over network structures is used to calculate posterior edge probabilities, which quantify evidence in favour of links, through time, between network components. The calculations required for model averaging are carried out exactly. This is done by exploiting a connection between variable selection and network inference for DBNs [Murphy, 2002], echoing

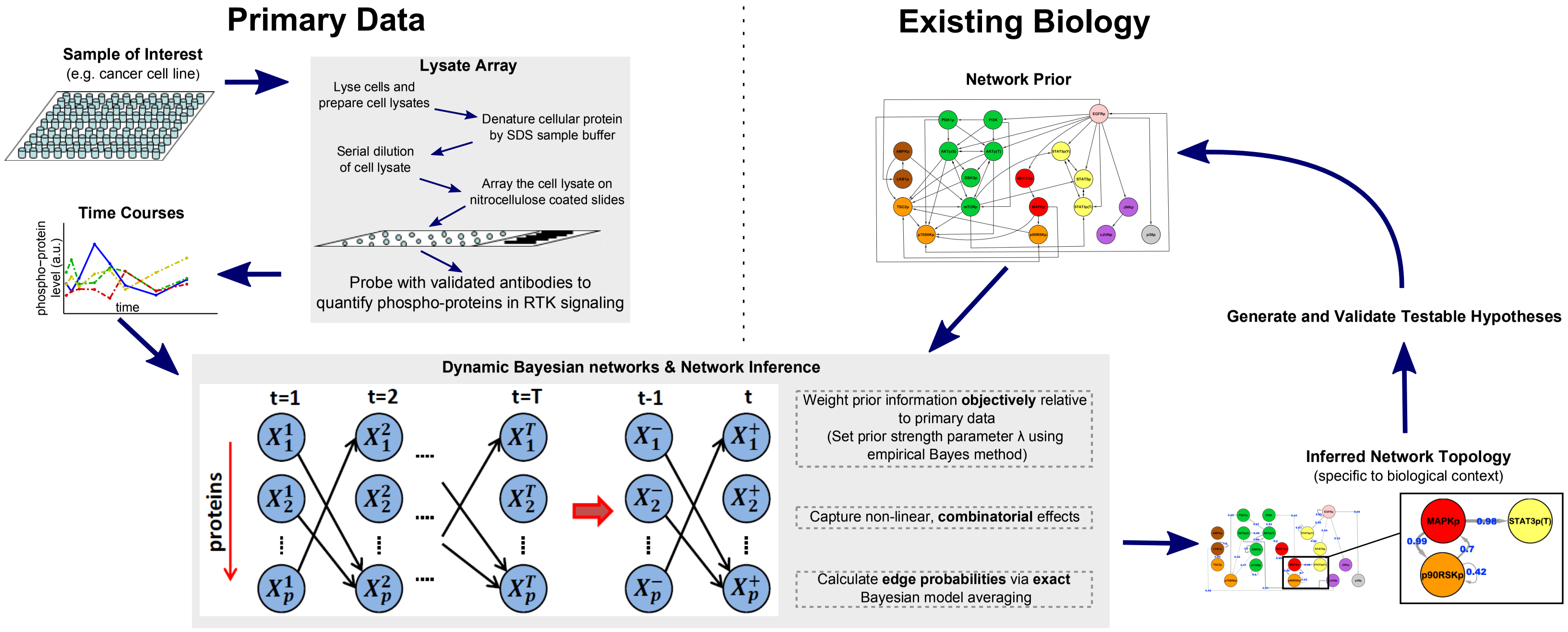

Primary Data
Existing Biology
Sample of Interest
(e.g. cancer cell line)
Lysate Array
Lyse cells and prepare cell lysates
Denature cellular protein by SDS sample buffer
Serial dilution of cell lysate
Array the cell lysate on nitrocellulose coated slides
Probe with validated antibodies to quantify phospho-proteins in RTK signaling
Time Courses
phospho−protein level (a.u.)
time
Network Prior
Generate and Validate Testable Hypotheses
Dynamic Bayesian networks & Network Inference
proteins
t=1
t=2
t=T
t-1
t
X_1^1
X_2^1
X_p^1
X_1^2
X_2^2
X_p^2
X_1^T
X_2^T
X_p^T
X_1^-
X_2^-
X_p^-
X_1^+
X_2^+
X_p^+
Weight prior information **objectively** relative to primary data
(Set prior strength parameter λ using empirical Bayes method)
Capture non-linear, **combinatorial** effects
Calculate **edge probabilities** via **exact** Bayesian model averaging
Inferred Network Topology
(specific to biological context)
MAPKp
STAT3p(T)
p90RSKp
0.98
0.99
0.7
0.42

Figure 4.1: **Data-driven characterisation of signalling networks.** Reverse-phase protein arrays interrogate signalling dynamics in samples of interest. Network structure is modelled using statistical models known as dynamic Bayesian networks. Inferences regarding network topologies are made by integrating primary phosphoproteomic data with existing biology, using informative priors objectively weighted by an empirical Bayes approach. Probability scores regarding network features then allow the generation and prioritisation of hypotheses for experimental validation. Novel biology elucidated in this way feeds back to refine current networks, in turn informing future experiment-modelling iterations.

recent work in undirected graphical models [Meinshausen and Bühlmann, 2006].

Informative Bayesian priors on network topology have been applied in several studies for inference of gene regulatory networks from a combination of gene expression data and prior biological knowledge regarding network structure. This knowledge can be either in the form of known interactions, curated from online databases and the literature [Imoto *et al.*, 2003; Froehlich *et al.*, 2007], or in the form of other data types such as transcription factor binding location data [Tamada *et al.*, 2003; Bernard and Hartemink, 2005; Werhli and Husmeier, 2007], which can be used to generate a prior network. Informative network priors have also been used to integrate protein signalling data with knowledge from databases and the literature [Werhli and Husmeier, 2007; Mukherjee and Speed, 2008].

However, in biological applications it is not obvious how to set the weight accorded to the network prior. Since network priors are often derived from known signalling maps, which are in turn based on published studies on normal cells, this issue is especially relevant in studying disease states like cancer, in which samples may exhibit altered phenotypes from the normal case [Pawson and Warner, 2007; Yuan and Cantley, 2008]. Then, it may be uncertain as to how relevant available prior information is for a given study. Following the approach used in Chapter 3, we use empirical Bayes to automatically weight the contribution of the network prior relative to proteomic data. This is related to the approach proposed by Werhli and Husmeier [2007] in which full Bayesian inference for the prior weighting is performed, with a flat prior over hyperparameters, and Markov chain Monte Carlo is used to sample from the joint posterior over networks and hyperparameters. We carry out a simpler maximum marginal likelihood empirical Bayes analysis, but do so within an exact framework that is computationally fast at the moderate data dimensions that are typical of phosphoproteomic data. A related maximisation approach was performed by Imoto *et al.* [2003], but instead of averaging over networks as the empirical Bayes approach does, it aimed to maximise (via a heuristic search) the joint posterior over networks and hyperparameters. In addition to empirical Bayes hyperparameter selection, we put forward simple diagnostics to check sensitivity to specification of the network prior.

As in Chapter 3, we avoid lossy thresholding of data and retain some ability to capture combinatorial interplay through the use of interaction terms. We also use the same variant of Bayesian parameter prior as used in Chapter 3 (Zellner's g-prior [Zellner, 1986]) to obtain a closed-form marginal likelihood score for the networks. In contrast to the widely-used 'BGe' score [Geiger and Heckerman, 1994; Friedman *et al.*, 2000; Werhli and Husmeier, 2007] (see Section 2.3.5.2) this gives a formulation that is essentially free of user set parameters and invariant to data rescaling. We explore the relationship between the two scoring methods further in Section 4.4 below. Also, a discussion of the choice between continuous versus discrete modelling can be found in Chapter 6.

A number of authors, including Sachs *et al.* [2005]; Ellis and Wong [2008] and Bender *et al.* [2010], have used statistical approaches to explore signalling network topology; the work presented in this Chapter is in this vein. However, we note that statistical network inference approaches typically use discrete or linear models and, as discussed in Chapter 1, these are a coarse approximation to the underlying chemical kinetics. In contrast, there is a rich body of work on modelling signalling using ordinary differential equations (ODEs) and related dynamical models [Schoeberl *et al.*, 2002; Chen *et al.*, 2009; Wang *et al.*, 2009a]. When network topology is known, ODEs offer a powerful modelling framework. Our work complements ODE-based approaches by providing a tractable way to explore large spaces of candidate network topologies in a data-driven manner, and thereby generate hypotheses regarding novel signalling links.

There are only a small number of recent studies in the literature that employ statistical approaches for structure learning of protein signalling networks in cancer [Guha *et al.*, 2008; Mukherjee and Speed, 2008; Ciaccio *et al.*, 2010; Bender *et al.*, 2010]. In particular, Bender *et al.* [2010] also propose a method for inference of cancer signalling networks from reverse-phase protein array time-series data, after external perturbation of network components. The present work is similar in spirit, but differs methodologically in that it uses DBNs, Bayesian model averaging, and network priors to incorporate existing biological knowledge.

Thus, we combine protein array technology with dynamic network inference to shed light on signalling networks in samples of interest. We apply these approaches to the breast cancer cell line MDA-MB-468. MDA-MB-468 is an adenocarcinoma, originally from a 51-year old patient, belonging to the well-characterised basal breast cancer subtype [Perou *et al.*, 2000; Neve *et al.*, 2006]; the line is EGFR amplified and PTEN, RB1, SMAD4 and p53 mutant. We learn a network model that is specific to this line, predicting a number of known and novel signalling links which we validate using independent inhibition experiments.

The remainder of the Chapter is organised as follows. In Section 4.2 we begin by referring to the relevant background information given in Chapter 2. We then describe the particular details of the structure learning approach used here, including the marginal likelihood score, network prior, exact model averaging for calculation of posterior edge probabilities and empirical Bayes analysis for automatic weighting of prior information. In Section 4.3.1 we show results of the structure learning approach on both simulated data and data from a synthetically constructed network in yeast [Cantone *et al.*, 2009]. The utility of prior information is investigated and comparisons are made to several other existing structure learning approaches for time series data. In Section 4.3.3 results on proteomic data from breast cancer cell line MDA-MB-468 are shown, including robustness analyses and independent validation experiments. The Chapter concludes with a discussion in Section 4.4.

4.2 Methods

The reader is referred to the following Sections for background information required for this Chapter: Bayesian linear models (Section 2.3.1); BNs (Section 2.3.4.1); and DBNs and score-based structure learning for BNs/DBNs (Section 2.3.5.2). The notation used here also follows that used in these Sections.

All computations were carried out in MATLAB R2010a using software that is freely available at `http://go.warwick.ac.uk/stevenhill/DynamicNetworkInference`.

4.2.1 Bayesian score

Recall from Section 2.3.5.2 that in a Bayesian score-based approach to DBN structure learning, we are interested in the posterior distribution over graphs G given data $\mathbf{X}$. This is given (up to proportionality) by Bayes' theorem in (2.43) and is also reproduced here,

$$P(G \,|\, \mathbf{X}) \propto p(\mathbf{X} \,|\, G) P(G) \tag{4.1}$$

where $p(\mathbf{X} \,|\, G)$ is the marginal likelihood, and $P(G)$ is a prior distribution over the space of graphs $\mathcal{G}$ that allows for the incorporation of existing signalling biology into inference. We call this the 'network prior' and discuss it further below. Therefore, we can calculate (up to proportionality) posterior probability scores for each graph structure, resulting from the integration of experimental data with existing domain knowledge.

4.2.1.1 Marginal Likelihood

We make a number of simplifying assumptions for DBNs, following previous work and as described in Section 2.3.5.2. Specifically, we make first-order Markov and stationarity assumptions and assume that edges are only permitted forwards in time. Recall that this allows the full "unrolled" graph structure (Figure 2.13(b)), in which each random variable X_j^t (protein j at time t) is represented by an individual node in the graph, to be reduced to the "collapsed" structure consisting of just two time slices representing adjacent time points (Figure 2.13(c)).

The above assumptions result in the likelihood given in (2.52), which we reproduce here (up to a multiplicative constant that does not depend on graph G),

$$p(\mathbf{X} \mid G, \mathbf{\Theta}) = \prod_{j=1}^{p} \prod_{t=2}^{T} p(X_j^t \mid X_{\pi_G(j)}^{t-1}, \theta_j) \tag{4.2}$$

$$= \prod_{j=1}^{p} p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-, \theta_j) \tag{4.3}$$

where $\mathbf{X}_j^+ = \left(X_j^2, \dots, X_j^T\right)^\mathsf{T}$ denotes all data for protein j in the second time slice of the "collapsed DBN" and $\mathbf{X}_{\pi_G(j)}^- = \left(X_{\pi_G(j)}^1, \dots, X_{\pi_G(j)}^{T-1}\right)^\mathsf{T}$ denotes all data for parents of variable j in the first time slice.

The marginal likelihood is obtained by integrating out model parameters $\{\theta_j\}$ from the likelihood (4.3). This has the effect of accounting for model complexity by penalizing complex models with many parameters and thereby helps to avoid over-fitting of the model to the data (see Section 2.3.2.2). The conditionals $p(X_j^t \mid X_{\pi_G(j)}^{t-1}, \theta_j)$ constituting the likelihood are taken to be Gaussian. These describe the dependence of child nodes on their parents and can be thought of as regression models, with parents and child corresponding to covariates and response respectively. We take these "local" models to be linear-in-the-parameters, but allow dependence on products of parents as well as parents themselves (i.e. a linear model with interaction terms as used in Chapter 3). The models are fully saturated, including products of distinct parents up to all parents. For each protein j, let $\mathbf{B}_j$ denote a $n \times (2^{|\pi_G(j)|} - 1)$ design matrix, with columns corresponding to each parent of j, and products of distinct parents up to and including the product over all parents, and where n is sample size. The sample size is the number of adjacent pairs of time points in the data. That is, if we have m time courses each consisting of T time points, then $n = m(T-1)$. Then we have

$$p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-, \theta_j) = \mathcal{N}(\mathbf{B}_j \boldsymbol{\beta}_j, \sigma_j^2 I_n) \tag{4.4}$$

where I_n is the $n \times n$ identity matrix. We note that $\mathbf{X}_j^+$ and each column of the design matrix $\mathbf{B}_j$ are standardised to have zero mean and unit variance.

The regression coefficients, forming a vector $\boldsymbol{\beta}_j$, and variance σ_j^2, constitute parameters θ_j. We use the same parameter priors as used for the Bayesian variable selection study in Chapter 3. That is, following Zellner [1986]; Smith and Kohn [1996]; and Nott and Green [2004], we use the reference prior $p(\sigma_j^2) \propto \sigma_j^{-2}$

for variances and a Normal$(\mathbf{0}, n\sigma_j^2({\mathbf{B}_j}^\mathsf{T}\mathbf{B}_j)^{-1})$ prior (Zellner's g-prior) for regression coefficients. As noted in Chapter 3 this formulation has attractive invariance properties under rescaling of the data and, in contrast to the widely-used 'BGe' score [Geiger and Heckerman, 1994] (see Section 2.3.5.2), has no free, user-set parameters.

Following Geiger and Heckerman [1994] we assume prior parameter independence. This yields the following integral for the marginal likelihood,

$$\begin{aligned} p(\mathbf{X} \mid G) &= \int p(\mathbf{X} \mid G, \mathbf{\Theta}) p(\mathbf{\Theta} \mid G) \; \mathrm{d}\Theta \\ &\propto \prod_{j=1}^{p} \iint p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-, \boldsymbol{\beta}_j, \sigma_j^2) p(\boldsymbol{\beta}_j \mid G, \sigma_j^2) p(\sigma_j^2) \; \mathrm{d}\boldsymbol{\beta}_j \mathrm{d}\sigma_j^2. \end{aligned} \tag{4.5}$$

This is a product of integrals of the same form as those that yielded the marginal likelihood for Bayesian variable selection in Chapter 3, given in (3.4). Therefore we have the following closed-form marginal likelihood (multiplicative constants that do not depend on G are omitted):

$$\begin{aligned} p(\mathbf{X} \mid G) \propto \prod_{j=1}^{p} (1+n)^{-(2^{|\pi_G(j)|}-1)/2} \Big(& {\mathbf{X}_j^+}^\mathsf{T} \mathbf{X}_j^+ \\ & - \frac{n}{n+1} {\mathbf{X}_j^+}^\mathsf{T} \mathbf{B}_j \left(\mathbf{B}_j^\mathsf{T} \mathbf{B}_j \right)^{-1} \mathbf{B}_j^\mathsf{T} \mathbf{X}_j^+ \Big)^{-\frac{n}{2}} . \end{aligned} \tag{4.6}$$

As was the case for Bayesian variable selection in Chapter 3 (see Section 3.2.1), a restriction on $|\pi_G(j)|$ (see Section 4.2.2 below) means we do not encounter problems with inversion of $\mathbf{B}_j^\mathsf{T}\mathbf{B}_j$ due to matrix singularity.

4.2.1.2 Network prior

The prior distribution $P(G)$ captures existing knowledge concerning signalling network structure. We follow Imoto *et al.* [2003]; Werhli and Husmeier [2007]; and Mukherjee and Speed [2008] and use a prior of the form

$$P(G) \propto \exp(\lambda f(G)) \tag{4.7}$$

where λ is a strength parameter, weighting the contribution of the prior, and $f(G)$ is a real-valued function over graphs, scoring the degree to which graphs concord with our prior beliefs. The objective selection of the strength parameter λ is described below. A value of $\lambda = 0$ results in a flat prior over graph space and as $\lambda \to \infty$, the prior becomes sharply peaked around the graphs that give maximum values of $f(G)$.

We use available canonical signalling maps, obtained from online reposito-

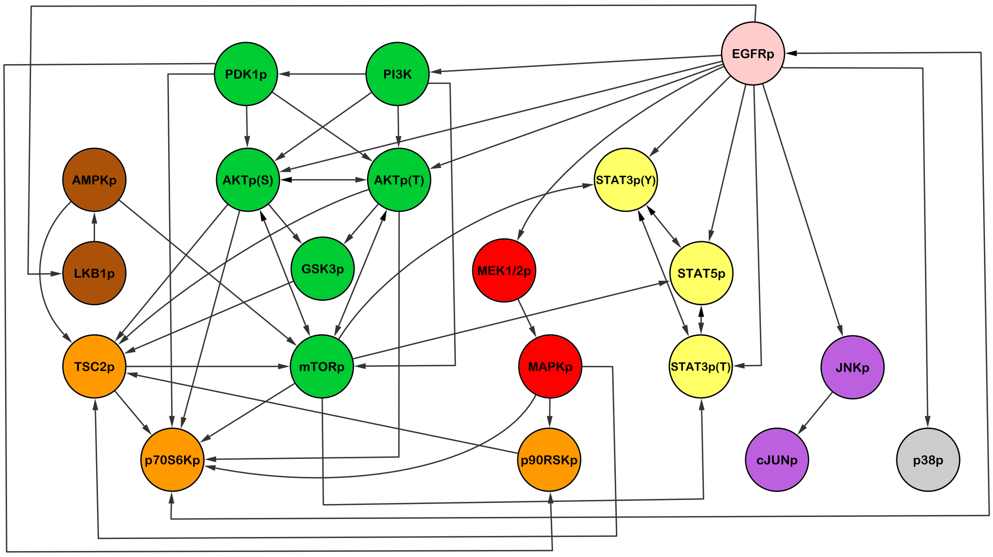


Figure 4.2: **Network prior.** Existing biology is captured and integrated during modelling using a prior probability distribution on graphs $P(G) \propto \exp(\lambda f(G))$, with $f(G) = -|E(G)\backslash E^*|$ where $E(G)$ is the set of edges contained in G and E^* is a set of *a priori* expected edges. The graph shows edge set E^*. Edges represent interactions through time. Each node also has a self-loop edge (i.e. an edge starting and finishing at the same node, these are not displayed). The edge set includes expected indirect edges which operate via components not included in our study.

ries (`cellsignal.com`, `stke.sciencemag.org`) and the literature [Oda *et al.*, 2005; Yarden and Sliwkowski, 2001], to obtain a set of edges we may expect to see in an inferred network. The edge set includes indirect edges via components not included in our study, edges between protein phosphoforms and isoforms, and self-loop edges (i.e. edges starting and finishing at same node). We denote this set of *a priori* expected edges by E^*. We let $f(G) = -|E(G)\backslash E^*|$ where $E(G)$ is the set of edges contained in G. That is, $f(G)$ is the number of edges in G that are not included in our expected edge set E^*. Therefore our prior does not actively promote any particular edge, but rather penalises unusual edges.

The edge set E^* is given in Figure 4.2 and includes links (i) from EGFR to AKTs, MEK, JNK, p70, LKB1, p38, PI3K and STAT3/5; (ii) from MEK to ERK, and from ERK to p70, p90 and TSC2; (iii) from mTOR, PDK1 and PI3K to AKTs, and from AKTs to GSK3, mTOR, p70 and TSC2 [Manning *et al.*, 2002]; (iv) from PI3K to PDK1 and mTOR; (v) from LKB1 to AMPK [Shaw *et al.*, 2004], and from AMPK to mTOR [Hardie, 2004] and TSC2; (v) from mTOR, PDK1 and TSC2

[Goncharova *et al.*, 2002] to p70, and from p70 to EGFR; (vi) from cJun to JNK; (vii) from mTOR to STAT3/5 [Yokogami *et al.*, 2000] and from TSC2 to mTOR [Inoki *et al.*, 2002]; (viii) from PDK1 to p90 [Jensen *et al.*, 1999], and from p90 and GSK3 to TSC2.

The prior incorporates existing knowledge in a "soft" probabilistic manner and does not restrict inferred edges to those included in the prior. This is an important feature in the cancer setting since cancer-specific networks may differ from the general biology upon which the prior is built. Empirical results investigating robustness to prior specification are reported below (Figure 4.7).

4.2.2 Exact inference by variable selection

We are interested in calculating posterior probabilities of edges $e = (a, b)$ in the graph G. Note that $a, b \in \{1, \ldots, p\}$, with a, b representing variables from the first and second time slices of the "collapsed" DBN respectively. For simplicity, we use $e = (a, b)$ in what follows and leave the time associated with the vertices implicit. The posterior probability of the edge is calculated by averaging over the space of all possible graphs $\mathcal{G}$, as given by (2.50) and reproduced here,

$$P(e \mid \mathbf{X}) = \sum_{G \in \mathcal{G}} \mathbb{1}_G(e) P(G \mid \mathbf{X}) \tag{4.8}$$

where $P(G \mid \mathbf{X})$ is the posterior distribution over graphs and $\mathbb{1}_G(e)$ is an indicator function evaluating to unity if and only if edge e is in graph G. For background information on Bayesian model selection and model averaging, see Sections 2.3.2.2 and 2.3.2.3.

For DBNs with p vertices in each time slice, the size of the graph space is 2^{p^2}, hence growing super-exponentially with p. This precludes explicit enumeration of the sum in (4.8) for even small-to-moderate p. However, since the DBNs used here have only edges forward in time and are therefore guaranteed to be acyclic, we can exploit a connection between network inference and variable selection [Murphy, 2002] for efficient and exact calculation of posterior edge probabilities, thereby increasing confidence in results while avoiding the need for expensive convergence diagnostics. We give full details below, but in brief, instead of averaging over full graphs G as in (4.8), we consider the simpler problem of variable selection for each protein. That is, for each protein j, we score subsets of potential parents $\pi(j) \subseteq \{1, \ldots, p\}$. Model averaging is then carried out in the variable selection sense, i.e. by averaging over subsets of parents rather than over full graphs.

Specifically, for each "response" variable $\mathbf{X}_j^+$ in the second time slice of the

DBN, we calculate posterior scores for subsets $\pi(j) \subseteq \{1, \ldots, p\}$ of potential predictors from the first time slice ('parent sets'),

$$\begin{aligned} P(\pi(j) \,|\, \mathbf{X}^-, \mathbf{X}_j^+) &\propto p(\mathbf{X}_j^+ \,|\, \mathbf{X}^-, \pi(j)) P(\pi(j) \,|\, \mathbf{X}^-) \\ &= p(\mathbf{X}_j^+ \,|\, \mathbf{X}_{\pi(j)}^-) P(\pi(j)) \end{aligned} \tag{4.9}$$

where $\mathbf{X}^- = \left(\mathbf{X}_1^-, \ldots, \mathbf{X}_p^-\right)$ denotes all data in the first time slice and $P(\pi(j))$ is a prior distribution over parent sets for variable j. The likelihood $p(\mathbf{X}_j^+ | \mathbf{X}_j^-, \pi\,(j)\,, \theta_j)$ is as in (4.4) above, with parameter priors for $\boldsymbol{\beta}_j$ and σ_j^2 also as described above. Integrating out parameters θ_j then results in the marginal likelihood $p(\mathbf{X}_j^+ | \mathbf{X}^-, \pi\,(j))$,

$$\begin{aligned} p(\mathbf{X}_j^+ \,|\, \mathbf{X}^-, \pi\,(j)) \propto (1+n)^{-(2^{|\pi(j)|}-1)/2} \Big(& {\mathbf{X}_j^+}^{\mathsf{T}} \mathbf{X}_j^+ \\ & - \frac{n}{n+1} {\mathbf{X}_j^+}^{\mathsf{T}} \mathbf{B}_j \left({\mathbf{B}_j}^{\mathsf{T}} \mathbf{B}_j\right)^{-1} {\mathbf{B}_j}^{\mathsf{T}} \mathbf{X}_j^+ \Big)^{-\frac{n}{2}} . \end{aligned} \tag{4.10}$$

This is a marginal likelihood of the form that appears in Chapter 3 for Bayesian variable selection. The parent sets $\pi(j)$ play the same role here as the inclusion indicator vector γ in (3.4).

We now discuss how model averaging in the variable selection sense can be used to make inference about DBN structure. We perform model averaging to calculate the posterior probability of a specific predictor variable X_a^- being in the model for response variable X_b^+. In terms of the DBN framework, we can think of this as an edge $e = (a, b)$. Then we have,

$$P(e \,|\, \mathbf{X}^-, \mathbf{X}_b^+) = \sum_{\pi(b)} \mathbb{1}_{\pi(b)}(a) P(\pi(b) \,|\, \mathbf{X}^-, \mathbf{X}_b^+) \tag{4.11}$$

where the summation is over all possible parent sets for variable $\mathbf{X}_b^+$. If the network prior $P(G)$ factorises into a product of local priors over parents sets $\pi_G(j)$ for each variable,

$$P(G) = \prod_{j=1}^{p} P(\pi_G(j)) \tag{4.12}$$

(the network prior used here satisfies this property; see below) then posterior edge probabilities calculated by averaging over parent sets (4.11) equal those calculated by averaging over the (much larger) space of graphs (4.8). This is essentially due to the modular form of the marginal likelihood in (4.6) and the guaranteed acyclicity of the DBNs employed here. Indeed, this equivalence holds for any modular scoring function and modular network prior used with DBNs (with edges only allowed forwards in time), as we now demonstrate.

