THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: ……

University ID number: ……

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ……..... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / NO (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / NO (*Please delete as appropriate*)

OR My thesis can be made publicly available only after…..[date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. **GRANTING OF NON-EXCLUSIVE RIGHTS**

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. **DECLARATIONS**

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. **LEGAL INFRINGEMENTS**

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: ..... Date: .....

Bayesian Inference for Protein Signalling Networks

Christopher James Oates
PhD Thesis
University of Warwick
Complexity Science Doctoral Training Centre

August 22, 2013

Contents

List of Figures

List of Tables

Acknowledgements

This doctoral research was funded by the Engineering and Physical Sciences Research Council (EPSRC) through the Complexity Science Doctoral Training Centre and Department of Statistics at the University of Warwick (Coventry, UK). I am grateful to the Department of Biochemistry at the Netherlands Cancer Institute (Amsterdam, NL) for hosting my visit in the 2011-2012 academic year. The students and staff at each institution have made this study a pleasure.

Many people have had a direct impact upon this research, including Tarmo Aijö, John Aston, Roderick Beijersbergen, Jen Bowskill, Thijn Brummelkamp, Quentin Caudron, Colm Connaughton, Frank Dondelinger, Joe Gray, Pantelis Hadjipantelis, Laura Heiser, Steven Hill, Kathy Jastrzebski, Steve Kiddle, Theo Knijnenburg, James Korkola, Robert MacKay, Gordon Mills, Sergio Morales, Chris Penfold, Tassos Perrakis, Anas Rana, Phil Richardson, Monica Rigat, Gareth Roberts, Jordi Vidal Rodriguez, Titia Sixma, Paul Spellman, Simon Spencer, Nicholas Städler, Vlad Vyshemirsky, Lodewyk Wessels, Rachel Wilkerson and many anonymous referees.

In addition I am grateful to Stafford Library for providing an excellent writing environment and to Wikipedia for hastening the pace of this research.

The most gratitude must go to Sach Mukherjee, Lucy Astley and my family, for giving me an excellent training in different respects.

Declaration

Parts of this thesis have been published or are in submission:

- Oates CJ, Mukherjee S (2012) Network Inference and Biological Dynamics. *Ann. Appl. Stat.* **6**(3):1209-1235.
- Oates CJ, Hennessy BT, Lu Y, Mills GB, Mukherjee S (2012) Network Inference Using Steady State Data and Goldbeter-Koshland Kinetics. *Bioinformatics* **28**(18):2342-2348.
- Oates CJ, Mukherjee S (2012) Causal Variable Selection Using Equilibrium Relations from Nonlinear Dynamics. *Workshop on Causal Structure Learning, Uncertainty in Artificial Intelligence (UAI'12).* Santa Catalina, CA, USA.
- Oates CJ, Dondelinger F, Bayani N, Korkola J, Gray JW, Mukherjee S (2013) Network Inference and Dynamical Prediction Using Biochemical Kinetics. *In submission.*
- Oates CJ, Korkola J, Gray JW, Mukherjee S (2012) Joint Estimation of Multiple Exchangeable Networks. *In revision.*

For Chapters 3, 4 the reader is recommended to contact CJO for the most up-to-date version of papers.

The thesis is CJO's own work except where it contains work based on collaborative research, in which case the nature and extent of CJO's contribution is indicated. This thesis has not been submitted for a degree at another university.

Summary

Cellular response to a changing chemical environment is mediated by a complex system of interactions involving molecules such as genes, proteins and metabolites. In particular, genetic and epigenetic variation ensure that cellular response is often highly specific to individual cell types, or to different patients in the clinical setting. Conceptually, cellular systems may be characterised as networks of interacting components together with biochemical parameters specifying rates of reaction. Taken together, the network and parameters form a predictive model of cellular dynamics which may be used to simulate the effect of hypothetical drug regimens.

In practice, however, both network topology and reaction rates remain partially or entirely unknown, depending on individual genetic variation and environmental conditions. Prediction under parameter uncertainty is a classical statistical problem. Yet, doubly uncertain prediction, where both parameters and the underlying network topology are unknown, leads to highly non-trivial probability distributions which currently require gross simplifying assumptions to analyse. Recent advances in molecular assay technology now permit high-throughput data-driven studies of cellular dynamics. This thesis sought to develop novel statistical methods in this context, focussing primarily on the problems of (i) elucidating biochemical network topology from assay data and (ii) prediction of dynamical response to therapy when both network and parameters are uncertain.

Notation

This thesis assumes knowledge of standard mathematical and statistical notation. Application-specific notation aims to follow the conventions listed below. When convenient, these may be explicitly overlooked in order to simplify presentation.

$\mathbb{N}_0$	non-negative integers
$\mathbb{R}_+$	non-negative reals
$\mathcal{J} = \{1, \ldots, J\}$	index set of individuals, possibly biological samples
$\mathcal{P} = \{1, \ldots, P\}$	index set of state variables
$\mathcal{X}_p$	chemical species associated with index $p \in \mathcal{P}$
$\mathcal{X}_p^*$	phosphorylated form of species $\mathcal{X}_p$
G	chemical reaction graph
N	directed network
$\boldsymbol{N}$	discrete state vector in $\mathbb{N}_0^P$
$\boldsymbol{X}$	continuous state vector in $\mathbb{R}_+^P$
$\boldsymbol{\theta}$	parameter vector
S, P, E	substrate, product, enzyme respectively
$\mathcal{E}_p$	set of kinases acting on species $\mathcal{X}_p$
$\mathcal{I}_{p,E}$	set of inhibitors for kinase $E \in \mathcal{E}_p$
K	Michaelis-Menten parameter
$\mathcal{F}_{\boldsymbol{X}}$	natural filtration of the stochastic process $\boldsymbol{X}$
$\mathcal{N}$	Gaussian density / space of directed networks (depending on context)
$\perp\!\!\!\perp$	statistical independence
$\mathcal{D}(\boldsymbol{v})$	diagonal matrix with diagonal entries $\boldsymbol{v}$
$\boldsymbol{y}$	data, possibly corrupted by measurement noise
$\mathbb{I}$	indicator function

Introduction

The last two decades have seen rapid advances in biotechnology, enabling increasingly precise quantitative measurement of molecular species in biological samples. In the 1990s the introduction of the DNA microarray facilitated the simultaneous and rapid quantification of RNA abundance for multiple genes (Fig. 1). Comparison of these *gene expression* data across multiple biological samples offered an unbiased approach to screen for genes which are statistically implicated (*differentially expressed*) in a biological context of interest, relative to control samples. Microarray technology revolutionised fundamental biological research by providing a mechanism by which to constrain experimental design, reducing the number of candidate genes for experimental investigation (e.g. knock-out or knock-down) [Crowther, 2002]. Translational research was also transformed, with gene expression data forming the basis for several signatures which are predictive of response to therapy [van 't Veer *et al.*, 2002]. Subsequent years saw the continued emergence of high-throughput biotechnologies, including array-comparative genomic hybridization (A-CGH), chromatin immunoprecipitation (ChIP) -on-chip DNA-binding assays, single-nucleotide polymorphism (SNP) arrays, high-throughput drug screening, protein microarrays and next generation sequencing. The increasing ease and decreasing cost of obtaining large amounts of data on a biological system have led to an emphasis on integrative, *systems level* analysis. This paradigm is central to the field of oncology, where it has become apparent that cancer is an emergent disease resulting from interplay between the functional effects of genetic or epigenetic mutations [Weinberg, 2007].

Multivariate biological data present significant challenges for modelling, computation and statistical interpretation. The need to analyse large biological datasets has sparked much interest in multivariate and high-dimensional statistics [Bühlmann and van de Geer, 2011]. The visual representation of interplay in a multivariate system which is afforded by a graph or network has proven extremely popular. A (standard) biological network consists of a set of nodes, representing molecular species such as genes, proteins or metabolites, and a set of edges which describe interactions or interplay between the nodes. Often attention is restricted to one particular form of molecular species (e.g. genes) and one form of interaction (e.g. transcriptional regulation). The type of molecular species which form the set of nodes and the biological mechanism which is encoded by the edges will lend its name to the network, so that we speak of gene regulatory networks, protein signalling networks or metabolic interaction networks, for example. Experimentalists have elucidated network topology for important biological processes, but the inherently combinatorial nature of networks provides a fundamental barrier to elucidating large amounts of topology on an edge-by-edge basis. The automatic characterisation of biological networks from high-throughput data obtained in a context of interest, such as a tissue type or a disease state, has become a prominent research goal in systems biology.

Cancer is a prevalent disease, with more than 1 in 3 people in the UK developing some form of cancer during their lifetime. Due in part to an ageing population, cancer incidence rates in Great Britain have risen by 22% in males and by 42% in females since the mid-1970s. Worldwide in 2008, there were estimated to be around 7.6 million cancer-related deaths and 12.7 million new cases. Intensive research on an international scale has led to advances in cancer therapeutics, such that cancer survival rates in the UK have doubled in the last 40 years. (All statistics taken from Cancer Research UK [2013] on 17/04/2013.)

One of the biggest scientific achievements of the last decade was the development of targeted anti-cancer drugs, which have demonstrated potential to revolutionise clinical treatment of the disease [Sudhakar, 2009]. For example Imatinib (Novartis Pharma AG [2006] trade name Glivec) has rendered a subset of otherwise terminal leukaemia into a manageable chronic condition by targeting a tyrosine kinase enzyme, known as BCR-ABL, which exists only in cancer cells and not in healthy cells [Moen *et al.*, 2007]. (Only a small minority of patients will acquire resistance to Imatinib [Mauro, 2006]). Recently more drugs than ever are entering into clinical trials, yet the rate at which drugs are approved

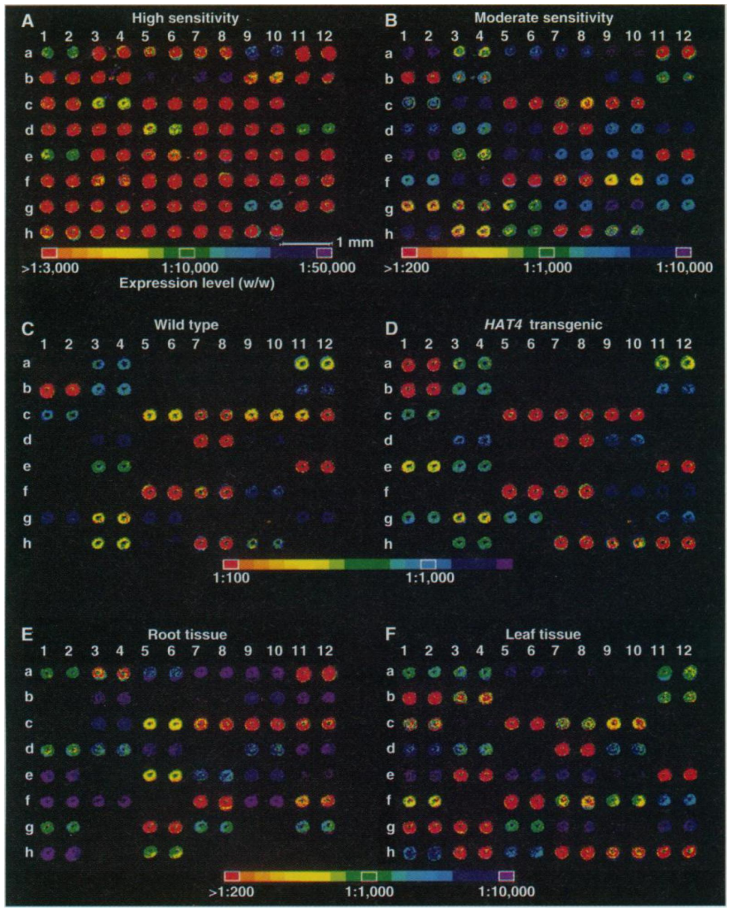


Figure 1: The first reported use of miniaturized microarrays for gene expression profiling appeared in Schena *et al.* [1995]. In total 45 genes were measured in *Arabidopsis*. Two years later Lashkari *et al.* [1997] reported an assay of 2,479 genes in the yeast *S. Cerevisiae*. Modern microarrays can contain up to 47,000 genes (e.g. Affymetrix GeneChip Human Genome U133 Plus 2.0). [Figure reproduced with permission from Schena *et al.* [1995].]

has dropped to an all time low [Silverman, 2012]. Experimental evidence suggests that many patients become resistant to therapy via activation of secondary *survival* pathways which were not targeted by the original treatment [Lee *et al.*, 2012]. Effective inhibition of these secondary pathways would be expected to have a significant benefit for patients [Wetterskog *et al.*, 2013]. The shift from mono-therapeutics to poly-therapeutics necessitated by the complex, multivariate nature of cancer has, in part, driven the move towards systems biology.

A systems-level understanding of biological signalling processes introduces major experimental, translational and theoretical challenges. For example in oncology, prior to treatment it is currently extremely difficult to predict which pathways will require targeting in order to achieve maximum efficacy. It is practically infeasible to assess all possible combinations of drugs in the laboratory using cultured cancer cells. Indeed, whilst the number of available drugs is now large, the number of pairs of drugs is considerably bigger. Moreover, each pair of drugs might be applied in a different sequential order, at different doses, at different times, for longer or shorter treatment durations etc. In principle these difficulties could be averted with access to an accurate computational model of cellular signalling dynamics, since then hypothetical drug regimens could be rapidly explored *in silico* [Hopkins, 2008]. However this raises several theoretical challenges and there is currently a methodological void for systems-level inference and prediction in cellular signalling systems.

Cellular signalling systems have been modelled in a variety of ways, including discrete logic models [Bender *et al.*, 2010], discrete time Markov processes such as dynamic Bayesian networks (DBNs) [Hill *et al.*, 2012a], Markov jump processes [Paulsson, 2005; Wilkinson, 2006], Gaussian processes [Honkela *et al.*, 2010], structural equation models (SEMs) [Liu *et al.*, 2008], ordinary differential equations (ODEs) [Chen *et al.*, 2009; Schoeberl *et al.*, 2002] and stochastic differential equations (SDEs) [Finkenstädt *et al.*, 2013]. Almost all models of cellular signalling are rooted in network representations, either explicitly as in DBNs and SEMs, or implicitly as in ODEs and SDEs [Sokol and Hansen, 2013]. Relating these models to data is often challenging. In this setting there are two main problems; (i) inference of model parameters, such as reaction rates, and (ii) uncovering a network structure which adequately describes interplay in the biological system under study. Classically, much effort has been directed at the first problem of estimating kinetic parameters, such as reaction rates, from noisy experimental data on the molecular species. The second problem, which is commonly referred to as *network inference*, has received relatively less theoretical attention. In many biological contexts the edge structure of the network may be uncertain (e.g. due to genetic or epigenetic alterations in disease states). Then, an important biological goal is to perform network inference in a context-specific manner [Ideker and Krogan, 2012], that is, using data acquired in the biological context of interest. The ability to accurately estimate context-specific network topology has potential to greatly accelerate progress within systems biology, pharmacology and related disciplines [Csermely *et al.*, 2013]. For example, protein signalling network structure has been shown determine the response of cells to certain therapeutic interventions [Lee *et al.*, 2012]. Advances in high-throughput data acquisition have led to much interest in such data-driven characterization of biological networks.

This thesis aims to contribute advances in the data-driven estimation of biological networks. Focussing primarily on inference for protein signalling networks, the novel contributions of this thesis can be summarised as follows:

- Chapter 2, *From Biological Dynamics to Network Inference*:
  - Motivated by tractable approximations of complex stochastic dynamical systems, a connection is drawn between several existing network inference algorithms in terms of a unified statistical model. This framework makes explicit the assumptions underlying each approach, with particular emphasis on time series data obtained at uneven sampling intervals.
  - A comprehensive empirical investigation assessed 32 different network inference algorithms from this unified family using both simulated and real datasets where the data-generating networks were known by design.
  - Our results highlight critical issues regarding the treatment of uneven sampling intervals, which are shown to significantly effect the algorithms' performance.
  - One statistical formulation is shown to perform favourably in most data generating regimes; this is taken as a basis for subsequent methodological development in Chapter 3.
- Chapter 3, *Network Inference and Prediction Using Chemical Kinetics*:

- A novel statistical framework is presented which integrates non-linear chemical kinetics into inference for protein signalling networks.
- For time course data, Monte Carlo computation of model selection criteria is leveraged to compute Bayes factors for non-linear dynamical systems defined on a network. Inference over networks is facilitated by Bayesian model averaging.
- Empirical investigations demonstrate improved network reconstruction on both simulated and real datasets in comparison to approaches rooted in linear dynamical formulations.
- The methodology is demonstrated to be able to predict the effect of held-out interventions, both *in silico* and *in vitro*. In particular the methodology facilitates prediction of cellular response in the challenging setting where neither the chemical reaction network, nor the corresponding parameters are known *a priori*.

- Chapter 4, *Joint Estimation of Multiple Networks from Time Course Data*:
  - It is often the case that data are collected on multiple individuals $j \in \mathcal{J}$ which may differ with respect to interplay between variables. For example, in biology, different cell lines may possess differing protein signalling networks. A hierarchical Bayesian framework is proposed for joint inference in this setting.
  - Unlike previous proposals, which were computationally prohibitive, an efficient, exact Bayesian algorithm is proposed for reporting posterior marginal inclusion probabilities in the hierarchical setting.
  - A comprehensive study of joint estimation is undertaken, demonstrating how joint models may yield improved network inference results both *in silico* and *in vitro*, using data obtained from a panel of breast cancer cell lines.

This thesis is organised as follows: Chapter 1 introduces key concepts in biology, experimentation, chemistry and statistics. Chapter 2 formally defines the network inference problem and describes potential pitfalls using a wide range of examples. Chapter 3 presents recent work on network inference and dynamical prediction rooted in non-linear models of chemical kinetics. In Chapter 4 we present efficient computational techniques for joint inference of multiple networks. Finally Chapter 5 contains a concise summary and discussion of open statistical challenges in bio-molecular signalling systems.

Chapter 1

Background Material

Scientific investigation of complex systems increasingly requires a broad tool-kit of analytic, computational and experimental techniques. This thesis assumes a background in both mathematics and statistics; in particular we will make use of differential equations, dynamical systems, stochastic processes, Bayesian statistics and Markov chain Monte Carlo. To a lesser extent we assume a basic understanding of cellular biology, including gene regulation and protein synthesis. In this Chapter we build on these bases in order to familiarise the reader with concepts necessary to follow the remainder. In particular we will discuss protein signalling mediated by phosphorylation, aberrant protein signalling in cancer, emerging experimental platforms, mathematical formalisms for chemistry, graphical models in statistics and a theory of inferred causation.

1.1 Biological Background

In this Section we introduce the fundamental biochemical process of protein signalling mediated by phosphorylation, discuss aberrant signalling in genetic diseases such as cancer, and describe some modern approaches to therapy which exploit the biochemistry of phosphorylation. Throughout we explicate these concepts in the context of well characterised signalling pathways in mammalian cells.

1.1.1 The (Not So) Central Dogma

Cellular response to a changing environment is mediated by a complex system of interactions involving molecules such as genes, proteins and metabolites. The *central dogma of molecular biology* provides a powerful constraint on the form of these interactions by specifying that certain information transferral processes are generally uni-directional [Crick, 1970]. In the language of graphical models, the central dogma postulates a set of conditional independences, as can be seen in Fig. 1.1. Specifically, the central dogma implies that (i) DNA may be transcribed into RNA but generally not *vice versa* (ii) RNA may be translated into protein molecules but generally not *vice versa* and (iii) proteins may regulate transcription of RNA by binding to promoter regions (such proteins are known as *transcription factors*).

Since Crick's description of the central dogma in 1970 it has become increasingly clear that many molecular interactions operate outside this paradigm; for example the post-translational modification of proteins (see Sec. 1.1.2) was not explicitly covered by the central dogma. This thesis focuses primarily on such interactions between protein molecules. However it is important to appreciate that these interactions are embedded in wider cellular signalling processes and are not generally causally sufficient (see Sec. 1.4.2).

1.1.2 Protein Signalling

Cell signalling is part of a complex system of communication that coordinates basic cellular activities. In brief, the role of cellular signalling is to receive both internal signals and external signals from the cellular membrane, to correctly process these signals and initiate a transcriptional response by modulating gene expression. The new gene expression profile may initiate other cellular processes e.g. cell division or apoptosis. Here we describe the chemical mechanisms underpinning the processes of signalling. In

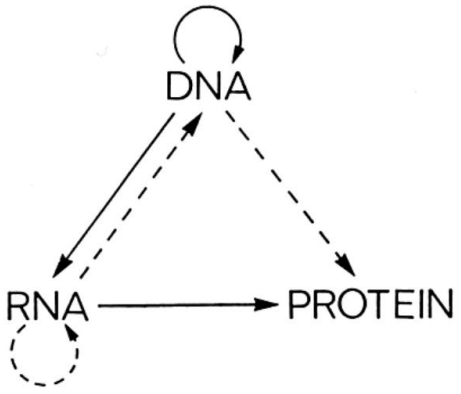


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

Figure 1.1: Central dogma of molecular biology, reproduced from the original paper of Crick [1970].

particular we focus on a particular form of chemical change, known as protein phosphorylation, which plays an important role in aberrant protein signalling in cancer (see Sec. 1.1.3).

Definition 1 (Phosphorylation)**.** Phosphorylation *is the addition of a phosphate (PO_4^{3-}) group from a high energy donor molecule, such as ATP, to a specific protein substrate, usually on the serine, threonine, or tyrosine amino acid (or* residue*). When there is no ambiguity regarding the residue, the phosphorylated protein is simply referred to as a* phosphoprotein*.*

Phosphorylation is an example of a post-translational modification (others being methylation, ubiquitylation, cleaving etc.). Post-translational modifications may alter a protein's function or activity through a conformational change, for example enzyme phosphorylation may modulate catalytic activity by exposing/blocking the active domain. Phosphorylation is reversible and many residues on a protein may be phosphorylated (the p53 protein contains more than 18 different phosphorylation sites).

Definition 2 (Kinase and phosphatase)**.** *Enzymes which catalyse phosphorylation are known as* kinases*, whilst enzymes which catalyse dephosphorylation are known as* phosphatases*.*

Both kinases and phosphatases are typically highly specific, thereby exerting very precise control over cellular function. In many cases phosphorylation induces an activation of functionality, though this is not true in general, with counter examples including the retinoblastoma protein Rb which becomes inactive when phosphorylated. Often kinases and phosphatases are themselves phosphorylated proteins, so that an interconnected network of protein phosphorylation operates. Certain sub-networks have received much attention from the biological community - these well studied systems are typically referred to as "pathways". Below we present a detailed example of a protein signalling pathway.

Example 1 (Mitogen activated protein kinase (MAPK) pathway)**.** *Receptor tyrosine kinases (RTKs) such as the epidermal growth factor receptor (EGFR) are specifically activated by extracellular ligands (Fig. 1.2). For example, binding of epidermal growth factor (EGF) to EGFR activates the intracellular kinase activity of this RTK. Subsequently, docking proteins such as GRB2 and SOS bind to the activated EGFR and activate members of the Ras subfamily via phosphorylation (most notably H-Ras and K-Ras). At this point a cascade of phosphorylation occurs from Ras to Raf to MEK to MAPK, where at each step the parent acts as a kinase to phosphorylate and activate the child. RAF and MAPK are both serine/threonine-selective protein kinases, whereas MEK is a tyrosine/threonine kinase. This cascade leads into the cell nucleus, where one effect of MAPK activation is to regulate the activities of several transcription factors, including ribosomal protein S6. By altering the levels and activities of transcription factors, MAPK is able to modulate transcription of genes that are important for the cellular function and can lead to diverse phenotypes such as differentiation, proliferation and apoptosis.*

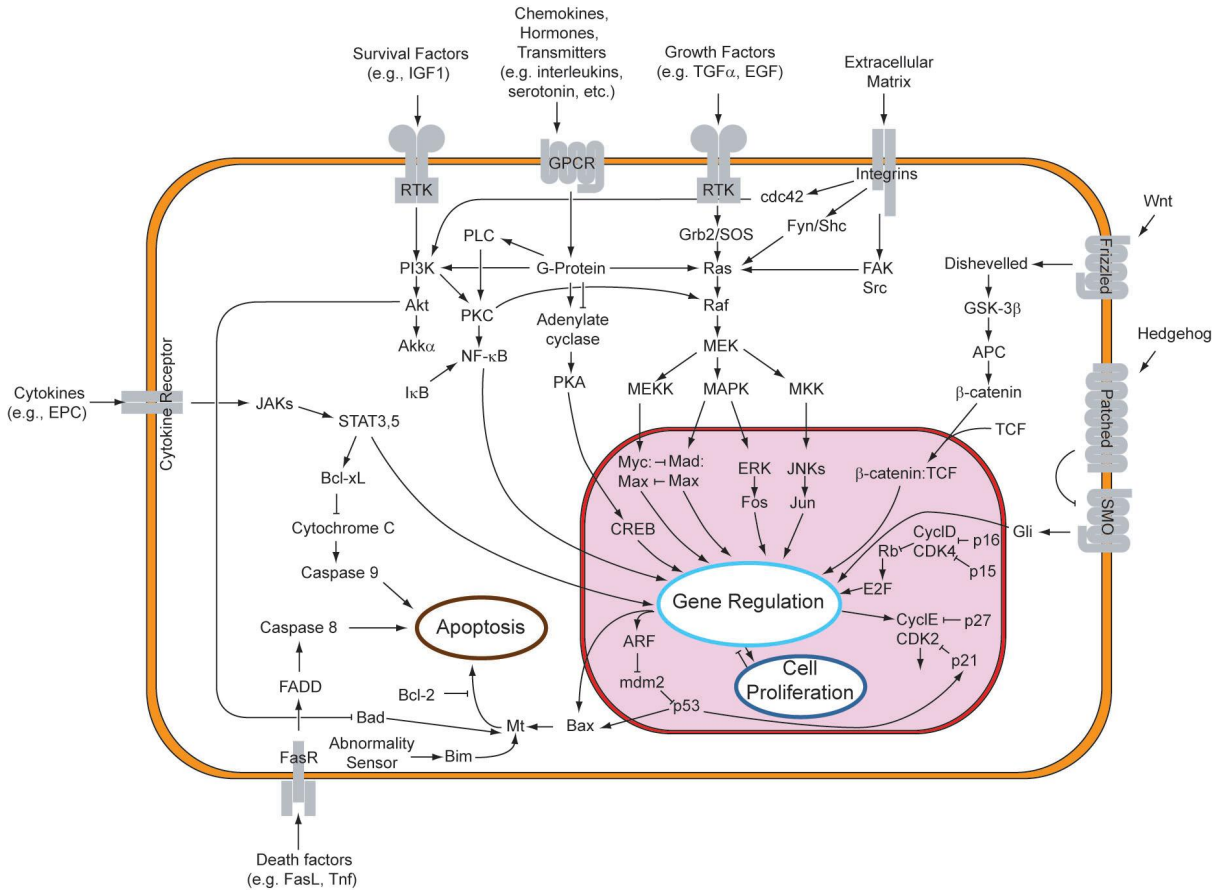


Figure 1.2: Schematic description of well-characterised signalling pathways in mammalian cells. Transmembrane receptors (grey) such as receptor tyrosine kinases (RTKs) receive an external chemical signal and transmit this through to the nucleus (purple) via a sequence of chemical reactions known as "signalling", involving kinases such as MAPK and Akt. ["Overview of signal transduction pathways." <http://en.wikipedia.org/wiki/File:Signal_transduction_pathways.png>, accessed Feb. 2013.]

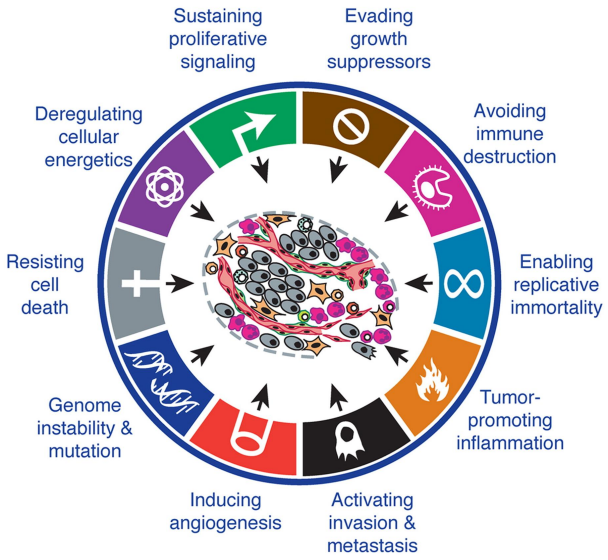


Figure 1.3: Hallmarks of Cancer. These 10 characteristics are believed to be necessary conditions for cancer to occur; hence each represents a unique opportunity for targeted anti-cancer therapies. [Figure adapted from Hanahan and Weinberg [2011].]

1.1.3 Protein Signalling and Cancer

Multiple studies have demonstrated the remarkable genomic heterogeneity of cancer [The 1000 Genomes Project Consortium, 2010; The Cancer Genome Atlas Network, 2012]. Nevertheless there are a collection of concepts which provide justification for a unified theory of cancer. In particular Hanahan and Weinberg [2011] defined ten "hallmarks" which represent necessary criteria for a disease to manifest as cancer in the clinical setting (Fig. 1.3). Several of these hallmarks, such as sustained proliferative signalling, evading growth suppressors and resisting cell death, may be facilitated by aberrant protein signalling, including signalling mediated by phosphorylation [Lee *et al.*, 2012; Moen *et al.*, 2007]. Thus protein phosphorylation plays a leading role in oncogenesis.

In the context of cancer, there are three classes of gene which have become paradigmatic [Vogelstein and Kinzler, 2004].

Definition 3 (Oncogene, tumour-suppressor and stability genes)**.** Proto-oncogenes*, when mutated, become constitutively active (*oncogenes*) or active under conditions in which the wild-type gene is not.* Tumour-suppressor *genes are targeted in the opposite way by genetic alterations; mutations reduce the activity of the gene product, leading to tumour development.* Stability genes *(or* caretaker *genes) keep genetic alterations to a minimum; thus when they are inactivated, mutations in other genes occur at a higher rate, including mutations in proto-oncogenes and tumour-suppressor genes.*

Example 2 (MAPK pathway)**.** *Uncontrolled growth is a prerequisite for the development of all cancers [Hanahan and Weinberg, 2011]. In many cancers, a defect in the MAPK pathway leads to that uncontrolled growth. For example, the proto-oncogene BRAF (whose role in MAPK signalling is becoming increasingly understood [Xu et al., 2010]) is known to be causally implicated in melanoma [Flaherty et al., 2010], with approximately 80% of cases involving a BRAF mutation.*

Example 3 (Akt pathway)**.** *A key hallmark of cancer is resistance to cell death. In "wild type" cells, programmed cell death (*apoptosis*) is induced by either extracellular signals (inc. toxins, hormones, growth factors etc.) or intrinsic signals (inc. DNA damage). Apoptosis is an important defence against abnormal cellular behaviour and is disabled in cancer states. Akt (Fig. 1.2) is a key inhibitor of apoptosis which must be phosphorylated in order to be active. Phosphorylation of Akt is in turn regulated by PI3K; a protein frequently constitutively active in breast cancer (see Example 4 or Korkola et al. [2013]). Through over activation, PI3K provides a route for cancer cells to evade apoptosis.*

Example 4 (Intrinsic sub-typing in breast cancer)**.** *Breast cancer is classically stratified according to a handful of genetic and histological markers. A genome-wide expression profile is used to cluster cancers*

into basal, luminal or claudin-low subgroups, whilst histological staining is used to explore the expression of human epidermal growth factor receptor 2 *(HER2),* oestrogen receptor *(ER) and* progesterone receptor *(PR) [Sotiriou and Pusztai, 2009]. In total there are five* "intrinsic" subtypes *of breast cancer;* luminal A*,* luminal B*,* HER2-enriched*,* basal-like *and* claudin-low *[Eroles et al., 2012], though this classification is disputed [Curtis et al., 2012]. In addition, genetic markers are used to further stratify biological samples. For example PIK3CA is a proto-oncogene which is mutated in 33% of breast cancer patients [Cizkova it et al., 2012]. Mutation renders its protein product PI3K constitutively active, meaning that it is no longer under the influence of receptor tyrosine kinases (RTKs; Fig. 1.2). HER2, another well known proto-oncogene, is amplified in approximately 30% of breast cancers. A frequently mutated tumour suppressor in many cancers is the TP53 gene [The Cancer Genome Atlas Network, 2012]. BRCA1 and BRCA2 are stability genes which assist in DNA repair pathways. Certain germ-line mutations in BRCA genes, common in certain population groups including Ashkenazi Jews, associate with an increased breast cancer risk. For example, women with an abnormal BRCA1 or BRCA2 gene have up to a 60% risk of developing breast cancer by age 90 [Breastcancer.org, 2012].*

1.1.4 Targeted Cancer Therapies

Molecular cancer therapies may broadly be divided into targeted and untargeted therapies. A targeted therapy is a type of medication which blocks the growth of cancer cells by specifically interfering with molecules needed for carcinogenesis and tumour growth. In contrast, an untargeted therapy interferes with all rapidly dividing cells (e.g. traditional chemotherapy). In practice, patients are often treated with a combination of both targeted and untargeted therapies [Carlson *et al.*, 2009].

Each of the hallmarks of cancer (Fig. 1.3) defines, in principle, a set of targets for therapeutic intervention. In this thesis we are primarily concerned with interventions which tackle sustained proliferative signalling and evasion of apoptosis; in particular interventions which tackle aberrant protein phosphorylation in the MAPK and Akt pathways (Examples 2,3). In this context there are two main molecular weapons; small molecule inhibitors and monoclonal antibodies.

Definition 4 (Small molecules)**.** Small molecules *are molecules with a low molecular weight (*$<$ 800 *Daltons) which enables them to rapidly diffuse across cell membranes in order to reach intracellular sites of action. In pharmacology, small molecules may bind to a protein and act as an effector, thereby altering the protein's activity or function.*

Example 5 (Small molecule kinase inhibitors)**.** *A* protein kinase inhibitor *is a type of small molecule inhibitor that specifically blocks the action of one or more protein kinases. Protein kinase inhibitors can be subdivided or characterised by the targets of the kinase whose activity is inhibited; most kinases act on both serine and threonine amino acids, the tyrosine kinases act on tyrosine, and a number (dual-specificity kinases) act on all three. Fig. 1.4 displays the 3D structure of EGFR inhibitor erlotinib, a reversible tyrosine kinase inhibitor.*

Definition 5 (Monoclonal antibodies)**.** *An* antibody *is produced by the immune system in order to identify and potentially neutralise foreign objects such as bacteria and viruses. Antibodies act by specifically binding to a target protein or cell type, thereby either tagging the target for attack by other parts of the immune system, or neutralising the target directly. Antibodies may be produced in large quantities in vitro for use in pharmacology.* Monoclonal *antibodies are antibodies derived from identical immune cells derived from a common ancestor.*

Example 6 (Monoclonal antibodies in cancer)**.** *It is possible to design antibodies specific to almost any cell surface target. Tumour cells can display cell surface receptors that are absent or present in smaller quantities on the surfaces of healthy cells; often these are responsible for activating cellular signal transduction pathways that cause the unregulated growth and division of the tumour cell. Thus antibodies can be used to destroy malignant tumour cells and prevent tumour growth by blocking specific cell receptors. Examples include HER2, a constitutively active cell surface receptor that is produced at abnormally high levels on the surface of approximately 30% of breast cancer tumour cells. The monoclonal antibody Trastuzumab has been clinically approved to block the HER2 receptor in HER2 positive breast cancer patients [McKeage and Perry, 2002].*

Example 7 (MAPK pathway)**.** *Many compounds have been designed to specifically inhibit biochemical interactions within the MAPK pathway. The first drug licensed to act on this pathway was Sorafenib,*

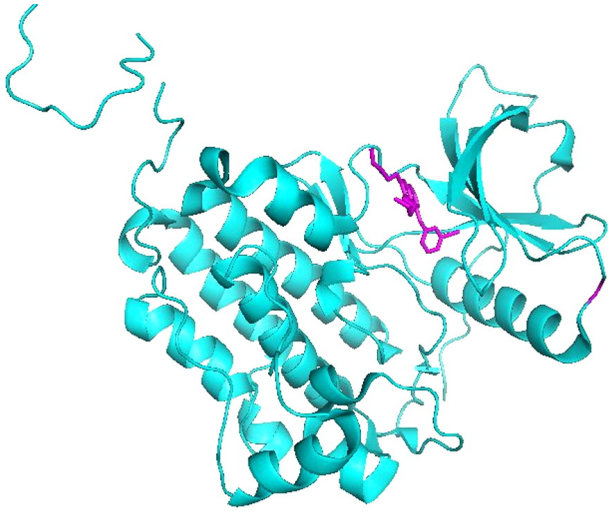


Figure 1.4: The 3D structure of EGFR inhibitor Erlotinib, a reversible tyrosine kinase inhibitor (source: www.rcsb.org; 1M17.pdb). Here the inhibitor (purple) binds to the epidermal growth factor receptor (EGFR; blue) in the ATP binding site, preventing catalytic activity.

an inhibitor of the Raf kinase. Now dozens of treatments for molecular players in this pathway are under clinical investigation [Roberts and Der, 2007]. For example, Fig. 1.4 shows how Erlotinib, a small molecule inhibitor of EGFR, can reduce catalytic activity by blocking the ATP binding site.

Example 8 (Small molecule inhibitors in breast cancer)**.** *Over-expression of HER2, ER or PR transmembrane receptor proteins generally indicates that breast cancer cells are dependent on signalling downstream of these receptors; in this case a natural strategy is to inhibit these receptors [Carlson et al., 2009]. For example, the small molecule inhibitor* Lapatinib *is used in combination therapy for HER2 positive breast cancer [Korkola et al., 2013]. Lapatinib belongs to a family of tyrosine kinase inhibitors, each of which specifically targets proteins involved in phosphorylation. Other family members involved in clinical trials to treat breast cancer include Gefitinib [ClinicalTrials.gov, 2013a], Cabozantinib [ClinicalTrials.gov, 2013b] and Neratinib [ClinicalTrials.gov, 2013c].*

1.2 Experimental Background

1.2.1 Cancer Cell Lines

There exist several experimental systems for the study of cancer, including real patients, mouse models [Frese and Tuveson, 2007], *ex vivo* tissue samples [Burdall *et al.*, 2003], cell lines [Neve *et al.*, 2006], *ex cellulo* assays [Hsieh *et al.*, 1997] and virtual screening [Shoichet, 2004]. This thesis restricts attention to cell line models of cancer, in particular cell lines derived from breast cancer patients. The use of cell lines offers a number of advantages over alternative model systems, in addition to several disadvantages. We discuss both below:

- Strengths:
  - *Cost.* Initial purchase of cell lines will typically cost in the region of £500 - £1,000 [ATCC, 2013]. Once acquired, cells may be cloned in unlimited quantity.
  - *Speed.* Human fibroblast cells, for example, take approximately 24 hours to divide, facilitating rapid experimentation. In contrast, mouse models require several months per generation, and often several generations of selective breeding to obtain a desired genetic profile.
  - *Convenience.* Basic laboratory equipment is sufficient to handle and maintain cell culture.
  - *Variety.* Due to the rapidly expanding catalogue of cell lines, it is possible to construct (or purchase) panels of lines which exhibits a relatively high degree of genetic heterogeneity.
  - *Robustness.* Cell lines are easily replaced from frozen stocks, providing a secure back-up against e.g. power failure or contamination.

– *Regulation.* Laboratory use of primary tissue culture requires patient approval [Burdall *et al.*, 2003], whereas cell lines may be used without the permission of the patient donor.

– *Reproducibility.* Cell lines (also mouse models) are standardised, with (in principle) identical cell cultures available to researchers globally. This level of reproducibility is not possible in patient studies, for example.

- Weaknesses:

  – *Model misspecification.* Cell lines (asim. *ex cellulo* assays and virtual screening) are far removed from the relevant *in vivo* setting. In particular, (i) cell cultures typically occupy only two spatial dimensions (although 3D is arriving, e.g. [Hsiao *et al.*, 2012; Souza *et al.*, 2010]) (ii) the media in which cells are grown may differ from the relevant tumour micro-environment [Arya *et al.*, 2012] (iii) cell lines are, by definition, immortalised; thus cell lines have been selected for a genetic profile which is amenable to immortalisation (iv) the experimental set-up is idealised, so that otherwise challenging clinical aspects such as drug delivery or immunological response to therapy are ignored.

  – *Lineage.* Many established breast cancer cell lines were not derived from primary breast tumours, but from tumour metastases. In particular, cell line catalogues tend to over-represent the more aggressive, metastatic, late-stage tumours, rather than the primary lesion. Since most drug therapies are directed against the primary tumour [Burdall *et al.*, 2003], this contributes to the problem of model misspecification.

  – *Contamination.* Cell line cross-contamination can be a problem for scientists working with cultured cells; indeed, studies suggest up to 15-20% of cells used in experiments have been misidentified or contaminated with another cell line [Cabrera *et al.*, 2006]. In particular the HeLa cell line (the first cell line to be developed in 1952, from a glandular cancer of the cervix) was notorious as a cross-contaminant [Nelson-Rees *et al.*, 1981]. More recently, of 252 new cell lines deposited at the German Cell Line Bank, 18% were found to be cross-contaminated [Masters, 2000]. This has led to a drive to define standardised procedures for verification of cell line identity [Masters, 2001].

  – *Mutation.* Cell lines are prone to genotypic and phenotypic drift during their continual culture. This is particularly common in older and more frequently used cell lines. Sub-populations may arise and cause phenotypic changes over time by the selection of specific, more rapidly growing clones within a population. It has been demonstrated that MCF-7 cells (the most commonly used breast cancer cell line) show markedly different karyotypes (number and appearance of chromosomes in the nucleus) between different UK laboratories [Bahia *et al.*, 2002; Osborne *et al.*, 1987]

  – *Growth.* Maintaining cells in culture is non-trivial, with precise control required over temperature, CO_2 levels, the growth medium and the plating density. Common pitfalls in this area include nutrient depletion, accumulation of dead cells, contact inhibition (where over-crowding induces the inhibition of signalling processes) and cellular differentiation.

Cell lines have been widely used to investigate aberrant signalling processes in cancer. These studies have included identification of bio-markers which are predictive of drug response [Heiser *et al.*, 2011; Korkola *et al.*, 2013], structure learning of signalling networks [Bender *et al.*, 2010; Hill *et al.*, 2012a] and the identification of optimal drug combinations [Iadevaia *et al.*, 2010; Nelander *et al.*, 2008]. Coordinated efforts to obtain data over hundreds of cancer cell lines [Barretina *et al.*, 2012] provide an excellent resource for such scientific enquiry.

1.2.2 Proteomics

Once an experimental system is available, it becomes necessary to accurately quantify the molecular profiles displayed by a phenotype of interest. For cell lines, there exist several platforms which can be used for analysis of protein phosphorylation, including Western blot [Burnette, 1981], enzyme-linked immunosorbent assay (ELISA; Engvall and Perlmann [1971]), mass spectrometry [Choudhary and Mann, 2010; Harsha and Pandey, 2010; Nita-Lazar *et al.*, 2008], flow cytometry [Herzenberg *et al.*, 2002; Perez and Nolan, 2002], live cell imaging [Baker, 2010], reverse phase protein arrays (RPPA; Paweletz *et al.*

[2001]) and Luminex [Du *et al.*, 2008]. An excellent review of these methods can be found in Hill [2012a]. This thesis restricts attention to RPPA, which we discuss in detail below.