If the marginal likelihood $P(\mathbf{X} \mid G)$ has a modular form (as in (4.6)) and the network prior $P(G)$ satisfies (4.12), then the posterior over graphs $P(G \mid \mathbf{X})$ is simply a product of posteriors over parent sets for each variable (4.9), as specified by the edge structure of G:

$$
\begin{aligned}
P(G \mid \mathbf{X}) &= \frac{P(\mathbf{X} \mid G)P(G)}{\sum_G P(\mathbf{X} \mid G)P(G)} && (4.13)\\
&= \frac{\prod_{j=1}^{p} p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-)P(\pi_G(j))}{\sum_{\pi_G(1)} \cdots \sum_{\pi_G(p)} \prod_{j=1}^{p} p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-)P(\pi_G(j))} && (4.14)\\
&= \prod_{j=1}^{p} \frac{p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-)P(\pi_G(j))}{\sum_{\pi_G(j)} p(\mathbf{X}_j^+ \mid \mathbf{X}_{\pi_G(j)}^-)P(\pi_G(j))} && (4.15)\\
&= \prod_{j=1}^{p} P(\pi_G(j) \mid \mathbf{X}^-, \mathbf{X}_j^+). && (4.16)
\end{aligned}
$$

We can now observe that edge probabilities calculated via averaging over the full graph space (4.8) equal those calculated from a variable selection approach (4.11). In particular, for an edge $e = (a, b)$,

$$
\begin{aligned}
P(e \mid \mathbf{X}) &= \sum_G \mathbb{1}_G(e)P(G \mid \mathbf{X}) && (4.17)\\
&= \sum_G \mathbb{1}_G(e) \prod_{j=1}^{p} P(\pi_G(j) \mid \mathbf{X}^-, \mathbf{X}_j^+) && (4.18)\\
&= \sum_{\pi_G(1)} \cdots \sum_{\pi_G(p)} \mathbb{1}_G(e) \prod_{j=1}^{p} P(\pi_G(j) \mid \mathbf{X}^-, \mathbf{X}_j^+) && (4.19)\\
&= \left(\sum_{\pi(b)} \mathbb{1}_{\pi(b)}(a)P(\pi(b) \mid \mathbf{X}^-, \mathbf{X}_b^+) \right) \prod_{\substack{1<j<p\\ j\neq b}} \left(\sum_{\pi(j)} P(\pi(j) \mid \mathbf{X}^-, \mathbf{X}_j^+) \right) && (4.20)\\
&= \sum_{\pi(b)} \mathbb{1}_{\pi(b)}(a)P(\pi(b) \mid \mathbf{X}^-, \mathbf{X}_b^+) && (4.21)\\
&= P(e \mid \mathbf{X}^-, \mathbf{X}_b^+) && (4.22)
\end{aligned}
$$

The prior used here satisfies the modular form of (4.12). The expected edge set E^* can be represented by p index sets $\pi^*(j) \subseteq \{1, \ldots, p\}$ for the parents of each "response" variable $\mathbf{X}_j^+$. That is, $\pi^*(j) = \{k \mid (k, j) \in E^*\}$. Then, (4.12) holds if we define the prior over parent sets to be $P(\pi_G(j)) \propto \exp(\lambda f_j(\pi_G(j)))$ where $f_j(\pi_G(j)) = -|\pi_G(j) \backslash \pi^*(j)|$.

Note that, for each variable $\mathbf{X}_j^+$, the space of possible parent sets is of size 2^p (as opposed to 2^{p^2} for the full graph space). Hence it is much more computationally efficient to calculate edge probabilities via (4.11) than (4.8). However, the problem is still exponential in p. Motivated by the fact that typically only a small number of key upstream regulators are critical for any given signalling component [Beard and Qian, 2008], and following related work in both gene regulation [Husmeier, 2003; Werhli and Husmeier, 2007] and protein signalling [Ellis and Wong, 2008], we enforce a maximum in-degree constraint and only consider up to d_{max} proteins jointly influencing a target. While, under this restriction, the full graph space size still grows quicker than exponential in p, the space of parent sets becomes polynomial in p. This enables exact calculation of edge probabilities via (4.11). For all experiments reported below, we set $d_{max} = 4$. We investigate sensitivity of our breast cancer proteomic data results to this constraint below (Figure 4.6).

The equivalence between posterior probabilities calculated via averaging over the full graph space and those calculated via a variable selection approach does not merely hold for single edge probabilities, but for any graph feature that can be fully specified at a local parent set level. That is, equivalence holds if an indicator function $\mathbb{1}_G(\zeta)$, specifying whether graph G has a feature of interest ζ or not, can be expressed as a product of local indicator functions over parent sets with features ζ_j,

$$\mathbb{1}_G(\zeta) = \prod_{j=1}^{p} \mathbb{1}_{\pi_G(j)}(\zeta_j). \tag{4.23}$$

For example, for single edge probailities (4.8) we have $\zeta = e = (a, b)$, $\zeta_b = a$ and $\zeta_j = \varnothing$ for $j \neq b$, where we define the convention that the indicator function always evaluates to unity when the feature of interest is $\varnothing$. Features satisfying this local factorisation include existence of sets of edges, non-existence of sets of edges and in-degree related features. However, the equivalence does not extend to arbitrary graph features. We also note that the equivalence does not hold for static BNs due to the acyclicity constraint; independently inferring parent sets for each variable $\mathbf{X}_j^+$ leads to a loss of the global edge structure information required to detect cycles.

4.2.3 Empirical Bayes

The prior strength λ controls the relative contribution of prior and data. We set this parameter using an objective, empirical Bayes approach (background information on empirical Bayes approaches can be found in Section 2.3.8). Specifically, we maximise the marginal likelihood

$$\begin{aligned} p(\mathbf{X} \mid \lambda) &= \mathbb{E}\left[p(\mathbf{X} \mid G)\right]_{P(G \mid \lambda)} \\ &= \sum_G p(\mathbf{X} \mid G) P(G \mid \lambda). \end{aligned} \tag{4.24}$$

Following similar arguments as above, (4.24) can be rewritten in terms of summations over parent sets $\pi(j)$ as follows,

$$p\left(\mathbf{X} \mid \lambda\right) = \prod_{j} \sum_{\pi(j)} p\left(\mathbf{X}_j^{+} \mid \mathbf{X}_{\pi(j)}^{-}\right) P\left(\pi\left(j\right) \mid \lambda\right). \tag{4.25}$$

This allows the marginal likelihood to be calculated efficiently within the exact inference framework used here. The marginal likelihood score is calculated over a grid of hyperparameter values and those resulting in the largest score are used in the analysis.

4.3 Results

4.3.1 Simulation study

Objective assessment of network structure learning performance is a non-trivial problem, since in the majority of applications the true underlying network structure is unknown. For results on experimental data, one option is to compare predicted links to those reported in the literature or appearing in online databases. However, this can be inconclusive or misleading due to uncertainty in the literature itself, or reported links may only apply to certain contexts. Moreover, the discovery of a link that is not supported by existing evidence does not allow any conclusions to be drawn; the link could be a novel discovery or could be a false positive. Another option is to perform independent inhibition experiments to validate predictions (we carry out such validations below), but this can only practically be done for a handful of links, not for the entire inferred network.

An alternative approach is to simulate data from a known network structure, infer a new network from the data and compare this new network to the original data-generating network. Thus, the data-generating graphs provided a "gold-standard" against which to assess the analyses. We carry out such a simulation study here.

Mirroring the protein signalling study below, we formed DBNs with 20 vertices (corresponding to proteins) in each time slice, and simulated 4 complete time courses of 8 time points each. We carried out inference as described above, and used the same network prior as for the proteomic data below (Figure 4.2).

Data-generating graphs were created so as to agree only partially with the prior used. This was done using a random, Erdös-Renyi-like approach. In particular, an edge set $E(G)$ for a data-generating graph G was created from the prior graph edge set E^{*} using a two-step process. First, edges contained in the prior edge set

E^* were removed at random, leaving only 20 of the original 74 edges. Second, 10 edges that were not originally in E^* were added; these were chosen uniformly at random. This process gave randomly generated graphs with 30 edges. For each such randomly generated graph, 10 of the edges were not in the prior, while 54 of the edges in the prior were not in the data-generating graph. This created a scenario in which a non-trivial proportion of the prior graph used did not agree with the data-generating graph.

Data were generated from a given graph by ancestral sampling (through time), using a Gaussian model with interaction terms (see (4.4)). The zero-order or bias term was always included, and the remaining terms were independently included with probability 0.5 (subject to each parent in the graph being represented by at least one term; this ensured that the data model was faithful to the graph). This meant that some dependencies were strictly linear, while others included interactions. Model regression coefficients were sampled from a uniform distribution over $[-1,-0.1] \cup [0.1,1]$ and were independent of time. Root nodes (initial time point) were also sampled from this uniform distribution and Gaussian noise was set at a variance of 0.5. This can be thought of as simulating data from a sparse vector autoregressive (VAR) model. For example, if protein j has two parents in the data-generating graph, proteins k_1 and k_2, then for the initial time point $t = 1$, we take $X_j^1 \sim \text{Uniform}([-1,-0.1] \cup [0.1,1])$ and for $t \in \{2,\dots,8\}$ we have

$$X_j^t = \beta_0 + \gamma_1\beta_1 X_{k_1}^{t-1} + \gamma_2\beta_2 X_{k_2}^{t-1} + \gamma_3\beta_3 X_{k_1}^{t-1} X_{k_2}^{t-1} + \epsilon_{jt} \tag{4.26}$$

where $\epsilon_{jt} \sim \mathcal{N}(0,0.5)$, regression coefficients $\beta_l \sim \text{Uniform}([-1,-0.1] \cup [0.1,1])$ and γ_l are independent Bernoulli random variables taking value one with probability 0.5, thereby selecting which terms are included. Sampled γ's are rejected if they are not faithful to the graph: as an example, for the three protein illustration above, $\gamma_1 = 1$ and $\gamma_2 = \gamma_3 = 0$ would be rejected as it would mean that, contrary to the graph, protein k_2 is not actually a parent of protein j.

We used exact network inference for DBNs as described above (using a model with interaction terms and an informative network prior with strength parameter set by empirical Bayes) to calculate posterior edge probabilities $P(e \mid \mathbf{X})$. Thresholding these probabilities at level τ produced edge set $E_\tau = \{e \mid P(e \mid \mathbf{X}) \geq \tau\}$, which was compared to the true data-generating graph to calculate number of true positives (correct edges called at threshold τ) and number of false positives (incorrect edges called at threshold τ). A receiver operating characteristic (ROC) curve was created by plotting number of true positives against number of false positives for varying threshold τ. Area under the ROC curve (AUC) provides a measure of network inference accuracy with higher values indicating better performance.

We generated a total of 25 random graphs, as described above. Empirical Bayes setting of prior strength parameter resulted in an average value of $\lambda = 3.54 \pm 0.34$ over the 25 experiments. Figure 4.3 shows average ROC curves and average AUC values obtained. We also show results for: DBN inference without interaction terms and/or using a flat prior over graph space (i.e. $P(G) = \text{constant}$), a baseline correlational analysis (thresholded absolute correlation coefficients between variables at adjacent time points), variable selection via ℓ_1-penalised (lasso) regression [Tibshirani, 1996] (see Section 2.3.3.3, implemented using Matlab package `glmnet` [Friedman *et al.*, 2010]), and several previously proposed network inference approaches for time course data. These approaches are: a Gaussian graphical model approach using functional data analysis and shrinkage-based estimation [Opgen-Rhein and Strimmer, 2006] (see Section 2.3.6, implemented using R package `GeneNet`); a non-Bayesian approach for inferring DBNs [Lèbre, 2009] (implemented using R package `G1DBN`); and a non-parametric Bayesian approach using Gaussian processes [Äijö and Lähdesmäki, 2009] (see Section 2.3.7.4, implemented using Matlab package `gp4grn`). (All Matlab and R packages were used with default settings). All the above methods resulted in a set of edge scores (e.g. posterior edge probabilities, absolute regression coefficients, absolute partial correlations) that were thresholded to produce ROC curves. Since Gaussian graphical models are undirected graphs, the inferred networks from this approach were compared with the data-generating graph with edge direction information removed. We note the lasso approach also provides a single graph (i.e. a point estimate) by selecting all those edges with non-zero regression coefficients.

Mean AUCs (± SD) for DBN inference with an informative prior and flat prior (and with interaction terms) were 0.93 ± 0.03 and 0.84 ± 0.05 respectively. The baseline correlational analysis, Lasso and previously proposed network inference approaches resulted in mean AUC values ranging from 0.54 to 0.75. Hence we see that the network prior provides significant gains in sensitivity and specificity, even though, by design of the simulation experiment, a non-trivial proportion of information in the prior is not in agreement with the data-generating model. We also note that, due to its inability to model combinatorial interplay, the proposed DBN method without interaction terms is outperformed by the method including interaction terms.

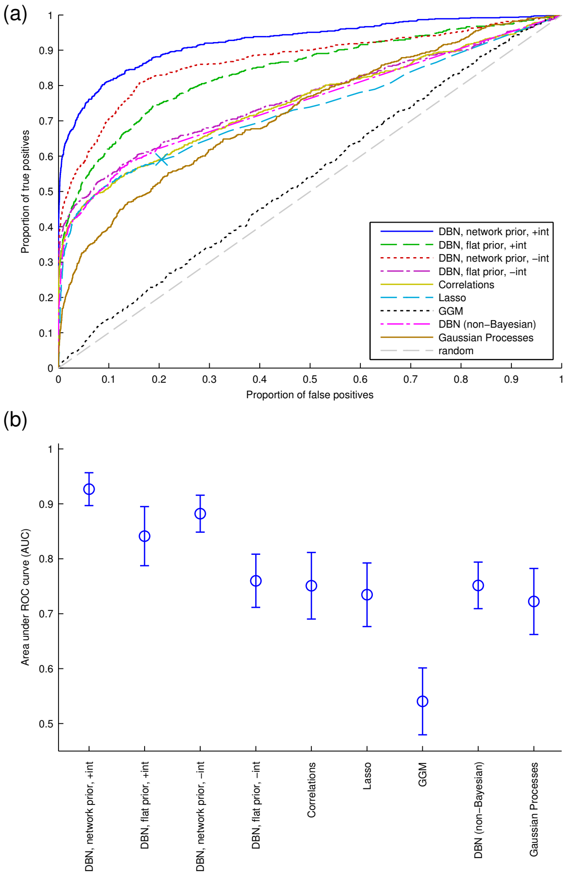


Figure 4.3: **Simulation study.** (a) Average receiver operating characteristic (ROC) curves. True positive rate (for network edges) plotted against false positive rate across a range of thresholds. Simulated data were generated from known graph structures by ancestral sampling. Graph structures were created to be in only partial agreement with the network prior shown in Figure 4.2 (see text for full details of simulation). Results shown are averages obtained from 25 iterations. Legend - "DBN, network prior": proposed DBN inference approach using a network prior, weighted objectively by empirical Bayes; "DBN, flat prior": proposed DBN inference approach using a flat prior over network space; "+int/-int": with/without interaction terms; "correlations": absolute Pearson correlations between proteins at adjacent time points; "Lasso": ℓ_1-penalised regression (curve produced by thresholding absolute regression coefficients, whilst marker 'X' is single graph obtained by taking all non-zero coefficients to be edges); "GGM": a Gaussian graphical model approach for time series data; "DBN (non-Bayesian)": a non-Bayesian method for DBN inference; "Gaussian Processes": a non-parametric Bayesian approach using Gaussian processes. (b) Average area under the ROC curve (AUC). AUC provides a measure of accuracy in network inference; higher values indicate better performance. Results shown are mean AUC±SD over 25 iterations.

4.3.2 Synthetic yeast network study

The objectivity of simulation studies depends on how biologically realistic the data-generating model is. In our simulations above, the data generating model is not biologically realistic and also matches the inference model. While conclusions can be made regarding, for example, the utility of incorporating prior information, performance on this data is not likely to reflect performance on real data, and the comparisons between methods are biased in favour of those that are based on the data-generating model. Due to these limitations, simulation strategies that are more realistic, for example based on systems of ODEs, can improve objectivity of assessments and comparisons [Husmeier, 2003].

Recent work by Cantone *et al.* [2009] provides another approach for assessing structure learning performance using biologically realistic data and a known gold-standard network. A gene regulatory network was synthetically constructed in the yeast *Saccharomyces cerevisiae.* This network, called the IRMA network, is composed of five genes and six regulatory interactions (plus a protein-protein interaction). These interactions include feedback mechanisms and the network was designed so that the five genes are negligibly affected by genes not in the network. Since the network is known, gene expression (mRNA) data obtained from the system can be used to assess performance of structure learning algorithms; we apply this approach here.

We use data from the "switch-off' experiments [Cantone *et al.*, 2009], which consists of 18 time points (every 10 minutes up to 3 hours) and is averaged over four replicates. As noted by Cantone *et al.* [2009], it is less likely that the protein-protein interaction can be recovered from the mRNA level data, so we assess performance based on the network with six edges. In order to investigate the effect of including prior information, we formed prior networks that partially agree with the IRMA network as follows. First, the prior edge set E^* is taken to include the six edges in the IRMA network and all self-loop edges. Second, three edges, chosen at random, are added to E^*. Third, two edges in the IRMA network, chosen at random, are removed from E^*. This results in prior networks containing seven edges (plus self-loop edges), four of which are in the IRMA network, and the IRMA network contains two edges that are not in the prior.

We applied exact network inference for DBNs as described above, using an informative network prior (generated as just described) with strength parameter set by empirical Bayes and linear model with interaction terms. Empirical Bayes resulted in an average value of $\lambda = 4.72 \pm 3.75$ over 25 different randomly generated prior networks. As in the simulation study, posterior edge probabilities were used

Method	AUC
DBN, network prior, +int	0.82±0.04
DBN, flat prior, +int	0.75
DBN, network prior, -int	0.74±0.04
DBN, flat prior, -int	0.67
Correlations	0.39
Lasso	0.50
GGM	0.44
DBN (non-Bayesian)	0.43
Gaussian Processes	0.75

Table 4.1: **Synthetic yeast network study.** Inference methods and regimes assessed on time series gene expression data generated from a 5-node synthetically constructed gene regulatory network in yeast [“switch-off ” experiment data; Cantone *et al.*, 2009]. The data-generating network is known, providing a gold-standard against which performance can be assessed. Network priors were generated to be in partial agreement with the true, underlying network structure (see text for details). Results shown are area under the ROC curve (AUC). Legend - “DBN, network prior”: proposed DBN inference approach using a network prior, weighted objectively by empirical Bayes; “DBN, flat prior”: proposed DBN inference approach using a flat prior over network space; “+int/-int”: with/without interaction terms; “correlations”: absolute Pearson correlations between proteins at adjacent time points; “Lasso”: ℓ_1-penalised regression; “GGM”: a Gaussian graphical model approach for time series data; “DBN (non-Bayesian)”: a non-Bayesian method for DBN inference; “Gaussian Processes”: a non-parametric Bayesian approach using Gaussian processes. The regimes using a network prior are mean AUC±SD over 25 randomly generated prior network structures.

to generate ROC curves and AUC scores. Table 4.1 shows AUC scores obtained for the same methods and regimes considered above in the simulation study. AUC scores for DBN inference with an informative prior and flat prior (and with interaction terms) were 0.82±0.04 (mean±SD) and 0.75 respectively. As in the simulation study above, we observe gains in accuracy through use of a network prior, even though the prior is only in partial agreement with the true network. We also see that inclusion of interaction terms provides an improvement in performance. The Gaussian processes method performs comparably to the DBN approach described in this Chapter, but is significantly more computationally intensive (for the five gene network, 10 iterations of the DBN approach takes approximately 1.5 seconds, compared to 160 seconds for the Gaussian processes method). The baseline correlational analysis, Lasso, Gaussian graphical model and non-Bayesian DBN inference approaches did not perform well, with AUC values ranging from 0.39 to 0.5 (i.e. no improvement over a random classifier).

4.3.3 Network model for breast cancer cell line MDA-MB-468

We used DBNs to model network topology using a combination of reverse-phase protein array (RPPA) phosphoproteomic data, from cell line MDA-MB-468, and existing knowledge of signalling topology, incorporated using an informative prior distribution over network structures. Time courses were carried out at eight time points (5, 15, 30, 60, 90, 120, 180, 240 minutes) in triplicate, under four growth conditions (0, 5, 10, 20ng/ml EGF), for 20 proteins (see Table A.3). For further details of RPPA protocol, see Section A.2.1.

The prior strength parameter λ was set in an objective manner using an empirical Bayes approach (Figure 4.4), resulting in a value of $\lambda = 3$. Using exact model averaging over network space (as described above) we calculated probability scores for each of the 400 possible edges between proteins. Figure 4.5a shows the inferred network (all edges with probability $p \geq 0.4$) with a corresponding heat map depicting all 400 edge probabilities (Figure 4.5b). Inference accounts for both fit-to-data and model complexity, and indeed the model learned is sparse (posterior expected number of edges is ~25). We note although PI3K is not presented as a phosphoprotein, we include it based on its known regulatory role in tyrosine kinase receptor signalling.

We discuss the inferred network further below. We first present an analysis of robustness of the inferred network to the maximum in-degree constraint, prior specification and perturbation of data points, followed by an empirical check of predictive capability and model fit.

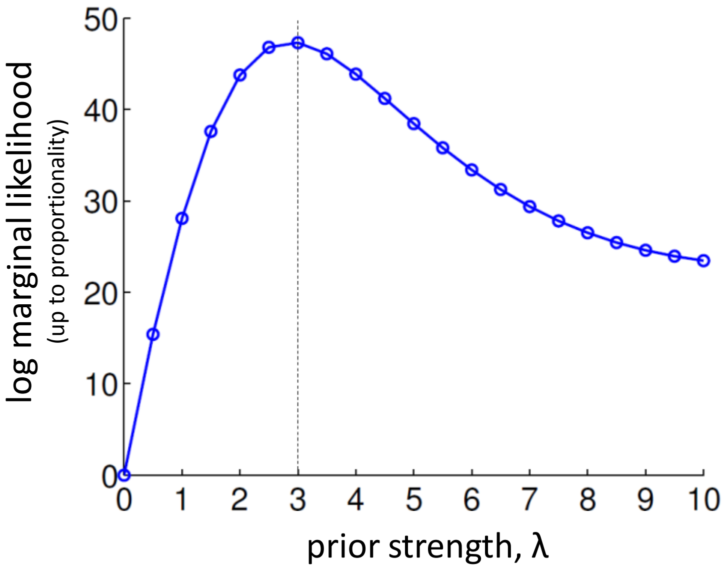


Figure 4.4: **Objective weighting of informative network prior.** Empirical Bayes marginal likelihood vs. prior strength λ. An informative prior on networks was used to integrate proteomic data with existing knowledge of signalling topology (derived from available signalling maps, see Figure 4.2 and text for details). Prior strength λ was set by an empirical Bayes approach. This was done by empirically maximising marginal likelihood $p(\text{data} \mid \lambda)$ as shown (in increments of 0.5); this gave $\lambda = 3$, the value used in the analyses.

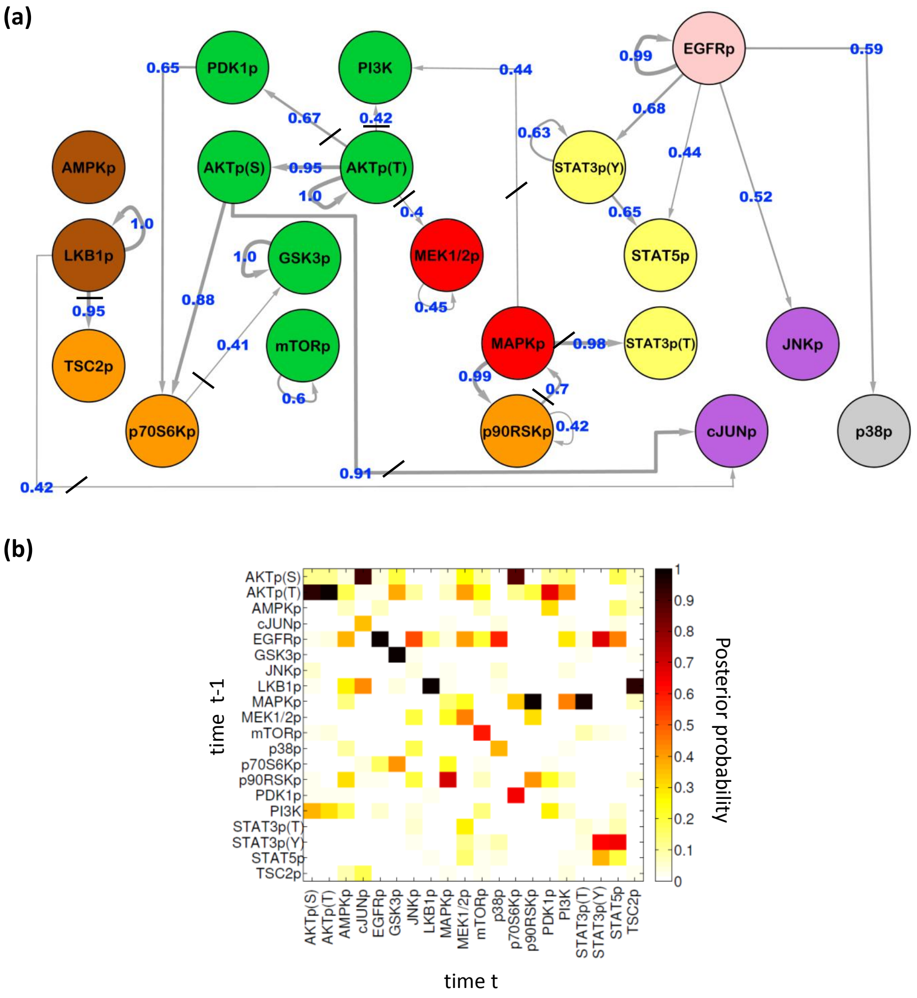


Figure 4.5: **Data-driven signalling topology in the breast cancer cell line MDA-MB-468.** (a) Receptor tyrosine kinase regulated signalling networks for the cell line MDA-MB-468. Reverse-phase protein arrays were used to interrogate the phosphoproteins shown, including key components of AKT, MAPK and STAT pathways, through time. Dynamic Bayesian networks were used to integrate the data with an objectively-weighted informative prior, derived from existing biology. Edges represent probabilistic relationships between proteins, through time. Edge labels indicate corresponding (posterior) probabilities (calculated exactly; edge thickness proportional to probability; all edges with probability ≥0.4 shown; strikethroughs "/" indicate edges not expected under the network prior; links indicate inferred influence, which may be positive or negative, i.e. excitatory or inhibitory, sign of edge not displayed; full list of proteins and associated antibodies given in Table A.3). (b) Heatmap showing all 400 posterior edge probabilities.

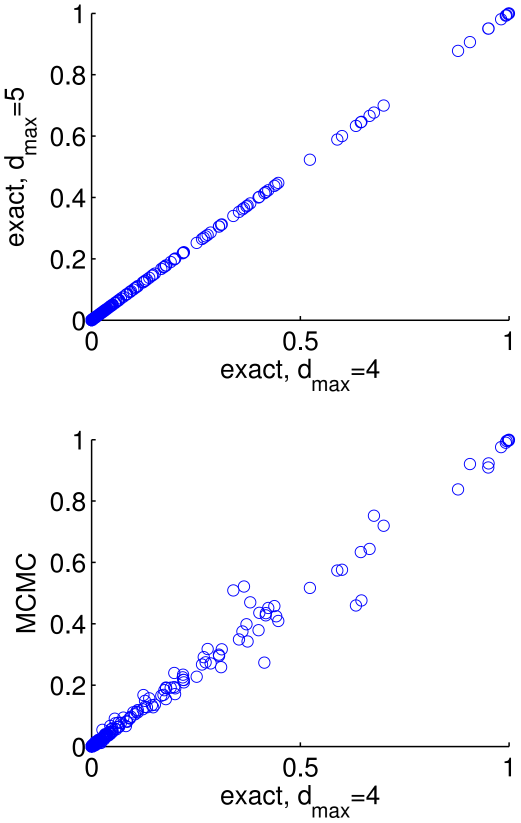


Figure 4.6: **Effect of maximum in-degree constraint $\mathbf{d_{max}}$.** Results reported in Figure 4.5 were obtained by exact inference with a maximum in-degree of $d_{max} = 4$. These results were compared with results obtained by exact inference with maximum in-degree increased to $d_{max} = 5$ (top), and by Markov chain Monte Carlo-based inference without any in-degree restriction (bottom).

4.3.3.1 Robustness to maximum in-degree constraint

We compared the results reported in Figure 4.5, obtained by exact calculation with maximum in-degree $d_{max} = 4$, with results obtained by (i) exact calculation with maximum in-degree $d_{max} = 5$, and (ii) a MCMC-based analysis with no restriction on in-degree. We found very close agreement between the regimes (Figure 4.6), showing that results were not dependent on the sparsity restriction.

4.3.3.2 Robustness to prior specification

We investigated the robustness of results reported to changes in the prior strength parameter λ (Figure 4.7a). This was done by comparing results over a range of λ

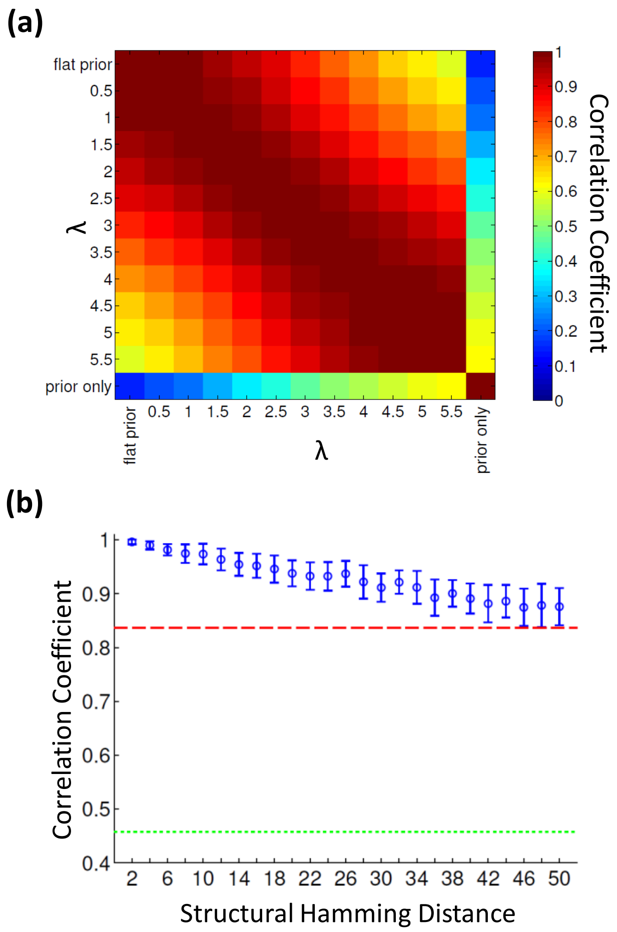


Figure 4.7: **Prior specification sensitivity analysis.** (a) Sensitivity to prior strength. To ensure robustness to specification of strength parameter λ, results were compared over a range of values of λ plus the flat prior (i.e. $\lambda = 0$) and prior only. Heatmap shows Pearson correlation coefficients (between all 400 posterior edge probabilities) for all pairs of prior regimes. Posterior results were not sensitive to precise value of λ and differed markedly from prior alone. (b) Sensitivity to prior graph. To investigate robustness to changes in the prior graph the prior graph was perturbed and results obtained compared to those reported. Correlation (as in (a) above) is shown between results as a function of number of edge changes in the prior graph ("Structural Hamming Distance", larger values indicate a greater change to the prior graph). Dashed red line and dotted green line show the correlation between reported results (with $\lambda = 3$) and those obtained with a flat prior and prior only respectively (as seen in (a)).

values, plus prior alone (i.e. no data; with $\lambda = 3$) and data alone (i.e. flat prior). Since the parameter is in the exponent, this covers a wide range of prior strength regimes. Results are not sensitive to the precise value of λ: different regimes agree well with each other, while differing somewhat from data alone and markedly from prior alone. This shows that inference did not simply recapitulate the prior but rather integrated prior and data.

We also investigated robustness to changes in the prior graph. This was done by perturbing the prior graph and comparing inferred posterior edge probabilities to those reported above (Figure 4.7b). Perturbations were made by making edge removals and additions, keeping the total number of edges constant. The size of the perturbation can be quantified by the number of edge differences from the original prior graph (“structural hamming distance” or SHD). Results are robust to changes in the prior graph: for example, changing one third of the edges (25 out of 74 edges; SHD equal to 50), gave edge probabilities that showed a correlation of 0.88 ± 0.03 with those reported (mean Pearson correlation ± SD; calculated from 25 perturbed prior graphs).

4.3.3.3 Robustness to data perturbation

We sought also to investigate the robustness of results reported to perturbation of the data. We did so by removing parts of the data and replacing with the average of adjacent time points. The data consists of four time series of eight time points each. We removed data, for all 20 proteins simultaneously, from between 1 and 4 time-point/condition combinations: this corresponded to removing between 1/32 and 1/8 of the data. These deletions represent a non-trivial change to a small data set. Figure 4.8a shows Pearson correlation coefficients between edge probabilties reported above (from unperturbed data) and those inferred from perturbed data. We observe good agreement between the edge probabilities, demonstrating that results are not overly sensitive to changes to the data. For example, perturbing 1/8 of the data resulted in a correlation coefficient of 0.83 ± 0.05 (calculated from 25 perturbed datasets).

We also considered the case of completely removing *all* data at one of the eight time points. This reduces the amount of data by almost 15% and also changes the interval between sampled time points. Removing each of the six intermediate time points in turn, we again compared results obtained to the edge probabilities reported above. We found an average correlation coefficient of 0.81 ± 0.06. This demonstrates that the results reported are robust, even when 1/8 of the data is removed and time intervals are substantially changed. Indeed, even when deleting two complete time

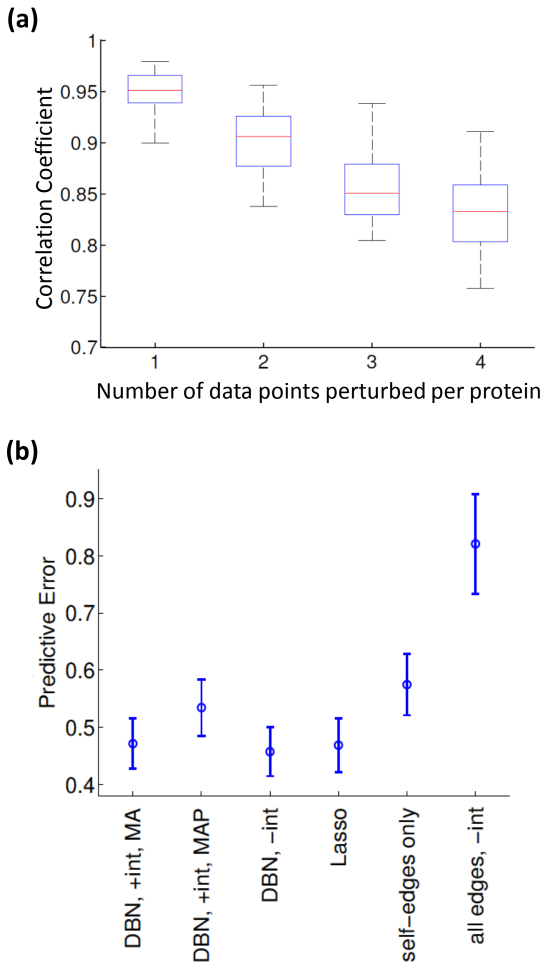


Figure 4.8: **(a) Sensitivity to data perturbation and (b) cross-validation.** (a) Data were removed, for each of the 20 proteins under study, from between 1 and 4 randomly selected time-point/condition combinations: this corresponded to removing between 1/32 and 1/8 of the data. Deleted data were replaced with the average of adjacent time ponts. Pearson correlation coefficients are shown between edge probabilities inferred from perturbed data and those obtained from the original, unperturbed data (Figure 4.5b). Results shown are over 25 iterations (except for "1", in which all possible deletions were carried out). (b) Predictive capability was empirically assessed by leave-one-out-cross-validation. Results shown are mean absolute predictive errors ±SEM for DBN network inference with interaction terms in the linear model and either exact model averaging ('DBN, +int, MA') or using the highest scoring graph ('DBN, +int, MAP); DBN network inference without interaction terms using exact model averaging ('DBN, -int'); variable selection via ℓ_1-penalised regression ('Lasso'); a baseline auto-correlative analysis ('self-edges only'); and a baseline, non-sparse linear model, with each variable predicted from all others ('all edges, -int').

points (randomly selected), i.e. 1/4 of the data, we found reasonable agreement with the results reported (correlation coefficient of 0.71 ± 0.08).

For both analyses above, the prior strength parameter was fixed to the value used in the original analysis (λ=3).

4.3.3.4 Predictive capability and model fit

As an empirical check of model fit and predictive capability, we carried out leave-one-out-cross-validation (LOOCV); see Section 2.3.2.1 for background information on cross-validation. We note that in the present setting LOOCV cannot be expected to guide detailed model choice or design. This is due to the fact that for small data sets LOOCV alone is limited in its ability to choose between models or regimes. Rather, our aim in using LOOCV was simply to highlight any egregious mismatch between data and model. In the time-course setting it is also difficult to interpret LOOCV results in terms of absolute error: rather we compared LOOCV error against baseline analyses.

LOOCV was carried out using inferred posterior parent set probabilties (4.9). At each iteration, one of the n samples was removed from the data and the exact inference procedure described above used to learn the posterior distributions over parent sets $P(\pi(j) \mid \mathbf{X}^-, \mathbf{X}_j^+)$ from the remaining $n-1$ training samples. Here, we have denoted the training data for variable j by $\mathbf{X}^-$ and $\mathbf{X}_j^+$ and we denote the corresponding held-out sample by $\mathbf{Z}^-$ and Z_j^+. Given the training data, the posterior predictive mean $\mathbb{E}\left[Z_j^+ \mid \mathbf{Z}^-, \mathbf{X}^-, \mathbf{X}_j^+\right]$ can be used to predict the value of held-out data Z_j^+ from $\mathbf{Z}^-$. This is identical to the prediction approach described in Section 3.2.5. Using our current notation, (3.10) becomes

$$\mathbb{E}\left[Z_j^+ \mid \mathbf{Z}^-, \mathbf{X}^-, \mathbf{X}_j^+\right] = \sum_{\pi(j)} \mathbb{E}\left[Z_j^+ \mid \mathbf{Z}^-, \mathbf{X}^-, \mathbf{X}_j^+, \pi(j)\right] P(\pi(j) \mid \mathbf{X}^-, \mathbf{X}_j^+) \qquad (4.27)$$

and (3.11) becomes

$$\mathbb{E}\left[Z_j^+ \mid \mathbf{Z}^-, \mathbf{X}^-, \mathbf{X}_j^+, \pi(j)\right] = \frac{n}{n+1} \tilde{\mathbf{B}}_j \left(\mathbf{B}_j{}^\mathsf{T} \mathbf{B}_j\right)^{-1} \mathbf{B}_j{}^\mathsf{T} \mathbf{X}_j^+ \qquad (4.28)$$

where $\mathbf{B}_j$ is the $(n-1) \times (2^{|\pi(j)|} - 1)$ training data design matrix for variable j, including products of parents, and $\tilde{\mathbf{B}}_j$ is the corresponding $1 \times (2^{|\pi(j)|} - 1)$ design matrix for the held-out sample. Full network inference, including empirical Bayes learning of the hyperparameter, is carried out at each cross-validation iteration.

Figure 4.8b shows these predictions compared with those from (i) the single highest-scoring graph under the posterior distribution over DBN structures (this is the *maximum a posteriori* or MAP counterpart to the Bayesian model averaging approach we propose); (ii) the proposed DBN inference method with a fully linear regression model without interaction terms; (iii) an ℓ_1-penalised regression (Lasso) approach in which parents are inferred via variable selection; (iv) a baseline autoregressive model in which each variable depends only on itself at the previous time point; and (v) a baseline non-sparse linear model in which each protein depends on all parents. For (i), (iii), (iv) and (v) predictions are made using Equation (4.28) only. The proposed method shows lower LOOCV error relative to the baseline models and to the MAP model, and performs comparably to DBN inference with a fully linear model and to Lasso regression.

4.3.3.5 Experimental validation

Many of the edges inferred recapitulate previously described (direct and indirect) links (including MAPK → p90RSKp and AKT → p70S6Kp). A number of other edges were unexpected, including signalling links which, to the best of our knowledge, have not previously been reported. We experimentally tested some of these predictions by inhibitor approaches. Edges were selected on the basis of posterior probability, biological interest and availability of selective inhibitors by which to carry out validation experiments.

The edge MAPKp → STAT3p(T727) appears with a high posterior probability of 0.98. This suggests the possibility of cross-talk between the MAPK and JAK/STAT pathways. To investigate this link, we used a MEK inhibitor (MEKi) and monitored the response of MAPKp and STAT3p(T727) through time (Figure 4.9a). Inhibition successfully reduced MAPK phosphorylation (paired t-test p-value for 10uM MEKi, calculated over 8 time points, $p = 5.0 \times 10^{-4}$): since MAPK is directly downstream of MEK, this showed that the inhibitor was effective. Moreover, in line with model predictions, we observed a corresponding decrease in STAT3p(T727) ($p = 3.3 \times 10^{-4}$). We note that the MEK to MAPK link does not appear in the inferred model; the MEKi data reported here suggest this is a false negative. We note also that these results do not imply that MAPK *directly* regulates STAT3 in MDA-MB-468, since an indirect influence, via one or more intermediate players that are not measured, would be consistent with both the model and the inhibition experiment (we return to the question of causal and mechanistic interpretations in Chapter 6).

The network model predicts a previously described edge AKTp → p70S6Kp and, of greater interest, two unexpected links AKTp → MEKp and AKTp → cJUNp. The former suggests possible cross-talk between the AKT and MAPK pathways, and

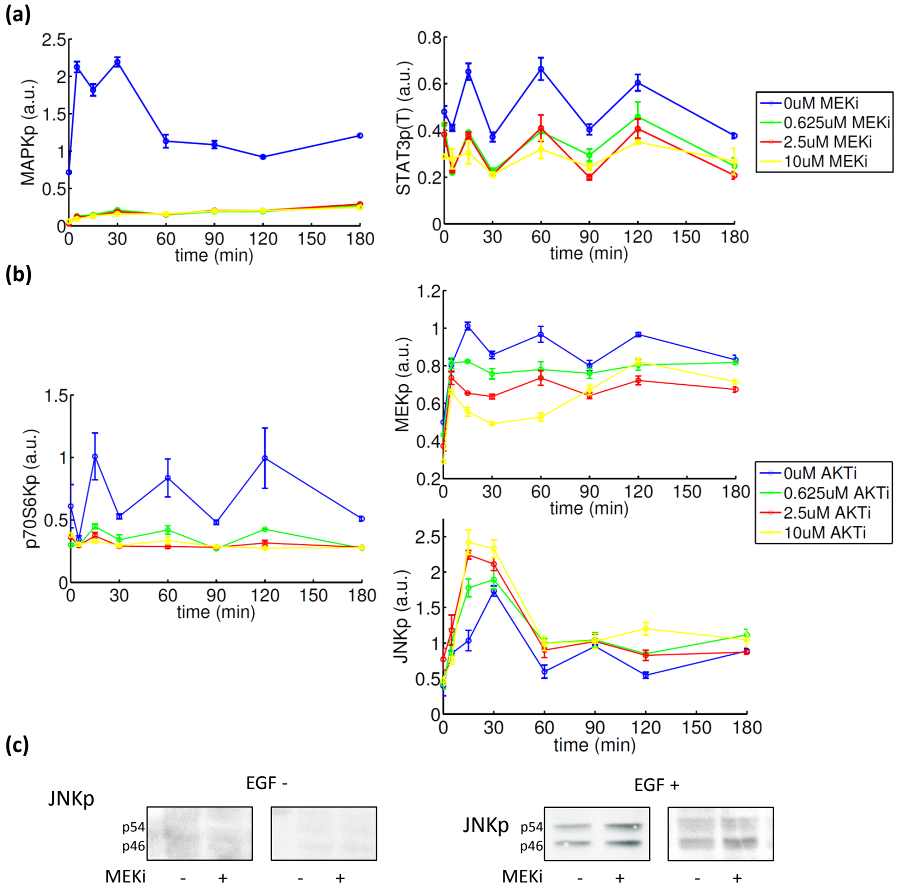


Figure 4.9: **Validation of predictions by targeted inhibition and phosphoprotein monitoring in breast cancer cell line MDA-MB-468.** (a) MAPK-STAT3 cross-talk. Network modelling (Figure 4.5a) predicted a novel link between phospho-MAPK (MAPKp) and STAT3p(T727) in the breast cancer cell line MDA-MB-468. The hypothesis of MAPK-STAT3 cross-talk was tested by MEK inhibition: this successfully reduced MAPK phosphorylation and resulted in a corresponding decrease in STAT3p(T727) (RPPA data; MEK inhibitor GSK2B at 0uM (i.e. no inhibition), 0.625uM, 2.5uM, 10uM; measurements taken 0,5,15,30,60,90,120,180 minutes after EGF stimulation; average values over 3 replicates shown, error bars indicate SEM). (b) AKTp → p70S6Kp, AKT-MAPK cross-talk and AKT-JNK/JUN cross-talk. AKTp is linked to p70S6kp, MEKp and cJUNp. In line with these model predictions, use of an AKT inhibitor reduced both p70S6K and MEK phosphorylation and increased JNK phosphorylation. (RPPA data; AKT inhibitor GSK690693B at 0uM (i.e. no inhibition), 0.625uM, 2.5uM, 10uM; measurements taken 0,5,15,30,60,90,120,180 minutes after EGF stimulation; average values over 3 replicates shown, error bars indicate SEM). (c) EGFR mediated JNK activation. EGFRp → JNKp is predicted by the network model. Cells were subjected to EGF stimulation; this resulted in activation of JNK (duplicate Western blots for p54 and p46 JNK isoforms; MEK inhibitor U0126 at 5uM).
[Validation data from Mills Lab at MD Anderson Cancer Center, Houston].

the latter suggests crosstalk between the AKT and JNK/JUN pathways. We tested these links using an AKT inhibitor (AKTi; Figure 4.9b). Phosphorylation of p70S6K was reduced by AKTi (paired t-test p-value for 10uM AKTi, calculated over 8 time points, $p = 5.0 \times 10^{-3}$), validating the edge predicted (and verifying the effect of the inhibitor). We also observe a clear decrease in MEKp levels ($p = 1.8 \times 10^{-3}$) and an increase in JNKp levels ($p = 0.047$), providing independent evidence in favour of the existence of cross-talk in both cases (JNK is known to be directly upstream of cJUN). We note that we do not observe an effect on cJUN itself in the validation experiments, only on JNK. This could be due to JNK and cJUN having different rates of (de)phosphorylation and, in particular, how these rates relate to the time scales of the experiment.

The observed differences in phosphorylation levels, between uninhibited and inibited regimes, in the MEKi and AKTi validation data are consistent both through time and between replicates at individual time points. Moreover, the time course data is non-longitudinal due to the destruction of samples in the measurement process. Therefore, this consistency is not simply a result of longitudinal correlation.

The inferred network contains several edges from EGFRp to downstream proteins in several pathways. We tested one of these predictions; the edge EGFRp → JNKp. We subjected the cell line to EGF stimulation and monitored JNK activation by Western blotting. We found that JNK was activated upon EGF stimulation (Figure 4.9c). The result was independent of MEK inhibition, suggesting that the effect is not via the MAPK pathway.

Further to our analysis regarding robustness of results to prior specification, Figure 4.10 shows the effects of prior strength and prior graph on each of the five validated predictions in Figure 4.9. The edge probabilities for the unexpected, novel edges (MAPKp→STAT3p(T), AKTp(T)→MEKp, AKTp(S)→cJUNp), which are not contained in the prior, remain high (relative to the average probability for edges not in the prior graph) across a wide range of prior strengths. In particular, the edge probability for MAPKp→STAT3p(T) shows no decrease with increasing prior strength, despite not being featured in the prior. The prior was not required for prediction of these three edges, nor the expected edge AKTp(S)→p70S6Kp. However, the validated edge EGFRp→JNKp would not have been highlighted without the prior. We observe also that edge probabilities for these links remain stable when structural changes are made to the edge set of the prior.

To further assess the utility of our approach, we considered the ability of the methods in Figure 4.8b to discover the five validated edges in Figure 4.9. We found that only the MAP model from the proposed DBN approach contains all five

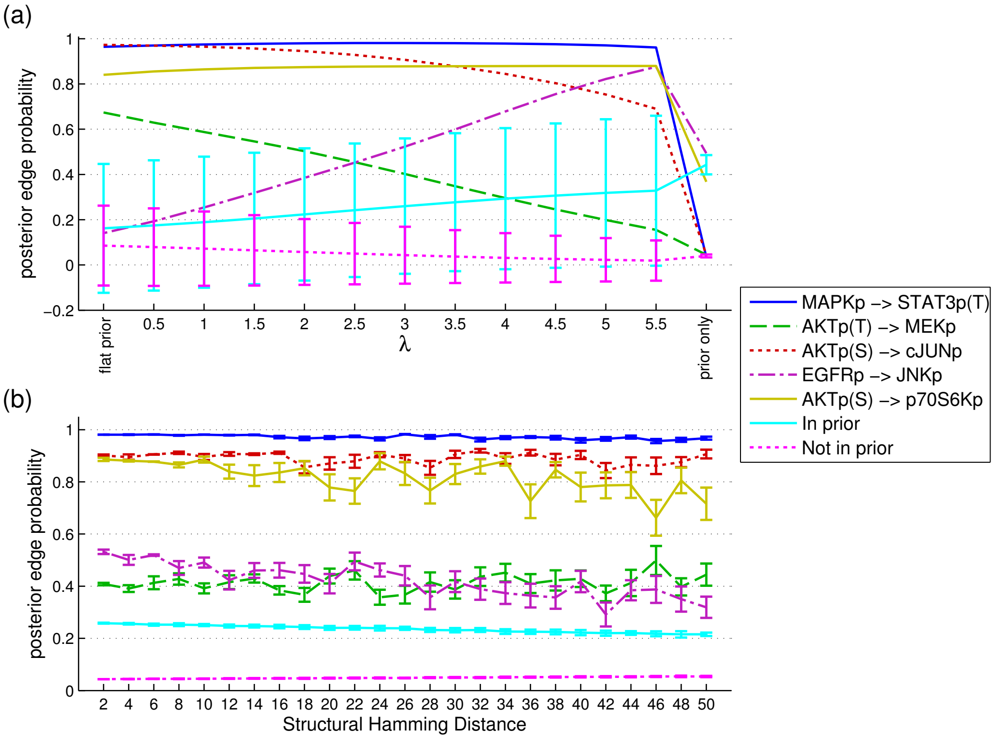


Figure 4.10: **Sensitivity of edge probabilities for validated links to prior specification.** (a) Sensitivity to prior strength. Posterior edge probabilities for validated predictions (see Figure 4.9) are shown for a range of values of prior strength parameter λ, plus the flat prior ($\lambda = 0$) and prior only. Also shown are average posterior edge probabilities (± SD) for the following edge sets: (i) all edges contained in the prior graph and (ii) all edges not contained in the prior graph. (b) Sensitivity to prior graph. Prior graphs were perturbed as described in text. Posterior edge probabilities for validated predictions are shown as a function of number of edge changes to prior graph ("Structural Hamming Distance"). Also shown are average posterior edge probabilities for the following edge sets: (i) all edges contained in the unperturbed prior graph and (ii) all edges not contained in the unperturbed prior graph. (Results reported are averages over 25 perturbed prior graphs; error bars indicate SEM).

edges. Yet, since this is a single graph, there are no edge weights to aid prioritisation of follow-up experiments. The DBN approach without interaction terms fails to discover one of the edges (AKTp(S)→MEKp), the Lasso approach only finds MAPKp→STAT3p(T), and the baseline linear model fails to find any of the edges (to permit a fair comparison for each method edge probabilities (or absolute regression coefficients) were thresholded to give a number of edges equal to the network reported in Figure 4.5a). However, we note that this assessment is biased since it is possible that the other approaches discovered edges that would also be validated in independent experiments, but were not found in our approach.

4.4 Discussion

In this Chapter, we brought together statistical network modelling and reverse-phase protein array technology to enable a data-driven analysis of signalling network topology. We combined ideas from graphical models, empirical Bayes and variable selection to yield an integrative analysis that was computationally tractable and essentially free of user-set parameters. Bayesian model averaging was used to calculate posterior edge probabilities, quantifying evidence in favour of individual links, and thereby aiding selection of specific links for experimental validation. Results were shown on both simulated data and data from a synthetically constructed network in yeast [Cantone *et al.*, 2009]. These results demonstrated that the proposed inference procedure performed favourably relative to several other structure learning approaches for time series data. An application to an individual breast cancer cell line (MDA-MB-468) enabled generation of hypotheses regarding specific links, which were subsequently validated in independent experiments.

Model averaging has previously been used to score edges in Bayesian network modelling of molecular networks [Friedman *et al.*, 2000; Husmeier, 2003; Ellis and Wong, 2008; Mukherjee and Speed, 2008]. In addition to providing a measure for the importance of individual edges, model averaging can help improve robustness of results over simply taking the MAP model, especially at small-to-moderate sample sizes (see Section 2.3.2.3). Indeed, we found that the model averaging approach offered an improved predictive capability over taking the MAP graph (see Figure 4.8b).

Network inference in general, and model averaging in particular, are often viewed as computationally burdensome. Certainly, this can often be the case (e.g. for static BNs with many nodes). However, for the DBNs employed here, using a variable selection approach as described above, network inference is relatively efficient. For datasets of moderate dimensionality this approach is arguably fast enough for routine exploratory use. For example, empirical Bayes analysis and inference of posterior edge probabilities for the 20 variables in our cancer study took under 20 seconds (on a standard single-core personal computer).

We took account of known signalling biology by means of a prior distribution on networks, weighted objectively using empirical Bayes. Prior information can aid inference regarding network structures from limited data [Geier *et al.*, 2007; Werhli and Husmeier, 2007; Mukherjee and Speed, 2008]. Indeed, our results on simulated data and on data from the synthetic yeast network also demonstrated that inclusion of prior information, in the form of a network prior, can improve inference accuracy, even when a non-trivial proportion of information in the prior is erroneous. The use of a prior incorporates existing knowledge in a 'soft' probabilistic manner that can be over-ridden by data. In contrast to hard constraints, this does not preclude discovery of unexpected edges. Indeed, the network model yielded novel biological predictions that were validated by targeted inhibition. We verified empirically that results reported were not overly sensitive to prior specification or data perturbation. Comparisons of predictive capability with baseline models suggested that the sparse models learned were indeed predictive.