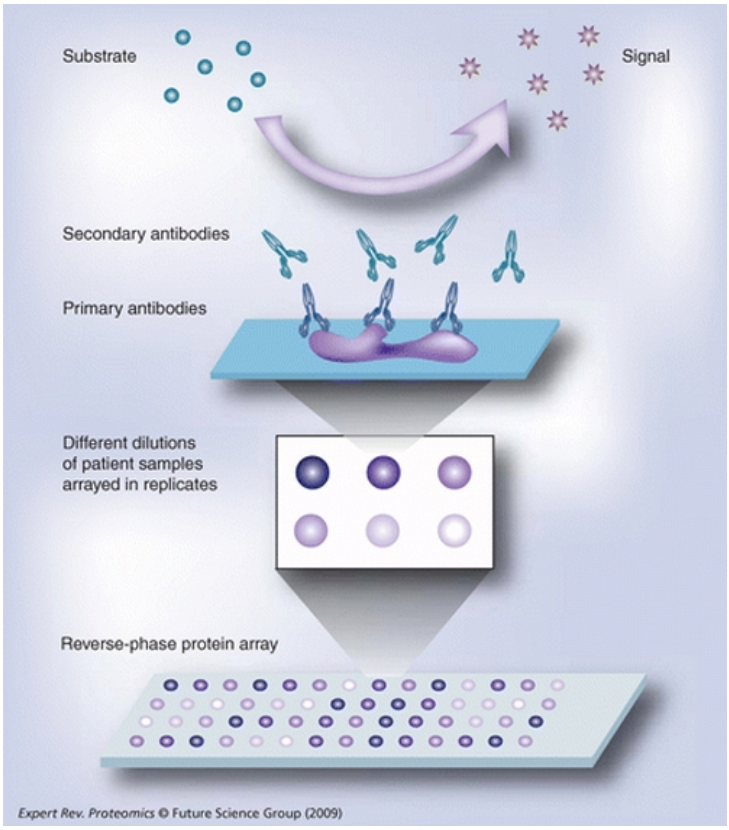
RPPA, first introduced in Paweletz *et al.* [2001], is an experimental platform for the quantitative measurement of (phospho)protein abundance in biological samples. The reverse phase format operates in three stages (Fig. 1.5(a)): Firstly, an individual test sample (e.g. a cell lysate) is immobilised in an individual array spot. Secondly, the slide is incubated with a primary antibody that binds specifically to the protein of interest. Finally, this antibody is detected using a labelled secondary antibody (as in ELISA and Western blot) and subsequent signal amplification.

The above description is deliberately simplified and in practice experimental designs will be more complex. A protein microarray slide contains many spots which are grouped into batches, with these batches arranged in a grid (Fig. 1.5(b)), allowing for multiple samples to be immobilised and tested simultaneously. Probing multiple arrays (spotted with the same lysate) with different antibodies provides the effect of generating a multiplex readout. In practice, however, bandwidth is reduced since an entire batch must be allocated to testing of a single biological sample, due to the need to obtain dilution series (discussed below), in addition to technical replicates. Full details of RPPA protocol can be found in [Hennessy *et al.*, 2010; Tibes *et al.*, 2006].

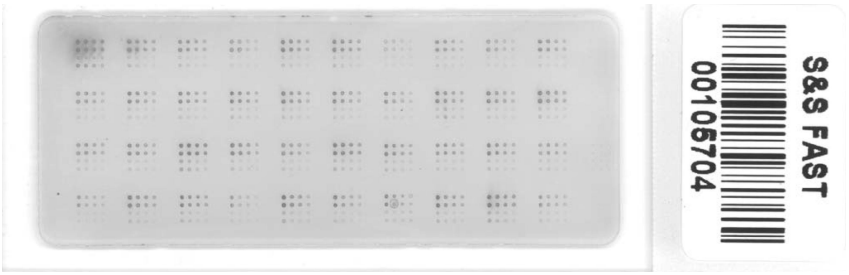
Protein concentrations and levels of phosphorylation can vary greatly, so accurate measurement over a wide *dynamic range* is required. The dynamic range of measurements is extended by diluting each sample several times (at a known dilution ratio) and spotting onto the array at each dilution step. Hence, if the protein concentration in the original undiluted sample is near saturation, it can still be detected in the diluted samples. Fig. 1.5(b) displays batches containing eight-step dilutions in duplicate. Dilution series also aid the accurate quantification of protein concentrations by increasing the effective statistical sample size. Quantification is usually carried out using response curves which relate observed signal intensities to 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* [Hu *et al.*, 2007]).

We briefly summarise the strengths and weaknesses of the RPPA platform in relation to alternative technologies:

- Strengths:
  - *Reproducibility.* Careful analysis of technical error by Hennessy *et al.* [2010] concluded that each of (i) between-batch, (ii) between-slide and (iii) between-run error variances were "low" compared to signal.
  - *High throughput.* Multiple slides can be used to probe for multiple (phospho)proteins; tens or hundreds of proteins are often measured in the same experiment, providing an advantage over low-bandwidth techniques such as flow cytometry and live cell imaging.
  - *Sensitivity.* RPPA is highly sensitive, requiring 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 [Ramaswamy *et al.*, 2005].
  - *Applicable.* Denatured lysates (proteins which have lost their three-dimensional conformation) may be assayed, allowing antibodies to bind that previously would not have been able to do so, providing an advantage over tissue microarrays.
- Weaknesses:
  - *Antibody availability.* The main limitation of RPPAs is the availability of sufficiently specific primary and secondary antibodies. Specificity is crucial for RPPA, since the signal from a spot could be due to cross-reactivity from unspecific binding and it is not possible to determine if this is the case from the data themselves. Therefore antibodies have to be carefully validated by Western blotting prior to their use in RPPA assays [Hennessy *et al.*, 2010]. The number of available validated antibodies is continuously growing, with around 150 available at the time of writing [M. D. Anderson RPPA Core Facility, 2013]. Antibodies must be designed to detect particular phosphoforms of a given protein. It is currently the case that many phosohoforms do not have a corresponding (validated) antibody, restricting the scope of RPPA analysis at present.



(a)



(b)

Figure 1.5: (a) Reverse phase protein arrays operate in three stages. Firstly, an individual test sample (e.g. a cell lysate) is immobilised in an individual array spot. Secondly, the slide is incubated with a primary antibody that binds specifically to the protein of interest. Finally, this antibody is detected using a labelled secondary antibody (as in ELISA and Western blot) and subsequent signal amplification. (b) A typical reverse-phase protein array with 40 samples shown as the 40 batches on the slide. Each batch represents one individual sample with 16 spots, which are the results of duplicates of eight-step dilutions. [(a) Reproduced from Stoevesandt *et al.* [2009]. (b) Reproduced from Telesca *et al.* [2011].]

– *Aggregate data.* Unlike flow cytometry and live cell imaging, RPPA provides no quantification of single-cell variation, since many ($\geqslant 10^3$) cell lysates are required to generate a read-out. Moreover, the population over which measurements are obtained contains cells which may not be synchronised with respect to signalling processes. Experimental protocol (Section B.6.1) partially synchronises cells by starvation followed by simultaneous stimulation, however the extent to which this strategy succeeds is unclear. Consequently, only population-average expression data is obtainable, which may compromise causal inference due to Simpsons'-type confounding.

– *Batch effects.* RPPA data are susceptible to batch effects; in particular, batch effects relating to a single slide are possible, so that a good experimental design will involve slide-slide control mechanisms.

– *Relative quantification.* Protein expression is quantified in relative terms between samples. It is therefore not possible to estimate absolute concentrations.

– *Destructive sampling.* Time course data is necessarily non-longitudinal due to the destructive observation process, leading to increased variability between temporally neighbouring samples.

– *Low frequency.* Due to manual preparation of the biological samples, it is difficult to achieve high temporal resolution using RPPA. For instance the time course data analysed in Chapter 3 have maximum time resolution of 30 minutes, although it is practically possible to sample up to 5 minute intervals. Compared to certain phosphorylation mechanisms, which can last mere seconds, this resolution may preclude identification of rapid signalling events.

The application of RPPA within cancer biology has recently been reviewed by Hill [2012a] and is reproduced below:

> RPPAs have been used to investigate cancer cell signalling, both in cancer cell lines [Hill *et al.*, 2012a; Tibes *et al.*, 2006] and in primary tumour samples [Malinowsky *et al.*, 2012; 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 [Quintás-Cardama *et al.*, 2012; Sheehan *et al.*, 2005; Telesca *et al.*, 2011] or between cancer subtypes [Boyd *et al.*, 2008; Gujral *et al.*, 2012; York *et al.*, 2012], the identification of signalling bio-markers that are predictive of response to certain anticancer agents [Boyd *et al.*, 2008], the identification of optimal drug combinations [Iadevaia *et al.*, 2010; Lavezzari *et al.*, 2012] and structure learning of signalling networks [Bender *et al.*, 2010; Pierobon *et al.*, 2012]. For further studies see, for example, [Hu *et al.*, 2007; Spurrier *et al.*, 2008; Zhang and Pelech, 2012] and references therein. RPPAs have promising utility in the development of personalised therapies [Pierobon *et al.*, 2012]; 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 [Lavezzari *et al.*, 2012], could provide information that guides the discovery and application of targeted therapies. Indeed, RPPAs have recently been involved in several clinical trials (e.g. Beasley *et al.* [2012]; Davies *et al.* [2012]; Mueller *et al.* [2010]).

1.3 Chemical Background

In this Section we formalise the idea of a system of chemical reactions, describe convenient approximations to the dynamics as the volume of the system increases and briefly survey the state-of-the-art statistical approaches to inference for such systems

1.3.1 Continuous Time Markov Processes

Let $N_i(t) \in \mathbb{N}_0$ denote the number of molecules of protein species $\mathcal{X}_i$, $i \in \mathcal{P} = \{1, \ldots, P\}$, present at time $t \in \mathbb{R}_+$. Then $\boldsymbol{N}(t) = [N_1(t) \ldots N_P(t)]$ is assumed to characterise the state of the system at time t.

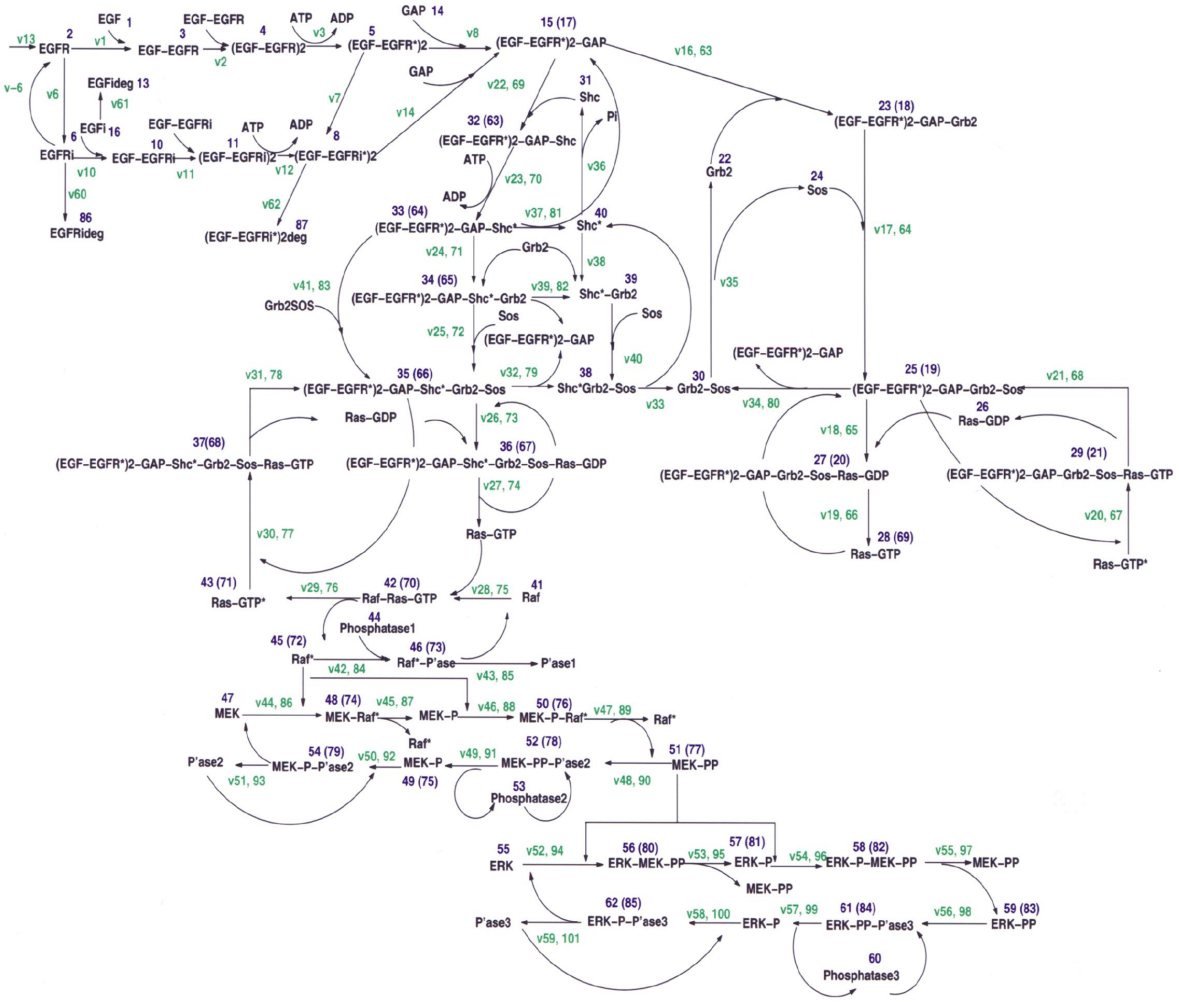


Figure 1.6: Chemical reaction graph G for the MAPK signalling pathway; reproduced from Schoeberl *et al.* [2002]. [Reaction rates vi are shown in green and reactants are shown in black. Hyphens are used to indicate chemical complexes and arrows indicate the reaction topology.]

Definition 6 (Chemical reaction graph)**.** *A* chemical reaction graph *is a system of v chemical reactions $\mathcal{R}_1, \ldots, \mathcal{R}_v$ with rate constants $k_1, \ldots, k_v$ and reaction coefficients $p_{ij}, q_{ij} \in \mathbb{N}_0$:*

$$\begin{aligned}
\mathcal{R}_1 &: p_{11}\mathcal{X}_1 + p_{12}\mathcal{X}_2 + \cdots + p_{1P}\mathcal{X}_P \xrightarrow{k_1} q_{11}\mathcal{X}_1 + q_{12}\mathcal{X}_2 + \cdots + q_{1P}\mathcal{X}_P \\
\mathcal{R}_2 &: p_{21}\mathcal{X}_1 + p_{22}\mathcal{X}_2 + \cdots + p_{2P}\mathcal{X}_P \xrightarrow{k_2} q_{21}\mathcal{X}_1 + q_{22}\mathcal{X}_2 + \cdots + q_{2P}\mathcal{X}_P \\
&\vdots \\
\mathcal{R}_v &: p_{v1}\mathcal{X}_1 + p_{v2}\mathcal{X}_2 + \cdots + p_{vP}\mathcal{X}_P \xrightarrow{k_v} q_{v1}\mathcal{X}_1 + q_{v2}\mathcal{X}_2 + \cdots + q_{vP}\mathcal{X}_P
\end{aligned}$$

Here the reaction coefficients p_{ij}, q_{ij} are non-negative integers, since only entire molecules may react. Collecting together reaction coefficients produces matrices $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{N}_0^{v \times P}$ whose transposed difference $\boldsymbol{S} = (\boldsymbol{Q} - \boldsymbol{P})^T \in \mathbb{N}_0^{P \times v}$ is known as the *stoichiometry* matrix. The ith column $\boldsymbol{s}_i$ of $\boldsymbol{S}$ is then the state change vector for reaction $\mathcal{R}_i$, quantifying the net change in protein quantities as a result of reaction $\mathcal{R}_i$ occurring. Fig. 1.6, reproduced from Schoeberl *et al.* [2002], contains a chemical reaction graph representation for the MAPK pathway. Note that the use of "graph" here is non-standard, motivated by a graphical representation of kinase-substrate reactions which we will exploit in Chapter 3.

Definition 7 (Continuous time Markov process)**.** *The continuous time stochastic process $\boldsymbol{N}(t)$ is* Markov

if for all $\boldsymbol{n}, \boldsymbol{n}' \in \mathbb{N}_0^P$ *there exists* $q_{\boldsymbol{n},\boldsymbol{n}'} \in \mathbb{R}_+$ *such that*

$$\mathbb{P}(\boldsymbol{N}(t+\delta t) = \boldsymbol{n}' | \boldsymbol{N}(t) = \boldsymbol{n}) = q_{\boldsymbol{n},\boldsymbol{n}'}\delta t + o(\delta t). \tag{1.1}$$

The $q_{\boldsymbol{n},\boldsymbol{n}'}$ *are known as* transition rates.

Definition 8 (Mass action kinetics)**.** *Under (stochastic)* mass action kinetics *the state vector* $\boldsymbol{N}$ *is a continuous time Markov process with transition rates given by*

$$q_{\boldsymbol{n},\boldsymbol{n}'} = \sum_i \mathbb{I}\{\boldsymbol{n} - \boldsymbol{n}' = \boldsymbol{s}_i\} h_i(\boldsymbol{n}) \tag{1.2}$$

where

$$h_i(\boldsymbol{N}) = k_i \prod_j \binom{N_i}{p_{ij}} \tag{1.3}$$

is the hazard *of reaction* $\mathcal{R}_i$ *occurring.*

Mass action kinetics assumes a well-mixed chemical population and sufficiently large numbers of reactants; these assumptions must be carefully assessed in real biological systems [Sayikli and Bagci, 2011]. Such dynamics are easily simulated using, for instance, the Gillespie algorithm [Wilkinson, 2006]. Using modern parallel processing technology, forward simulation is possible at computational complexity $\mathcal{O}(t \log(P))$ where t is the duration of the simulation and P is the number of biochemical species [Li and Petzold, 2008].

1.3.2 Chemical Langevin Equation

Inference for continuous time Markov processes from discrete, noisy data is extremely challenging [Wilkinson, 2006]. One popular solution is to approximate the discrete variables N_i by continuous variables ΩX_i where Ω is the volume of the system and X_i is the density or concentration of $\mathcal{X}_i$. Two well known approximations of this form are the chemical Langevin equation (CLE) and the linear noise approximation (LNA); we derive both in the following Sections.

Theorem 1 (Chemical Langevin equation)**.** *The continuous time Markov process* $\boldsymbol{N}(t)$ *can be approximated by* $\Omega\boldsymbol{X}(t)$ *where* $\boldsymbol{X} \in \mathbb{R}_+^P$ *satisfies the stochastic differential equation*

$$d\boldsymbol{X} = \sum_i \overline{h}_i(\boldsymbol{X})\boldsymbol{s}_i dt + \frac{1}{\sqrt{\Omega}} \sum_i \sqrt{\overline{h}_i(\boldsymbol{X})}\boldsymbol{s}_i dB_i. \tag{1.4}$$

where $\overline{h}_i(\boldsymbol{X}) = \lim_{\Omega\to\infty} \Omega^{-1} h_i(\lceil \Omega \boldsymbol{X} \rceil)$.

Sketch Proof: Consider a time interval $I = [t, t+\delta t)$ where δt is sufficiently small that hazards $h_i(\boldsymbol{N}(s))$ are approximately constant for $s \in I$. Then the number R_i of reactions $\mathcal{R}_i$ which occur during the interval may be modelled using a Poisson random variable $R_i \approx\sim Po(\lambda_i)$ with mean $\lambda_i = h_i(\boldsymbol{N}(t))\delta t$. The diffusion approximation proceeds by using instead a Gaussian $R_i \approx\sim \mathcal{N}_i(\lambda_i, \lambda_i)$ whose mean and variance are chosen to match those of the Poisson distribution. Thus we obtain

$$\begin{aligned} \boldsymbol{N}(t+\delta t) - \boldsymbol{N}(t) &\approx\sim \sum_i \mathcal{N}_i(h_i(\boldsymbol{N}(t))\delta t, h_i(\boldsymbol{N}(t))\delta t)\boldsymbol{s}_i & (1.5)\\ &= \sum_i h_i(\boldsymbol{N}(t))\boldsymbol{s}_i\delta t + \sum_i \sqrt{h_i(\boldsymbol{N}(t))}\boldsymbol{s}_i\mathcal{N}_i(0,\delta t). & (1.6) \end{aligned}$$

Close to the thermodynamic limit ($\Omega^{-1}h_i(\boldsymbol{N}) \approx \overline{h}_i(\boldsymbol{X})$) we may rewrite Eqn. 1.6 as

$$\boldsymbol{X}(t+\delta t) - \boldsymbol{X}(t) \approx\sim \sum_i \overline{h}_i(\boldsymbol{X}(t))\boldsymbol{s}_i\delta t + \frac{1}{\sqrt{\Omega}} \sum_i \sqrt{\overline{h}_i(\boldsymbol{X}(t))}\boldsymbol{s}_i\mathcal{N}_i(0,\delta t). \tag{1.7}$$

Taking $\delta t \to 0$ we then arrive at the chemical Langevin equation (CLE).

The Gaussian approximation to a Poisson density relies on the parameters $\lambda_i = h_i\delta t$ being sufficiently large. However the initial assumption of constant hazard rate over the interval I required that the width

δt of the interval is small. Thus it is not clear *a priori* whether such a regime exists. However it has been proven that the CLE is a good approximation to the stochastic dynamics whenever the system is sufficiently close to the thermodynamic limit (Section 1.3.4) [Gillespie, 2009; Wallace *et al.*, 2012]. □

The CLE relaxes the assumption of discrete state space whilst preserving important behavioural features of the original continuous time Markov process, including conserved quantities such as total molecular concentrations. However the quality of the CLE approximation may deteriorate in situations where low concentrations are encountered, in which case the CLE underestimates the effect of stochastic fluctuations. Inference using the CLE has been studied by [Golightly and Wilkinson, 2011] who exploit efficient particle MCMC sampling strategies.

1.3.3 Linear Noise Approximation

Inference for SDEs remains challenging despite several recent advances in this area (e.g. Kalogeropoulos *et al.* [2010]; Papaspiliopoulos *et al.* [2012] and references therein), since in general the likelihood function is unavailable in closed form [Wilkinson, 2006]. An attractive approach is to develop a closed form approximation to the CLE; the LNA which we describe below is one well known example.

Noting that the CLE (Eqn. 1.4) differs to the macroscopic rate equation (Eqn. 1.12) by a term of order $1/\sqrt{\Omega}$, we take the ansatz $\boldsymbol{X}(t) \approx \boldsymbol{\mu}(t) + \boldsymbol{\xi}(t)/\sqrt{\Omega}$ where $\boldsymbol{\mu}$ is the deterministic solution to the macroscopic rate equation.

Theorem 2 (Linear noise approximation [van Kampen, 1976])**.** *The solution $\boldsymbol{X}(t)$ of the CLE may be approximated by $\boldsymbol{\mu}(t) + \boldsymbol{\xi}(t)/\sqrt{\Omega}$ where $\boldsymbol{\mu}$ is the deterministic solution to the macroscopic rate equation*

$$\frac{d\boldsymbol{\mu}}{dt} = \sum_i \overline{h}_i(\boldsymbol{\mu})\boldsymbol{s}_i, \quad \boldsymbol{\mu}(0) = \boldsymbol{x}_0. \tag{1.8}$$

and $\boldsymbol{\xi}$ satisfies the SDE

$$d\boldsymbol{\xi} = \sum_i D_{\boldsymbol{\mu}}\overline{h}_i(\boldsymbol{\xi})\boldsymbol{s}_i dt + \sum_i \sqrt{\overline{h}_i(\boldsymbol{\mu})}\boldsymbol{s}_i dB_i, \quad \boldsymbol{\xi}(0) = \mathbf{0} \tag{1.9}$$

where $D_{\boldsymbol{\mu}}\overline{h}_i(\boldsymbol{\xi}) = d\overline{h}_i(\boldsymbol{\mu})/d\boldsymbol{\mu} \cdot \boldsymbol{\xi}$ denotes the directional derivative of $\overline{h}_i$, evaluated at $\boldsymbol{\mu}$, in the direction $\boldsymbol{\xi}$.

Sketch Proof: Using a linear expansion of the hazards $\overline{h}_i(\boldsymbol{\mu}(t) + \boldsymbol{\xi}(t)/\sqrt{\Omega})$ about $\boldsymbol{\mu}(t)$ results in

$$\overline{h}_i\left(\boldsymbol{\mu}(t) + \frac{\boldsymbol{\xi}(t)}{\sqrt{\Omega}}\right) = \overline{h}_i(\boldsymbol{\mu}(t)) + \frac{1}{\sqrt{\Omega}}\sum_j f_{ij}(t)\xi_j(t) + \mathcal{O}\left(\frac{1}{\Omega}\right) \tag{1.10}$$

where $f_{ij} = d\overline{h}_i(\boldsymbol{\mu})/d\mu_j$. Upon substitution of our ansatz $\boldsymbol{X}(t) \approx \boldsymbol{\mu}(t) + \boldsymbol{\xi}(t)/\sqrt{\Omega}$ into the CLE (Eqn. 1.4) and using Eqn. 1.10 we obtain, up to $\mathcal{O}(1/\sqrt{\Omega})$, Eqn. 1.9. □

The law of $\boldsymbol{X}$ under the LNA is encoded in the solution to Eqn. 1.9. Recently Wallace *et al.* [2012] described how to solve Eqn. 1.9 exactly; specifically, $\boldsymbol{\xi}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(t))$ is Gaussian in distribution where covariance matrix $\boldsymbol{\Sigma}$ satisfies the following system of linear ODEs

$$\frac{d\boldsymbol{\Sigma}}{dt} = \boldsymbol{S}\boldsymbol{f}\boldsymbol{\Sigma} + (\boldsymbol{S}\boldsymbol{f}\boldsymbol{\Sigma})^T + \boldsymbol{S}\mathrm{diag}(\overline{\boldsymbol{h}}(\boldsymbol{\mu}))\boldsymbol{S}^T \tag{1.11}$$

subject to the initial condition $\boldsymbol{\Sigma}(0) = \mathbf{0}$, where $\mathrm{diag}(\boldsymbol{h})$ represents the diagonal matrix with diagonal equal to $\boldsymbol{h}$. Thus we may augment the macroscopic rate equation (Eqn. 1.8) with the covariance equations (Eqn. 1.11) and jointly solve the system in order to obtain an exact distribution for the LNA of $\boldsymbol{X}(t)$.

The LNA has recently received attention from the statistical and applied mathematics communities: Komorowski *et al.* [2011] proposed using the LNA to approximate the Fisher information matrix for stochastic chemical kinetics, thereby investigating sensitivity, robustness and identifiability of chemical systems. Mugler *et al.* [2011] reverse-engineered biochemical networks which process signals according to some given functional form; here the LNA is employed to obtain a tractable statistical framework.

In a similar way Finkenstädt *et al.* [2013]; Komorowski *et al.* [2009] used the LNA to uncover the rate parameters governing the expression of a single gene. Furthermore the theory has been extended in several directions: Pahlajani *et al.* [2011] investigated extensions to the LNA for cases where reaction rates induce separable time scales, overcoming potential stiffness of the associated ODEs. Challenger *et al.* [2012] introduced spatial heterogeneity by extending the LNA to compartmentalised models of chemical interaction. Stathopoulos and Girolami [2012] demonstrated how to exploit manifold MCMC techniques for efficient inference under the LNA.

1.3.4 Thermodynamic Limit

In situations where the volume Ω of the system is large, concentrations of molecular species are not too low, and the system is approximately well mixed, it may be desirable to model the process $\boldsymbol{X}(t)$ as fully deterministic. The *thermodynamic limit* allows molecular quantities N_i and the system volume Ω to approach infinity together in such a way that concentrations $X_i = N_i/\Omega$ remain constant. In this limit the concentrations may be shown to satisfy the continuous solution of the macroscopic rate equation

$$\frac{d\boldsymbol{X}}{dt} = \sum_i \overline{h}_i(\boldsymbol{X})\boldsymbol{s}_i, \quad \boldsymbol{X}(0) = \boldsymbol{x}_0. \tag{1.12}$$

Mass action kinetics do not permit analytic solution, meaning that exact inference is typically facilitated using forward-simulation ("likelihood free") approaches, e.g. [Chen *et al.*, 2009; Toni *et al.*, 2009]. Nevertheless such simulation-intensive approaches do not lend themselves to rapid, interactive inference.

In cellular biology the topology of chemical reaction graphs may be highly structured with an emphasis on *motifs* which confer certain dynamical properties such as stability, feedback, or switch-like behaviour [Alon, 2007]. For several such motifs there exist a number of well-studied analytic approximations to the dynamics which may assist in modelling efforts. Below we describe some examples from enzyme kinetics which are central to this thesis.

Example 9 (Michaelis-Menten kinetics)**.** *Michaelis-Menten kinetics is an approximation to mass actions kinetics which describes the conversion of a substrate $\mathcal{X}_S$ into a product $\mathcal{X}_P$ under the catalytic activity of an enzyme $\mathcal{X}_E$ [Michaelis and Menten, 1913]. Specifically we seek to approximate the dynamics arising from the chemical reaction motif*

$$\mathcal{X}_S + \mathcal{X}_E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathcal{X}_E\mathcal{X}_S \overset{k_2}{\to} \mathcal{X}_E + \mathcal{X}_P \tag{1.13}$$

where standard shorthand notation encodes a system of $v = 3$ chemical reactions; see Def. 6. Under mass action kinetics (below) the dynamical system corresponding to Eqn. 1.13 does not permit closed form solution:

$$\frac{dX_S}{dt} = -k_1 X_E X_S + k_{-1} X_{ES} \tag{1.14}$$

$$\frac{dX_E}{dt} = -k_1 X_E X_S + (k_{-1} + k_2) X_{ES} \tag{1.15}$$

$$\frac{dX_{ES}}{dt} = k_1 X_E X_S - (k_{-1} + k_2) X_{ES} \tag{1.16}$$

$$\frac{dX_P}{dt} = k_2 X_{ES} \tag{1.17}$$

Michaelis-Menten kinetics state that the rate of production of P is given approximately by

$$\frac{dX_P}{dt} \approx \frac{V X_{E_0} X_S}{X_S + K} \tag{1.18}$$

where X_{E_0} denotes the total concentration of enzyme (including molecules involved in the complex X_{ES}), V is the maximal reaction rate *and K is a* Michaelis-Menten *parameter.*

Eqn. 1.18 is an attractive alternative to the system of Eqns. 1.14-1.17 since only two parameters are required to characterise the dynamics. Moreover unlike mass action kinetics (Eqns. 1.14-1.17), Michaelis-Menten kinetics confer an analytic solution expressed in terms of the Lambert $\mathcal{W}$ function [Schnell and

Mendoza, 1997]

$$X_P(t) \approx X_{S_0} K \mathcal{W}\left(\frac{X_{S_0}}{K_M}\exp\left(\frac{-Vt + X_{S_0}}{K_M}\right)\right), \tag{1.19}$$

although Eqn. 1.19 is itself rarely used due to a large computational burden associated with evaluation of the $\mathcal{W}$ function.

Example 10 (Derivation of Michaelis-Menten kinetics)**.** *The Michaelis-Menten approximation (Eqn. 1.18) may be obtained from mass action kinetics (Eqn. 1.13) under two alternative assumptions:*

1. Equilibrium assumption: *When the substrate is in instantaneous chemical equilibrium with the complex, i.e. $dX_S/dt \approx 0$, it follows from Eqn. 1.15 that $k_1 X_S X_E = k_{-1} X_{ES}$ and hence $X_E = k_{-1}X_{ES}/(k_1 X_S)$. Substituting this into the conservation equation $X_E + X_{SE} = X_{E_0}$ we obtain $X_{ES} = X_{E_0} X_S/(X_S + K_d)$ where $K_d = k_{-1}/k_1$.*

2. Quasi-steady-state assumption: *When the concentration of the intermediate complex does not change on the time-scale of product formation, i.e. $dX_{ES}/dt \approx 0$, it follows from Eqn. 1.15 that $k_1 X_E X_S = (k_{-1} + k_2)X_{ES}$ and hence $X_E = (k_{-1} + k_2)X_{ES}/(k_1 X_S)$. Substituting this into the equation for conservation of enzyme $X_E + X_{ES} = X_{E_0}$ we obtain $X_{ES} = X_{E_0} X_S/(X_S + K_m)$ where $K_m = (k_{-1} + k_2)/k_1$.*

Since the rate of product formation is proportional to X_{ES}, the Michaelis-Menten equation (Eqn. 1.18) follows from either assumption. The equilibrium assumption is satisfied when $k_2 << k_{-1}$, whereas the quasi-steady-state assumption is satisfied when $X_{E_0} << X_{S_0} + K_m$. There has been much work to understand the applicability and limitations of the Michaelis-Menten approximation and we refer the reader to Sanft et al. [2011].

In applications the Michaelis-Menten approximation has been used to describe, amongst other processes, the behaviour of gene regulation by transcription factors [Cantone *et al.*, 2009] and the kinetics of protein phosphorylation networks [Xu *et al.*, 2010]. The formulation easily extends to include multiple enzymes, multiple substrates, substrate and product inhibition, linear, hyperbolic and parabolic inhibitors of enzyme activity and co-operative binding of substrate. Chapter 4 of Leskovac [2003] describes automatic algorithms that facilitate the construction of appropriate rate equations from mass action descriptions. In Chapter 3 we exploit Michaelis-Menten kinetics with linear inhibition to facilitate inference of protein signalling networks from time course data on protein phosphorylation processes. Linear inhibition is given particular attention in Chapter 5 of Leskovac [2003].

Example 11 (Goldbeter-Koshland kinetics)**.** *The* Goldbeter-Koshland *formula [Goldbeter and Koshland, 1981] describes the concentration of a chemical species subject to enzymes with opposite effects:*

$$\mathcal{X}_S + \mathcal{X}_E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathcal{X}_{ES} \overset{k_2}{\to} \mathcal{X}_E + \mathcal{X}_P \tag{1.20}$$

$$\mathcal{X}_P + \mathcal{X}_{E'} \underset{k'_{-1}}{\overset{k'_1}{\rightleftharpoons}} \mathcal{X}_{E'P} \overset{k'_2}{\to} \mathcal{X}_{E'} + \mathcal{X}_S \tag{1.21}$$

where, for example, $\mathcal{X}_S$ and $\mathcal{X}_P$ may represent unphosphorylated and phosphorylated forms of a protein, $\mathcal{X}_E$ a kinase and $\mathcal{X}_{E'}$ a phosphatase. By equating two Michaelis-Menten approximations via $dX_P/dt = 0$ we obtain

$$\frac{V X_E X_S}{X_S + K} - \frac{V' X_{E'} X_P}{X_P + K'} = 0 \tag{1.22}$$

Goldbeter and Koshland manipulated Eqn. 1.22 to obtain the formula

$$X_P = X_{S_0}\left(1 - \frac{2v'J}{B + \sqrt{B^2 - 4(v - v')v'J}}\right) \tag{1.23}$$

where $X_{S_0} = X_S + X_P$ is the total amount of the chemical species, $v = V X_E$, $v' = V' X_{E'}$, $J = K/X_{S_0}$, $J' = K'/X_{S_0}$ and $B = v - v' + J'v + Jv'$. In their original paper, Goldbeter and Koshland showed that

this solution can be exquisitely sensitive to the parameters J and J'; so called ultra-sensitivity. *Oates et al. [2012] exploited Goldbeter-Koshland kinetics to facilitate inference of protein signalling networks from equilibrium data on protein phosphorylation processes.*

1.4 Statistical Background

In this Section we introduce the concept of a causal graphical model, which will be the central object of interest in this thesis. For completeness, we provide a short introduction to causal theory and make explicit the causal assumptions underlying this work. We then expound these ideas in the context of protein signalling and motivate the statistical challenge of causal network inference in this setting.

1.4.1 Graphical Models

A graphical model is a collection $\boldsymbol{X}$ of random variables accompanied by a graph G describing a factorisation of the joint density $p_{\boldsymbol{X}}(\boldsymbol{x})$. There are many types of graphical model, including Bayesian networks, maximal ancestral graphs [Richardson and Spirtes, 2002], Gaussian graphical models [Wainwright and Jordan, 2008], factor graphs [Kschischang *et al.*, 2001] and chain event graphs (Smith and Anderson [2008]; see also references therein). Graphical models have become increasingly popular in systems biology due to their ability to succinctly describe many interactions within a complex, multivariate stochastic process. For this thesis we restrict attention to Bayesian networks.

A directed acyclic graph (DAG) G comprises of a set $\mathcal{P}$ of vertices and a set $E \subset \mathcal{P} \times \mathcal{P}$ of directed edges, such that G contains no directed cycles. We write $G_p \subset \mathcal{P}$ for the parents of vertex $p \in \mathcal{P}$ according to G, so that formally the graph G factorises as $G = G_1 \times \cdots \times G_P$ where $\mathcal{P} = \{1, \ldots, P\}$.

Definition 9. *A multivariate random variable $\boldsymbol{X} = (X_1, \ldots, X_P)$ is said to be a* Bayesian network *with respect to G if its joint density factorises as*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{p \in \mathcal{P}} p_{X_p|\boldsymbol{X}_{G_p}}(x_p) \tag{1.24}$$

and such that no proper sub-graph of G satisfies Eqn. 1.24 [Pearl, 2009, p. 15].

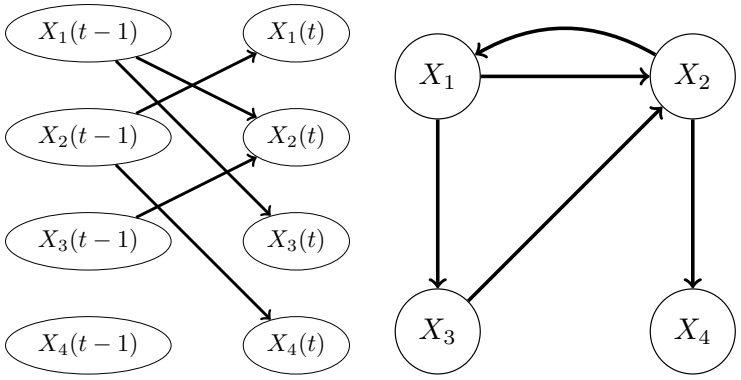
Here we have used the notation that $\boldsymbol{X}_{G_p}$ contains only the components of $\boldsymbol{X}$ which correspond to the parents G_p of variable p. Note that Bayesian networks do not necessarily imply a causal structure among the variables, since any random variable $\boldsymbol{X} = (X_1, X_2)$ where X_1 and X_2 are not independent is necessarily a Bayesian network with respect to both the graph $X_1 \to X_2$ and the graph $X_2 \to X_1$ (see the next Section).

Efficient inference in general (discrete) Bayesian networks relies on the *belief propagation* algorithm [Pearl, 1982] and related extensions [Kschischang *et al.*, 2001]. In Chapter 4 we exploit belief propagation for efficient inference over a hierarchical system of DBNs.

Bayesian networks have been widely used to model biological processes at steady state, including gene regulation [Maathuis *et al.*, 2010] and protein signalling [Sachs *et al.*, 2005]. However the framework is (naively) limited by the requirement that no directed cycles are permitted; in particular this restricts their utility in protein signalling systems, where feedback control mechanisms are crucial [Avraham and Yarden, 2011]. One solution is to model the temporal evolution of the system using Bayesian networks, thereby allowing for feedback regulation under a temporal constraint (arrows can only point forward in time).

Example 12 (Dynamic Bayesian networks)**.** *In a* dynamic Bayesian network *(DBN) the random variable $\boldsymbol{X}(t)$ has an explicit (discrete) time index $t \in \mathbb{N}_0$. This thesis restricts attention to DBNs which additionally satisfy the first order Markov assumption $\boldsymbol{X}(t) \perp\!\!\!\perp \{\boldsymbol{X}(t-\tau) : \tau \geqslant 2\} | \boldsymbol{X}(t-1)$. These conditional independence relationships are sufficient to guarantee $\boldsymbol{X}(t)$ is a Bayesian network with respect to some* time-slice *graph (Fig. 1.7(a)). We further restrict attention to* feed forward *DBNs, where $X_p(t) \perp\!\!\!\perp X_q(t) | \boldsymbol{X}(t-1)$ for all $p \neq q$; in other words there are no within-time-slice edges. The conditional independence relations underlying feed forward DBNs are conveniently summarised as a (static) network G with exactly P vertices (Fig. 1.7(b)); note that this latter network need not be acyclic.*

For the remainder of this thesis we use DBN to refer only to "first order Markov, feed forward" DBNs, though this terminology is non-standard. DBNs have emerged as popular tools for the analysis



(a) DBN representation. (b) "static" representation.

Figure 1.7: Dynamic Bayesian networks (DBNs). (a) A "time-slice" dynamic Bayesian network (DBN) is a bipartite graph with vertices corresponding to variables at successive time points. (b) The corresponding "static" representation with exactly one vertex for each variable.

of multivariate time course data due to (i) the fact that no acyclicity assumption is required on the (static) network, (ii) the full network topology is identifiable from observational data (see Ex. 13), (iii) computational tractability results from a factorization of the likelihood function over variables $p \in \mathcal{P}$, and (iv) in special cases there exist closed form expression for Bayes factors [Hill *et al.*, 2012a].

1.4.2 Causal Inference

Variable selection approaches have been widely studied within the statistical literature where they typically serve two related yet distinct goals; (i) improvement of prediction performance (e.g. in regression or classification); (ii) selection of variables which actually drive the response in a control-theoretic sense. The distinction between variable selection problems (i) and (ii) mirrors the general distinction between regression and causal inference, as discussed extensively by Pearl [2009] and others. Problem (i) is classical and has been studied extensively. However problem (ii) is attracting increasing attention since in many settings the results of variable selection are used to prioritise interventional experiments. For instance in biological applications the selected variables may be subjected to knock-down or knock-out treatments. In this thesis we are principally concerned with problem (ii).

In this Section we formalise a theory of inferred causality and briefly discuss some statistical techniques for estimating causal variables from data. In particular we define a *causal Bayesian network* which reconciles Definition 12 with a causal calculus. We restrict attention to DAGs G and random variables $\boldsymbol{X}$. For convenience below we use $\boldsymbol{X}$ to denote both the random variable and its law.

Assumption 1 (Causal sufficiency)**.** *The set of measured variables $\boldsymbol{X}$ includes all of the common causes of pairs (X_p, X_q). i.e. there are no unobserved confounders.*

The remainder of this thesis will assume causal sufficiency; in Chapter 3 this is at the level of the chemical reaction graph G and in Chapter 4 this is at the level of the biological network N (see Section 1.4.4 for a discussion of this distinction). For protein signalling both assumptions are difficult to justify on anything but pragmatic grounds, since we typically can only ever access a small fraction of the relevant molecular species. Moreover, as discussed in Section 1.1.1, protein signalling is itself embedded in wider signalling processes such as genetic regulation. However, since biological networks $N(G)$ are coarse summaries of chemical reaction graphs G, a causal sufficiency assumption on the latter is weaker. An extended discussion of this point is reserved for Section 1.4.4 but recent work on latent variables including Colombo *et al.* [2012] and Chandrasekaran *et al.* [2012] may help to relax the causal sufficiency assumption.