Approximate inference methods such as Markov chain Monte Carlo (MCMC) [Robert and Casella, 2004] are often used for inference in BNs and DBNs [Madigan *et al.*, 1995; Friedman and Koller, 2003; Husmeier, 2003; Mukherjee and Speed, 2008; Ellis and Wong, 2008; Grzegorczyk and Husmeier, 2008] (see also Section 2.3.5.2). In contrast, we used a variable selection approach and sparsity constraints to calculate posterior edge probabilities exactly, thereby removing Monte Carlo uncertainty (and the need for associated diagnostics). The exact approach also facilitates the empirical Bayes analysis. In high dimensions, where the exact approach becomes intractable, the fully Bayesian MCMC approach proposed in Werhli and Husmeier [2007] can be used to sample from the joint posterior over networks and hyperparameters. We note that the variable selection approach also provides benefits for model averaging with MCMC-based inference since it factorises the problem and also allows computations to be trivially run in parallel.

The exact model averaging approach is possible due to the guaranteed acyclicity of DBNs with edges permitted forward in time only. Under a modular posterior scoring function, summations over the entire graph space decompose into products of summations over parent sets. Thus, posterior edge probabilities can be calculated by model averaging in a variable selection sense. Such a decomposition has previously been exploited in the context of (static) BN structure learning, when an ordering over nodes is assumed, which also guarentees acyclicity (node orders were introduced in Section 2.3.5.2). Buntine [1991] first noted the decomposition and Cooper and Herskovits [1992] proposed the K2 algorithm, which exploits the decomposition (using maximisations instead of summations) to find the MAP BN structure under a given ordering. The order MCMC method proposed by Friedman and Koller [2003] and exact order-space method of Koivisto [2006] (see Section 2.3.5.2) use the decomposition to perform model averaging under a given order, and consider the whole space of orders. It is also used by Werhli and Husmeier [2007] to approximate the normalisation constant of a network prior. We note that the restriction on edge directionality in the DBNs considered here can be regarded as a known ordering over nodes, in which all nodes in the second time slice appear after all nodes in the first time slice, along with additional restrictions to preclude edges within a time slice.

As discussed in Chapter 3, the parameter priors employed here (i.e. the g-prior formulation for regression coefficients and improper reference prior for the variance) are a special (and limiting) case of the NIG prior described in Section 2.3.1.2. The priors are placed on the parameters of the local conditionals (4.4), which is in contrast to the widely-used BGe score (outlined in Section 2.3.5.2), where the Normal-Wishart prior is placed on the parameters of the joint Gaussian distribution over all variables. However, the two formulations are closely related; a Wishart prior on the precision matrix of the joint Gaussian is equivalent to a NIG prior on the local conditionals, with hyperparameters determined by those of the Wishart prior [see e.g. Dobra *et al.*, 2004]. Further work is required to determine whether the parameterisation of the (limiting) NIG prior used here is coherent in the sense that it corresponds to a well-defined prior on the joint Gaussian distribution. We note that this coherency is more relevant for BNs than for the DBNs used here, since structure learning for DBNs reduces to independent variable selection problems with the set of response variables (nodes in second time slice) being disjoint from the set of predictor variables (nodes in first time slice). It is not necessary during inference to consider the global network structure and so it is natural to only consider local conditionals and local priors on those conditionals.

The DBN model in this work makes a widely-used assumption of homogeneity of parameters and network structure through time. However, these assumptions are likely to be unrealistic for cellular protein signalling, e.g. in the event that accumulating epigenetic alterations fundamentally alters the state of the cells and thereby the underlying dynamical system. The softening of these homogeneity assumptions can lead to a rapid increase in numbers of parameters and/or the size of graph space, resulting in statistical (and computational) challenges, especially when the number of time points observed is small. Recently, non-homogeneous DBN methods have been proposed in the literature that ameliorate these effects [Robinson and Hartemink, 2010; Grzegorczyk and Husmeier, 2011b].

For several reasons, it is not expected or indeed possible, that data-driven characterisations of signalling network structure can reveal the 'correct' context-specific structure. These reasons, some of which we discuss further in Chapter 6, include: paucity of data, random stochasticity in data, challenges in experimental design such as selection of appropriate time points, possible low information level in observational data (i.e. as opposed to interventional data), issues regarding the interpretation of inferred links as being causal, information loss due to data discretisation, biologically unrealistic models, and unidentifiability of models [Craciun and Pantea, 2008; Oates and Mukherjee, 2012]. Therefore, results should be regarded as a summary of the given dataset, that can be used to generate hypotheses for independent validation.

In order to get an idea of the accuracy of our inferred breast cancer cell line signalling network we performed a systematic comparison of the inferred network structure (Figure 4.5(a)) with a network generated from independent, published siRNA data [Lu *et al.*, 2011]. Application of an siRNA (partially) knocks down a specific gene and therefore prevents synthesis of the corresponding protein. The data shows the effects of siRNAs, targeting specific proteins, on phosphorylation levels of several signalling proteins. Large effects are evidence in favour of links between the protein being targeted by siRNA and affected proteins. The intersection of the siRNA dataset and our data allows a network with 135 possible edges to be compared. A network with 50 edges is produced from the siRNA data. Of the 10 edges that appear in our inferred network, 7 are also in the siRNA network. This agreement is more than would be expected by chance and is significant at the 5% level. However, the siRNA data does not provide a gold-standard reference for systematic validation of inferred links due to only partial knockdowns of the target and off-target effects.

We predicted and validated novel links, suggesting existence of cross-talk between signalling pathways, and links that have been previously well documented. Overall, our results suggest that statistical approaches, such as those presented here, can usefully integrate proteomic data with existing knowledge to generate hypotheses regarding context-specific signalling links; here, specific to an individual breast cancer cell line. By applying these approaches to many individual cancers, we could probe signalling heterogeneity across, and even within, cancer subtypes, and thereby shed light on therapeutic heterogeneity. Thus the methods reported here could help guide development of personalised cancer therapies in the future, that are targeted to specific cancer subtypes or even individual cancers. However, the sheer complexity of cancer signalling is daunting and so the present work is only a first step in the direction of characterising signalling networks that are context-specific.

Chapter 5

Network-based clustering with mixtures of sparse Gaussian graphical models

5.1 Introduction

Clustering of high-dimensional data has been the focus of much statistical research over the past decade. The increasing prevalence of high-throughput biological data has been an important motivation for such efforts. For example, clustering methods have been applied to gene expression data to find sets of coregulated genes [Eisen *et al.*, 1998] and discover disease subtypes [Golub *et al.*, 1999] (further references can be found in Section 2.3.9). In this Chapter we focus on the latter application; that is, to cluster a small-to-moderate number of high-dimensional samples (e.g. tissue samples or cell lines) with the aim to discover disease subtypes. Numerous clustering algorithms have been used in biological applications, notably for gene expression data. The reader is referred to Section 2.3.9 for background information on clustering and, in particular, on the following widely-used clustering methods: K-means, hierarchical clustering and model-based clustering. As noted in Section 2.3.9, model-based clustering [McLachlan and Basford, 1987; Fraley and Raftery, 1998; McLachlan and Peel, 2000; Fraley and Raftery, 2002] with Gaussian mixture models is a popular approach to clustering that is rooted in an explicit statistical model.

In this Chapter, we bring together clustering and structural inference for graphical models. Background information on graphical models and graphical model structure learning can be found in Sections 2.3.4-2.3.6. In particular, we develop a model-based clustering approach with components defined by graphical models.

This allows simultaneous recovery of cluster assignments and learning of cluster-specific graphical model structure. Our work is of particular relevance to questions concerning undiscovered heterogeneity at the level of network structure. Such questions arise in diverse molecular biology applications. The edge structure of biological networks can differ depending on context, e.g. disease state or other subtype [Pawson and Warner, 2007; Yuan and Cantley, 2008], in ways that may have implications for targeted and personalised therapies [Pe'er and Hacohen, 2011]. When such heterogeneity is well-understood, samples can be partitioned into suitable subsets prior to network structure learning [Altay *et al.*, 2011] (or other supervised network-based approaches [Chuang *et al.*, 2007]). However, molecular classifications that underpin such stratifications are in their infancy and often data may harbour hitherto unknown subtypes. Moreover, if subtypes differ with respect to underlying network structure, clustering and structural inference become related tasks: clustering methods that do not model cluster-specific covariance structure (including K-means, hierarchical clustering or model-based clustering with diagonal covariance matrices [de Souto *et al.*, 2008]), may be unable to discover the correct clustering, thereby compromising also the ability to elucidate network structure.

A specific motivation for the work we present comes from the molecular biology of cancer. Cancers show a remarkable degree of biological heterogeneity [TCGA-Network, 2011] that has key therapeutic implications. The identification of subsets of cancers that share underlying biology is of great interest in both basic and translational cancer research, as it may suggest ways to rationally target therapies to responsive sub-populations. As discussed in Chapter 1 and Section 2.1, biological networks called signalling networks play a central role in cancer and it is components of these networks that are targets for many new therapeutic agents. An attractive idea is therefore to cluster cancer samples into groups that show evidence of shared signalling network structure. Recently, biochemical assays that permit interrogation of signalling proteins, post-translationally modified by phosphorylation on specific sites, have reached a level of maturity [Hennessy *et al.*, 2010; Ciaccio *et al.*, 2010; Bendall *et al.*, 2011] that starts to allow such approaches to be pursued (see also Section 2.2). We show an application to such data below.

As cluster-specific network models, we use sparse Gaussian graphical models. Recall from Section 2.3.4.2 that these are multivariate Gaussian models in which an undirected graph is used to represent conditional independence relationships between variables, and that inferring the edge set of a Gaussian graphical model is equivalent to identifying the location of non-zero entries in the precision matrix. Background information regarding structure learning of sparse Gaussian graphical models (i.e sparse precision matrix estimation) can be found in Section 2.3.6. In particular, maximum penalised likelihood estimators with an ℓ_1 penalty applied to the precision matrix have been proposed by Yuan and Lin [2007]; Friedman *et al.* [2008]; Rothman *et al.* [2008] and D'Aspremont *et al.* [2008]. Analogous to the lasso (see Section 2.3.3.3), where sparse models are encouraged by shrinking some regression coefficients to be exactly zero, the ℓ_1 penalty on the precision matrix encourages sparsity by estimating some matrix entries as exactly zero. Since a sparse precision matrix corresponds to a sparse Gaussian graphical model structure, ℓ_1-penalised estimation is well-suited for inference of molecular networks, where sparsity is often a valid assumption. Moreover, regularisation enables estimation in the challenging 'large p, small n' regime that is ubiquitous in these settings, but renders standard covariance estimators inapplicable or ill-behaved.

Our work adds to the literature in three main ways. First, the penalised mixture-model formulation we propose extends previous work. Mukherjee and Hill [2011] put forward a related 'network clustering' approach, but this is not rooted in a formal statistical model and estimation is carried out using a heuristic, K-means-like algorithm with 'hard' cluster assignments. We show empirically that likelihood-based inference via an EM algorithm formulation confers benefits over this approach. EM algorithms for penalised likelihoods have previously been proposed for finite mixture of regression models [Khalili and Chen, 2007; Städler *et al.*, 2010] and for penalised model-based clustering [Pan and Shen, 2007; Zhou *et al.*, 2009]. The approach in Zhou *et al.* [2009] is similar to the one here. However, our ℓ_1 penalty takes a more general form, allowing also for dependence on mixing proportions at the level of the full likelihood. We show that at smaller sample sizes in particular, the ℓ_1 penalty we propose offers substantial gains. Furthermore, while we are interested in simultaneous clustering and cluster-specific network structure learning, Zhou *et al.* [2009] focus on clustering in combination with variable selection.

Second, we present empirical results investigating the performance of penalisation regimes. A penalty parameter controls the extent to which sparsity is encouraged in the precision matrix and corresponding graphical model. The choice of method for setting the penalty parameter together with the different forms of the ℓ_1 penalty itself result in several possible regimes that can be difficult to choose between *a priori*. Our results show that the choice of regime can be influential and suggest general recommendations.

Third, we present an application in cancer biology. We analyze data from key signalling proteins, post-translationally modified by phosphorylation on specific sites, in a panel of breast cancer cell lines. It remains unclear whether breast cancers display substantive heterogeneity at the level of signalling network structure. We find that ℓ_1-penalised network-based clustering recapitulates, from phosphoproteomic data alone, a biological classification that was previously established using independent gene expression data [Perou *et al.*, 2000; Sørlie *et al.*, 2001]. Moreover, clustering methods that do not take cluster-specific covariance structure into account fail to do so. This is a striking finding because it suggests that breast cancer subtypes do indeed differ at the level of signalling network structure.

The remainder of this Chapter is organised as follows. In Section 5.2 we introduce ℓ_1-penalised estimation for Gaussian graphical models and then go on to describe the proposed mixture model. In Section 5.3.1 we present an empirical comparison, on synthetic data, of several regimes for the ℓ_1 penalty term and tuning parameter selection. This is followed in Section 5.3.2 by the application to breast cancer data. In Section 5.4 we close with a discussion of our findings and suggest areas for future work.

5.2 Methods

5.2.1 Penalised estimation of Gaussian graphical model structure

Let $\mathbf{X} = (X_1, \ldots, X_p)^{\mathsf{T}}$ denote a random vector having p-dimensional Gaussian density $f(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A Gaussian graphical model $G = (V, E)$ describes conditional independence relationships between the random variables $X_1, \ldots, X_p$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ denote the inverse covariance or precision matrix. Then, as explained in Section 2.3.4.2, non-zero entries in $\boldsymbol{\Omega} = (\omega_{ij})$ correspond to edges in the graphical model, that is $\omega_{ij} \neq 0 \iff (i,j) \in E$. Thus, inferring the edge set of a Gaussian graphical model is equivalent to identifying the location of non-zero entries in the precision matrix.

Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathsf{T}}$ is a random sample from $f(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ denote sample mean and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathsf{T}}$ sample covariance. The precision matrix $\boldsymbol{\Omega}$ may be estimated by maximum likelihood. The log-likelihood function is given, up to a constant, by

$$l(\boldsymbol{\Omega}) = \log|\boldsymbol{\Omega}| - \operatorname{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}) \tag{5.1}$$

where $|\cdot|$ and $\operatorname{tr}(\cdot)$ denote matrix determinant and trace respectively. The maximum likelihood estimate is given by inverting the sample covariance matrix, $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Sigma}}^{-1}$. However for $n < p$, $\hat{\boldsymbol{\Sigma}}$ is singular and so cannot be used to estimate $\boldsymbol{\Omega}$. Even when $n \geq p$, $\hat{\boldsymbol{\Omega}}$ can be a poor estimator for large p and does not in general yield sparse precision matrices.

Sparse estimates can be encouraged by placing an ℓ_1 penalty on the entries of the precision matrix $\mathbf{\Omega}$. This results in the following penalised log-likelihood:

$$l_p(\mathbf{\Omega}) = \log|\mathbf{\Omega}| - \operatorname{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}) - \lambda \left\|\mathbf{\Omega}\right\|_1 \tag{5.2}$$

where $\left\|\mathbf{\Omega}\right\|_1 = \sum_{i,j} |\omega_{i,j}|$ is the elementwise ℓ_1 matrix norm and λ is a non-negative tuning parameter controlling sparsity of the estimate. The maximum penalised likelihood estimate is obtained by maximising (5.2) over symmetric, positive-definite matrices. This is a convex optimisation problem and several procedures have been proposed to obtain solutions. Yuan and Lin [2007] used the maxdet algorithm, while D'Aspremont *et al.* [2008] proposed a more efficient semi-definite programming algorithm using interior point optimisation. Rothman *et al.* [2008] offered a fast approach employing Cholesky decomposition and the local quadratic approximation, and Friedman *et al.* [2008] proposed the even faster graphical lasso algorithm, based on the coordinate descent algorithm for the lasso (see Section 2.3.3.3). We use the graphical lasso algorithm in our investigations and refer the interested reader to the references for full details.

5.2.2 Mixture of penalised Gaussian graphical models

For background information on Gaussian mixture models and model-based clustering, see Section 2.3.9.3. Notation used below also follows that used in the aforementioned Section. In particular, the log-likelihood for a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a finite Gaussian mixture distribution is given by (2.66), and is reproduced here,

$$l(\mathbf{\Theta}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right). \tag{5.3}$$

where the mixing proportions π_k satisfy $0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1$, f_k is the p-dimensional multivariate Gaussian density with component-specific mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{\Sigma}_k$, and $\mathbf{\Theta} = \{(\pi_k, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) : k = 1, \ldots, K\}$ is the set of all unknown parameters.

In model-based clustering each mixture component corresponds to a cluster. In the present setting, since each cluster (or component) is Gaussian distributed with a cluster-specific (unconstrained) covariance matrix, each cluster represents a distinct Gaussian graphical model.

In the Gaussian mixture model with cluster-specific covariance matrices, the number of parameters is of order Kp^2. Estimation is more challenging than for a single precision matrix (or Gaussian graphical model) and so, as described above, in settings where number of variables p is moderate-to-large in relation to sample size n, overfitting and invalid covariance estimates are a concern. We employ an ℓ_1 penalty on each of the K precision matrices to promote sparsity and ameliorate these issues. Such ℓ_1 penalties have previously been proposed for clustering with Gaussian graphical models [Zhou *et al.*, 2009; Mukherjee and Hill, 2011].

We propose the following penalised log-likelihood,

$$l_p(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k \left(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k \right) \right) - \frac{n}{2} p_{\lambda,\gamma}(\mathbf{\Theta}) \tag{5.4}$$

where the penalty term is given by

$$p_{\lambda,\gamma}(\mathbf{\Theta}) = \lambda \sum_{k=1}^{K} \pi_k^{\gamma} \left\| \mathbf{\Omega}_k \right\|_1 \tag{5.5}$$

and γ is a binary parameter controlling the form of the penalty term. Setting $\gamma = 0$ results in the conventional penalty term, as used in Zhou *et al.* [2009], with no dependence on the mixing proportions π_k. Setting $\gamma = 1$ weights the penalty from each cluster by its corresponding mixing proportion. While this form of penalty is novel in this setting, an analogous penalty has been proposed by Khalili and Chen [2007] and Städler *et al.* [2010] for ℓ_1-penalised finite mixture of regression models. In this work, we empirically compare these two forms of penalty term for clustering with, and estimation of, Gaussian graphical models.

5.2.3 Maximum penalised likelihood

As with the unpenalised log-likelihood (5.3), the penalised likelihood (5.4) can be maximised using an EM algorithm, which we now describe. The EM algorithm for the unpenalised likelihood was described in Section 2.3.9.3. Our algorithm is similar to that of Zhou *et al.* [2009], but they consider only the $\gamma = 0$ regime and also penalise the mean vectors to perform variable selection.

Let z_i be a latent variable satisfying $z_i = k$ if observation $\mathbf{x}_i$ belongs to cluster k (the responsibility of cluster k for sample $\mathbf{x}_i$). Then we have $P(z_i = k) = \pi_k$ and $p(\mathbf{x}_i \mid z_i = k) = f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$. The penalised log-likelihood for the complete data $\{\mathbf{x}_i, z_i\}_{i=1}^{n}$ is

$$l_{p,c}(\mathbf{\Theta}) = \sum_{i=1}^{n} \log(\pi_{z_i}) + \log \left(f_{z_i} \left(\mathbf{x}_i \mid \boldsymbol{\mu}_{z_i}, \mathbf{\Sigma}_{z_i} \right) \right) - \frac{n}{2} p_{\lambda,\gamma}(\mathbf{\Theta}). \tag{5.6}$$

In the E-step of the EM algorithm, given current estimates of the parameters

$\boldsymbol{\Theta}^{(t)}$, we compute

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) &= \mathbb{E}\left[l_{p,c}(\boldsymbol{\Theta}) \mid \{\mathbf{x}_i\}_{i=1}^n , \boldsymbol{\Theta}^{(t)} \right] \\ &= \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(t)} \left[\log(\pi_k) + \log\left(f_k\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right)\right)\right] - \frac{n}{2} p_{\lambda,\gamma}(\boldsymbol{\Theta}) \end{aligned} \tag{5.7}$$

where $\tau_{ik}^{(t)}$ is the posterior probability of observation $\mathbf{x}_i$ belonging to cluster k,

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t)} f_k \left(\mathbf{x}_i \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)} \right)}{\sum_{j=1}^K \pi_j^{(t)} f_j \left(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)} \right)} \tag{5.8}$$

and can be thought of as a ‘soft’ cluster assignment.

In the M-step we seek to maximise $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\Theta}$ to give new estimates for the parameters $\boldsymbol{\Theta}^{(t+1)}$. When $\gamma = 0$ the mixture proportions π_k do not appear in the penalty term $p_{\lambda,\gamma}(\boldsymbol{\Theta})$ and so we use the standard EM algorithm update for unpenalised Gaussian mixture models, given in (2.70), and reproduced here,

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{n}. \tag{5.9}$$

For $\gamma = 1$, since π_k appears in the penalty term, maximisation of $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ with respect to π_k is non-trivial. We follow Khalili and Chen [2007] and use the standard update (5.9). If the standard update improves $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ then this is sufficient to obtain (local) maxima of (5.4). An improvement is not guaranteed here, but as found in Khalili and Chen [2007], the method works well in practice.

Since the penalty term is independent of $\boldsymbol{\mu}_k$, we again use the standard update, given in (2.71), and reproduced here,

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n \tau_{ik}^{(t)}}. \tag{5.10}$$

The update for $\boldsymbol{\Sigma}_k$, or equivalently $\boldsymbol{\Omega}_k$, is given by

$$\begin{aligned} \boldsymbol{\Omega}_k^{(t+1)} &= \underset{\boldsymbol{\Omega}_k}{\arg\max} \left[\sum_{i=1}^n \tau_{ik}^{(t)} \left(\log|\boldsymbol{\Omega}_k| - \mathrm{tr}(\boldsymbol{\Omega}_k \mathbf{S}_k^{(t)}) \right) - n\lambda \left(\pi_k^{(t+1)} \right)^\gamma \|\boldsymbol{\Omega}_k\|_1 \right] \\ &= \underset{\boldsymbol{\Omega}_k}{\arg\max} \left[\log|\boldsymbol{\Omega}_k| - \mathrm{tr}(\boldsymbol{\Omega}_k \mathbf{S}_k^{(t)}) - \tilde{\lambda}_k^{(t)} \|\boldsymbol{\Omega}_k\|_1 \right] \end{aligned} \tag{5.11}$$

where

$$\mathbf{S}_k^{(t)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^{\mathsf{T}}}{\sum_{i=1}^n \tau_{ik}^{(t)}} \tag{5.12}$$

is the standard EM algorithm update for $\mathbf{\Sigma}$ (see also (2.72)) and

$$\tilde{\lambda}_k^{(t)} = n\lambda \frac{\left(\pi_k^{(t+1)}\right)^{\gamma}}{\sum_{i=1}^n \tau_{ik}^{(t)}}. \tag{5.13}$$

The optimisation problem in (5.11) is of the form of that in (5.2) with $\hat{\mathbf{\Sigma}}$ replaced by $\mathbf{S}_k^{(t)}$ and a scaled tuning parameter $\tilde{\lambda}_k^{(t)}$. Hence we can use the efficient graphical lasso algorithm [Friedman *et al.*, 2008] to perform the optimisation.

From (5.9) we have

$$\tilde{\lambda}_k^{(t)} = \begin{cases} \frac{\lambda}{\pi_k^{(t+1)}} & \text{if } \gamma = 0 \\ \lambda & \text{if } \gamma = 1 \end{cases} \tag{5.14}$$

Hence, when $\gamma = 0$, $\tilde{\lambda}_k^{(t)}$ is a cluster-specific parameter inversely proportional to effective cluster sample size, whereas $\gamma = 1$ simply yields λ. We note that, even though $\gamma = 1$ gives a cluster-specific tuning parameter in the penalised log-likelihood (5.4) while $\gamma = 0$ does not, the converse is actually true for the EM algorithm updates (5.11).

Our overall algorithm is as follows:

1. Initialise $\mathbf{\Theta}^{(0)}$: Randomly assign each observation $\mathbf{x}_i$ into one of K clusters, subject to a minimum cluster size $n_{\min}$. Set $\pi_k^{(0)} = n_k/n$ where n_k is the number of observations assigned to cluster k, set $\boldsymbol{\mu}_k^{(0)}$ to sample mean of cluster k, and set $\mathbf{\Omega}_k^{(0)}$ to the maximum penalised likelihood estimate for the cluster k precision matrix (using (5.2)).

2. E-step: Calculate posterior probabilities ('soft' assignments) $\tau_{ik}^{(t)}$ using (5.8).

3. M-step: Calculate updated parameter estimates $\mathbf{\Theta}^{(t+1)}$ using (5.9)-(5.12).

4. Iterate or terminate: Increment t. Repeat steps 2 and 3, or stop if one of the following criteria is satisfied:

   - A maximum number of iterations T is reached; $t > T$.
   - A minimum cluster size $n_{\min}$ is reached; $\sum_{i=1}^n \tau_{ik}^{(t)} < n_{\min}$ for some k.

- Relative change in penalised log-likelihood is below a threshold δ; $\left|l_p(\mathbf{\Theta})^{(t)}/l_p(\mathbf{\Theta})^{(t-1)} - 1\right| \le \delta$.

In all experiments below we set $T = 100$, $n_{\min} = 4$ and $\delta = 10^{-4}$. Since the EM algorithm may only find local maxima, we perform 25 random restarts and select the one giving the highest penalised log-likelihood. 'Hard' cluster assignments are obtained by assigning observations to the cluster k with largest probability τ_{ik}.

5.2.4 Tuning parameter selection

Two approaches are commonly used to set the tuning parameter: multifold cross-validation (CV) and criteria such as BIC. CV and BIC were introduced in Sections 2.3.2.1 and 2.3.2.2 respectively. In multifold CV, the data samples are partitioned into S data subsets, denoted by $\mathbf{X}^{(s)}$ for $s = 1, \ldots, S$. Let $\hat{\mathbf{\Theta}}_\lambda^{(-s)} = \{(\pi_{k\lambda}, \boldsymbol{\mu}_{k\lambda}, \mathbf{\Sigma}_{k\lambda}) : k = 1, \ldots, K\}$ denote the penalised likelihood estimate, obtained using tuning parameter λ and by application of the EM algorithm described above to all data save that in subset $\mathbf{X}^{(s)}$ (training data). Performance of this estimate is assessed using the predictive log-likelihood; that is, Equation (5.3) applied to subset $\mathbf{X}^{(s)}$ (test data). This is repeated S times, allowing each subset to play the role of test data. The CV score is

$$\mathrm{CV}(\lambda) = \sum_{s=1}^{S} \sum_{i:\mathbf{x}_i \in \mathbf{X}^{(s)}} \log\left(\sum_{k=1}^{K} \hat{\pi}_{k\lambda}^{(-s)} f_k\left(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_{k\lambda}^{(-s)}, \hat{\mathbf{\Sigma}}_{k\lambda}^{(-s)}\right)\right). \tag{5.15}$$

Then we choose λ that maximises $\mathrm{CV}(\lambda)$, where the maximisation is performed via a grid search. Finally, the selected value is used to learn penalised likelihood estimates from all data.