We now proceed to establish sufficient conditions for the identification of causation from data. For some of the definitions below it will be convenient to refer to a variable X_i simply by its index i.

Definition 10. *A path is a sequence of consecutive edges (of any directionality) [Pearl, 2009, p. 12].*

Definition 11. *A path ρ in G is* d-separated *by a set A of vertices if and only if either (i) ρ contains a chain $i \to m \to j$ or fork $i \leftarrow m \to j$ such that $m \in A$, or (ii) ρ contains a collider $i \to m \leftarrow j$ such*

that $m \notin A$ and no descendant of m is in A. A set A is said to d-separate *sets B and C if and only if A d-separates all paths from vertices in B to vertices in C [Pearl, 2009, p. 16].*

Definition 12 (Markov)**.** *$\boldsymbol{X}$ is* Markov *with respect to G if $\boldsymbol{X}_B \perp\!\!\!\perp \boldsymbol{X}_C | \boldsymbol{X}_A$ whenever A d-separates B and C in G, for all disjoint sets A, B, C [Pearl, 2009, p. 16].*

Note that Def. 12 is sometimes referred to as the "global" Markov property. In this Chapter we use the conditional independence notation $\perp\!\!\!\perp$ due to Dawid [1980].

Definition 13 (Faithful)**.** *$\boldsymbol{X}$ is* faithful *to G if A d-separates B and C in G whenever $\boldsymbol{X}_B \perp\!\!\!\perp \boldsymbol{X}_C | \boldsymbol{X}_A$, for all disjoint sets A, B, C [Pearl, 2009, p. 48].*

Note that Def. 13 is sometimes referred to as "stability", "DAG-isomorphism" or "perfect-mapness".

Definition 14 (Causal graph)**.** *The* causal graph *G for a set of random variables $\boldsymbol{X}$ is constructed by drawing directed arrows from nodes to their direct effects.*

Def. 14 is deliberately imprecise, since the philosophical foundations of causality are beyond the scope of this thesis. The reader is invited to refer to Section 2.3 of Pearl [2009], which formulates a pragmatic definition of "direct effects" via Occam's Razor.

Assumption 2. *The causal graph G for the random variables $\boldsymbol{X}$ satisfies (i) $\boldsymbol{X}$ is Markov with respect to G, and (ii) $\boldsymbol{X}$ is faithful to G.*

An intuitive approach to causal inference, based on Assumption 2, is to identify all graphs G for which data support the Markov and faithfulness conditions. However these two conditions together are insufficient, in general, to determine a unique graph structure. For example the graphs $X_1 \to X_2$ and $X_2 \to X_1$ both satisfy the Markov and faithfulness conditions for any random variable $\boldsymbol{X} = (X_1, X_2)$ where X_1 and X_2 are not independent. This is an example of non-identifiability known as *observational equivalence*.

Definition 15 (Observational equivalence)**.** *Two graphs are* observationally equivalent *if every probability distribution $\boldsymbol{X}$ which is Markov with respect to one of the graphs is also Markov with respect to the other [Pearl, 2009, p. 19].*

Theorem 3. *Two graphs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures [Pearl, 2009, p. 19]. (The* skeleton *of a graph is the undirected graph obtained by removing all arrowheads. A* v-structure *is two converging arrows whose tails are not connected by an arrow.)*

Recent research has asked under what conditions is it possible to obtain a unique causal graph: Peters *et al.* [2011] proposed to replace the faithfulness assumption by an assumption of an identifiable model class. Hauser and Bühlmann [2012] proposed a refined partition of the space of DAGs by considering identifiability under intervention. Peters and Bühlmann [2012] proved that Gaussian structural equation models are uniquely identifiable under an assumption of equal error variances. This thesis focusses primarily on DBNs, which benefit from a uniqueness result:

Example 13. *For a DBN G, if G' is observationally equivalent to G then we must have $G = G'$.*

Proof. Recall that this thesis restricts attention only to feed forward DBNs. Thus given the skeleton of G we can add all of the arrowheads by requiring that arrows point forward in time. Thus if G and G' have the same skeletons, they must be the same directed graph. □

In settings where it is not possible to uniquely identify the causal graph from observational data, one approach is to actively manipulate the data-generating process:

Definition 16 (Intervention)**.** *An intervention on a subset A of variables has the effect of fixing these variables to a (known) constant value $\boldsymbol{x}_A$. The resulting probability density is written $p_{\boldsymbol{X}}(\boldsymbol{x} | do(\boldsymbol{X}_A = \boldsymbol{x}_A))$ [Pearl, 2009, Sec. 3.2].*

It is possible to define more general interventions which change the nature of the probabilistic dependence (e.g. Eaton and Murphy [2007]), but we do not explore these ideas in this thesis. Moreover, we restrict attention to a particular class of causal graphical models known as causal Bayesian networks:

Definition 17 (Causal Bayesian network Pearl [2009] p. 23)**.** *A Bayesian network* $\boldsymbol{X}$ *with respect to* G *is* causal *if, for any intervention* $do(\boldsymbol{X}_A = \boldsymbol{x}_A)$ *the following factorisation holds:*

$$p_{\boldsymbol{X}}(\boldsymbol{x}|do(\boldsymbol{X}_A = \boldsymbol{x}_A)) = \prod_{p \notin A} p_{X_p|\boldsymbol{X}_{G_p}}(x_p) \tag{1.25}$$

Example 14 (Causal DBN)**.** *A DBN* $\boldsymbol{X}$ *is causal with respect to* G *if, for any intervention* $do(\boldsymbol{X}_A(t-1) = \boldsymbol{x}_A)$ *the following factorisation holds:*

$$p_{\boldsymbol{X}(t)}(\boldsymbol{x}(t)|do(\boldsymbol{X}_A(t-1) = \boldsymbol{x}_A)) = \prod_{p \notin A} p_{X_p(t)|\boldsymbol{X}_{G_p}(t-1)}(x_p(t)) \tag{1.26}$$

The *do* operator tells us how, according to the DAG G, the data-generating distribution will change under intervention. In particular this allows us in Chapters 3 and 4 to integrate both observational and interventional data into inference using a unified statistical framework consistent with a theory of causation. The use of interventional data may help to practically discriminate between candidate causal DAGs which are observationally equivalent [Hauser and Bühlmann, 2012].

There exist a number of generic techniques for identifying causal Bayesian networks (up to observational equivalence) from data, including the PC algorithm [Spirtes *et al.*, 2000], the Fast Causal Inference (FCI) algorithm [Spirtes *et al.*, 2000; Zhang, 2008] and the Really Fast Causal Inference (RFCI) algorithm [Colombo *et al.*, 2012]. It may also be desired to estimate the causal effect of an unseen intervention [Pearl, 2009], which may be achieved using the Interventional-calculus when the DAG is Absent (IDA) algorithm [Maathuis *et al.*, 2010]. Each of these algorithms are based on repeated tests of conditional independence. This has two main advantages; (i) each algorithm is provably consistent, and (ii) each algorithm is generic, not depending on any particular parametric formulation. However this approach to inference may be sub-optimal when model-specific information is available, for example when the data-generating system is a DBN [Hill *et al.*, 2012b].

Causal Bayesian networks have been used to analyse biological signalling processes: Sachs *et al.* [2005] inferred a causal Bayesian network using flow cytometry measurements of phosphorylated protein concentrations in single cells. [Maathuis *et al.*, 2010] used causal Bayesian networks to predict (from observational data) the causal effects of gene deletion in yeast, validating their findings on 267 mutant strains.

1.4.3 Causality in Protein Signalling

In atypical situations, such as ectopic over-expression of protein species, there is potential for molecules to interact which would not normally interact in any meaningful way. This may be due to simply an abundance of a reactant leading to an increase in product formation an expansion of the spatial territory of species in the cell. Alternatively, a mutation may lead to constitutive activation of a protein species, rendering it independent of its canonical regulatory architecture. Thus in protein signalling, data collected from multiple individuals $j \in \mathcal{J}$ may differ with respect to interplay between variables, such that corresponding causal graphs G^j may be individual-specific. Interplay in protein signalling networks can depend on the genetic and epigenetic state of the individuals, such that even for a well-defined system, such as signalling downstream of a certain receptor class, details may differ between even closely related samples [Ideker and Krogan, 2012]. Even when belonging to the same lineage, samples differ with respect to signalling network connections [Lee and Tzou, 2009]. Continuing reduction in the unit cost of biochemical assays has led to an increase in experimental designs that include panels of potentially heterogeneous individuals [Barretina *et al.*, 2012; Cao *et al.*, 2011; Maher, 2012; The Cancer Genome Atlas Network, 2012]. In such settings, given individual specific data $\mathbf{y}^j$, scientific interest often focuses on the individual-specific networks G^j and their similarities and differences. The data-driven characterisation of context-specific protein signalling networks is an active area of both statistical and experimental research.

In brief, this thesis aims to use model-based techniques to elucidate causal Bayesian networks from RPPA measurement of phosphorylated protein concentrations. It this Section we make this problem precise and discuss the principal challenges.

Definition 18 (Biological network)**.** *A* biological network N *has chemical species* $\mathcal{X}_i$ *as vertices, with an edge* $\mathcal{X}_i \to \mathcal{X}_j$ *denoting that species* $\mathcal{X}_i$ *is a reactant in at least one chemical reaction* $\mathcal{R}_k$ *which produces*

$\mathcal{X}_j$ *as a product.*

Example 15 (Phosphorylation network)**.** *A* phosphorylation network *is a particular type of biological network known as a* protein signalling network *in which vertices are phosphoproteins* $\mathcal{X}_i^*$ *and edges* $\mathcal{X}_i^* \to \mathcal{X}_j^*$ *denote that* $\mathcal{X}_i^*$ *is an enzyme catalysing the conversion of unphosphorylated* $\mathcal{X}_j$ *to phosphorylated* $\mathcal{X}_j^*$*. Fig. 1.2 displays a phosphorylation network uncovered by extensive biochemistry. (Note that a handful of interactions in Fig. 1.2 refer to more general protein-protein interaction and not specifically to phosphorylation.)*

In biological settings, the problem of identifying biological networks from data is often referred to as *network inference*. This thesis pursues an approximate description of protein signalling networks, which need not be acyclic, in terms of (static) causal DBNs (Fig. 1.7(b) and Ex. 14, see also Voortman *et al.* [2010]). In this way biological network inference may be cast, modulo approximations, as inference for causal Bayesian networks.

As discussed in Section 1.4.2 there exist several generic algorithms for identification of causal Bayesian networks from data. In this thesis we sought to exploit domain-specific knowledge in order to achieve improved estimation in the context of protein phosphorylation networks. Before proceeding, we make explicit the link between *fine grain* chemical reaction graphs and *coarse grain* protein signalling networks.

1.4.4 Causal Graphs and Biological Networks

An important technical point for causal inference in biological signalling systems is that the level of description which we seek is often substantially coarser than the relevant level for the dynamics. For example, in Chapter 3 we define a protein signalling network on six species {4EBP1,Akt,EGFR,GSK3ab,MEK,S6}, yet none of these species are thought to directly interact, in the sense of forming a biochemical complex (except possibly Akt and GSK3ab). Regulation of Akt by EGFR, for instance, (typically) occurs indirectly via phosphatidylinositol-3-kinases (Fig. 1.2). The missing variable issue for biological networks is arguably more severe than in, say, economics or epidemiology. Indeed, the variables which are quantifiable on a single assay may represent only a small fraction of the minimal causally sufficient state vector. Moreover it is often the case that little specific insight is available into the nature of the missing variables or their relationship to observations.

In this thesis we use G to denote a fine scale representation of multivariate systems, whereas N will be reserved for coarse scale representations. To be concrete, in Chapter 3 we use G to denote chemical reaction graphs and N to denote phosphorylation networks. In this setting G differs from N by containing, in addition to phosphorylated protein species, unphosphorylated protein species and enzyme-substrate complexes. Fig. 1.6 displays the chemical reaction graph G for MAPK phosphorylation; compare this with the coarse biological network of Fig. 1.2, where MAPK phosphorylation is represented by a sub-network of only 9 vertices and 8 edges. Thus the biological network representation $N = N(G)$ is a coarse summary of the underlying chemical reaction graph G.

Coarse representations of biological processes may assist with issues of identifiability. It is known that stoichiometries $\boldsymbol{S}$ defining the molecular-level kinetics are in general non-identifiable from observational data in the thermodynamic limit. Indeed, the algebraic structure of the set of reaction $\mathcal{R}_1, \ldots, \mathcal{R}_v$ is in general non-identifiable; moreover even for fixed reactions $\mathcal{R}_1, \ldots, \mathcal{R}_v$ the corresponding rates $k_1, \ldots, k_v$ are in general unidentifiable [Craciun and Pantea, 2008]. However, mainstream descriptions of biological networks, e.g. protein signalling networks (Definition 18), are coarser summaries of the underlying dynamics. Such networks are useful because they are closely tied to validation experiments in which interventions (e.g. RNA interference or inhibitors) target network vertices. For example, inference of an edge in a gene regulatory network corresponds to the qualitative prediction that intervention on the parent will influence the child (via transcription factor activity [Maathuis *et al.*, 2010]). It remains unclear to what extent such coarse biological network structure can be usefully identified from various kinds of data.

1.4.5 Network Inference

Network inference approaches are now widely used in biological applications to probe regulatory relationships between molecular components such as genes or proteins. Many specific methods have been proposed, in the statistical literature as well as in bioinformatics and bioengineering, with some popular approaches reviewed in Bansal *et al.* [2007]; Bonneau [2008]; Hecker *et al.* [2009]; Lee and Tzou [2009];

Markowetz and Spang [2007]. Graphical models play a prominent role in this literature, as does variable selection. A distinction is often made between statistical and "mechanistic" approaches [Ideker and Krogan, 2012]. The former is usually used to refer to "black box" models that are built on conventional regression formulations and variants thereof, while the latter usually refers to "white box" models that are explicitly rooted in chemical kinetics, e.g. systems of coupled ordinary differential equations (ODEs). Of course this distinction is somewhat artificial, since it is possible in principle to carry out formal statistical network inference based on mechanistic models (e.g. systems of ODEs). In practice, however, the construction of a "grey box" algorithm which benefits from both the efficiency of statistical approaches and the interpretability of mechanistic approaches remains challenging [Xu *et al.*, 2010].

Many network inference schemes are based on formulations that are closely related in terms of the underlying statistical model. For example, vector autoregressive (VAR) models (including Granger causality-related approaches as special cases; Bolstad *et al.* [2011]; Morrissey *et al.* [2010]; Opgen-Rhein and Strimmer [2007]; Zou and Feng [2009]), linear dynamic Bayesian networks [Hill *et al.*, 2012a; Kim *et al.*, 2003], and certain ODE-based approaches [Bansal and di Bernardo, 2007; Li and Petzold, 2008; Nam *et al.*, 2007] are intimately related, being based on linear regression, but with potentially differing approaches to variable selection. In recent years, several empirical comparisons of competing network inference schemes have emerged, including Altay and Emmert-Streib [2010]; Bansal *et al.* [2007]; Hache *et al.* [2009]; Smith *et al.* [2001]; Werhli *et al.* [2006]. Assessment methodology has received attention, including attempts to automate the generation of large scale biological network models for automatic benchmarking of performance [Marbach *et al.*, 2009; Van den Bulcke *et al.*, 2006]. In particular the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges [Marbach *et al.*, 2012; Prill *et al.*, 2010] have provided an opportunity for objective empirical assessment of competing approaches. At the same time developments in synthetic biology have led to the availability of gold standard data from hand-crafted biological systems, such that the underlying network is known by design [Camacho and Collins, 2009; Cantone *et al.*, 2009; Minty *et al.*, 2009].

1.5 Discussion

In this Chapter we have introduced the biological problem of inference for protein signalling networks, surveyed the stochastic chemical kinetic literature and introduced key concepts from graphical models and causality. The remainder of this thesis seeks to combine these ideas within a Bayesian framework for model-based inference of protein signalling networks mediated by phosphorylation. Specifically; in Chapter 2 we explore existing approaches to inference in this setting; Chapter 3 proposes a novel methodology for network inference which is rooted in non-linear chemical kinetics; Chapter 4 extends these ideas to account for variation between biological samples and finally Chapter 5 suggests directions for further research.

Chapter 2

From Biological Dynamics to Network Inference

In the previous Chapter we introduced the statistical problem of network inference and explored its significance within the wider scientific context of biological signalling processes. As discussed in Section 1.4.5, many methods have been proposed for inference of biological networks. However the connections and differences between their statistical formulations have received less attention. In this Chapter, we show how a broad class of statistical network inference methods, including a number of existing approaches, can be described in terms of variable selection for the linear model. This reveals some subtle but important differences between the methods, including the treatment of time intervals in discretely observed data. In developing a general formulation, we also explore the relationship between single-cell stochastic dynamics and network inference on averages over cells. This clarifies the link between biochemical networks as they operate at the cellular level and network inference as carried out on data that are averages over populations of cells. We present empirical results, comparing thirty-two network inference methods that are instances of the general formulation we describe, using two published dynamical models. Our investigation sheds light on the applicability and limitations of network inference and provides guidance for practitioners and suggestions for experimental design. Since aspects of biological dynamics may not be identifiable at steady-state, time-varying data is usually preferred, and this is the setting we focus on here.

2.1 Introduction

Network inference methods can be viewed as generating hypotheses about cell biology. Yet the link between biochemical networks at the cellular level and network inference as applied to bulk or aggregate data (i.e. data that are averages over large numbers of cells) from assays such as microarrays remains unclear. In applications to noisy time-varying data there is uncertainty in the predictor variables of the same order of magnitude as uncertainty in the responses, yet often only the latter is explicitly accounted for. Moreover, the treatment of time intervals in discretely observed data remains unclear, with contradictory approaches appearing in the literature. Most high-throughput assays, including array based technologies (e.g. gene expression or protein arrays), as well as single-cell approaches (e.g. FACS-based) involve destructive sampling, i.e. cells are destroyed to obtain the molecular measurements. The impact of the resulting non-longitudinality upon inference does not appear to have been investigated.

The contributions of this Chapter as follows:

1. First, we explore the connection between biological networks at the cellular level and the discrete time linear statistical models that are widely used for inference. Starting from a description of stochastic dynamics at the single-cell level we describe a general statistical approach rooted in the linear model. This makes explicit the assumptions that underlie a broad class of network inference approaches. This also clarifies the relationship between "statistical" and "mechanistic" approaches to biological networks.

2. Second, we explore how a number of published network inference algorithms can be recovered as special cases of the model we arrive at. This sheds light on the differences between them, including

how different assumptions lead to quite different treatments of the time step.

3. Third, we present an empirical study comparing 32 different approaches that are special cases of the general model we describe. To do so, we simulate stochastic dynamics at the single-cell level from known networks, under global perturbation of two published dynamical models. In many applications the data $\mathbf{y}$ arise from *global perturbation* of the cellular system, for example by varying culture conditions or stimuli. The extent to which networks can be characterized using global perturbations remains poorly understood, since it is likely that such data expose only a subspace of the phase space associated with cellular dynamics. Our investigation enables a clear assessment of the network inference methods in terms of estimation bias and consistency, since the true data-generating network is known. Furthermore, the simulation accounts for both averaging over cells, non-longitudinality due to destructive sampling and the fact that only a subspace of the dynamical phase space is explored. Using this approach, we investigate a number of data regimes, including both even and uneven sampling intervals, longitudinal and non-longitudinal data and the large sample, low noise limit. We find that the net effect of predictor uncertainty, non-longitudinality and limited exploration of the dynamical phase space is such that certain network estimators fail to converge to the data-generating network even in the limits of large datasets and low noise. However, we point to a simple formulation which might represent a default choice, delivering promising performance in a number of regimes. This formulation forms the basis for subsequent work in Chapter 3.

4. A key implication of our analysis is that uneven time steps may pose inferential problems, even when using models that apparently handle the sampling intervals explicitly. We therefore investigate this case by carrying out network inference on unevenly sampled data using a variety of statistical models. We find that the ability to reconstruct the data-generating network is much reduced in all cases, with some approaches faring better than others. Since biological data are often unevenly resolved in time, this observation has important implications for experimental design.

Note that this Chapter focusses exclusively on inference procedures rooted in discrete time. Whilst several statistical approaches now exist to facilitate efficient inference in continuous time descriptions of dynamical systems (e.g. Campbell and Steele [2012]; Dattner and Klaassen [2013]; Dondelinger *et al.* [2013]; Girolami and Calderhead [2011]), it remains the case that network inference algorithms are almost exclusively rooted in poorly understood discrete time formulations.

The remainder of this Chapter is organized as follows. We begin in Section 2.2 with a description of stochastic dynamics in single cells and show how a series of assumptions allow us to arrive at a statistical framework rooted in the linear model. Section 2.3 contains an empirical comparison of several inference schemes, addressing questions of performance and consistency in a number of data-generating regimes. In Section 2.4 we discuss our results and point to several specific areas for future work.

2.2 Methods

The cellular dynamics that underlie network inference are subject to stochastic effects [Elowitz *et al.*, 2002; Kou *et al.*, 2005; McAdams and Arkin, 1997; Paulsson, 2005; Swain *et al.*, 2002]. We therefore begin our description of the data-generating process at the level of single cells before discussing the relationship to aggregate data of the kind acquired in high-throughput biochemical assays. We then develop a general statistical approach, rooted in the linear model, for data from such a system observed discretely in time. We discuss inference and show how a number of existing approaches can be recovered as special cases of the general model we describe. Our exposition clarifies a number of technical but important distinctions between published methodologies, which until now have received little attention.

2.2.1 Data-Generating Process

2.2.1.1 Stochastic Dynamics in Single Cells

Let $\mathbf{X} = (X_1, \ldots, X_P) \in \mathbb{R}_+^P$ denote a state vector describing the abundance of molecular quantities of interest. The components of the state vector (e.g. mRNA, protein or metabolite levels) are identified with the vertices of a network N that describes the biological network of interest. In this Chapter the

expression levels $\mathbf{X}(t)$ of a single cell at time $t \in \mathbb{R}_+$ are modelled as continuous random variables that we assume satisfy a time-homogeneous stochastic delay differential equation (SDDE)

$$d\mathbf{X} \quad = \quad \mathbf{f}(\mathcal{F}_{\mathbf{X}})dt + \mathbf{g}(\mathcal{F}_{\mathbf{X}})d\mathbf{B} \tag{2.1}$$

where $\mathbf{f}, \mathbf{g}$ are drift and diffusion functions respectively, $\mathcal{F}_{\mathbf{X}}(t) = \{\mathbf{X}(s) : s \leqslant t\}$ is the natural filtration (the history of the state vector $\mathbf{X}$) and $\mathbf{B}$ denotes a standard Brownian motion. One well known example is the CLE, as defined in Section 1.4. Note that this Chapter does not consider finite state space models; this is thought to be reasonable for the biological systems considered here, but in general the stochasticity due to low copy number will need to be encoded into inference (see Section 1.3.1 and Paulsson [2005]). Note also that this Chapter does not consider chemical reaction graphs G, only the coarser biological networks N. We do not restrict attention to phosphorylation networks, but biological networks in general. This set-up most closely matches the approach of published network inference algorithms, whose behaviour are the subject of this Chapter. The edge structure E of the biological network N is defined by the drift function $\mathbf{f}$, such that $(i,j) \in E \iff f_j(\mathbf{X})$ depends on X_i. Recent work by Sokol and Hansen [2013] formalises these ideas using a causal interpretation of stochastic differential equations. Also related are the frameworks of Dash [2003]; Iwasaki and Simon [1994].

We further assume that the functions $\mathbf{f}, \mathbf{g}$ are sufficiently regular and depend only on recent history $\mathcal{F}_{\mathbf{X}}([t-\tau, t])$. For example in the context of gene regulation τ might be the time required for one cycle of transcription, translation and binding of a transcription factor to its target site; the characteristic time scale for gene regulation. This is a finite memory requirement and can be considered a generalization of the Markov property. Equivalently, this property codifies causal sufficiency for SDDEs. It is common practice to take $\tau = 0$, in which case the process defined by Eqn. 2.1 is Markovian. This stochastic dynamical system with phase space $\{(\mathbf{f}(\mathcal{F}_{\mathbf{X}}), \mathbf{X}) : \mathbf{X} \in \mathbb{R}_+^P\}$ forms the basis of the following exposition.

2.2.1.2 Aggregate Data

A variety of experimental techniques, including notably microarrays and related assays, capture average expression levels $\mathbf{X}^{(K)} := \sum_{k=1}^{K} \mathbf{X}^k/K$ over cells, where $\mathbf{X}^k$ denotes the expression levels in cell k. We do not consider effects due to inter-cellular signalling, which are typically assumed to be negligible. Averaging sacrifices the finite memory property (a generalization of the fact that the sum of two independent Markov processes is not itself Markovian). However it is usually possible to construct a finite memory approximation of the form

$$d\mathbf{X}^{(K)} \quad = \quad \mathbf{f}^{(K)}(\mathcal{F}_{\mathbf{X}^{(K)}})dt + \mathbf{g}^{(K)}(\mathcal{F}_{\mathbf{X}^{(K)}})d\mathbf{B}^{(K)} \tag{2.2}$$

using a so-called *system size expansion* [van Kampen, 2007]. (The CLE is a specific example of a system size expansion.) Approximations of this kind derive from a coarsening of the underlying state space, assuming that the new state vector $\mathbf{X}^{(K)}$ is causally sufficient. The statistical models discussed in this Chapter rely upon coarsening assumptions in order to control the dimensionality of state space.

Using the mild regularity conditions upon cellular stochasticity $\mathbf{g}$ the strong law of large numbers gives that in the large sample limit the sample average $\mathbf{X}^{\infty} := \lim_{K\to\infty} \mathbf{X}^{(K)} = \mathbb{E}(\mathbf{X})$ equals the expected state of a single cell (almost surely). We note that the relationship between the single-cell dynamics as it appears in Eqn. 2.1 and this deterministic limit may be complicated, since in general $\mathbb{E}(\mathbf{f}(\mathcal{F}_{\mathbf{X}})) \neq \mathbf{f}(\mathcal{F}_{\mathbb{E}(\mathbf{X})})$. However for linear $\mathbf{f}$, say for simplicity $\mathbf{f} \equiv \mathbf{f}(\mathbf{X}) = \mathbf{A}\mathbf{X}$, we have

$$\begin{aligned} d\mathbf{X}^{(K)} = \frac{1}{K}\sum_{k=1}^{K} d\mathbf{X}^k \quad &= \quad \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{f}(\mathcal{F}_{\mathbf{X}^k})dt + \mathbf{g}(\mathcal{F}_{\mathbf{X}^k})d\mathbf{B}^k\right) \\ &= \quad \frac{1}{K}\sum_{k=1}^{K}\mathbf{A}\mathbf{X}^k dt + \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}(\mathcal{F}_{\mathbf{X}^k})d\mathbf{B}^k \\ &= \quad \mathbf{A}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{X}^k\right)dt + \mathbf{R}^{(K)} \\ &= \quad \mathbf{A}\mathbf{X}^{(K)}dt + \mathbf{R}^{(K)} = \mathbf{f}(\mathcal{F}_{\mathbf{X}^{(K)}})dt + \mathbf{R}^{(K)} \end{aligned} \tag{2.3}$$

where $\mathbf{R}^{(K)} := \sum_k \mathbf{g}(\mathcal{F}_{\mathbf{X}^k})d\mathbf{B}^k/K \to \mathbf{0}$ almost surely as $K \to \infty$ and so $d\mathbf{X}^{\infty}/dt = \mathbf{f}(\mathcal{F}_{\mathbf{X}^{\infty}})$. In other

words, the average over large numbers of cells shares the same drift function as the single cell, so that inference based on averaged data applies directly to single cell dynamics. Otherwise this may not hold. This has implications when using non-linear forms, such as Michaelis-Menten or Hill kinetics, to describe the behaviour of a large sample average; these non-linear functions are derived from single cell biochemistry and may not apply equally to the large sample average $\mathbf{X}^\infty$. The error entailed by commuting drift and expectation may be assessed using the multivariate Feynman-Kac formula [Øksendal, 1998].

In practice the observation process may be complex and indirect, for example measurements of gene expression may be relative to a *housekeeping* gene, assumed to maintain constant expression over the course of the experiment. Moreover the details of the error structure will depend crucially on the technology used to obtain the data. To limit scope, this Chapter assumes the averaged expression levels $\mathbf{X}^\infty(t)$ are observed at discrete times $t = t_j$ $(0 \leqslant j \leqslant n)$ with additive zero-mean measurement error as $\mathbf{Y}(t_j) = \mathbf{X}^\infty(t_j) + \mathbf{w}_j$, where the $\mathbf{w}_j$ are independent, identically distributed uncorrelated Gaussian random variables.

2.2.2 Discrete Time Models

Network inference is usually carried out using coarse-grained models (Eqn. 2.2) that are simpler and more amenable to inference than the process described by Eqn. 2.1. Here, informed by the foregoing treatment of cellular dynamics, we develop a simple network inference model for data observed discretely in time. We clarify the assumptions of the statistical model, and show how several published approaches can be recovered as special cases.

2.2.2.1 Approximate Discrete Time Likelihood

Network inference entails statistical comparison of networks $N \in \mathcal{N}$, where $\mathcal{N}$ denotes the space of candidate networks. The space $\mathcal{N}$ may be large (naively, there are $2^{P \times P}$ possible networks on P vertices), although biological knowledge may provide constraints. Network comparisons require computation of a model selection score for each network that is considered, which in turn entails use of the likelihood (e.g. maximization of information criteria, or integration over the likelihood in the Bayesian setting). Therefore, exploration over large model spaces is often only feasible given a closed-form expression for the likelihood (or preferably for the model score itself).

However the likelihood for a SDDE model (Eqn. 2.2) is not generally available in closed form. There has been recent research into computationally efficient approximate likelihoods for fully observed, noiseless diffusions [Hurn *et al.*, 2007], but it remains the case that the simplest (though least accurate) closed-form approximate likelihood is based on the Euler-Maruyama discretisation scheme for stochastic differential equations (SDEs), which in the more general SDDE case may be written as (henceforth dropping the superscript K)

$$\mathbf{X}(t_j) \quad \approx \quad \mathbf{X}(t_{j-1}) + \Delta_j \mathbf{f}(\mathcal{F}_{\mathbf{X}}(t_{j-1})) + \mathbf{g}(\mathcal{F}_{\mathbf{X}}(t_{j-1}))\Delta\mathbf{B}_j \tag{2.4}$$

where $\Delta\mathbf{B}_j \sim \mathcal{N}(\mathbf{0}, \Delta_j \mathbf{I})$ and $\Delta_j = t_j - t_{j-1}$ is the sampling time interval. Incorporating measurement error into this *Riemann-Itô* likelihood [Fuchs, 2013] requires an integral over the hidden states $\mathbf{X}$ which would destroy the closed-form approximation. Therefore the observed (non-longitudinal) data $\mathbf{y}$ are directly substituted for the latent states $\mathbf{X}$, yielding the (triply) approximate likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}, N) \quad \approx \quad \prod_{j=1}^{n} \mathcal{N}(\mathbf{y}(t_j); \boldsymbol{\mu}(t_j), \boldsymbol{\Sigma}(t_j)) \tag{2.5}$$

where $\boldsymbol{f}$, $\boldsymbol{g}$ are the drift and diffusion functions associated with the network N, whose parameters are denoted by $\boldsymbol{\theta}$. Here $\boldsymbol{\mu}(t_j) = \mathbf{y}(t_{j-1}) + \Delta_j \mathbf{f}(\mathcal{F}_{\mathbf{y}}(t_{j-1}))$, $\boldsymbol{\Sigma}(t_j) = \Delta_j \mathbf{g}(\mathcal{F}_{\mathbf{y}}(t_{j-1}))\mathbf{g}(\mathcal{F}_{\mathbf{y}}(t_{j-1}))'$, and $\mathcal{N}(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Implicit here is that the functions $\mathbf{f}, \mathbf{g}$ depend on $\mathcal{F}_{\mathbf{y}}$ only through time lags which coincide with the measurement times t_{j-1}.

Thus an approximate likelihood may be obtained from a state-space approximation to the original SDDE model (Eqn. 2.2). Despite reported weaknesses with the Riemann-Itô likelihood [Fuchs, 2013; Hurn *et al.*, 2007] and the poorly characterized error incurred by plugging in non-longitudinal observations, this form of approximate likelihood is widely used to facilitate network inference (Eqn. 2.5 corresponds to a Gaussian DBN for the observations $\mathbf{y}$, as described in Example 12, generalized to allow dependence on history). This is due both to the possibility of parameter orthogonality, allowing

inference to be performed for each network node separately, and the possibility of conjugacy, leading to a closed-form marginal likelihood $p(\mathbf{y}|N) = \int p(\mathbf{y}|\boldsymbol{\theta}, N)p(\boldsymbol{\theta}|N)d\boldsymbol{\theta}$. (Note that in the remainder of this thesis, for brevity, we use notation which does not distinguish random variables from their arguments.)

2.2.2.2 Linear Dynamics

Kinetic models have been described for many cellular processes [Cantone *et al.*, 2009; Schoeberl *et al.*, 2002; Swat *et al.*, 2004; Wilkinson, 2009]. However, statistical inference for these often non-linear models may be challenging [Bonneau, 2008; Wilkinson, 2006, 2009; Xu *et al.*, 2010]. Moreover, there is no guarantee that conclusions drawn from cellular averages will apply to single cells, because as noted above the deterministic behaviour seen in averages may not coincide with the single cell drift. However, linear dynamics satisfy $\mathbb{E}(\mathbf{f}(\mathcal{F}_{\mathbf{X}})) = \mathbf{f}(\mathcal{F}_{\mathbb{E}(\mathbf{X})})$ exactly, so that conclusions drawn from averages apply directly to single cells. For notational simplicity consider the Markovian $\tau = 0$ regime. A Taylor approximation of the cellular drift $\mathbf{f}$ about the origin gives

$$\mathbf{f}(\mathbf{X}) \quad \approx \quad \mathbf{f}(\mathbf{0}) + D\mathbf{f}|_{\mathbf{x}=\mathbf{0}}\,\mathbf{X} \tag{2.6}$$

where $D\mathbf{f}$ is the Jacobian matrix of $\mathbf{f}$. The constant term can be omitted ($\mathbf{f}(\mathbf{0}) = \mathbf{0}$), since absent any molecules there can be no change in expression levels. Then, the Jacobian $D\mathbf{f}$ captures the dynamics approximately under a linear model. Furthermore, the absence of an edge in the network N implies a zero entry in the Jacobian, that is $(i,j) \notin E \Rightarrow (D\mathbf{f})_{ji} = 0$. Conversely, however, obtaining the Jacobean at $\mathbf{x} = \mathbf{0}$ does not imply complete knowledge of the edge sparsity structure E. We note that the general SDDE case is similar but with additional differentiation required for the additional dependencies of $\mathbf{f}$. Henceforth we write equations for the simpler Markovian model, although they hold more generally.

One may ask whether the restriction to linear drift functions allows the computational difficulties associated with inference for continuous time models to be avoided, since in the Markovian ($\tau = 0$) case both the SDE (Eqn. 2.1) and limiting ordinary differential equation (ODE) have exact closed form solutions. In the ODE case, for example, $\mathbf{X}(t) = \exp(\mathbf{A}t)\mathbf{X}_0$ and under Gaussian measurement error the likelihood has a closed form as products of terms $\mathcal{N}(\mathbf{y}(t_j); \exp(\mathbf{A}t_j)\mathbf{X}_0, \mathbf{M})$ where the parameters $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{X}_0, \mathbf{M})$ include the model parameters $\mathbf{A}$, initial state vector $\mathbf{X}_0$ and the measurement error covariance $\mathbf{M}$. Unfortunately evaluation of the matrix exponential is computationally demanding and inference for the entries of $\mathbf{A}$ must be performed jointly since in general $\exp(\mathbf{A})$ does not factorize usefully. It therefore remains the case that inference for continuous time models is computationally burdensome, even when the models are linear.

2.2.2.3 The Dynamical System as a Regression Model

The Jacobian $D\mathbf{f}$ with entries $(D\mathbf{f})_{i,j} = \partial f_i/\partial x_j|_{\mathbf{x}=\mathbf{0}}$ is now the object of inference. We can identify the Jacobian with the unknown parameters in a linear regression problem by modelling the expression of variable p using

$$\begin{bmatrix} \dot{X}_1(t_1) & \dots & \dot{X}_P(t_1) \\ \vdots & & \vdots \\ \dot{X}_1(t_n) & \dots & \dot{X}_P(t_n) \end{bmatrix} \approx \begin{bmatrix} X_1(t_0) & \dots & X_P(t_0) \\ \vdots & & \vdots \\ X_1(t_{n-1}) & \dots & X_P(t_{n-1}) \end{bmatrix} \begin{bmatrix} (Df)_{1,1} & \dots & (Df)_{P,1} \\ \vdots & & \vdots \\ (Df)_{1,P} & \dots & (Df)_{P,P} \end{bmatrix} \tag{2.7}$$

where the gradients $\dot{X}_p(t_j)$ are approximated by finite differences, in this case $(X_p(t_j) - X_p(t_{j-1}))/\Delta_j$. More generally for processes with memory the matrix may be augmented with columns corresponding to lagged state vectors and the vector $(D\mathbf{f})_{p,\bullet}$ augmented with the corresponding derivatives of the drift function $\mathbf{f}$ with respect to these lagged states. To avoid confusion we write $\mathbf{A}$ for $D\mathbf{f}$ when discussing parameters, since the drift $\mathbf{f}$ is unknown. Similarly, design matrices will be denoted by $\mathbf{B}$ to suppress the dependence on the random variables $\mathbf{X}$. So the columns of Eqn. 2.7 may be written compactly as

$$\dot{\mathbf{X}}_p \quad \approx \quad \mathbf{B}\boldsymbol{A}'_{p,\bullet}. \tag{2.8}$$

Inference for the parameters $\boldsymbol{A}_{p,\bullet}$ may be performed independently for each variable p. Whilst Eqn. 2.8 is fundamental for inference, one can equivalently consider the dynamically intuitive expression given by

the rows of Eqn. 2.7:

$$\dot{\mathbf{X}}(t_j) \quad \approx \quad \mathbf{A}\boldsymbol{B}'_{j,\bullet}. \tag{2.9}$$

An interesting issue arises from the dual interpretation of the regression model as a dynamical system (Eqn. 2.9), because there are natural restrictions on $\mathbf{A}$ to avoid the solution tending to infinity. For instance if the sampling interval Δ is constant then we require $\mathbb{R}(\lambda) \leqslant 0$ for each eigenvalue λ of $\mathbf{A} + \Delta\mathbf{I}$. The inference schemes which we discuss do not account for this, because the condition forces a non-trivial coupling between rows $\boldsymbol{A}_{p,\bullet}$, jeopardizing parameter orthogonality.

Finally, the generative model is specified by substituting noisy, non-longitudinal observables $\mathbf{Y}$ for latent variables $\mathbf{X}$ into Eqn. 2.9 and stating the dependence of the approximation error on the sampling interval Δ_j. Under a further approximation of uncorrelated Gaussian error we arrive at a model

$$\dot{\mathbf{Y}}(t_j) \quad \sim \quad \mathcal{N}(\mathbf{A}\dot{B}'_{j,\bullet}, h(\Delta_j)\mathcal{D}(\sigma_1^2, \ldots, \sigma_P^2)) \tag{2.10}$$

where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a variance function that must be specified and $\mathcal{D}(\mathbf{v})$ represents the diagonal matrix with diagonal $\mathbf{v}$. The justification for assuming uncorrelated measurement error over a given time interval Δ must be made on a context-specific basis. In general this assumption cannot hold for all Δ, unless the data-generating process itself induces uncorrelated trajectories. As a consequence, care must be taken when working with unevenly sampled data, or data obtained at regular time intervals for which the assumption of uncorrelated errors does not hold. We return to this point in Section 2.4.

There are a number of ways in which this regression is non-standard. For example, the substitution of observations for latent variables is clearly unsatisfactory because the linear regression framework does not explicitly allow for uncertainty in the predictor variables $\mathbf{B}$. It is unclear whether this introduces bias or leads to an overestimate of the significance of results. Moreover, it is unclear how to choose the variance function h, since the Euler-Maruyama approximation (Eqn. 2.4) is only valid for small sampling intervals Δ_j, but in this regime the responses $\dot{\mathbf{Y}}(t_j)$ are dominated by measurement error, such that the data may carry little information. These issues are investigated empirically in Sections 2.3 and 2.4 below.

2.2.3 A Unifying Framework

Eqn. 2.10 describes a class of models with specific instances characterized by choice of design matrix $\mathbf{B}$ and variance function h. Since any such model corresponds to the linear regression Eqn. 2.7, the task of determining the edge structure of the network, or equivalently the location of non-zero entries in the Jacobian $\mathbf{A}$, can be cast as a variable selection problem.

A number of specific network inference schemes can now be recovered by fixing the design matrix and variance function and coupling the resulting model with a variable selection technique. A selection of published network inference schemes that can viewed in this way is presented in Table 2.1. One might see these schemes classed as VAR models [Bolstad *et al.*, 2011; Morrissey *et al.*, 2010; Opgen-Rhein and Strimmer, 2007; Zou and Feng, 2009], DBNs [Hill *et al.*, 2012a; Kim *et al.*, 2003], or ODE-based approaches [Bansal and di Bernardo, 2007; Li and Petzold, 2008; Nam *et al.*, 2007], although as we have demonstrated this classification disguises their shared foundation in the linear model.

As shown in Table 2.1, the variance functions h, and therefore sampling intervals Δ_j, are not treated in a consistent way in the literature. In the special case of even sampling times $\Delta_j = \Delta$, a model is characterized only by its design matrix. If the standard design matrix is used then the entire family of models

$$\frac{\mathbf{Y}(t_j) - \mathbf{Y}(t_{j-1})}{\Delta} \sim \mathcal{N}(\mathbf{A}\mathbf{Y}(t_{j-1}), h(\Delta)\mathcal{D}(\sigma_1^2, \ldots, \sigma_P^2)) \tag{2.11}$$

reduces to a linear VAR(1) model

$$\mathbf{Y}(t_j) \sim \mathcal{N}(\bar{\mathbf{A}}\mathbf{Y}(t_{j-1}), \mathcal{D}(\bar{\sigma}_1^2, \ldots, \bar{\sigma}_P^2)) \tag{2.12}$$

where $\bar{\mathbf{A}} = \Delta\mathbf{A} + \mathbf{I}$ and $\bar{\sigma}_p^2 = \Delta^2 h(\Delta)\sigma_p^2$. More generally the VAR(q) model is prevalent in the literature (see Table 2.1), yet it does not explicitly handle uneven sampling intervals. This is a potentially important issue since uneven sampling is commonplace in global perturbation experiments, with high frequency sampling used to capture short term cellular response and low frequency sampling to capture the approach to equilibrium. We discuss the importance of modelling using a variance function and whether a natural

Design matrix $\mathbf{B}$	Variance function $h(\Delta)\propto$	Variable selection	Example
Standard	Δ^{-2}	Ridge regression	Bansal and di Bernardo [2007] "TSNIB"
Standard with lagged predictors	$\varnothing$	Group LASSO	Bolstad *et al.* [2011]
Quadratic	$\varnothing$	Conjugate Bayesian with network prior	Hill *et al.* [2012a]
Standard	$\varnothing$	Information criteria	Kim *et al.* [2003],
Non-linear (Hill) basis functions	1	AIC with backstepping	Li and Petzold [2008]
Standard	1	Conditional independence tests	Li *et al.* [2011] "DELDBN"
Standard	$\varnothing$	Semi-conjugate Bayesian	Morrissey *et al.* [2010]
Standard	Δ^{-2}	SVD and pseudoinverse	Nam *et al.* [2007] "LEARNe"
Standard	$\varnothing$	Multi-stage analytic shrinkage approach	Opgen-Rhein and Strimmer [2007]
Standard and non-linear with lagged predictors	$\varnothing$	Granger causality	Zou and Feng [2009]

Table 2.1: A non-exhaustive list of network inference schemes rooted in the linear model. The examples from literature demonstrate the statistical features indicated, but may differ in some aspects of implementation. The symbol $\varnothing$ denotes the VAR(q) model which lacks a variance function.

choice for such a function exists in Section 2.4 below. Section 2.3 explores whether inference may be improved through the use of either non-linear basis functions or lagged predictors to capture respectively non-linearity and memory in the underlying drift function.