In the larger sample case, an alternative to multifold CV is to partition the data into two and perform a single train/test iteration, selecting λ that maximises the predictive log-likelihood on the test data with penalised parameter estimates from the training data.

We define the following BIC score for our penalised mixture model:

$$\mathrm{BIC}(\lambda) = -2l(\hat{\mathbf{\Theta}}_\lambda) + \mathrm{df}_\lambda \log(n) \tag{5.16}$$

where $l(\cdot)$ is the unpenalised log-likelihood (5.3), $\hat{\mathbf{\Theta}}_\lambda$ is the penalised likelihood estimate obtained with tuning parameter λ and df_λ is degrees of freedom. Yuan and Lin [2007] proposed an estimate of the degrees of freedom for ℓ_1-penalised precision matrix estimation, which generalises to our penalised Gaussian mixture model setting to give

$$\mathrm{df}_\lambda = K(p+1) - 1 + \sum_{k=1}^{K} \#\left\{(j,j') : j \leq j', (\hat{\omega}_{k\lambda})_{jj'} \neq 0\right\}. \tag{5.17}$$

where $(\hat{\omega}_{k\lambda})_{jj'}$ is element (j,j') in $\hat{\mathbf{\Omega}}_{k\lambda}$, the penalised likelihood estimate for the cluster k precision matrix, using tuning parameter λ. Using a grid search, we choose λ that minimises BIC(λ).

BIC is often preferred over CV as it is less computationally intensive. However, we note that, even BIC can be computationally expensive when used within clustering since each λ value in the grid search requires a full application of EM-based clustering. Hence, to reduce computation time, we also consider a heuristic, approximate version of these approaches. The heuristic we propose relies on the notion that the optimal tuning parameter value does not depend strongly on cluster assignments but rather largely on general properties of the data (such as p and n). The approach proceeds as follows. First, observations are randomly assigned to clusters, producing K pseudo-clusters each with mean size n/K. Second, parameter estimates are obtained for the pseudo-clusters. $\hat{\pi}_k$ is taken to be the proportion of samples in pseudo-cluster k and $\hat{\boldsymbol{\mu}}_k$ is the sample mean of pseudo-cluster k. Then, for varying λ, we obtain penalised estimates $\hat{\mathbf{\Omega}}_{k\lambda}$ by optimising (5.2) for each pseudo-cluster with the graphical lasso. This can be done efficiently using the `glassopath` algorithm in R [Friedman *et al.*, 2008] which obtains penalised estimates for all considered values of λ simultaneously. Third, using these estimates, CV (BIC) scores are calculated and maximised (minimised) to select λ. These three steps are repeated multiple times and λ values obtained are averaged to produce a final value.

5.3 Results

5.3.1 Simulated data

In this section we apply the ℓ_1-penalised Gaussian graphical model clustering approach to simulated data. We consider a number of combinations of ℓ_1 penalty term and tuning parameter scheme (as described in Section 5.2 above) and assess their performance in carrying out three related tasks. First, recovery of correct cluster assignments. Second, estimation of cluster-specific graphical model structure (i.e. location of non-zero entries in cluster-specific precision matrices). Third, estimation of cluster-specific precision matrices (i.e. estimation of matrix elements, not just locations of non-zero entries). We note that this latter task is of less interest here since we are mainly concerned with clustering and inference of cluster-specific network structure.

5.3.1.1 Data generation

In our simulation we considered p-dimensional data consisting of $K = 2$ clusters, each with a known and distinct Gaussian graphical model structure (i.e. sparse precision matrix). Sparse precision matrices were created using an approach based on that used by Rothman *et al.* [2008] and Cai *et al.* [2011]. In particular, we created a symmetric $p \times p$ matrix B_1 with zeros everywhere except for p randomly chosen pairs of symmetric, off-diagonal entries, which took value 0.5. A second matrix B_2 was created from B_1 by selecting half of the p non-zero symmetric pairs at random and relocating them to new randomly chosen symmetric positions. We then set $\mathbf{\Omega}_k = B_k + \delta_k I$, where δ_k is the minimal value such that $\mathbf{\Omega}_k$ is positive-definite with condition number less than p. Finally, the precision matrices $\mathbf{\Omega}_k$ were standardised to have unit diagonals. This resulted in cluster-specific Gaussian graphical models each with p edges, half of which were shared by both network structures. Data were generated from $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}_1)$ and $\mathcal{N}\left(\frac{\alpha}{\sqrt{p}}\mathbf{1}, \mathbf{\Omega}_2\right)$ for clusters 1 and 2 respectively, where $\mathbf{1}$ is the vector of ones. The mean of cluster two is defined such that the parameter α sets the Euclidean distance between the cluster means. In the experiments below we consider $p = 25, 50, 100$ and cluster sample sizes of $n_k = 15, 25, 50, 100, 200$. We set $\alpha = 3.5$, resulting in individual component-wise means for cluster two of 0.70, 0.50 and 0.35 for $p = 25, 50$ and 100 respectively. This reflects the challenging scenario where clusters do not have substantial differences in mean values, but display heterogeneity in network structure while also sharing some network structure across clusters.

5.3.1.2 Cluster assignment

We assessed ability to recover correct cluster assignments from 50 simulated datasets, under the following four regimes for the penalty term $p_{\lambda,\gamma}(\mathbf{\Theta})$ in (5.5): $\gamma = 0$ or 1 and λ set by BIC or a train/test scheme, maximising the predictive log-likelihood on an independent test dataset with cluster sample size matching the training dataset. These regimes are described fully above and summarised in Table 5.1. We also compared with (i) K-means; (ii) standard non-penalised full-covariance Gaussian mixture models estimated using EM algorithm; and (iii) 'network clustering', an ℓ_1-penalised Gaussian graphical model clustering approach proposed by Mukherjee and Hill [2011]. This is similar to the approach employed here but uses a heuristic, K-means-like algorithm with 'hard' cluster assignments rather than a mixture-model

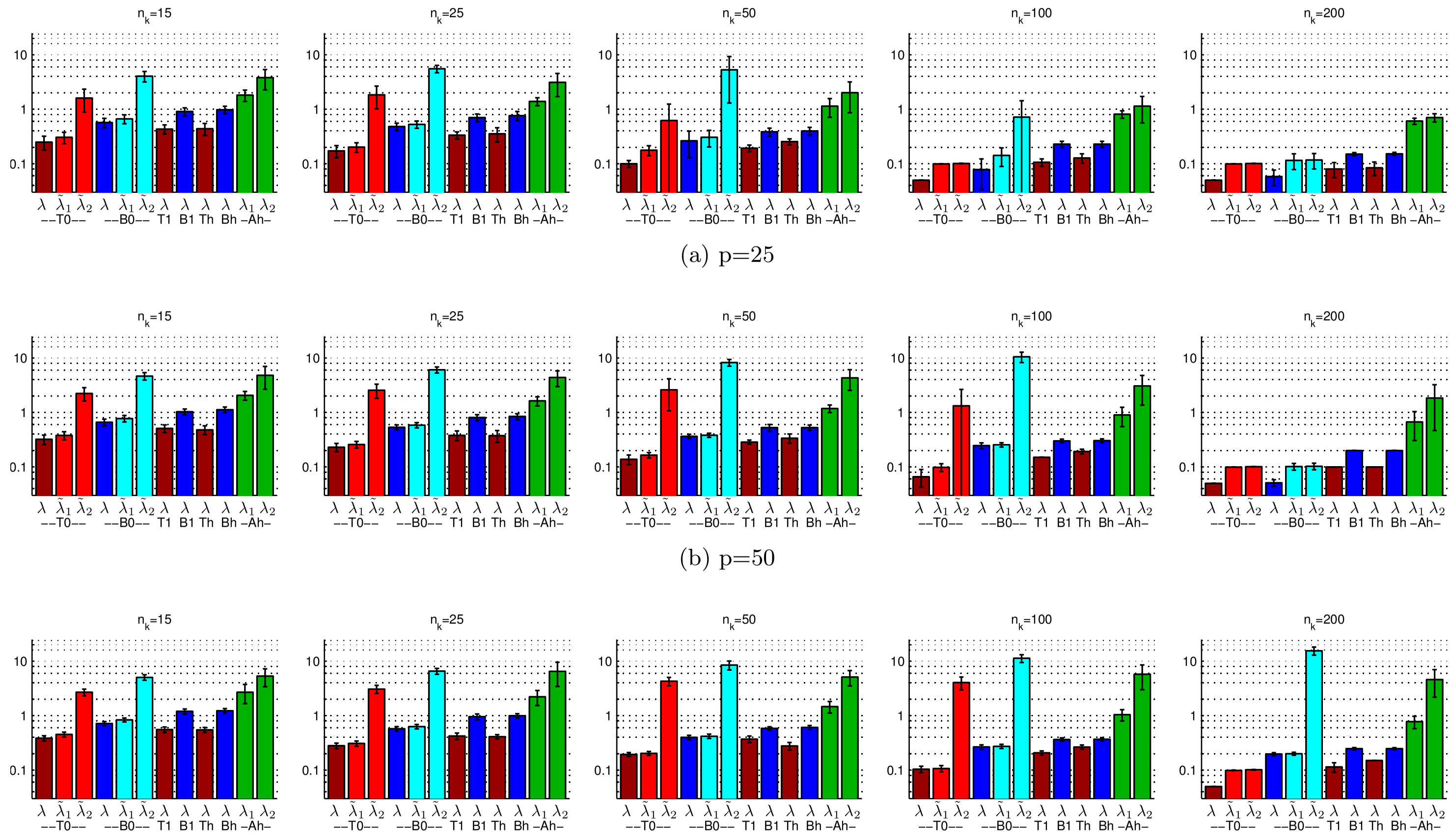
Method	**Penalty term**	**Tuning parameter selection**	**Abbrev.**
mixture of ℓ_1-penalised Gaussian graphical models with EM ('soft' assignments)	$p_{\lambda,\gamma}(\boldsymbol{\Theta})$ with $\gamma = 0$: $\lambda \sum_{k=1}^{K} \|\boldsymbol{\Omega}_k\|_1$	Train/test	T0
		BIC	B0
	$p_{\lambda,\gamma}(\boldsymbol{\Theta})$ with $\gamma = 1$: $\lambda \sum_{k=1}^{K} \pi_k \|\boldsymbol{\Omega}_k\|_1$	Train/test	T1
		BIC	B1
ℓ_1-penalised Gaussian graphical models ('hard' assignments) Mukherjee and Hill [2011]	$\lambda\|\boldsymbol{\Omega}_k\|_1$, $k = 1, \ldots, K$	Train/test	Th
		BIC	Bh
	$\lambda_k\|\boldsymbol{\Omega}_k\|_1$, $k = 1, \ldots, K$	Analytic[1]	Ah
K-means	n/a	n/a	KM
non-penalised Gaussian mixture model with EM	n/a	n/a	NP

[1] following Banerjee *et al.* [2008] (see text for details)

Table 5.1: **Clustering methods and regimes investigated, with corresponding abbreviations.**

formulation with EM algorithm. For (i) we used the `kmeans` function in the MATLAB statistics toolbox with K=2 and 1000 random initialisations and for (iii) we used MATLAB function `network_clustering` [Mukherjee and Hill, 2011]. For (ii) and (iii) we used the same stopping criteria as described in Methods above (namely, $T = 25$, $n_{\min} = 4$ and $\tau = 10^{-4}$) and again carried out 25 random restarts. Method (iii) requires maximisation of K penalised log-likelihoods of form (5.2) above (one for each cluster). For setting penalty parameters for this method, we considered either a single tuning parameter λ shared across both clusters and set by BIC or train/test, or cluster-specific tuning parameters λ_k, set analytically before each call to the penalised estimator using the equation proposed by Banerjee *et al.* [2008, Equation 3]. All computations were carried out in MATLAB R2010a, making an external call to the R package `glasso` [Friedman *et al.*, 2008]. Table 5.1 gives abbreviations for all methods and regimes investigated, which are used below and in figures.

Figure 5.1 shows average tuning parameter values selected by each regime. A grid search was used over values between 0.05 and 1.5, with increments of 0.05. Since, for $\gamma = 0$, the EM algorithm update tuning parameters $\tilde{\lambda}_k$ in (5.11) differ from λ, we also show $\tilde{\lambda}_k$ for these regimes. Using BIC to set the tuning parameter results in higher values than with train/test and, as expected, λ values increase with p and



(a) p=25

(b) p=50

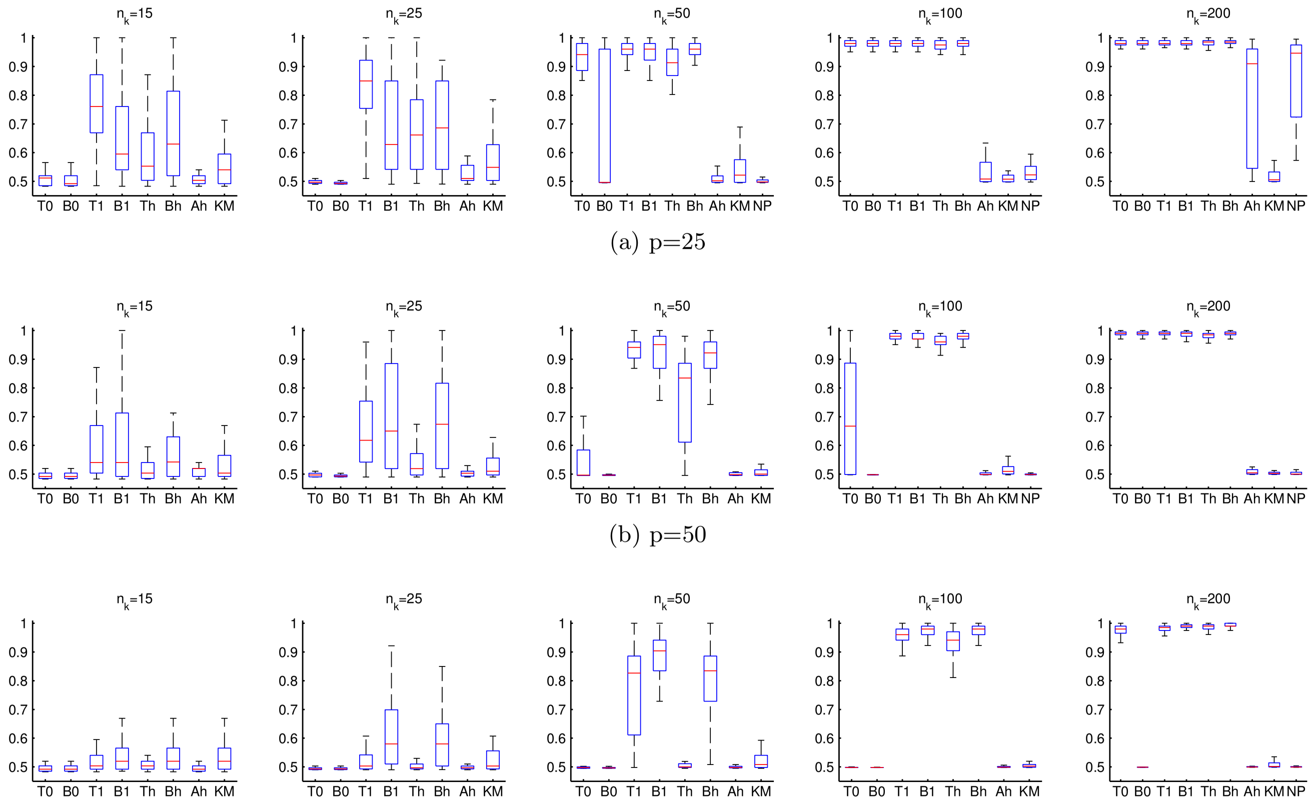
(c) p=100

Figure 5.1: **Simulated data; tuning parameter values selected for the methods and regimes in Table 5.1.** Three data dimensions ($p = 25, 50, 100$) and five per-cluster sample sizes ($n_k = 15, 25, 50, 100, 200$) were considered. Red and blue bars denote regimes employing train/test and BIC respectively. For the $\gamma = 0$ regimes (T0/B0) EM algorithm update tuning parameters $\tilde{\lambda}_k$, given in (5.14), are also shown (light red and light blue bars). (Results shown are mean values over 50 simulated datasets for each (p, n_k) regime, and are displayed on a log scale; error bars show standard deviations; cluster-specific tuning parameter λ_1 and EM algorithm update parameter $\tilde{\lambda}_1$ correspond to the largest cluster.)

decrease with n_k.

Figure 5.2 shows Rand indices (with respect to the true cluster assignments) obtained from clustering the simulated data. The Rand index [Rand, 1971] is a measure of similarity between cluster assignments, taking values between 0 and 1, where 0 indicates complete disagreement and 1 complete agreement. It is defined as follows. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n objects (here, p-dimensional samples). Let $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{K_u}\}$ and $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{K_v}\}$ be partitions of $\mathbf{X}$ (i.e. $\bigcup_{k=1}^{K_u} \mathbf{u}_k = \mathbf{X}$ and $\mathbf{u}_k \cap \mathbf{u}_{k'} = \varnothing$ for $k \neq k'$, and likewise for $\mathbf{V}$). Without loss of generality, let $\mathbf{U}$ and $\mathbf{V}$ be the true clustering and inferred clustering respectively, with K_u and K_v denoting number of clusters (in the simulation here and application below we have $K_u = K_v$). Let a be the number of pairs of objects that are in the same cluster in $\mathbf{U}$ and also in the same cluster in $\mathbf{V}$. Let b be the number of pairs of objects that are in different clusters in $\mathbf{U}$ and also in different clusters in $\mathbf{V}$. Let c be the number of pairs of objects that are in the same cluster in $\mathbf{U}$ but in different clusters in $\mathbf{V}$, and let d be the number of pairs of objects that are in different clusters in $\mathbf{U}$ but in the same cluster in $\mathbf{V}$. Then $a + b$ can be seen as the number of agreements and $c + d$ as number of disagreements. The Rand index is defined as $\frac{a+b}{a+b+c+d}$.

Box plots are shown over 50 simulated datasets for each (p, n_k) regime. The ℓ_1-penalised mixture model regimes with $\gamma = 1$ in the penalty term (T1/B1) consistently provide the best clustering results. At the largest sample sizes both train/test (T1) and BIC (B1) offer good clustering performance, with high Rand indices reported. However, for smaller sample sizes, train/test outperforms BIC at the lowest data dimensionality ($p = 25$), while the converse is true at higher dimensions ($p = 50, 100$). The non-mixture ℓ_1-penalised method (Th/Bh) also performs well, but the corresponding mixture model approaches with $\gamma = 1$ (T1/B1) are, for the most part, more effective at smaller sample sizes (see e.g. $n_k = 50$, $p = 50, 100$). This difference in performance is likely due to a combination of both differences in tuning parameter (Figure 5.1) and less accurate parameter estimation for the non-mixture approaches because they do not take uncertainty of assignment into account. Interestingly, the mixture model with conventional penalty term ($\gamma = 0$; T0/B0) shows poor performance relative to $\gamma = 1$ except at larger sample sizes, with consistently poor clustering accuracy for $n_k \leq p$. Similar performance is observed for the non-mixture method with analytic tuning parameter selection (Ah). The poor performance of these three regimes appears to be related to the fact that they all use cluster-specific tuning parameters (λ_k for Ah and $\tilde{\lambda}_k$ within EM algorithm for T0/B0), resulting in considerable differences in cluster-level penalties (see Figure 5.1). We comment further on this finding in Section 5.4 below. Due to



(a) p=25

(b) p=50

(c) p=100

Figure 5.2: **Simulated data; cluster assignment results.** Boxplots over the Rand index, a measure of similarity between inferred and true cluster labels (higher scores indicate better agreement, with a score of unity indicating perfect agreement), are shown for the methods and regimes in Table 5.1 at varying data dimensions p and per-cluster sample sizes n_k. (Results shown are over 50 simulated datasets for each (p, n_k) regime (see text for details); abbreviations for methods are summarised in Table 5.1; the non-penalised approach (NP) could not be used for $n_k \leq p$ due to small sample sizes resulting in invalid covariance estimates.)

its inability to capture the cluster-specific covariance (network) structure, K-means does not perform well, even at the largest sample size. Conventional non-penalised mixture models did not yield valid covariance estimates for sample sizes $n_k \leq p$, and for $n_k > p$ we only observe gains relative to K-means in the large sample $p = 25$, $n_k = 200$ case.

5.3.1.3 Estimation of graphical model structure

Figure 5.3 shows results for estimation of cluster-specific network structures for the methods and regimes in Table 5.1. For K-means, clustering is followed by an application, to each inferred cluster, of ℓ_1-penalised precision matrix estimation (see (5.2)) with tuning parameter set by either BIC or train/test.

Ability to reconstruct cluster-specific networks is assessed by calculating the true positive rate (TPR), false positive rate (FPR) and Matthews Correlation Coefficient (MCC),

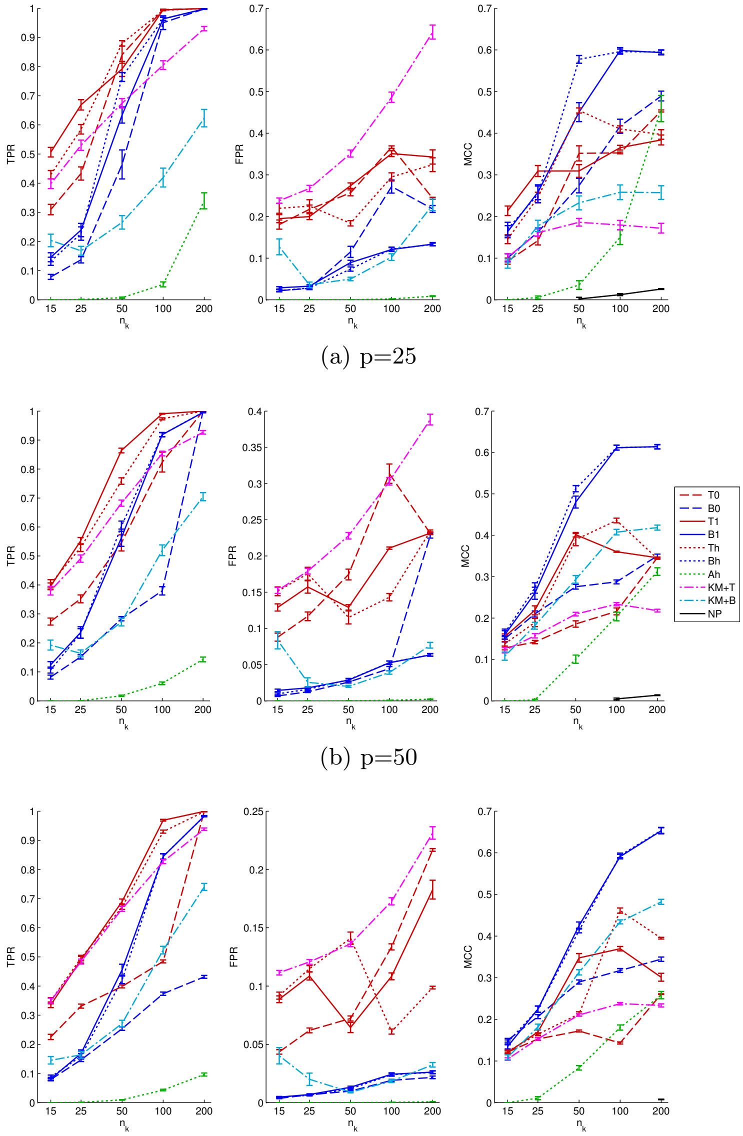
$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN} \tag{5.18}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.19}$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives (with respect to edges) respectively. MCC summarises these four quantities into one score and is regarded as a balanced measure; it takes values between -1 and 1, with higher values indicating better performance (see e.g. Baldi *et al.* [2000] for further details). Since the convergence threshold in the `glasso` algorithm is 10^{-4}, we take entries $\hat{\omega}_{ij}$ in estimated precision matrices to be non-zero if $|\hat{\omega}_{ij}| > 10^{-3}$. Since cluster assignments can only be identified up to permutation, in all cases labels were permuted to maximise agreement with true cluster assignments before calculating these quantities.

Figure 5.3 shows TPR, FPR and MCC plotted against per-cluster sample size n_k. Due to selection of larger tuning parameter values, BIC discovers fewer non-zeroes in the precision matrices than train/test, resulting in both fewer true positives and false positives. Under MCC, BIC, with either the $\gamma = 1$ mixture model (B1) or the non-mixture approach (Bh), leads to the best network reconstruction, outperforming all other regimes (except at small sample sizes with $p = 25$).

In general, train/test is not competitive relative to BIC; at larger sample sizes the best train/test regimes (T1/Th) are only comparable with the worst performing



(a) p=25

(b) p=50

(c) p=100

Figure 5.3: **Simulated data; estimation of graphical model structure.** True Positive Rate (TPR), False Positive Rate (FPR) and Matthews Correlation Coefficient (MCC) are shown as a function of per-cluster sample size n_k for the methods and regimes in Table 5.1 and at data dimensions $p = 25, 50, 100$. K-means clustering was followed by ℓ_1-penalised estimation of Gaussian graphical model structure with penalty parameter set by train/test ('KM+T') or BIC ('KM+B'). (Mean values shown over 50 simulated datasets for each (p, n_k) regime, error bars show standard errors; non-penalised approach (NP) only shown for MCC and could not be used for $n_k \leq p$ due to small sample sizes resulting in invalid covariance estimates.)

BIC regimes (B0/KM+B). We note that the non-penalised mixture approach (NP), with sample size sufficiently large to provide valid covariance estimates, does not yield sparse precision matrices (MCC scores are approximately zero).