2.2.4 Inference

An appealing feature of the discrete time model is that parameters corresponding to different variables are orthogonal in the sense that the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}, N)$ factorises over $(\boldsymbol{A}_{p,\bullet}, \sigma_p)$ for $p \in \mathcal{P}$. As a consequence network inference over $\mathcal{N}$ may be decomposed into P independent variable selection problems. For definiteness we focus on just two approaches to variable selection, the Bayesian marginal likelihood and AIC, both of which have been applied in this context previously (Table 2.1). We note that many other approaches are available, including notably the Bayesian Information Criterion (BIC), and can be applied here in analogy to what follows. Below we assume the response vector $\dot{\mathbf{y}}_p h^{-1/2}$ and the columns of the design matrix $\mathbf{B}h^{-1/2}$ are standardized to have zero mean and unit variance, but for clarity subsume this into unaltered notation.

2.2.4.1 Bayesian Variable Selection

For simplicity, the variance function is initially taken to be constant ($h = 1$). We set up a Bayesian linear model conditional on a network N using Zellner's g-prior [Zellner, 1986], that is with priors $\boldsymbol{A}_{p,\bullet}|\sigma_p^2 \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 n(\mathbf{B}_p{}'\mathbf{B}_p)^{-1})$ and $p(\sigma_p^2)\propto 1/\sigma_p^2$ where $\mathbf{B}_p$ is the design matrix $\mathbf{B}$ with non-predictors removed according to N. We note that while the g-prior is a common choice, alternatives may offer some advantages [Deltell *et al.*, 2012; Friedman *et al.*, 2000].

Let m_p be the number of predictors for variable p in the network N. Integrating the likelihood (induced by Eqn. 2.10) against the prior for $(\boldsymbol{A}_{p,\bullet}, \sigma_p^2)$ produces the following closed-form marginal likelihood

$$p(\mathbf{y}|N) \quad \propto \quad \prod_p \left(\frac{1}{1+n}\right)^{m_p/2} \left[\dot{\mathbf{y}}_p'\dot{\mathbf{y}}_p - \left(\frac{n}{1+n}\right)\hat{\dot{\mathbf{y}}}_p'\hat{\dot{\mathbf{y}}}_p\right]^{-n/2} \tag{2.13}$$

where $\hat{\dot{\mathbf{y}}}_p = \mathbf{B}_p(\mathbf{B}_p{}'\mathbf{B}_p)^{-1}\mathbf{B}_p{}'\dot{\mathbf{y}}_p$. These formulae extend to arbitrary variance functions h by substituting $\mathbf{B} \mapsto \mathbf{B}h^{1/2}$, $\dot{\mathbf{y}} \mapsto \dot{\mathbf{y}}h^{1/2}$. Network inference may now be carried out by Bayesian model averaging, using the posterior probability of a directed edge from variable i to variable j:

$$\mathbb{P}((i,j)|\boldsymbol{y}) \quad = \quad \frac{\sum_N p(\boldsymbol{y}|N)p(N)\mathbb{I}\{(i,j)\in N\}}{\sum_N p(\boldsymbol{y}|N)p(N)}. \tag{2.14}$$

In experiments below we take a network prior which, for each variable p, is uniform over the number of predictors m_p up to a maximum permissible in-degree $d_{\max}$, that is $p(N) \propto \prod_p \binom{P}{m_p}^{-1} \mathbb{I}\{m_p \leqslant d_{\max}\}$, but note that richer subjective network priors are available in the literature [Mukherjee and Speed, 2008]. Finally, a network estimator $\hat{N}$ is obtained by thresholding posterior edge probabilities: $(i,j) \in \hat{N} \Leftrightarrow \mathbb{P}((i,j)|\boldsymbol{y}) > \epsilon$. For small maximum in-degree $d_{\max}$, exact inference by enumeration of variable subsets may be possible. Otherwise, Markov chain Monte Carlo (MCMC) methods can be used to explore an effectively smaller model space [Ellis and Wong, 2008; Friedman and Koller, 2003]. In the experiments below we use exact inference by enumeration.

Posterior marginals $\mathbb{P}((i,j)|\boldsymbol{y})$ close to unity indicate that the corresponding edge is very likely present in the data-generating network, modulo the assumptions of the statistical model. In general, interpretation of the network estimator $\hat{N}$ is necessarily context-specific and in some cases may be difficult. For instance, in settings where variables in the dynamical system are highly correlated, the statistical phenomenon of multicollinearity, whereby posterior mass is shared between correlated predictor variables, requires a joint interpretation of the posterior marginals. For applications where posterior marginals are themselves covariates, e.g. classification or regression, such problems may be minor.

2.2.4.2 Variable Selection by Corrected AIC

Again, consider a constant variance function ($h = 1$); rescaling as described above recovers the general case. The usual maximum likelihood estimates $\hat{\boldsymbol{A}}_{p,\bullet} = (\mathbf{B}_p{}'\mathbf{B}_p)^{-1}\mathbf{B}_p{}'\dot{\mathbf{y}}_p$ and $\hat{\sigma}_p^2 = \frac{1}{n}\sum_j (\dot{\mathbf{y}}_p(t_j) - \hat{\dot{\mathbf{y}}}_p(t_j))^2$ induce closed forms $C_p\hat{\sigma}_p^{-n}$ for the maximized factors of the likelihood function, where C_p is a constant not depending on the choice of predictors. Corrected AIC scores [Burnham and Anderson, 2002] for each variable p are then

$$AIC_c(p,N) = n\log(\hat{\sigma}_p^2) + 2m_p + \frac{2m_p(m_p+1)}{n-m_p-1}. \tag{2.15}$$

Again we consider all models with maximum permissible in-degree $d_{\max}$. Lowest scoring models are chosen for each variable in turn, inducing a network estimator $\hat{N}$.

2.3 Results

In this Section, we present empirical results investigating the performance of a number of network inference schemes that are special cases of the general formulation described by Eqn. 2.10. Objective assessment of network inference is challenging [Prill *et al.*, 2010], since for most biological applications the true data-generating network is unknown. We therefore exploit two published dynamical models of biological processes, namely Cantone *et al.* [2009] and Swat *et al.* [2004], described in detail in Appendix A.1. The first is a synthetic gene regulatory network built in the yeast *Saccharomyces cerevisiae*. This five gene network and associated delay differential equations (DDEs) has received attention in computational biology [Camacho and Collins, 2009; Minty *et al.*, 2009], and has been shown to agree with gold-standard data (at least under an $\mathbb{E}(\mathbf{f}(\mathcal{F}_{\mathbf{X}})) \approx \mathbf{f}(\mathcal{F}_{\mathbb{E}(\mathbf{X})})$ assumption). Cantone *et al.* consider two experimental conditions; "switch-on" and "switch-off". Here the switch-on parameter values were used to generate data. The Swat model is a gene-protein network governing the G_1/S transition in mammalian cells. The model has a nine dimensional state vector and, unlike Cantone, is Markovian. We note that this model has not been directly verified in the manner of Cantone but is based on a theoretical understanding of cell cycle dynamics. There is undoubtedly bias from this essentially arbitrary choice of dynamical systems but a comprehensive sampling of the (vast) space of possible networks and dynamics is beyond the scope of this thesis.

2.3.1 Experimental Procedure

2.3.1.1 Simulation

We consider global perturbation data by initializing the dynamical systems from out of equilibrium conditions. This is a common setting for network inference approaches, but the limitations of inference from such data remain unclear. For each dynamical system $\mathbf{f}$, trajectories $\mathbf{X}^k$ of single cell expression levels were obtained as solutions to the SDDE Eqn. 2.1 with drift $\mathbf{f}$ and uncorrelated diffusion $\mathbf{g}(\mathbf{X}) =$

$\sigma_{\text{cell}}\mathcal{D}(\mathbf{X})$ (representing multiplicative cellular noise). Trajectories were obtained by numerically solving SDDEs with heterogeneous initial conditions using the Euler-Maruyama discretisation scheme (Eqn. 2.4). Whilst the dynamical systems $\boldsymbol{f}$ are guaranteed to produce non-negative trajectories, the same need not be true of the corresponding SDDE. To mitigate this issue, the Euler-Maruyama scheme was modified to reflect integration steps in the $X_p = 0$ axes where necessary, in order to preserve positivity.

To mimic destructive sampling and consequent non-longitudinality, solutions were regenerated at each time point. We are interested in data that are averages over a large number N of single-cell trajectories. However, the computational cost of solving $N \times n$ SDDEs to produce each data set is prohibitive. Therefore, only a smaller number $N^* << N$ of cells were simulated and a larger sample N then obtained by bootstrapping, i.e. re-sampling from the N^* trajectories uniformly with replacement. In practice N^* was taken sufficiently large such that a negligible change in experimental outcome results from further increase in N^*. Initial conditions for single cell trajectories varied with standard deviation σ_{cell}. Finally, uncorrelated Gaussian noise of magnitude σ_{meas} was added to simulate a measurement process with additive error. (Whilst this may result in negative data values, the regression models considered here do not rely on positivity.) In the experiments presented below, $N = 10,000$, $N^* = 30$ and $n = 20$ time points are taken within the dynamically interesting range (0-280 minutes for Cantone and 0-100 minutes for Swat). Measurement error and cellular noise are set to give signal-to-noise ratios $\langle \mathbf{X} \rangle / \sigma_{\text{meas}} \approx 10$, $\langle \mathbf{X} \rangle / \sigma_{\text{cell}} \approx 10$ (here $\langle \mathbf{X} \rangle$ represents the average expression levels of the variables $\mathbf{X}$ over all generated trajectories).

Fig. 2.1 shows typical datasets for the two dynamical systems. Whilst the data-generating procedure described above only captures a handful of the features of real experimental data, it is substantially richer than the majority of simulation studies which currently appear in the relevant literature. It is therefore interesting to assess existing statistical procedures in this context, where several modelling assumptions are likely to be violated.

2.3.1.2 Inference Schemes

The inference schemes which were assessed consisted of combinations from the following set of specifications:

Variable Selection	{ Bayesian, AIC_c }
Design matrix	{ Standard, Quadratic }
Lagged predictors	{ No, Yes }
Variance function $h(\Delta) \propto \Delta^{-\alpha}$	α = { 0, 1, 2 , $\varnothing$ }

For the design matrix "quadratic" refers to the augmentation of the predictor set by the pairwise products of predictors, the simplest non-linear basis functions. For the variance function the symbol $\varnothing$ is used to denote the VAR(q) model, which formally lacks a variance function. "Lagged predictors = Yes" indicates augmentation of the predictor set with lagged observations (a lag of ≈ 28 mins is used for Cantone and ≈ 10 mins for Swat). There are heuristic justifications for each of the candidate variance functions. For example the function with $\alpha = 2$ appears for small Δ_j when an exact Euler approximation and additive measurement error are assumed [Bansal and di Bernardo, 2007], whereas $\alpha = 1$ is reminiscent of the Euler-Maruyama discretisation Eqn. 2.4.

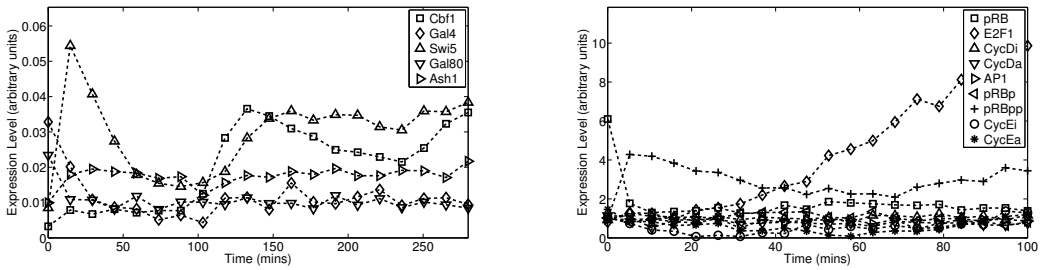
2.3.1.3 Empirical Assessment

The performance of each inference scheme is quantified by the area under the receiver operating characteristic (ROC) curve (AUR), averaged over 20 datasets [Fawcett, 2005]. This metric, equivalent to the probability that a randomly chosen true edge is preferred by the inference scheme to a randomly chosen false edge, summarizes, across a range of thresholds, the ability to select edges in the true data-generating graph. Results presented below use a computationally favourable in-degree restriction $d_{\max} = 2$. In order to check robustness to $d_{\max}$ experiments were repeated using $d_{\max} = 3$, with no substantial changes in observed outcome (Fig. 2.3(b)).

2.3.2 Empirical Results

2.3.2.1 Even Sampling Interval

Fig. 2.2(a) displays box-plots over AUR scores for the Cantone dynamical system under even sampling intervals. Note that under even sampling, for an otherwise identical scheme, changing variance function



(a) Data generated from Cantone *et al.* [2009]

(b) Data generated from Swat *et al.* [2004]

Figure 2.1: Two published dynamical models of cellular processes were used to generate datasets. Single cell trajectories were generated from an SDDE model (Eqn. 2.1) and averaged under measurement noise and non-longitudinality due to destructive sampling. (a) Data generated from (a model due to) Cantone *et al.* [2009], describing a synthetic network built in yeast. (b) Data generated from Swat *et al.* [2004], a theory-driven model of the G_1/S transition in mammalian cells.

does not affect the model, leading to identical AUR scores for schemes which differ only in variance function. (An exception to this is the VAR model, since the parameters $\mathbf{A}$ carry a subtly different meaning, which under a Bayesian formulation leads to a translation of the prior distribution and in the information criteria case changes the definition of the predictor set.)

Despite the presence of non-linearities and memory in the cellular drift $\mathbf{f}$, neither the use of quadratic basis functions nor the inclusion of lagged predictors appear to improve performance in terms of AUR. In order to verify that quadratic predictors are sufficiently non-linear and that lagged predictors are sufficiently delayed, we repeated the investigation using both cubic predictors and using a delay twice as long. Results (Figs. 2.2(c), 2.2(d)) demonstrate that no improvement to the AUR scores is achieved in this way.

Corresponding results for the Swat model are shown in Fig. 2.2(b). Here we find that none of the methods performs well. Note that these results are specific to the choice of experimental sampling interval Δ; in particular if data were generated using a different interval $\Delta' \neq \Delta$ then the statistical models place different assumptions on the data-generating process, which could lead to the selection of different edges in the network estimator $\hat{N}$. This point is discussed further in Section 2.4.1.

We also performed inference using biochemical data from the experimental system reported in Cantone *et al.* [2009] (specifically the switch-on dataset). AUR scores obtained using this data (Fig. 2.3(a)) were in close agreement with those obtained using synthetic data (Fig. 2.2(a)), suggesting that the results of the simulations may be relevant to real world studies.

2.3.2.2 Uneven Sampling Intervals

Many biological time-course experiments are carried out with uneven sampling intervals. We therefore repeated the analysis above with sampling times of 0, 1, 5, 10, 15, 20, 30, 40, 50, 60, 75, 90, 105, 120, 140, 160, 180, 210, 240 and 280 minutes. Fig. 2.3(c) displays the AUR scores so obtained. We find that all the methods perform worse in the uneven sampling regime, with no method performing significantly better than random. Corresponding results for the Swat model are shown in Fig. 2.3(d). Again, here we find that none of the methods performs well.

2.3.2.3 Consistency

Fig. 2.4(a) displays AUR scores for Cantone for a large number of evenly sampled time points ($n = 100$), and the limiting case of zero measurement noise and zero cellular heterogeneity ($\sigma_{\text{meas}} = 0$, $\sigma_{\text{cell}} = 0$, even sampling intervals). Consistency (in the sense of asymptotic convergence of the network estimate to the data-generating network) may be unattainable due to non-identifiability resulting from limited exploration of the dynamical phase space. However, as we have seen, network inference can nonetheless be informative. From Fig. 2.4(a) we see that the Bayesian schemes using linear predictors approach AUR equal to unity, and in this sense show empirical consistency with respect to network inference. However, some of the other methods do not converge to the correct graph even in this limit.

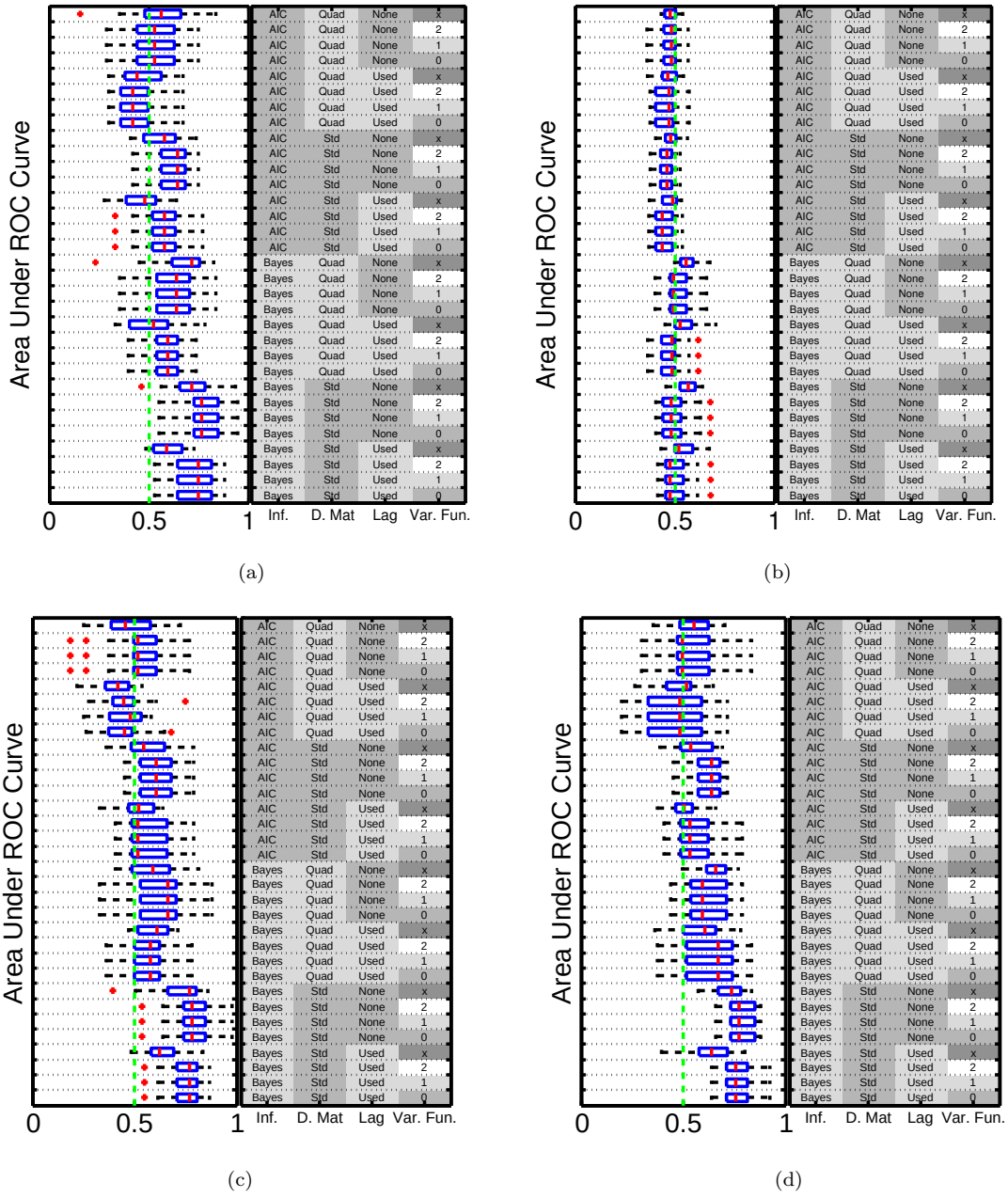


Figure 2.2: An empirical comparison of network inference schemes. Simulated experiments based on the published dynamical systems of Cantone *et al.* [2009]; Swat *et al.* [2004] allow benchmarking of performance in terms of area under ROC curves (AUR; higher scores correspond to better network inference performance). (a) AUR for Cantone *et al*, even sampling times. (b) AUR for Swat *et al*, even sampling times. (c) AUR for Cantone *et al*, even sampling times, with cubic predictors. (d) AUR for Cantone *et al*, even sampling times, with a delay of double duration.