5.3.1.4 Precision matrix estimation

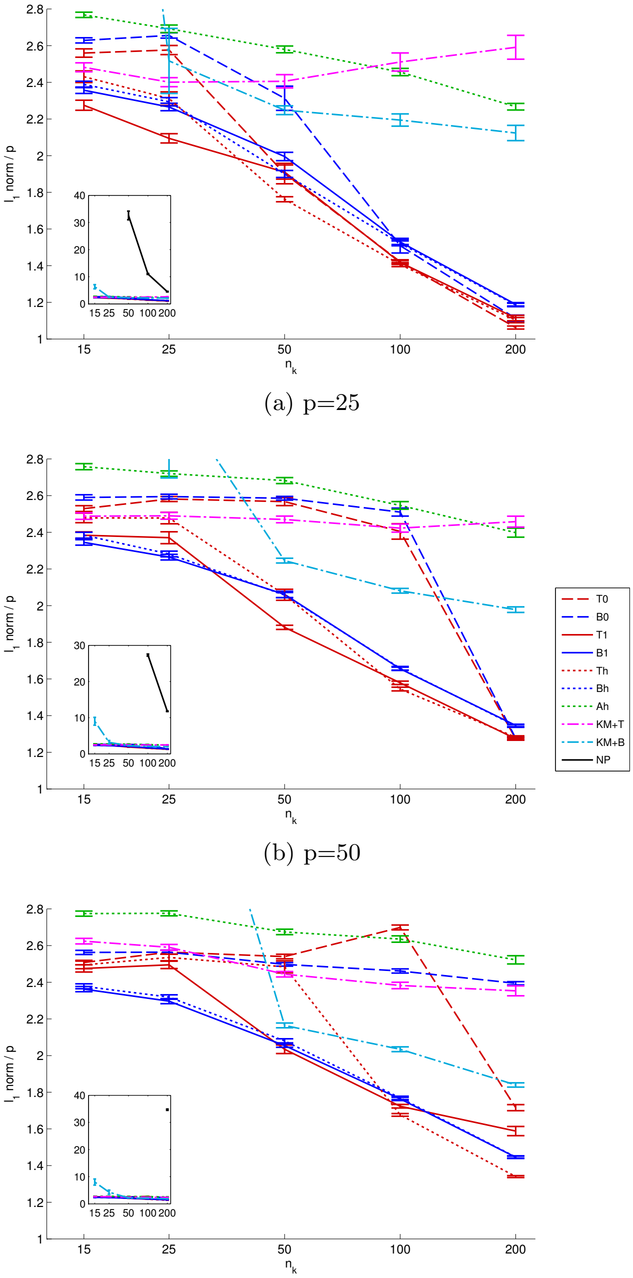
We also assessed ability to accurately estimate underlying cluster-specific precision matrices (i.e. values of the matrix elements rather than only locations of non-zeros). Accuracy is assessed using the elementwise ℓ_1 norm, $\sum_{k=1}^{K} \left\| \hat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k \right\|_1$, with inferred clusters matched to true clusters as described above. Results are shown in Figure 5.4. In contrast to clustering and Gaussian graphical model estimation, where BIC regimes B1/Bh mainly provide the best performance, the train/test methods T1/Th are mostly similar or better than B1/Bh for precision matrix estimation (the exception being small n_k, high p settings). Due to poor clustering performance, the mixture model approach with $\gamma = 0$ does not perform well unless n_k is sufficiently large. Neither K-means clustering (followed by ℓ_1-penalised precision matrix estimation), the penalised non-mixture approach with analytic tuning parameter selection (Ah), nor the non-penalised approach (NP) perform well, even at the largest sample size.

5.3.1.5 Approximate tuning parameter selection

We applied the heuristic method for setting the tuning parameter, described in Methods above, to the overall best-performing mixture model approach (regime B1; $\gamma = 1$, BIC). Figure 5.5 compares average λ values obtained using the heuristic method with those resulting from the full approach; we also show average Rand indices and computational timings. The λ values obtained via the heuristic scheme are well-behaved in the sense that they increase with p and decrease for larger n_k. We observe some bias relative to the full approach as the values obtained from the heuristic method are consistently higher. However, Rand indices remain in reasonable agreement and the heuristic offers some substantial computational gains; e.g. for $p = 25$ we see reduction of about 90% in computation time. This suggests that the heuristic approach could be useful for fast, exploratory analyses.

5.3.2 Application to breast cancer data

We applied the methods described to proteomic data obtained from a panel of breast cancer cell lines. Data consisted of $p = 39$ proteins, phosphorylated on specific sites, collectively covering a broad range of signalling pathways (see Table A.4), assayed

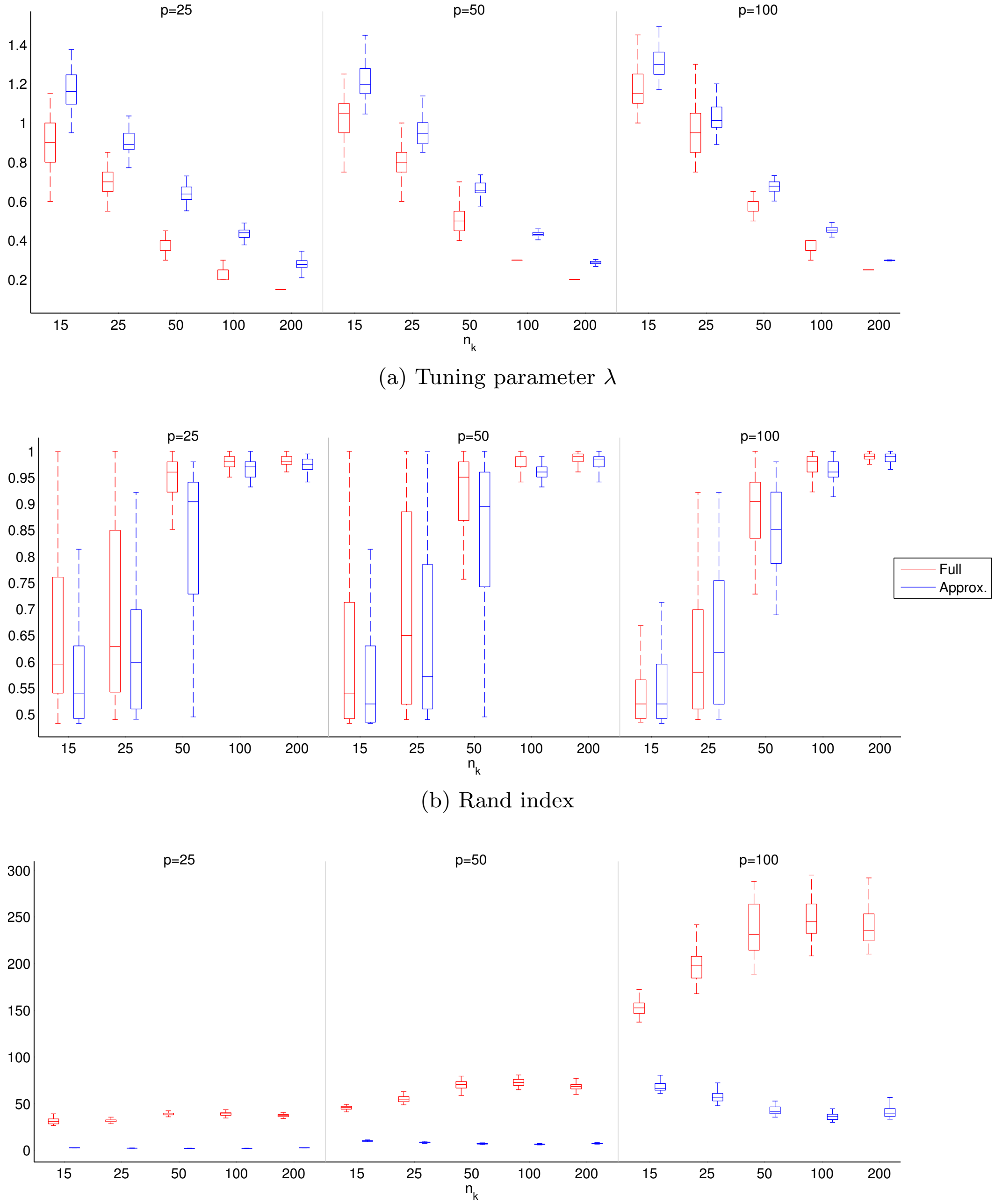


(a) p=25

(b) p=50

(c) p=100

Figure 5.4: **Simulated data; precision matrix estimation.** Elementwise ℓ_1 matrix norm shown for the methods and regimes in Table 5.1 as a function of per-cluster sample size n_k and at data dimensions $p = 25, 50, 100$ (smaller values indicate better agreement between true and inferred precision matrices). K-means clustering was followed by ℓ_1-penalised estimation of Gaussian graphical model structure with penalty parameter set by train/test ('KM+T') or BIC ('KM+B'). Inset: zoomed out version of main plot. (Mean matrix norm, normalised by p, over 50 simulated datasets per (p, n_k) regime, error bars show standard errors; non-penalised approach (NP) could not be used for $n_k \leq p$ due to invalid covariance estimates.)



(a) Tuning parameter λ

(b) Rand index

(c) Computation time (seconds)

Figure 5.5: **Simulated data; heuristic approach for tuning parameter selection.** (a) Boxplots over the tuning parameters selected by the heuristic method (see text for details) under regime B1 (mixture model with $\gamma = 1$ and BIC) are shown (blue boxes), together with the corresponding values obtained with the full, non-approximate approach (red boxes). (b) Resulting Rand indices and (c) computational time required to set the parameter are also shown. (All results are over 50 simulated datasets for each (p, n_k) regime).

Method/Regime	**Rand index**	**Tuning parameter**		
		λ	$\tilde{\lambda}_1/\lambda_1$	$\tilde{\lambda}_2/\lambda_2$
CV0	0.49 (0.01)	0.10 (0.00)	0.25 (0.03)	2.29 (0.26)
B0	0.49 (0.00)	0.23 (0.03)	0.25 (0.03)	2.34 (0.28)
CV1	0.59 (0.08)	0.15 (0.02)	-	-
B1	0.94 (0.02)	0.31 (0.02)	-	-
CVh	0.54 (0.08)	0.14 (0.03)	-	-
Bh	0.77 (0.15)	0.34 (0.04)	-	-
Ah	0.50 (0.03)	-	1.97 (0.15)	5.01 (1.33)
KM	0.61 (0.00)	-	-	-

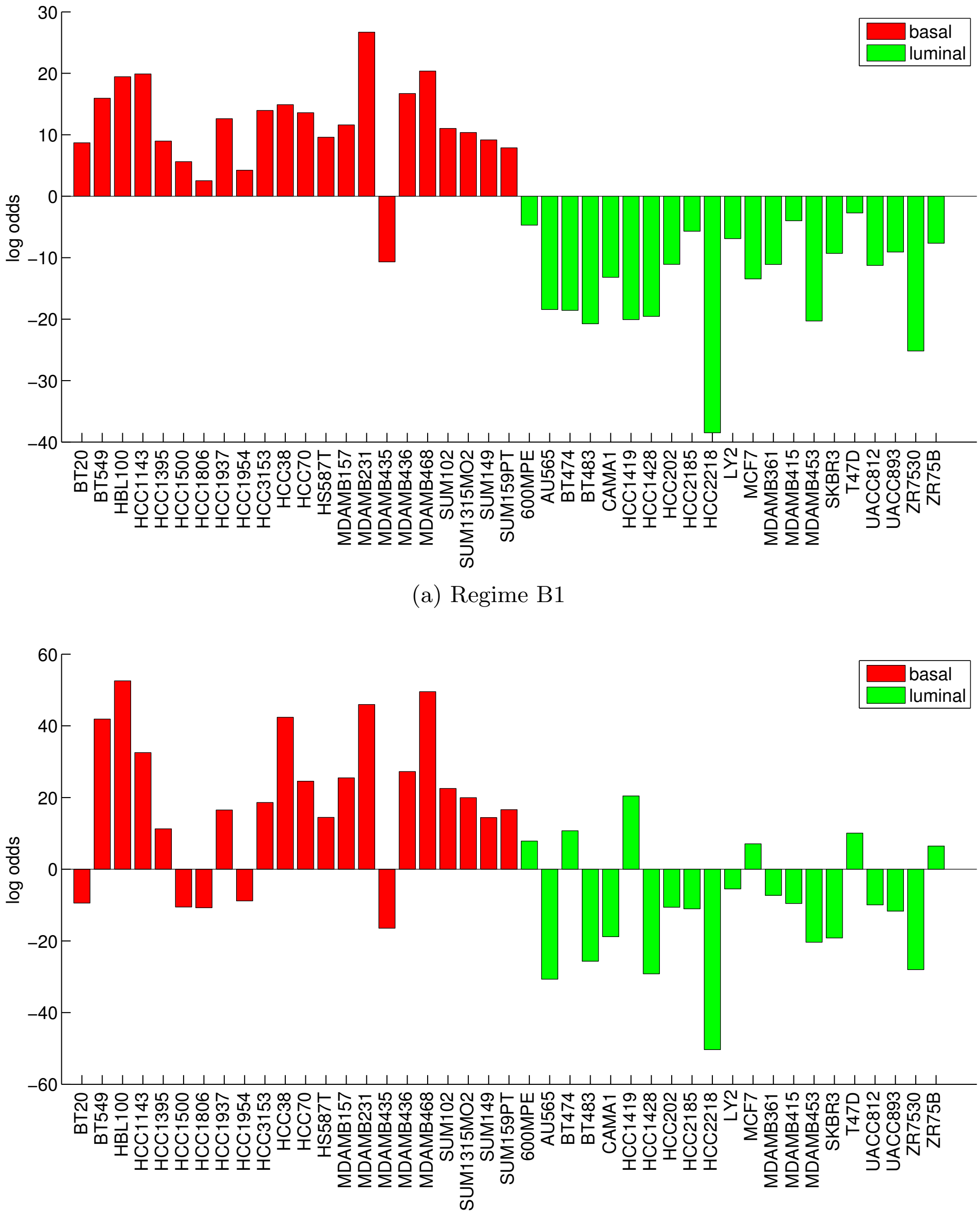
Table 5.2: **Breast cancer data; clustering results.** Rand indices with respect to an independent biological classification of the samples (see text for details) are shown for the methods and regimes in Table 5.1 (absent an independent test dataset, 6-fold cross-validation was used instead of train/test; regimes T0,T1,Th are therefore renamed CV0,CV1,CVh). Penalty parameter values are also shown. (Mean Rand indices over 10 iterations, each with 100 random initialisations, standard deviations given in parentheses; cluster-specific tuning parameter λ_1 and EM algorithm update parameter $\tilde{\lambda}_1$ correspond to the largest cluster.)

using reverse-phase protein arrays (RPPA, see Section 2.2.5; further details of RPPA protocol can be found in Section A.3.1). The $n = 43$ cell lines under study have been shown to reflect much of the biological heterogeneity of primary tumors [Neve *et al.*, 2006] (see Table A.5). In biological applications it is often difficult to objectively assess the correctness of cluster assignments. However, the cell lines we study have been assigned, using independent gene expression data, into two biological categories ("basal" and "luminal") (see Neve *et al.* [2006], classification based on findings due to Perou *et al.* [2000]; Sørlie *et al.* [2001], as described in Chapter 1). This enabled us to compare clustering results to a known classification. To challenge the analysis, three standard biomarkers (ER, PR and HER2) that are known to discriminate between basal and luminal subtypes at the transcriptional level, were not included.

Table 5.2 reports Rand indices with respect to the known classification along with tuning parameter values selected for the methods and regimes given in Table 5.1. Since no independent validation dataset is available, 6-fold cross-validation is used to set the tuning parameter (as described in Section 5.2.4), instead of a single train/test iteration (we denote these regimes CV0, CV1 and CVh). To assess robustness of results, 10 clustering iterations were performed, each with 100 random restarts. The mixture of ℓ_1-penalised Gaussian graphical models with $\gamma = 1$ and BIC (B1) is able to recapitulate the known labels and outperforms the other approaches. Figure 5.6(a) shows, for the highest penalised likelihood result over 10 iterations of regime B1, log odds in favour of basal subtype for each cell line; only one cell line is incorrectly assigned. For comparison, the corresponding result for regime CV1 ($\gamma = 1$ and cross-validation) is shown in Figure 5.6(b). Echoing the simulation results, BIC shows gains over cross-validation for both the mixture model with $\gamma = 1$ (B1/CV1; see Figure 5.6(b)) and non-mixture approach (Bh/CVh), while the mixture model approaches with $\gamma = 0$ (B0/CV0) show poor performance, with one cluster again being penalised substantially more than the other. K-means does not perform well. A mixture model with ℓ_1-penalised estimation but with a shared covariance matrix (i.e. identical network structure) for both mixture components also failed to cluster the data successfully (mean Rand indices for BIC and CV methods <0.55). These results suggest that penalised estimation itself is insufficient to discover the underlying subtypes and that taking cluster-specific covariance structure into account is crucial. We note that the conventional unpenalised Gaussian graphical model is not applicable here; due to high dimensionality relative to sample size, it does not yield valid covariance estimates. We note also that the approximate approach for tuning parameter selection with regime B1 performed well with a mean Rand index of 0.93 ± 0.02 and reduced the computation time by 83% relative to the full approach. Figure 5.7 shows the estimated cluster-specific precision matrices (i.e. network structures) resulting from regime B1 (for the clustering shown in Figure 5.6(a)).

5.4 Discussion

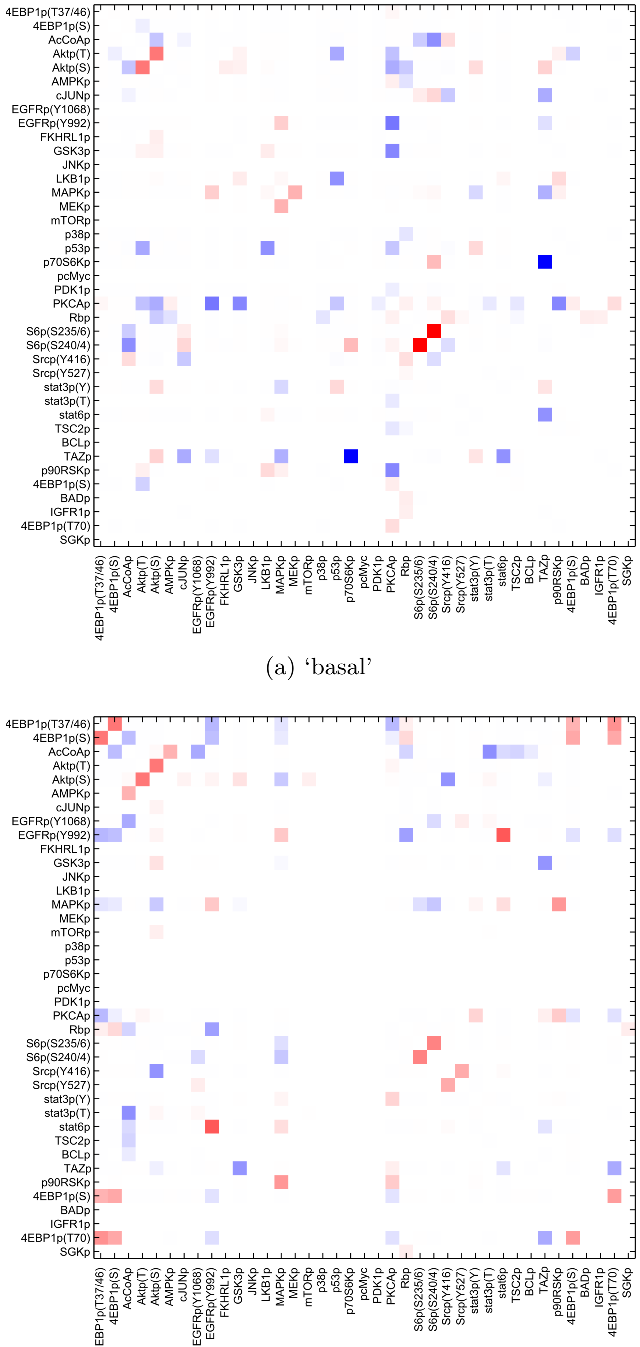
We presented a study of model-based clustering with mixtures of ℓ_1-penalised Gaussian graphical models. We found that performance is dependent on choice of penalty term and method for setting the tuning parameter. Along with the standard ℓ_1 penalty ($\gamma = 0$ in (5.5)) we considered an alternative penalty term, following recent work in penalised finite mixture of regression models [Khalili and Chen, 2007; Städler *et al.*, 2010], that is dependent on the mixing proportions π_k ($\gamma = 1$ in (5.5)). From our simulation study and application to breast cancer data, we draw some broad conclusions and recommendations, as follows. The combination of the $\gamma = 1$ penalty term (incorporating mixing proportions), together with the BIC criterion for selecting the tuning parameter (regime B1), appears to provide the most accurate clustering and estimation of graphical model structure. The only exception is in settings where both dimensionality and sample size are small; here, the smaller tuning parameter values selected by train/test (or cross-validation) provide superior results (regimes T1/CV1). For estimation of the precision matrix itself (as opposed to estimation of sparsity structure only), we again recommend the penalty term with $\gamma = 1$ and find that the less sparse estimates provided by train/test (or cross-validation) provide slight gains over BIC, except where dimensionality is large



(a) Regime B1

(b) Regime CV1

Figure 5.6: **Breast cancer data; clustering result for (a) regime B1 (mixture model with $\gamma = 1$ and BIC) and (b) regime CV1 (mixture model with $\gamma = 1$ and cross-validation).** Red and green indicate cell lines independently classified as basal and luminal respectively (see text for details); log odds in favour of basal subtype are shown for each cell line. (Results shown for highest penalised likelihood obtained over 10 iterations.)



(a) 'basal'

(b) 'luminal'

Figure 5.7: **Breast cancer data; cluster-specific precision matrices (networks) for the clustering shown in Figure 5.6(a) (obtained using regime B1).** (a) Precision matrix for the 'basal' cluster (cell lines with positive log odds in Figure 5.6(a)). (b) Precision matrix for the 'luminal' cluster (cell lines with negative log odds in Figure 5.6(a)). (Red and blue indicate negative and positive values respectively.)

relative to sample size.

The deleterious effect of the standard ℓ_1 penalty term ($\gamma = 0$), at all but the largest sample sizes, is intriguing. As described above, this is due to the fact that the standard penalty term leads to cluster-specific penalties in the EM algorithm update for the precision matrices. Indeed, we observed similar results when setting cluster-specific penalties analytically in a non-mixture model setting (regime Ah). These cluster-specific penalties are inversely proportional to the mixing proportions π_k: in itself this behavior seems intuitively appealing since clusters with small effective sample sizes are then more heavily regularised. However, we observed that a substantially higher penalty is applied to one cluster over the other, indicating that samples were mostly being assigned to the same cluster. This is likely due to the 'unpopular' cluster having a poor precision matrix estimate due to a large penalty. We note that this behavior is not due to (lack of) EM algorithm convergence; the penalised likelihood scores from these incorrect clusterings were higher than those obtained using the true cluster labels.

The related non-mixture model approach proposed by Mukherjee and Hill [2011] also performed well in our studies, but clustering results (both from simulated and real data) indicate that a mixture model with EM algorithm (and $\gamma = 1$ in the penalty term) offers more robust results.

Analysis of breast cancer proteomic data was able to successfully recapitulate an established biological classification of the samples. It is important to note that the basal/luminal breast cancer classification is based on differences in gene expression profiles [Perou *et al.*, 2000; Sørlie *et al.*, 2001]. It remains an open question as to whether such cancer subtypes differ substantively with respect to signalling network structure. Our results suggest that basal and luminal breast cancer subtypes do indeed display such heterogeneity; network-based clustering can recover the subtype classification from phosphoproteomic data alone and, moreover, methods that did not model cluster-specific graphical model structure failed to do so.

While our application focussed on protein signalling networks and cancer subtypes, the approaches we discuss can be applied in other settings where unknown clusters may differ with respect to underlying network structures; for example, gene regulatory networks, social networks or image classification. We also note that while for simplicity and tractability we focussed on the $K = 2$ clusters case, the methods we discuss are immediately applicable to the general K-cluster case. Moreover, since the approach we propose is model-based, established approaches for model selection in clustering, including information criteria, train/test and cross-validation, can be readily employed to select K.

Our results demonstrate the necessity of some form of regularisation to enable the use of Gaussian graphical models for clustering in settings of moderate-to-high dimensionality; indeed, we see clear benefits of penalisation already in the $p = 25$ case. The ℓ_1-penalty is an attractive choice since it encourages sparsity at the level of graphical model structure, and estimation with the graphical lasso algorithm [Friedman *et al.*, 2008] is particularly efficient, which is important in the clustering setting, where multiple iterations are required. Alternatives include shrinkage estimators [Schäfer and Strimmer, 2005b] and Bayesian approaches [Dobra *et al.*, 2004; Jones *et al.*, 2005]. However, it has been shown that the ℓ_1-penalised precision matrix estimator (5.2) is biased [Lam and Fan, 2009]. Alternative penalties have been proposed in a regression setting to ameliorate this issue; for example, the non-concave SCAD penalty [Fan and Li, 2001] and adaptive ℓ_1 penalty [Zou, 2006]. These penalties have recently also been applied to sparse precision matrix estimation [Fan *et al.*, 2009]. They are generally computationally more intensive, but it remains an open question whether they improve clustering accuracy relative to the ℓ_1 penalty considered here.

Graphical models based on direct acyclic graphs (DAGs) are frequently used for network inference, especially in biological settings where directionality may be meaningful (see Section 4.1). A natural extension to the ideas discussed here would be to develop a clustering approach based on DAGs rather than undirected models.

There are several recent and attractive extensions to graphical Gaussian model estimation that could be exploited to improve and extend the methods we discuss. For example, the time-varying Gaussian graphical model approach of Zhou *et al.* [2010] could be employed, or prior knowledge of network structure could be taken into account [Anjum *et al.*, 2009]; such information is abundantly available in biological settings, as discussed in Chapter 3. The joint estimation method for Gaussian graphical models proposed by Guo *et al.* [2011] explicitly models partial agreement between network structures that correspond to *a priori* known clusters. Such partial agreement could be incorporated in the current setting where clusters are not known *a priori*.

Chapter 6

Discussion and Outlook

Recent years have seen significant advances in high-throughput protocols that are capable of interrogating many cellular signalling proteins at once. Protein signalling plays an important role in the physiological functioning of cellular processes and dysregulation in signalling can lead to carcinogenesis. Therefore, the analysis of high-throughput proteomic data to investigate context-specific signalling networks and mechanisms is an important goal in molecular biology and oncology. This thesis aimed to facilitate such analyses by exploiting multivariate statistical approaches, rooted in graphical models. In particular, we focussed on structure learning of sparse graphical model structure. Below we give a brief summary of the thesis and then go on to discuss points that are relevant to the thesis as a whole.

In Chapter 3 we described a Bayesian variable selection method for the discovery of subsets of signalling proteins that jointly influence drug response. The Bayesian method allows for the incorporation of ancillary biological information, such as signalling pathway and network structures, via prior distributions. Prior information was automatically weighted relative to primary data using an empirical Bayes approach. We developed examples of informative pathway- and network-based priors and applied the approach to synthetic response data. The results demonstrated that empirical Bayes can aid prior elicitation and, in particular, help guard against mis-specified priors. Moreover, for discovery of the influential predictor subset, the proposed approach performed favourably in comparison to an alternative prior formulation and to penalised regression using the lasso. An application was also made to cancer drug response data, obtaining biologically plausible results. Overall the procedure is computationally efficient and has very few user-set parameters. Moreover, since it eschews MCMC in favour of an exact formulation, there is no inherent Monte Carlo error.

In Chapter 4 we used DBNs to learn the structure of context-specific signalling networks from phosphoproteomic time series data. The approach calculates posterior edge scores via exact Bayesian model averaging, by exploiting a connection between DBN structure learning and variable selection, and by using biochemically motivated sparsity constraints. Existing signalling biology is incorporated via an informative network prior, weighted objectively relative to primary data by empirical Bayes. Again, the overall approach is exact, fast and essentially free of user-set parameters. We performed an empirical investigation, applying the described approach to both simulated data and gene expression data generated from a synthetically constructed network in yeast [Cantone *et al.*, 2009]. The results showed that incorporation of prior knowledge can aid inference, even when a non-trivial proportion of information contained in the prior is erroneous. Moreover, the proposed approach is observed to have favourable performance relative to several other structure learning approaches. An application was made to a specific breast cancer cell line (MDA-MB-468). The inferred network allowed cell-line specific hypotheses to be generated regarding both previously reported and novel links. These links were validated in independent inhibition experiments. Thus, results suggest that the approach can usefully probe signalling network structure in specific contexts.

In Chapter 5 we described a network-based clustering method, combining model-based clustering with ℓ_1-penalised GGM structure learning, to discover cancer subtypes that differ in terms of subtype specific network structure. Estimation of cluster assignments and cluster-specific network structure is performed simultaneously. The described approach builds upon recent work by Zhou *et al.* [2009] and Mukherjee and Hill [2011]. We performed an empirical investigation to compare several specific penalisation regimes, including different forms for the penalisation term and different methods to set the penalisation tuning parameter. Results were shown on both simulated data and high-throughput breast cancer phosphoproteomic data and allowed general recommendations to be made regarding penalisation regime. The application to breast cancer data successfully recapitulated a known subtype classification from phosphoproteomic data alone, even though the classification was originally based on differences in gene expression profile. Moreover, methods that do not take cluster-specific network structure into account fail to recover the subtypes. This suggests that heterogeneity at the transcriptional level is reflected in a substantive way at the signalling network level.

Throughout this thesis we used continuous linear models, following previous work in Bayesian variable selection [Lee *et al.*, 2003; Nott and Green, 2004; Ai-Jun and Xin-Yuan, 2010] and graph structure learning [Grzegorczyk *et al.*, 2008; Friedman *et al.*, 2008; Bender *et al.*, 2010; Rau *et al.*, 2010]. However, discretised data has also previously been used in these settings [Mukherjee *et al.*, 2009; Husmeier, 2003; Sachs *et al.*, 2005; Ellis and Wong, 2008; Guha *et al.*, 2008]. Discrete models are often employed due to their natural ability to model nonlinear interactions. However, data discretisation is usually lossy and, moreover, it can be difficult to determine appropriate thresholds. To take the example of kinase activity in protein signalling, a binarisation threshold might correspond to the level of protein abundance sufficient to trigger kinase activity. However, this level depends on the proteins involved and kinetic parameters that are usually unknown in practice, and in general may differ from the marginal statistics (e.g. median or other percentiles) of the observed data that are often used for discretisation. The number of discretisation levels can be increased to reduce information loss, but this leads to an increase in model complexity. Geier *et al.* [2007] demonstrated on data simulated from non-linear ODE models, that linear Gaussian DBNs offer an improved performance over discrete DBNs. In Chapters 3 and 4, we use continuous linear models, but retain the possibility of capturing nonlinear interplay by including interaction terms.

As discussed in Section 2.3.5.2, inferred links between signalling proteins or between a protein and response of interest can not in general be interpreted as causal. Hidden or latent variables, when taken into account, may explain away the inferred interaction. Further, for BN structure learning, it is not possible to determine the directionality of some links due to the existence of equivalence classes. Several methods have been proposed in the literature for structure learning of molecular networks from interventional data [Cooper and Yoo, 1999; Pe'er *et al.*, 2001; Markowetz *et al.*, 2005; Dojer *et al.*, 2006; Eaton and Murphy, 2007b; Bender *et al.*, 2010; Luo and Zhao, 2011]. Comparative studies by Werhli *et al.* [2006] and Geier *et al.* [2007] have demonstrated the utility of using interventional data and taking the interventions into account in modelling. Interventions can break the symmetry within an equivalence class, thereby aiding elucidation of edge directionality. They can also help improve accuracy of inference by perturbing the state of the cell in specific ways, allowing dynamics to be observed that may be informative for inference, but are not possible to observe from observational data alone. The breast cancer data we consider in Chapter 4 is observational data (only global excitatory perturbations are used to initiate signalling processes). Adapting the proposed DBN inference approach to take account of interventions and applying the method to time series phosphoproteomic data with interventions (such data is becoming increasingly available) is likely to improve accuracy of results.

A vast number of protein species (including transcriptional and post-translational variants) may be involved in protein signalling. At present it is not possible to assay any more than a small fraction of these players. Therefore, data-driven studies of signalling confront a severe missing variable problem. As just discussed, this necessarily limits the causal or mechanistic interpretation of results. For example, an inferred edge from one protein to another, or from a protein to a response of interest, may operate via one or more unmeasured intermediates. Statistical models that permit the inclusion of unobserved, latent variables may help, but network inference with latent variables remains challenging [Knowles and Ghahramani, 2011]. We note that the hidden variable issue applies also to validation by inhibition. Thus, the novel links reported in Chapter 4 will require further work, including biochemistry and dynamical modelling, to better understand the mechanisms involved.

As discussed in Chapter 1 and Section 4.1, the statistical models employed in this thesis are not a realistic representation of the underlying biochemical mechanisms. However, the models are analytically tractable which in turn allows large spaces of network structures to be explored. ODE models offer a powerful and more realistic modelling framework, but typically the network structure is assumed to be known. In principle, statistical network inference can be explicitly based on biochemically plausible ODE models, but the model is then no longer solvable analytically. For example, for DBN structure learning in Chapter 4, the marginal likelihood would no longer be available in closed form and would have to be evaluated using approximate methods. Therefore, due to severe computational constraints such approaches are currently limited to investigating only a handful of hypothesised networks [Xu *et al.*, 2010] and not the large number of possible networks we consider. As computational processing power continues to advance and ever larger datasets become available, biologically realistic models are likely to become more widely-used within network structure learning approaches. To exploit these advances fully, new statistical and computational methods will also be needed.

Results obtained using the methods described in this thesis could be combined with those obtained from alternative approaches, with the aim to improve the overall reliability of results. This idea has been previously proposed in the context of clustering [Swift *et al.*, 2004] ('consensus clustering' method) and in the context of signalling network structure learning [Prill *et al.*, 2011] ('crowdsourcing network inference').

We note that, while we apply our methods in the setting of cancer protein signalling, they are also applicable to other molecular data types (e.g. gene microarray data) and to biological contexts other than cancer. However, since regulation occurs in the cell at many levels, including the genome, transcriptome, proteome and metabolome, to obtain a more complete understanding of cellular processes it will be necessary to include as many of these levels as possible in inference. Integration of multiple datasets is a challenging research area that is receiving a growing amount of attention. For example, methods have been proposed in the literature that perform network structure learning from gene expression data, and incorporate other data types, such as genome-wide binding data, within a Bayesian prior [Jensen *et al.*, 2007; Yeung *et al.*, 2011]. A recent example in the context of cancer is a clustering method for discovery of subtypes based on both transcriptional data and genetic copy number data [Yuan *et al.*, 2011].

On the experimental side, advances continue to be made in technology that allow larger or higher quality datasets to be produced. Significant advances in recent years include RNA-sequencing for transcriptional data [Wang *et al.*, 2009b] and mass cytometry for measuring cellular components, such as phosphoproteins, at the single-cell level [Bendall *et al.*, 2011]. Mass cytometry is similar to flow cytomety (see Section 2.2.4) but does not suffer from the issue with spectral overlap, allowing more proteins to be measured simultaneously. An important area for future work is that of experimental design. For example, phosphoproteomic time series data typically consist of time points unevenly sampled through time, covering several hours, as is the case for the breast cancer data used in Chapter 4. However, further work is needed to guide and optimise choice of time points. Signalling activations can be transient and can happen in a short space of time after stimulation of the cell. Hence, rapidly sampled time points immediately after stimulation may improve results. Another, important question concerns interventions. As discussed above, interventions can improve results, but given a finite amount of time and resources, it is not clear how to apportion efforts between proteins, time points, interventions and biological samples. In addition, it is also unclear which interventions should be performed to obtain data that is maximally informative with respect to network topology. The optimal design of interventional experiments is an active area of research [see e.g. He and Geng, 2008].

It is clear that there are many limitations associated with the methods and data used in this thesis. Indeed, in Section 4.4 we gave several reasons why statistical data-driven stucture learning approaches can not be expected to produce the true, underlying structure, and should only be used as a hypothesis generating tool. We further discussed some of these points above in this Section. However, as our results have shown, these methods can still usefully interrogate proteomic data to probe important questions in cancer signalling, with possible implications for personalised cancer therapy.

The translation of research findings into the clinic is the ultimate challenge for personalised medicine. To date, success has been very limited. An improved understanding is needed of the high levels of molecular heterogeneity observed in cancer, a key challenge being the identification of robust biomarkers or signatures that can accurately predict patient response to particular therapeutic agents [Weigelt *et al.*, 2012]. Hence, there is still much work to be done before personalised cancer therapy is truly a reality. Multivariate statistical approaches that are rooted in sparse graphical models are likely to be able to play an important role in this work.

Appendix A

Experimental data

A.1 Chapter 3 - Proteomics and drug response data

A.1.1 Proteomics

Cell lysates: For preparation of protein lysates cells were grown to 60-80% confluency in appropriate media [Neve *et al.*, 2006]. Cultures were placed on ice, media aspirated and washed in ice cold PBS containing 1mM phenylmethylsulfonyl fluoride (PMSF) and then with a buffer containing 50mM HEPES (pH7.5), 150mM NaCl, 25mM b-glycerophsphate, 25mM NaF, 5mM EGTA, 1mM EDTA, 15mM pyrophosphate, 2mM sodium orthovanadate, 10mM sodium molybdate, leupeptin (10mg/ml), aprotinin (10mg/ml) and 1mM PMSF. Cells were extracted in the same buffer containing 1%Nonidet-P40. Lysates were then clarified by centrifugation and frozen at $-80\,^{\circ}$C. Protein concentrations were determined using the Bio-Rad protein assay kit. Phopshoproteome analysis was performed at Kinexus (http://www.kinexus.ca/) on their KinetWorks$^{\mathrm{TM}}$ platform.

1	PDK1 (S244)	27	MNK1 (T209/T214)
2	S6 (S235)	28	MEK1 (S297)
3	ACC (S80)	29	MEK1 (T291)
4	Adducin-α (S726)	30	MEK1 (T385)
5	Adducin-γ (S693)	31	MEK1/2 (S217/S221)
6	BAD (S99)	32	MEK2 (T394)
7	BRCA1 (S1497)	33	MAPKAPK2 (T222)
8	CREB1 (S133)	34	MYPT1 (T696)
9	CDK1/2 (T14/Y15)	35	NR1 (S896)
10	CDK1/2 (T161/Y160)	36	p70S6K-α (T389)
11	ErbB2 (Y1248)	37	p70S6K-α (T421/S424)
12	eIF2B-ϵ (S540)	38	p85S6K-β (T444/S447)
13	Erk1 (T202/Y204)	39	PRK2 (T816)
14	FAK (S722)	40	AKT1 (S473)
15	FAK (S910)	41	AKT1 (T308) (S729)
16	FAK (Y397)	42	PKC-α (S657)
17	FAK (Y576)	43	PKC-$\alpha/\beta2$ (T638/T641)
18	GSK3-α (S21)	44	PKC-ϵ
19	GSK3-β (S9)	45	PKC-ζ/ι (T410/T403)
20	HSP27 (S15)	46	PP1/Ca (T320)
21	Histone H3 (S10)	47	Raf1 (S259)
22	Histone H3 (S28)	48	RB (S259)
23	IR (Y999)	49	RB (S780)
24	JNK (T183/Y185)	50	RB (S807/S811)
25	mTOR (S2448)	51	SHC1 (Y349/Y350)
26	MEK3/6 (S189/S207)	52	SRC (Y529)

Table A.1: **Phosphoproteins involved in our cancer drug response study.**

1	600MPE	19	MCF7
2	AU-565	20	MDAMB134VII
3	BT-20	21	MDMB157
4	BT-474	22	MDAMB175
5	BT-483	23	MDAMB231
6	BT-549	24	MDAMB361
7	CAMA-1	25	MDAMB436
8	HCC1143	26	SK-BR-3
9	HCC1187	27	SUM149PT
10	HCC1500	28	SUM159PT
11	HCC1569	29	SUM185PT
12	HCC202	30	SUM225CWN
13	HCC38	31	SUM52PE
14	HCC70	32	T-47D
15	Hs587T	33	UACC-893
16	LY2	34	ZR-75-1
17	MCF10A	35	ZR-75-8
18	MCF12A		

Table A.2: **Breast cancer cell lines involved in our cancer drug response study.**

A.2 Chapter 4 - Proteomics and validation experiments

Short Name (as used in main text)	**Antibody Name**	**Company**	**Catalogue**
AKTp(S)	AKT pS473	Cell Signaling	9271
AKTp(T)	AKT pT308	Cell Signaling	9275
AMPKp	AMPK pT172	Cell Signaling	2535
cJUNp	c-Jun pS73	Cell Signaling	9164
EGFRp	EGFR pY1173	Millipore	05-483
GSK3p	GSK3 pS21/9	Cell Signaling	9331
JNKp	JNK pT183 Y185	Cell Signaling	9251
LKB1p	LKB1 pS428	Cell Signaling	3051
MAPKp	MAPK pT202 Y204	Cell Signaling	9101
MEK1/2p	MEK 1/2 pS217	Cell Signaling	9121
mTORp	mTOR pS2448	Cell Signaling	2971
p38p	p38 pT180	Cell Signaling	9211
p70S6Kp	p70S6K pT389	Cell Signaling	9205
p90RSKp	p90RSK pT359	Cell Signaling	9344
PDK1p	PDK1 pS241	Cell Signaling	3061
PI3K	PI3K	Epitomics	1683
STAT3p(T)	STAT3 pT727	Cell Signaling	9134
STAT3p(Y)	STAT3 pY705	Cell Signaling	9131
STAT5p	STAT5 pY964	Cell Signaling	9351
TSC2p	TSC2 pT1462	Cell Signaling	3611

Table A.3: **Validated primary antibodies used in the MDA-MB-468 cell line study.**

A.2.1 Reverse phase protein arrays

Reverse phase protein array (RPPA) assays were carried out as previously described [Tibes *et al.*, 2006; Hennessy *et al.*, 2010]. Breast cancer cell line MDA-MB-468 was cultured in its optimal media to a logarithm growth phase. Time courses were carried out at eight time points (5, 15, 30, 60, 90, 120, 180, 240 minutes) in triplicate, under four growth conditions (0, 5, 10, 20ng/ml EGF). Cellular proteins were denatured by 1% SDS (with beta-mercaptoethanol) and diluted in five 2-fold serial dilutions in dilution buffer (lysis buffer containing 1% SDS). Serial diluted lysates were arrayed on nitrocellulose-coated FAST slides (Whatman, Inc) by Aushon 2470 Arrayer (Aushon BioSystems). Total 5808 array spots were arranged on each slide including the spots corresponding to positive and negative controls prepared from mixed cell lysates or dilution buffer, respectively.

Each slide was probed with a validated primary antibody (Table A.3) plus a biotin-conjugated secondary antibody. Only antibodies with a Pearson correlation coefficient between RPPA and western blotting of greater than 0.7 were used in reverse phase protein array study. Antibodies with a single or dominant band on western blotting were further assessed by direct comparison to RPPA using cell lines with differential protein expression or modulated with ligands/inhibitors or siRNA for phospho- or structural proteins, respectively. Extensive validation data for the antibodies used are presented in Hennessy *et al.* [2010].

The signal obtained was amplified using a DakoCytomation-catalysed system (Dako) and visualised by DAB colorimetric reaction. The slides were scanned, analysed, and quantified using a customised-software Microvigene (VigeneTech Inc.) to generate spot intensity.

Each dilution curve was fitted with a logistic model ("Supercurve Fitting" developed by the Department of Bioinfomatics and Computational Biology in MD Anderson Cancer Center,
`http://bioinformatics.mdanderson.org/OOMPA`). This fits a single curve using all the samples (i.e., dilution series) on a slide with the signal intensity as the response variable and the dilution steps are independent variable. The fitted curve is plotted with the signal intensities both observed and fitted - on the y-axis and the log2-concentration of proteins on the x-axis for diagnostic purposes. The protein concentrations of each set of slides were then normalised by median polish, which was corrected across samples by the linear expression values using the median expression levels of all antibody experiments to calculate a loading correction factor for each sample. Logged averages over RPPA triplicates (Figure A.1) were used for all network analyses.

A.2.2 Validation experiments

A.2.2.1 RPPA (Figure 4.9a,b)

The breast cancer cell line MDA-MB-468 was seeded at 90% confluency in 96-well plates at a density of 10,000 cells per well with 8% FBS-RPMI medium and allowed to attach. Cells were depleted of serum for 12 hours prior to treatment with the MEK inhibitor GSK2B (GlaxoSmithKline Inc.) at 10uM or AKT inhibitor GSK690693B (GlaxoSmithKline Inc.) at 10uM for 4 hours in each case. Cells were stimulated (EGF 20 ng/mL) prior to lysis in RPPA lysis buffer. Phosphoprotein profiling was carried out 0,5,15,30,60,90,120,180 minutes after EGF stimulus using RPPA as described above.

A.2.2.2 Western blots (Figure 4.9c)

The breast cancer cell line MDA-MB-468 was seeded at 90% confluency in 8% FBS-RPMI medium and allowed to attach. Cells were depleted of serum for 12 hours prior to treatment with the MEK inhibitor UO126 (EMD Chemicals Inc., Gibbstown, NJ) at 5uM for 4 hours. Cells were stimulated (EGF 20 ng/mL) for 15 minutes prior to lysis in RPPA lysis buffer.

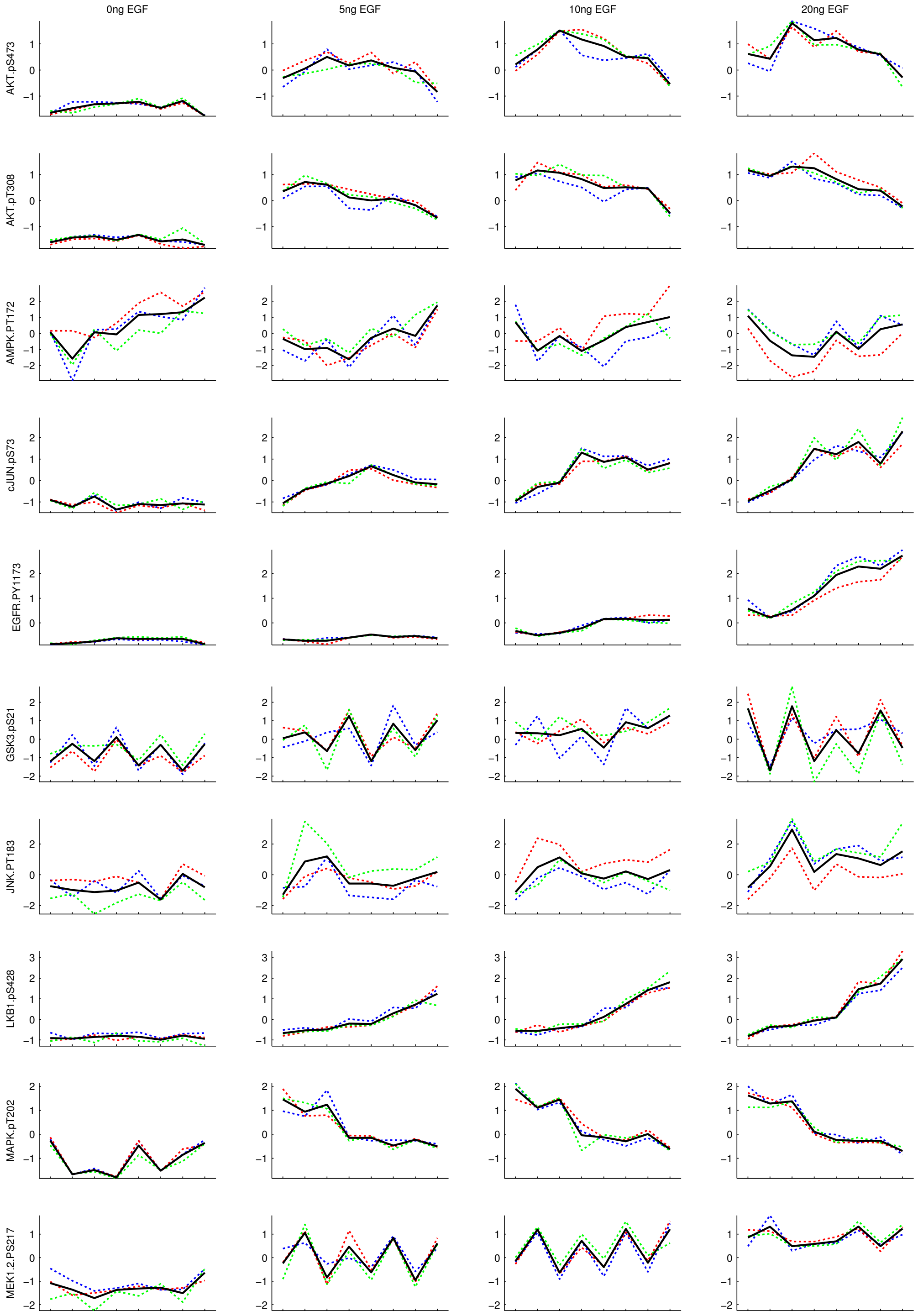


Figure A.1: **Time courses for MDA-MB-468.** (Coloured lines are raw triplicates; black lines are averages. (Time courses are standardised to have zero mean, unit variance across all conditions for each protein).

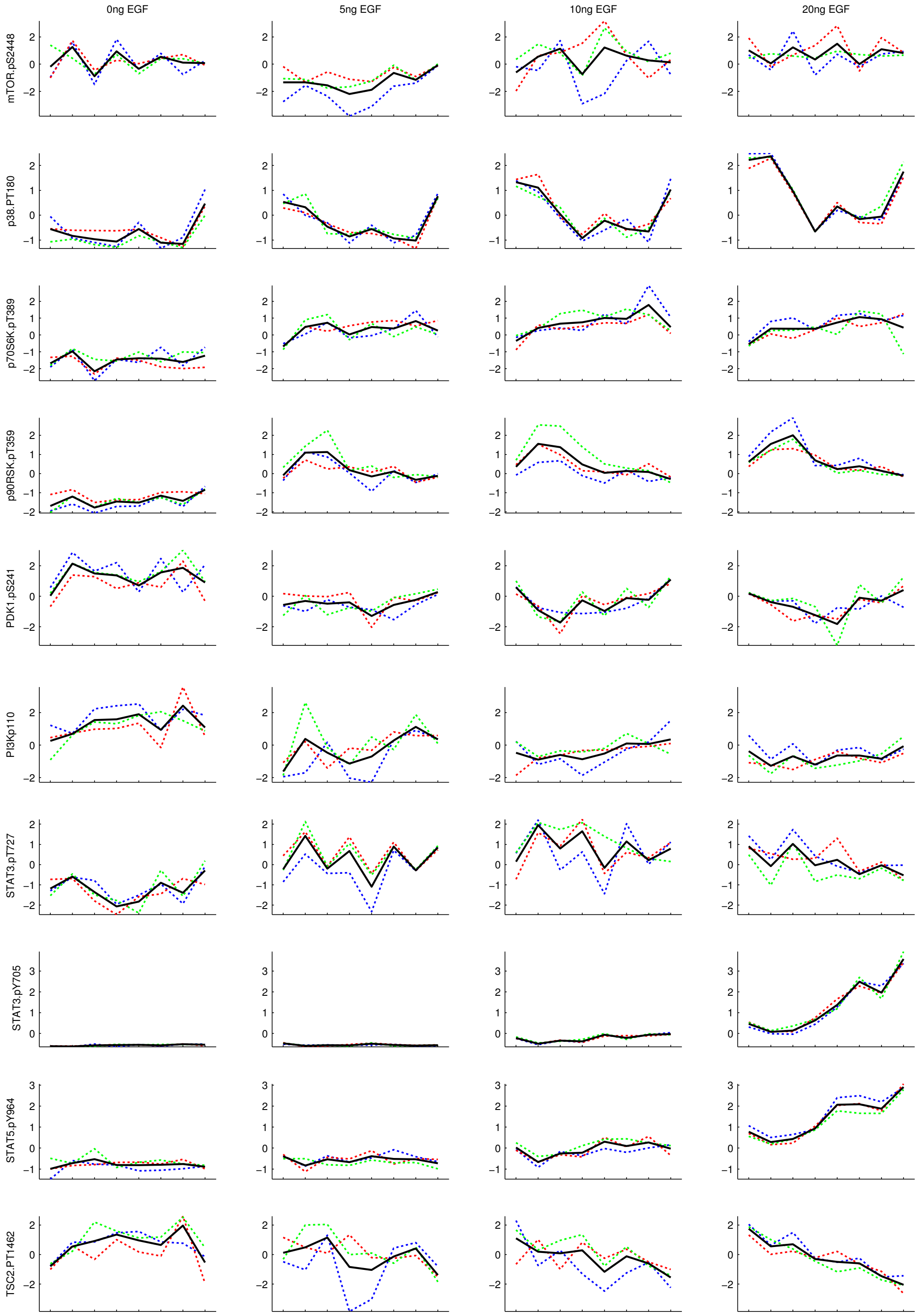


Figure A.1: **Time courses for MDA-MB-468.**

A.3 Chapter 5 - Proteomic data

Antibody Name	Source	Catalog Number
Phospho-4EBP1 (T37/46)	Cell Signaling	9459
Phospho-4EBP1 (Ser65)	Cell Signaling	9451
Phospho-Acetyl-CoA Carboxylase (Ser79)	Cell Signaling	3661
Phospho-Akt (T308)	Cell Signaling	9275
Phospho-Akt (S473)	Cell Signaling	9271
Phospho-AMPK (T172)	Cell Signaling	2535
Phospho-cJUN (S73)	Cell Signaling	9164
Phospho-EGFR (Y1068)	Cell Signaling	2234
Phospho-EGFR (Y992)	Cell Signaling	2235
Phospho-FKHRL1 (S318/321)	Cell Signaling	9465
Phospho-GSK3 alpha/belta (S21/9)	Cell Signaling	9331
Phospho-JNK (T183/Y185)	Cell Signaling	4671
Phospho-LKB1 (S428)	Cell Signaling	3051
Phospho-MAPK (p44/42 ERK1/2)(T202/Y204)	Cell Signaling	9101
Phospho-MEK1/2 (S217/221)	Cell Signaling	9121
Phospho-mTOR (S2448)	Cell Signaling	2971
Phospho-p38 MAPK (T180/Y182)	Cell Signaling	9211
Phospho-p53 (S15)	Cell Signaling	9284
Phospho-70S6K (T389)	Cell Signaling	9205
Phospho-c-Myc (Thr58/Ser62)	Cell Signaling	9401
Phospho-PDK1 (S241)	Cell Signaling	3061
Phospho-PKC alpha (S657)	Millipore	06-822
Phospho-Rb (S807/811)	Cell Signaling	9308
Phospho-S6 Ribosomal Protein (S235/236)	Cell Signaling	2211
Phospho-S6 Ribosomal Protein (S240/244)	Cell Signaling	2215
Phospho-Src (Y416)	Cell Signaling	2113
Phospho-Src (Y527)	Cell Signaling	2105
Phospho-Stat3 (Y705)	Cell Signaling	9131
Phospho-Stat3 (T727)	Cell Signaling	9134
Phospho-Stat6 (Y641)	Cell Signaling	9361
Phospho-TSC2 (T1462)	Cell Signaling	3617
Phospho-BCL (T70)	Cell Signaling	2871
Phospho-TAZ (S79)	Santa Cruz	sc-17610
Phospho-p90RSK (T359/S363)	Cell Signaling	9344
Phospho-4EBP1 (S65)	Cell Signaling	9456
Phospho-BAD (pS112)	Cell Signaling	9291
Phospho-IGFR1 (Y1135)	Cell Signaling	3024
Phospho-4EBP1 (T70)	Cell Signaling	9455
Phospho-SGK (S78)	Cell Signaling	3271

Table A.4: **Phosphoproteins analysed in our breast cancer study.**

A.3.1 Proteomics

Reverse phase protein array (RPPA) assays were carried out as previously described [Tibes *et al.*, 2006; Hennessy *et al.*, 2010]. See also Section A.2.1 above.

Cell line	Subtype
BT20	basal
BT549	basal
HBL100	basal
HCC1143	basal
HCC1395	basal
HCC1500	basal
HCC1806	basal
HCC1937	basal
HCC1954	basal
HCC3153	basal
HCC38	basal
HCC70	basal
HS587T	basal
MDAMB157	basal
MDAMB231	basal
MDAMB435	basal
MDAMB436	basal
MDAMB468	basal
SUM102	basal
SUM1315MO2	basal
SUM149	basal
SUM159PT	basal
600MPE	luminal
AU565	luminal
BT474	luminal
BT483	luminal
CAMA1	luminal
HCC1419	luminal
HCC1428	luminal
HCC202	luminal
HCC2185	luminal
HCC2218	luminal
LY2	luminal
MCF7	luminal
MDAMB361	luminal
MDAMB415	luminal
MDAMB453	luminal
SKBR3	luminal
T47D	luminal
UACC812	luminal
UACC893	luminal
ZR7530	luminal
ZR75B	luminal

Table A.5: **Cell lines analysed in our breast cancer study.**

Bibliography

Ai-Jun, Y. and Xin-Yuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **26**:215–222.

Äijö, T. and Lähdesmäki, H. (2009). Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, **25**:2937–2944.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**:716–723.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell.* Garland Science, New York, fifth edition.

Alizadeh, A.A., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**:503–511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**:6745–6750.

Altay, G., Asim, M., Markowetz, F., and Neal, D. (2011). Differential C3NET reveals disease networks of direct physical interactions. *BMC Bioinf.*, **12**:296.

Altay, G. and Emmert-Streib, F. (2010). Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, **26**:1738–1744.

Alvarez, R.H., Valero, V., and Hortobagyi, G.N. (2010). Emerging targeted therapies for breast cancer. *J. Clin. Oncol.*, **28**:3366–3379.

Anjum, S., Doucet, A., and Holmes, C.C. (2009). A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, **25**:2929–2936.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**:412–424.

Banerjee, O., El Ghaoui, L., and D'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**:485–516.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**:art. 78.

Bansal, M., Gatta, G.D., and di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**:815–822.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**:382–390.

Beard, D.A. and Qian, H. (2008). *Chemical Biophysics: Quantitative Analysis of Cellular Systems.* Cambridge University Press, Cambridge, UK.

Bendall, S.C., *et al.* (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**:687–696.

Bender, C., Henjes, F., Fröhlich, H., Wiemann, S., Korf, U., and Beißbarth, T. (2010). Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics*, **26**:i596–i602.

Bernard, A. and Hartemink, A.J. (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. In *Pac. Symp. Biocomput. 2005*, pp. 459–470. World Scientific, Singapore.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* John Wiley & Sons, New York.

Binder, H. and Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinf.*, **10**:18.

Bolstad, W.M. (2010). *Understanding Computational Bayesian Statistics.* John Wiley & Sons, Hoboken, New Jersey.

Boyd, Z.S., Wu, Q.J., O'Brien, C., Spoerke, J., Savage, H., Fielder, P.J., Amler, L., Yan, Y., and Lackner, M.R. (2008). Proteomic analysis of breast cancer molecular subtypes and biomarkers of response to targeted kinase inhibitors using reverse-phase protein microarrays. *Mol. Cancer Ther.*, **7**:3695–3706.

Brazil, D.P. and Hemmings, B.A. (2001). Ten years of protein kinase B signalling: a hard Akt to follow. *Trends Biochem. Sci.*, **26**:657–664.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, **24**:123–140.

Brown, P.J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. B*, **64**:519–536.

Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 52–60.

Burgering, B.M.T. and Coffer, P.J. (2002). Protein kinase B (c-Akt) in phosphatidylinositol-3-OH kinase signal transduction. *Nature*, **376**:599–602.

Burnette, W.N. (1981). "Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal. Biochem.*, **112**:195 – 203.

Butte, A. and Kohane, I. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **4**:18–29.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, **106**:594–607.

Cancer Research UK (2012). CancerStats - Cancer Statistics for the UK. http://info.cancerresearchuk.org/cancerstats. [Online; accessed 03-01-2012].

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M.P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**:172–181.

Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis.* Chapman & Hall, 3rd edition.

Chaussepied, M. and Ginsberg, D. (2004). Transcriptional regulation of AKT activation by E2F. *Mol. Cell*, **16**:831–837.

Chen, W.W., Schoeberl, B., Jasper, P.J., Niepel, M., Nielsen, U.B., Lauffenburger, D.A., and Sorger, P.K. (2009). Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.*, **5**:239.

Chickering, D., Geiger, D., and Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. In *Proceedings of the Fifth International Workshop on Artifical Intelligence and Statistics*, pp. 112–128.

Chickering, D.M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 87–98. Morgan Kaufmann, San Francisco, CA.

Chickering, D.M. (2002). Learning equivalence classes of Bayesian network structures. *J. Mach. Learn. Res.*, **2**:445–498.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Can. J. Stat.*, **24**:17–36.

Chipman, H., George, E.I., McCulloch, R.E., Clyde, M., Foster, D.P., and Stine, R.A. (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes-Monograph Series*, **38**:65–134.

Choudhary, C. and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.*, **11**:427–439.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**:140.

Ciaccio, M.F., Wagner, J.P., Chuu, C.P., Lauffenburger, D.A., and Jones, R.B. (2010). Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nat. Methods*, **7**:148–155.

Citri, A. and Yarden, Y. (2006). EGF–ERBB signalling: Towards the systems level. *Nat. Rev. Mol. Cell. Biol.*, **7**:505–516.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press.

Clyde, M. and George, E.I. (2004). Model uncertainty. *Stat. Sci.*, **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.

Cooper, G.F. and Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 116–125.

Craciun, G. and Pantea, C. (2008). Identifiability of chemical reaction networks. *J. Math. Chem.*, **44**:244–259.

D'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, **30**:56–66.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**:459–466.

de Souto, M., Costa, I., de Araujo, D., Ludermir, T., and Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinf.*, **9**:497.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, **28**:157–175.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**:1–38.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Wiley, London.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivar. Anal.*, **90**:196–212.

Dojer, N., Gambin, A., Mizera, A., Wilczynski, B., and Tiuryn, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinf.*, **7**:249.

Drton, M. and Perlman, M.D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, **91**:591–602.

Eaton, D. and Murphy, K. (2007a). Bayesian structure learning using dynamic programming and MCMC. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 101–108.

Eaton, D. and Murphy, K. (2007b). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pp. 107–114.

EBCTCG (Early Breast Cancer Trialists' Collaborative Group) (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, **351**:1451–1467.

Edwards, D. (2000). *Introduction to Graphical Modelling.* Springer, New York.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.*, **32**:407–499.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC press, Boca Raton, FL.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**:14863–14868.

Ellis, B. and Wong, W.H. (2008). Learning causal Bayesian network structures from experimental data. *J. Am. Stat. Assoc.*, **103**:778–789.

Emde, A., Köstler, W.J., and Yarden, Y. (2011). Therapeutic strategies and mechanisms of tumorigenesis of HER2-overexpressing breast cancer. *Crit. Rev. Oncol. Hematol., In Press.*

Engvall, E. and Perlmann, P. (1971). Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry*, **8**:871 – 874.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**:e8.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.*, **3**:521–541.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**:1348–1360.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.*, **41**:578–588.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**:611–631.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**:302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**:432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.*, **33**:1–22.

Friedman, N. (1998). The Bayesian structural EM algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 129–138.

Friedman, N. and Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach. Learn.*, **50**:95–125.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comp. Bio.*, **7**:601–620.

Friedman, N., Murphy, K., and Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 139–147. Morgan Kaufmann, San Francisco, CA.

Friedman, N., Nachman, I., and Pe'er, D. (1999). Learning Bayesian network structures from massive datasets: The sparse candidate algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 196–205.

Froehlich, H., Fellmann, M., Sueltmann, H., Poustka, A., and Beissbarth, T. (2007). Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinf.*, **8**:386.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**:102–105.

Geier, F., Timmer, J., and Fleck, C. (2007). Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst. Biol.*, **1**:11.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 235–243. Morgan Kaufmann, San Francisco, CA.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, Florida, second edition.

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. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**:881–889.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Stat. Sin.*, **7**:339–373.

Giudici, P. and Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Mach. Learn.*, **50**:127–158.

Golub, T.R., *et al.* (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**:531–537.

Goncharova, E.A., *et al.* (2002). Tuberin regulates p70 S6 kinase activation and ribosomal protein S6 phosphorylation. a role for the TSC2 tumor suppressor gene in pulmonary lymphangioleiomyomatosis (LAM). *J. Biol. Chem.*, **277**:30958–30967.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, **4**:117.

Grzegorczyk, M. and Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Mach. Learn.*, **71**:265–305.

Grzegorczyk, M. and Husmeier, D. (2011a). Improvements in the reconstruction of time-varying gene regulatory networks: Dynamic programming and regularization by information sharing among genes. *Bioinformatics*, **27**:693–699.

Grzegorczyk, M. and Husmeier, D. (2011b). Non-homogeneous dynamic Bayesian networks for continuous data. *Mach. Learn.*, **83**:355–419.

Grzegorczyk, M., Husmeier, D., Edwards, K.D., Ghazal, P., and Millar, A.J. (2008). Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics*, **24**:2071–2078.

Guha, U., *et al.* (2008). Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS. *Proc. Natl. Acad. Sci. USA*, **105**:14112–14117.

Guillemot, V., Tenenhaus, A., Le Brusquet, L., and Frouin, V. (2011). Graph constrained discriminant analysis: A new method for the integration of a graph into a classification process. *PLoS ONE*, **6**:e26146.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, **98**:1–15.

Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *J. Am. Stat. Assoc.*, **95**:795–806.

Hanahan, D. and Weinberg, R. (2011). Hallmarks of cancer: The next generation. *Cell*, **144**:646–674.

Hanahan, D. and Weinberg, R.A. (2000). The Hallmarks of Cancer. *Cell*, **100**:57–70.

Hardie, D.G. (2004). The AMP-activated protein kinase pathway - new players upstream and downstream. *J. Cell Sci.*, **117**:5479–5487.

Harsha, H.C. and Pandey, A. (2010). Phosphoproteomics in cancer. *Mol. Oncol.*, **4**:482–495.

Hartemink, A.J. (2005). Reverse engineering gene regulatory networks. *Nat. Biotech.*, **23**:554–555.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, **6**:422–433.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2003). *The Elements of Statistical Learning.* Springer, New York.

He, Y. and Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.*, **9**:2523–2547.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models - a review. *Biosystems*, **96**:86–103.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M.I. Jordan (ed.), *Learning in graphical models*, pp. 301–354. Kluwer Academic, Dordecht, Amsterdam.

Heckerman, D., Geiger, D., and Chickering, D.M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, **20**:197–243.

Heiser, L.M., *et al.* (2011). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. USA (to appear).*

Hennessy, B., *et al.* (2010). A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin. Proteom.*, **6**:129–151.

Herzenberg, L.A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L.A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: A view from Stanford. *Clin. Chem.*, **48**:1819–1827.

Hill, S.M., Neve, R.M., Bayani, N., Kuo, W.L., Ziyad, S., Spellman, P.T., Gray, J.W., and Mukherjee, S. (2012). Integrating biological knowledge into variable selection: An empirical Bayes approach with an application in cancer biology. *BMC Bioinf.*, **13**:94.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Stat. Sci.*, **14**:382–417.

Hu, J., He, X., Baggerly, K.A., Coombes, K.R., Hennessy, B.T., and Mills, G.B. (2007). Non-parametric quantification of protein lysate arrays. *Bioinformatics*, **23**:1986–1994.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**:2271–2282.

Iadevaia, S., Lu, Y., Morales, F.C., Mills, G.B., and Ram, P.T. (2010). Identification of optimal drug combinations targeting cellular networks: Integrating phosphoproteomics and computational network analysis. *Can. Res.*, **70**:6704–6714.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.*, **2**:343–372.

Ideker, T. and Lauffenburger, D. (2003). Building with a scaffold: emerging strategies for high-to low-level cellular modeling. *Trends Biotechnol.*, **21**:255–262.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB'03)*, pp. 104–113. IEEE Computer Society.

Inoki, K., Li, Y., Zhu, T., Wu, J., and Guan, K.L. (2002). TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling. *Nat. Cell. Biol.*, **4**:648–657.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, New York, third edition.

Jensen, C.J., Buch, M.B., Krag, T.O., Hemmings, B.A., Gammeltoft, S., and Frödin, M. (1999). 90-kDa ribosomal S6 kinase is phosphorylated and activated by 3-phosphoinositide-dependent protein kinase-1. *J. Biol. Chem.*, **274**:27168–27176.

Jensen, S.T., Chen, G., and Stoeckert, C.J. (2007). Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Stat.*, **1**:612–633.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *Ann. Stat.*, **35**:1487–1511.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Stat. Sci.*, **20**:388–400.

Jordan, M.I. (2004). Graphical models. *Statist. Sci.*, **19**:140–155.

Jordan, V.C. (2006). Tamoxifen (ICI46,474) as a targeted therapy to treat and prevent breast cancer. *Br. J. Pharmacol.*, **147**:S269–S276.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**:613–636.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, **90**:773–795.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.*, **90**:928–934.

Kerr, G., Ruskin, H.J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Comput. Biol. Med.*, **38**:283–293.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *J. Am. Stat. Assoc.*, **102**:1025–1038.

Khoury, G.A., Baliban, R.C., and Floudas, C.A. (2011). Proteome-wide post-translational modification statistics: Frequency analysis and curation of the Swiss-Prot database. *Sci. Rep.*, **1**:90.

Kim, S.Y., Imoto, S., and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings Bioinf.*, **4**:228–235.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, **295**:1662–1664.

Knowles, D.A. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Ann. Appl. Stat., In Press.*

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput.*, **11**:313–322.

Koivisto, M. (2006). Advances in exact Bayesian structure discovery in Bayesian networks. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 241–248.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, MA.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.*, **5**:21.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.*, **37**:4254–4278.

Lauffenburger, D.A. and Linderman, J.J. (1993). *Receptors: Models for Binding, Trafficking, and Signaling.* Oxford University Press, New York.

Lauritzen, S.L. (1996). *Graphical Models.* Oxford University Press, New York.

Lèbre, S. (2009). Inferring dynamic genetic networks with low order independencies. *Stat. Appl. Genet. Mol. Biol.*, **8**:9.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, **10**:603 – 621.

Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., and Mallick, B.K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**:90–97.

Lehmann, E. and Romano, J. (2008). *Testing Statistical Hypotheses.* Springer, New York, NY.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**:1175–1182.

Li, F. and Zhang, N.R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics . *J. Am. Stat. Assoc.*, **105**:1202–1214.

Li, S.Z. (2009). *Markov Random Field Modeling in Image Analysis.* Springer-Verlag, London.

Li, Z., Li, P., Krishnan, A., and Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics*, **27**:2686–2691.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.*, **103**:410–423.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. R. Stat. Soc. B*, **34**:1–41.

Lu, Y., *et al.* (2011). Kinome siRNA-phosphoproteomic screen identifies networks regulating AKT signaling. *Oncogene*, **30**:4567–4577.

Luo, R. and Zhao, H. (2011). Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *Ann. Appl. Stat.*, **5**:725–745.

Mackay, D.J.C. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Netw. Comput. Neural. Syst.*, **6**:469–505.

Madeira, S.C. and Oliveira, A.L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **1**:24–45.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.*, **89**:1535–1546.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev*, **63**:215–232.

Majewski, I.J. and Bernards, R. (2011). Taming the dragon: Genomic biomarkers to individualize the treatment of cancer. *Nat. Med.*, **17**:304–312.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**:661–675.

Manning, B.D., Tee, A.R., Logsdon, M.N., Blenis, J., and Cantley, L.C. (2002). Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberin as a target of the phosphoinositide 3-kinase/Akt pathway. *Mol. Cell*, **10**:151–162.

Margaritis, D. and Thrun, S. (2000). Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems 12*, pp. 505–511. MIT Press.

Markowetz, F., Grossmann, S., and Spang, R. (2005). Probabilistic soft interventions in conditional Gaussian networks. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pp. 214–221.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks - a review. *BMC Bioinf.*, **8**:S5.

McLachlan, G.J. and Basford, K.E. (1987). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker.

McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**:413–422.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models.* John Wiley & Sons, New York.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, **34**:1436–1462.

Ménard, S., Pupa, S.M., Campiglio, M., and Tagliabue, E. (2003). Biologic and therapeutic role of HER2 in cancer. *Oncogene*, **22**:6570–6578.

Monni, S. and Li, H. (2010). Bayesian methods for network-structured genomics data. In M.H. Chen, P. Müller, D. Sun, K. Ye, and D.K. Dey (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, pp. 303–315. Springer, New York, NY.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Am. Stat. Assoc.*, **78**:47–55.

Mueller, C., Liotta, L.A., and Espina, V. (2010). Reverse phase protein microarrays advance to use in clinical trials. *Mol. Oncol.*, **4**:461–481.

Mukherjee, S. and Hill, S.M. (2011). Network clustering: Probing biological heterogeneity by sparse graphical models. *Bioinformatics*, **27**:994–1000.

Mukherjee, S., Pelech, S., Neve, R.M., Kuo, W.L., Ziyad, S., Spellman, P.T., Gray, J.W., and Speed, T.P. (2009). Sparse combinatorial inference with an application in cancer biology. *Bioinformatics*, **25**:265–271.

Mukherjee, S. and Speed, T.P. (2008). Network inference using informative priors. *Proc. Natl. Acad. Sci. USA*, **105**:14313–14318.

Mukherjee, S., Speed, T.P., and Hill, S.M. (2010). Model averaging for biological networks with prior information. In F. Emmert-Streib and M. Dehmer (eds.), *Medical Biostatistics for Complex Diseases*, pp. 353–378. Wiley-VCH, Weinheim.

Murphy, K. and Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA.

Murphy, K.P. (2002). *Dynamic Bayesian networks: representation, inference and learning.* PhD thesis, Computer Science, University of California, Berkeley, CA.

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 Comput. Biol.*, **3**:e129.

Neve, R.M., *et al.* (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**:515–527.

Nevins, J.R. (2001). The Rb/E2F pathway and cancer. *Hum. Mol. Genet.*, **10**:699–703.

Nielsen, T.O., *et al.* (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *The Lancet*, **359**:1301 – 1307.

Nita-Lazar, A., Saito-Benz, H., and White, F.M. (2008). Quantitative phosphoproteomics by mass spectrometry: Past, present, and future. *Proteomics*, **8**:4433–4443.

Nott, D.J. and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *J. Comput. Graph. Stat.*, **13**:141–157.

Oates, C. and Mukherjee, S. (2012). Network inference and biological dynamics. *Ann. Appl. Stat. (to appear)*.

Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.*, **1**:2005.0010.

Opgen-Rhein, R. and Strimmer, K. (2006). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology, WCSB 2006*, pp. 73–76.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.*, **8**:1145–1164.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**:681–686.

Paweletz, C.P., Charboneau, L., Bichsel, V.E., Simone, N.L., Chen, T., Gillespie, J.W., Emmert-Buck, M.R., Roth, M.J., Petricoin III, E., and Liotta, L.A. (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**:1981–1989.

Pawson, T. and Warner, N. (2007). Oncogenic re-wiring of cellular signaling pathways. *Oncogene*, **26**:1268–1275.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Francisco, CA.

Pearl, J. and Verma, T. (1991). A theory of inferred causation. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pp. 441–452. Morgan Kaufman, San Francisco, CA.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, **144**:864–873.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**:S215–S224.

Perez, O.D. and Nolan, G.P. (2002). Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat. Biotech.*, **20**:155–162.

Perou, C.M., *et al.* (2000). Molecular portraits of human breast tumours. *Nature*, **406**:747–752.

Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alche Buc, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**:ii138–ii148.

Pourret, O., Naïm, P., and Marcot, B. (2008). *Bayesian Networks: A Practical Guide to Applications.* John Wiley & Sons, Chichester, UK.

Prill, R.J., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K., and Stolovitzky, G. (2011). Crowdsourcing network inference: The DREAM predictive signaling network challenge. *Sci. Signal.*, **4**:mr7.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.*, **92**:179–191.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**:846–850.

Rau, A., Jaffrézic, F., Foulley, J.L., and Doerge, R.W. (2010). An empirical Bayesian method for estimating biological networks from temporal microarray data. *Stat. Appl. Genet. Mol. Biol.*, **9**:Article 9.

Ray, L.B. (2010). Why cell signaling is noisy. *Sci. Signal.*, **3**:ec298.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proc. of the Third Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pp. 157–164. Univ. California Press, Berkeley, CA.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer, New York, 2nd edition.

Robinson, J.W. and Hartemink, A.J. (2010). Learning non-stationary dynamic Bayesian networks. *J. Mach. Learn. Res.*, **11**:3647–3680.

Robinson, R.W. (1973). Counting labeled acyclic digraphs. In F. Harary (ed.), *New Directions in the Theory of Graphs*, pp. 239–273. Academic Press, New York.

Rothman, A.J., Bickel, P.J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, **2**:494–515.

Rubbi, L., *et al.* (2011). Global phosphoproteomics reveals crosstalk between Bcr-Abl and negative feedback mechanisms controlling Src signaling. *Sci. Signal.*, **4**:ra18.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* Chapman & Hall/CRC Press.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**:523–529.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**:754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**:32.

Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., and Müller, G. (2002). Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.*, **20**:370–375.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**:461–464.

Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liau, L.M., Cloughesy, T.F., Mischel, P.S., and Nelson, S.F. (2003). Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, **22**:4918–4923.

Shaw, R.J., Kosmatka, M., Bardeesy, N., Hurley, R.L., Witters, L.A., DePinho, R.A., and Cantley, L.C. (2004). The tumor suppressor LKB1 kinase directly activates AMP-activated kinase and regulates apoptosis in response to energy stress. *Proc. Natl. Acad. Sci. USA*, **101**:3329–3335.

Shaywitz, A.J., Dove, S.L., Greenberg, M.E., and Hochschild, A. (2002). Analysis of phosphorylation-dependent protein-protein interactions using a bacterial two-hybrid system. *Sci. STKE*, **2002**:l11.

Sheehan, K.M., *et al.* (2005). Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol. Cell. Proteomics*, **4**:346–355.

Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., and McGuire, W.L. (1987). Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, **235**:177–182.

Slawski, M., zu Castell, W., and Tutz, G. (2010). Feature selection guided by structural information. *Ann. Appl. Stat.*, **4**:1056–1080.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**:317–343.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**:Art.3.

Sørlie, T., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**:10869–10874.

Speed, T.P. and Kiiveri, H.T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Stat.*, **14**:138–150.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search.* Springer-Verlag, New York, NY.

Spurrier, B., Ramalingam, S., and Nishizuka, S. (2008). Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat. Protocols*, **3**:1796–1808.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *TEST*, **19**:209–256.

Staunton, J.E., *et al.* (2001). Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA*, **98**:10787–10792.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the Third Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pp. 197–206. Univ. California Press, Berkeley, CA.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, **18**:S231–S240.

Stingo, F.C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, **27**:495–501.

Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., and Kellam, P. (2004). Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.*, **5**:R94.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**:ii227–ii236.

Tavazoie, S. (1999). Systematic determination of genetic network architecture. *Nat. Genet.*, **22**:281–285.

TCGA-Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**:609–615.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G.C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**:2405–2412.

Tibes, R., Qiu, Y.H., Lu, Y., Hennessy, B., Andreeff, M., Mills, G., and Kornblau, S. (2006). Reverse phase protein array: Validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.*, **5**:2512–2521.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**:267–288.

Toh, H. and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**:287–297.

Tsamardinos, I., Aliferis, C.F., and Statnikov, E. (2003). Algorithms for large scale Markov blanket discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pp. 376–380. AAAI Press.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–268. Elsevier Science, Amsterdam, NL.

Waddell, P.J. and Kishino, H. (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inf.*, **11**:129–140.

Wang, C.C., Cirit, M., and Haugh, J.M. (2009a). PI3K-dependent cross-talk interactions converge with Ras as quantifiable inputs integrated by Erk. *Mol. Syst. Biol.*, **5**:246.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**:57–63.

Wasserman, L. (2000). Bayesian model selection and model averaging. *J. Math. Psychol.*, **44**:92 – 107.

Wei, Z. and Li, H. (2008). A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Stat.*, **2**:408–429.

Weigelt, B., Pusztai, Lajos Ashworth, A., and Reis-Filho, J.S. (2012). Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol.*, **9**:58–64.

Weinberg, R.A. (2006). *The Biology of Cancer.* Garland Science, New York.

Werhli, A.V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**:2523–2531.

Werhli, A.V. and Husmeier, D. (2007). Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.*, **6**:15.

Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D.A., and White, F.M. (2007). Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. USA*, **104**:5860–5865.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**:714–721.

Xu, T.R., *et al.* (2010). Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.*, **3**:ra20.

Yang, L., *et al.* (2004). Akt/protein kinase B signaling inhibitor-2, a selective small molecule inhibitor of Akt signaling with antitumor activity in cancer cells overexpressing Akt. *Cancer Res.*, **64**:4394–4399.

Yarden, Y. and Sliwkowski, M.X. (2001). Untangling the ErbB signalling network . *Nat. Rev. Mol. Cell Biol.*, **2**:127–137.

Yeung, K.Y., Dombek, K.M., Lo, K., Mittler, J.E., Zhu, J., Schadt, E.E., Bumgarner, R.E., and Raftery, A.E. (2011). Construction of regulatory networks using expression time-series data of a genotyped population. *Proc. Natl. Acad. Sci. USA*, **108**:19436–19441.

Yokogami, K., Wakisaka, S., Avruch, J., and Reeves, S.A. (2000). Serine phosphorylation and maximal activation of STAT3 during CNTF signaling is mediated by the rapamycin target mTOR. *Curr. Biol.*, **10**:47–50.

Yu, L.R., Issaq, H.J., and Veenstra, T.D. (2007). Phosphoproteomics for the discovery of kinases as cancer biomarkers and drug targets. *Proteomics Clin. Appl.*, **1**:1042–1057.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Am. Stat. Assoc.*, **100**:1215–1225.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**:19–35.

Yuan, T.L. and Cantley, L.C. (2008). PI3K pathway alterations in cancer: variations on a theme. *Oncogene*, **27**:5497–5510.

Yuan, Y., Savage, R.S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**:e1002227.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti*, pp. 233–243. North-Holland, Amsterdam.

Zhang, X., Zhao, X.M., He, K., Lv, L., Cao, Y., Liu, J., Hao, J.K., Liu, Z.P., and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by PC-algorithm based on conditional mutual information. *Bioinformatics*, **28**:98–104.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Statist.*, **3**:1473–1496.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Mach. Learn.*, **80**:295–319.

Zhu, Y., Shen, X., and Pan, W. (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinf.*, **10**:S21.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**:1418–1429.

Zou, M. and Conzen, S.D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**:71–79.