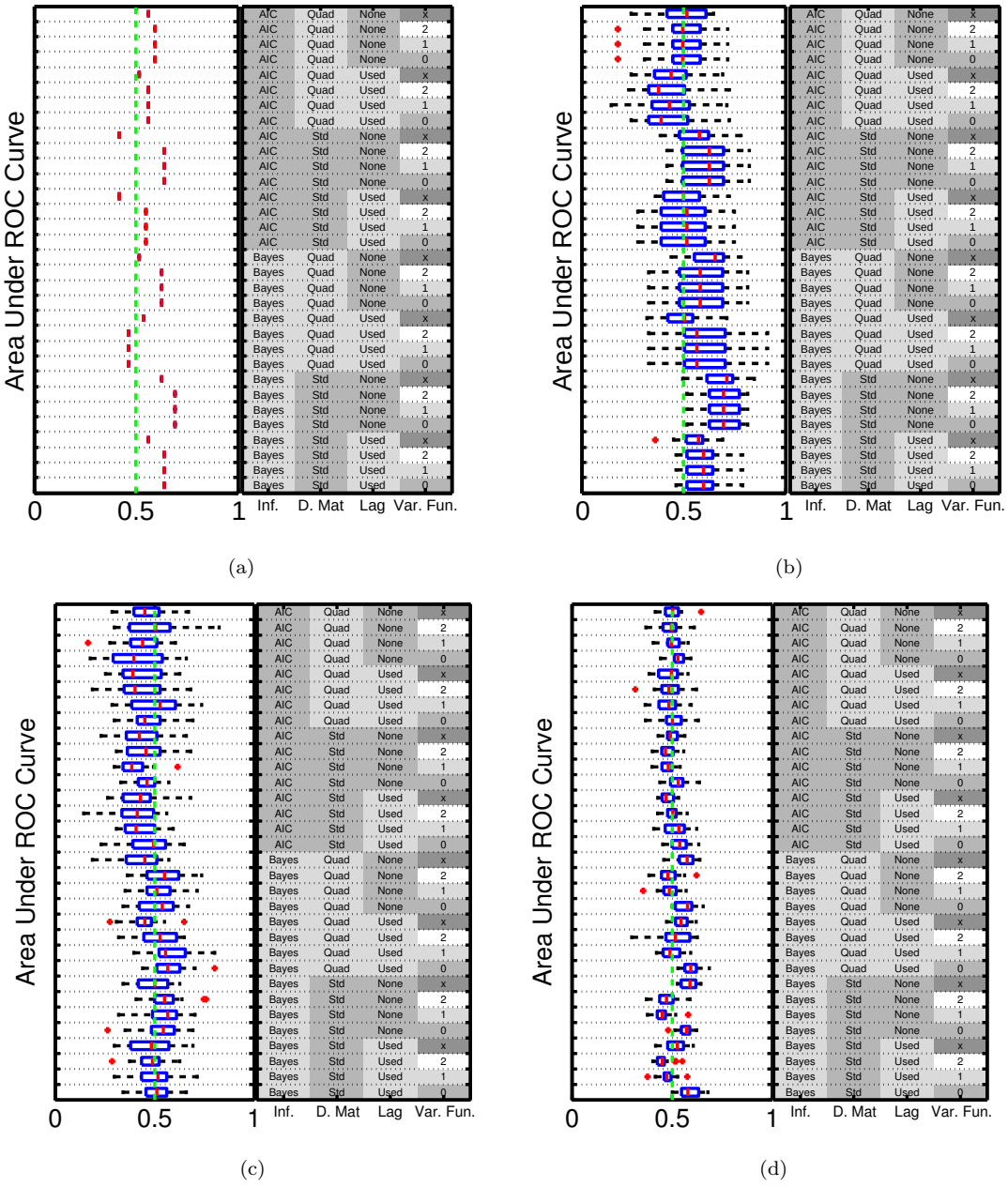


Figure 2.3: An empirical comparison of network inference schemes. Experiments based on the published dynamical systems of Cantone *et al.* [2009]; Swat *et al.* [2004] allow benchmarking of performance in terms of area under ROC curves (AUR; higher scores correspond to better network inference performance). (a) AUR for Cantone *et al*, even sampling times, based on *in vivo* data. (b) AUR for Cantone *et al*, even sampling times, using $d_{\max} = 3$. (c) AUR for Cantone *et al*, uneven sampling times. (d) AUR for Swat *et al*, uneven sampling times.

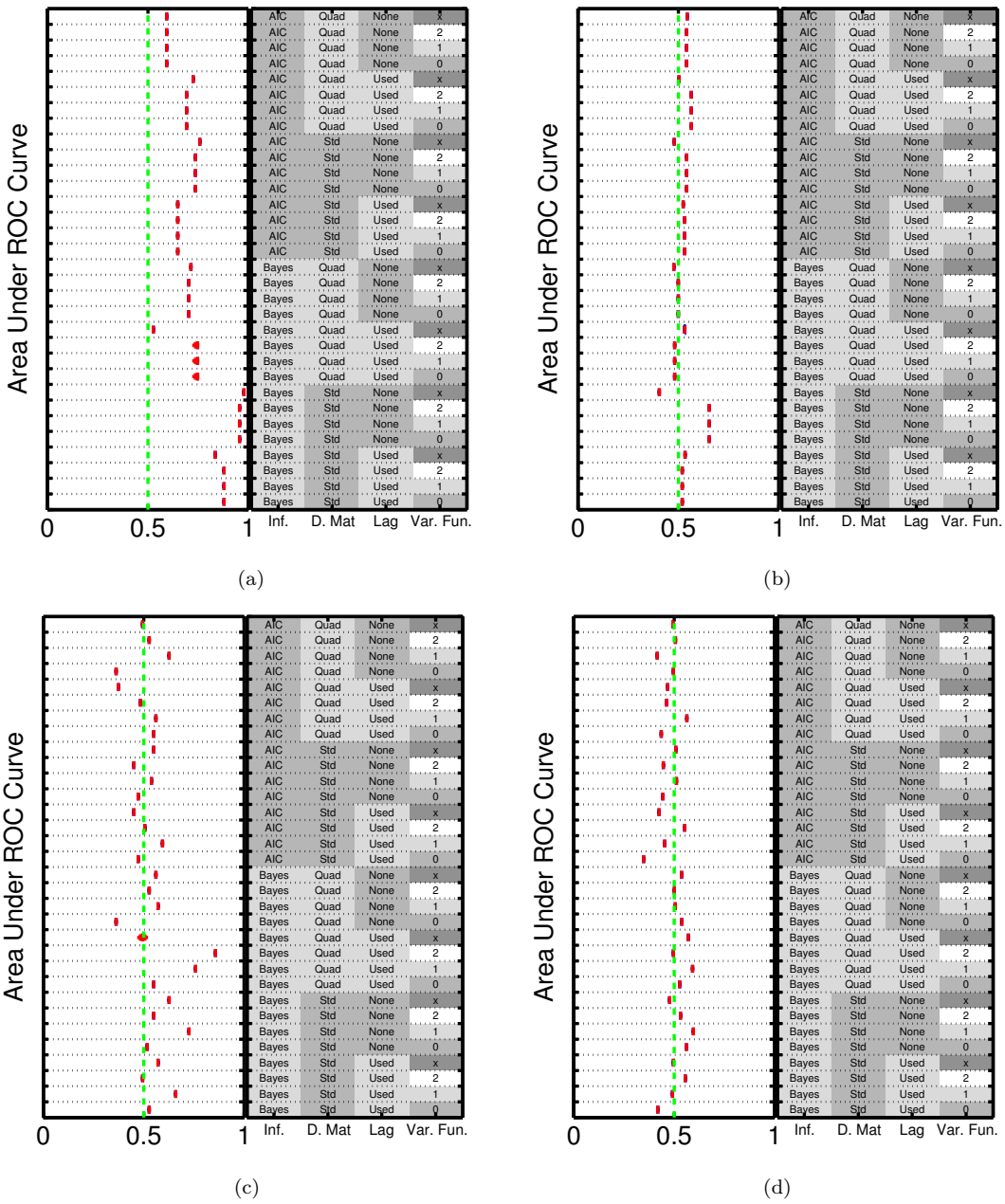


Figure 2.4: An empirical comparison of network inference schemes. Simulated experiments based on the published dynamical systems of Cantone *et al.* [2009]; Swat *et al.* [2004] allow benchmarking of performance in terms of area under ROC curves (AUR; higher scores correspond to better network inference performance). (a) AUR for Cantone *et al*, even sampling times. (b) AUR for Swat *et al*, even sampling times. (c) AUR for Cantone *et al*, uneven sampling times. (d) AUR for Swat *et al*, uneven sampling times.

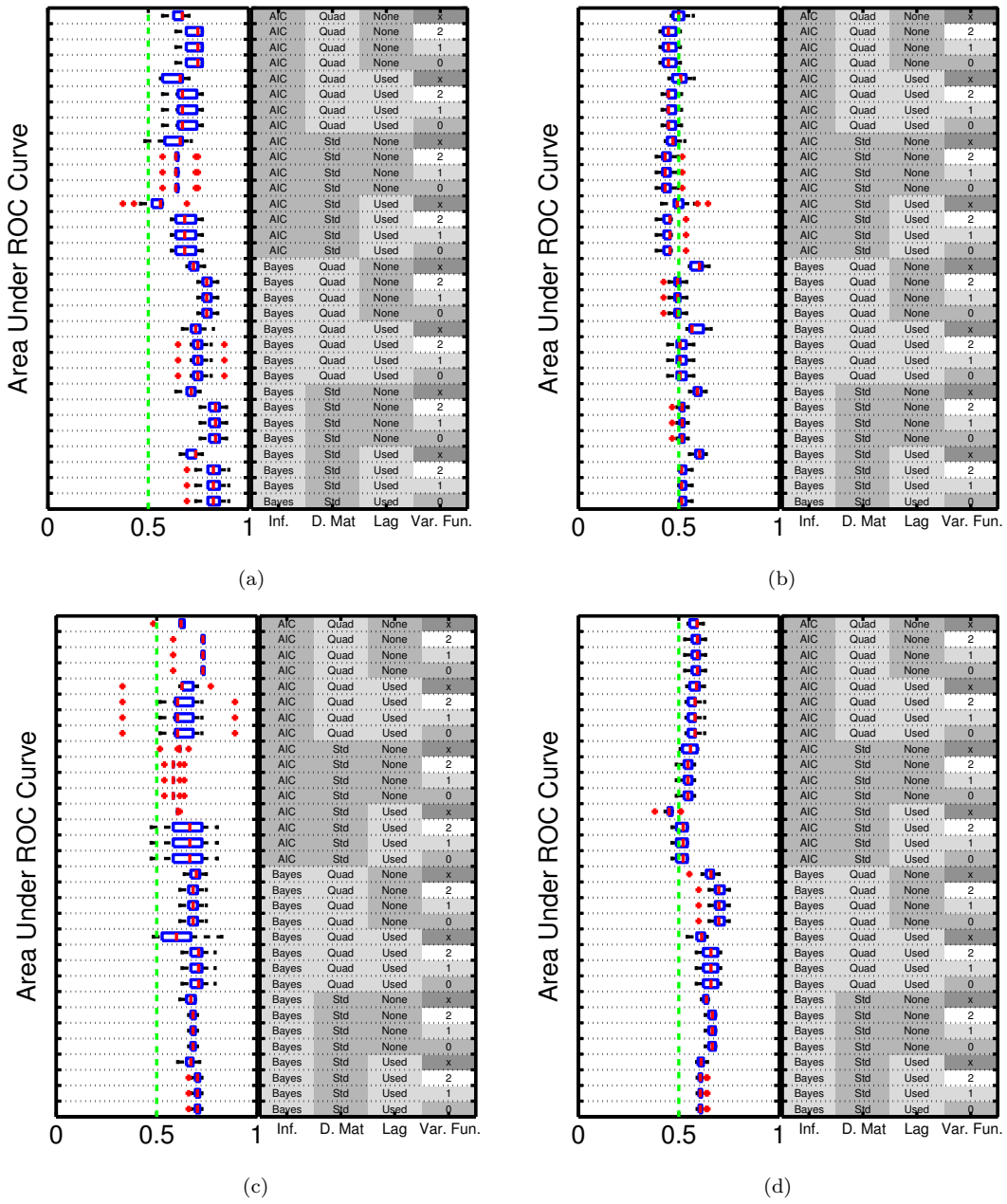


Figure 2.5: An empirical comparison of network inference schemes. Simulated experiments based on the published dynamical systems of Cantone *et al.* [2009]; Swat *et al.* [2004] allow benchmarking of performance in terms of area under ROC curves (AUR; higher scores correspond to better network inference performance). (a) AUR for Cantone *et al*, even sampling times, single cell data. (b) AUR for Swat *et al*, even sampling times, single cell data. (c) AUR for Cantone *et al*, even sampling times, inhibition data. (d) AUR for Swat *et al*, even sampling times, inhibition data.

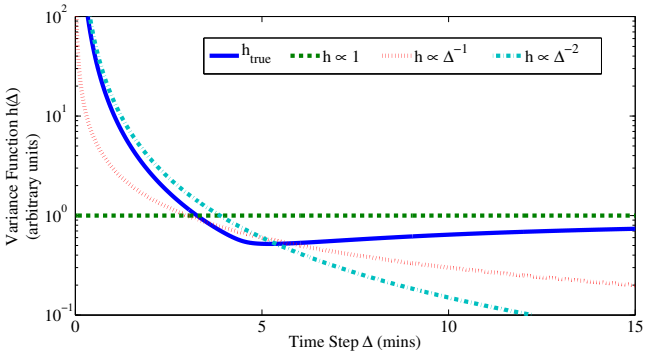


Figure 2.6: Variance functions used in literature provide partial approximation to the "true" functional form for Cantone *et al.* [2009]. For small time steps a power law $\Delta^{-\alpha}$ provides a good approximation, but for larger time steps a constant variance function may be more appropriate. In practice the precise form of h_{true} will be unknown.

2.4 Discussion

The analyses presented here were aimed at better understanding statistical network inference for biological applications. We showed how a broad class of approaches, including VAR models, linear DBNs and certain ODE-based approaches, are related to stochastic dynamics at the cellular level. We discuss a number of these aspects below and close with some views on future perspectives for network inference, including a statistical basis for more advanced work in Chapter 3.

2.4.1 Statistical Models for Longitudinal Data

In this Chapter we focussed on the popular approach of using DBNs to model time-varying dynamics. We found that uneven sampling intervals posed problems, even for methods that explicitly accounted for variation in the sampling interval. Further insight may be gained from uncertainty propagation analysis of the approximations indicated in Section 2.2.2: Assuming the true large sample process obeys $d\mathbf{X}^{\infty}/dt = \mathbf{F}(\mathbf{X}^{\infty})$, we have that under an observation process with independent additive Gaussian measurement error $\mathbf{Y}(t) \sim \mathcal{N}(\mathbf{X}^{\infty}(t), \mathbf{M})$ an expansion for the variance $\mathbb{V}(\dot{\mathbf{Y}} - \mathbf{F}(\mathbf{Y}))$ over a time interval Δ is given by

$$\mathbf{M}\Delta^{-2} + (\mathbf{I}\Delta^{-1} + D\mathbf{F})\mathbf{M}(\mathbf{I}\Delta^{-1} + D\mathbf{F})' + \dots \tag{2.16}$$

(see Section A.2.1 for details). Recall that the model family in Eqn. 2.10 approximates this variance by $h(\Delta)\mathcal{D}(\sigma_1^2, \dots, \sigma_P^2)$ where $h(\Delta) = \Delta^{-\alpha}$. From this perspective it is clear that each variance function we considered captures only partial variation due to Δ. It is therefore not surprising that performance suffers in the uneven sampling regime, which requires the variance function to apply equally to large Δ as to small Δ. Moreover, a natural choice of variance function driven by Eqn. 2.16 is not possible, since this would require knowledge of the unknown process $\mathbf{F}$. The implication for experimental design is that absent specific reasons for uneven sampling, it may be preferable to collect data at regular intervals.

Fig. 2.6 displays an approximation to the true variance function for the Cantone model (see Section A.2.3). Observe that for small sampling intervals Δ the true variance is best captured by a functional approximation of the form $h(\Delta) \propto \Delta^{-\alpha}$ with $\alpha = 1, 2$, whereas for intervals larger than 10 mins (which are more common in practice) the flat approximation $h(\Delta) \propto 1$ correctly captures the asymptotic behaviour. In applications where high frequency sampling is infeasible the flat variance function might be a sensible choice. To understand whether difficulties pertaining to sampling intervals disappear in the large sample limit, we repeated the empirical consistency analysis under uneven sampling (Figs. 2.4(c), 2.4(d)). Interestingly, we found that none of the methods appeared to be empirically consistent, and that the choice of variance function is influential. However, unevenly sampled data are common in biology and it may be the case that in some settings, the existence of multiple time scales (e.g. signalling, transcription, accumulating epigenetic alterations) mean that unevenly sampled data are nonetheless useful. Our findings suggest that care should be taken in the uneven sampling regime.

We focussed attention on DBNs together with regression models due to their prevalence in the bioinformatics literature (Table 2.1). Yet the use of DBNs together with regression models is problematic in several respects: (i) It is not clear how to enforce dynamic invariants; (ii) additional conditions required for probabilistically bounding the trajectories (i.e. $\mathbb{R}(\boldsymbol{A} + \Delta \boldsymbol{I}) < 0$) can lead to intractability of the likelihood; (iii) if the incremental covariance $\mathrm{Cov}(\dot{\boldsymbol{Y}}(t_j) - \boldsymbol{A}\dot{B}'_{j,\bullet})$ is defined to be diagonal on a characteristic time scale Δ then, unless $\boldsymbol{A}$ is a diagonal drift matrix, the incremental covariance will be non-diagonal on a time scale $\Delta' \neq \Delta$. Problem (iii) is particularly concerning, since for data which are unevenly sampled in time, the uncorrelated statistical models considered here do not capture the Δ-dependent covariance of the data-generating process. Moreover, even in the favourable case of even sampling, inference requires the major assumption that data were generated on a time scale such that the incremental distribution is uncorrelated. In practice, this means a change in the experimental sampling time interval Δ can lead to the inclusion of different edges in the network estimator $\hat{N}$, even in the favourable limit of large data. There exist alternative statistical methodologies which do not suffer from problems (i-iii); these include Continuous Time Bayesian Networks [Nodelman *et al.*, 2002] and Multiregression Dynamic Models [Queen and Smith, 1993]. It would be interesting to investigate whether these alternative approaches offer gains in the uneven sampling regime considered above.

2.4.2 Interventional Data

The Cantone data are favourable in the sense that trajectories show interesting time-varying behaviour under global perturbation, exploring a large proportion of the dynamical phase space. However such behaviour is dependent on the specific dynamical system and is not displayed by the Swat model, which has a much larger phase space, being a nine-dimensional dynamical system. This may help explain the poor performance of all the methods on this latter model using global perturbation data and perhaps reinforces the intuitive notion that dynamics that are favourable (in this informal sense) facilitate network inference. In some cases, perturbation data are available in which individual variables are inhibited (e.g. by RNA interference, gene knock-outs or inhibitor treatments). Such data have the potential to explore much more of the dynamical phase space, including regions which cannot be accessed without direct inhibition of specific molecular components. This is an important consideration because the statistical estimators described in Section 2.2.4 take the form

$$\hat{\mathbf{A}} = \langle D\mathbf{f}(\mathcal{F}_{\mathbf{X}}) \rangle_{\mathbf{X}} \tag{2.17}$$

where the average is over the region in state space visited during the experiments. Clearly if this region is only a small subspace of phase space then the estimate Eqn. 2.17 will be poor compared to one based on the entire phase space.

To investigate the added value of interventional treatments for network inference, we repeated both the Cantone and Swat analyses with an ensemble of datasets obtained by inhibiting each variable in turn; this gave 5 and 9 datasets for Cantone and Swat respectively. Whilst no improvement to the Cantone AUR scores was observed (Fig. 2.5(c)), there was improved performance for Swat (Fig. 2.5(d)). This suggests that global perturbations are insufficient to explore the Swat dynamical phase space, and supports the intuitive notion that intervention experiments may be essential for inference regarding larger dynamical systems. Nevertheless AUR scores remain far from unity. This may be because the Swat drift function contains complex interaction terms which single interventions alone fail to elucidate. An important problem in experimental design will be to estimate how much (possibly combinatorial) intervention is required to achieve a certain level of network inference performance. Theoretical work including Hauser and Bühlmann [2012] quantifies the extent to which interventions are necessary to distinguish between competing causal models; we focussed on the practical challenge of constructing statistical estimators for this purpose.

We considered precise artificial intervention of single components *in silico*. However, biological interventions may be imprecise and imperfect. For example, RNA interference achieves only incomplete silencing of the target and small-molecule inhibitors (Section 1.1.4) may have off-target effects. Moreover, at present such interventions are not instantaneous nor truly exogenous. This means that in many cases the system itself may be changed by the intervention, rendering resulting predictions inaccurate for the native system of interest. There remains a need for novel statistical methodology capable of analysing time-course data under biological interventions. Existing literature in causal inference [Pearl, 2009] and related work in graphical models [Eaton and Murphy, 2007] are relevant, but in biological applications

it may also be important to consider the mechanism of action of specific interventions.

2.4.3 Non-linear Models

We focused on linear statistical models. Clearly, linear models are inadequate in many cases. For example Rogers *et al.* [2007] demonstrate the benefit of a non-linear model based on Michaelis-Menten chemical kinetics for inference of transcription factor activity. In Chapter 3 we attempt to integrate non-linear models of enzyme kinetics into inference for protein phosphorylation networks. Alternatively Äijö and Lähdesmäki [2009] consider the use of a non-parametric Gaussian process (GP) interaction term in the regression, which is naturally more flexible than linear regression using finitely many basis functions. This may help to overcome the linearity restriction, but introduces additional degrees of freedom, including the GP covariance function and associated hyper-parameters. Whilst a thorough comparison of such approaches was beyond the scope of this thesis, the potential utility of non-parametric interaction terms is worthy of investigation. In this study we observed that neither the use of predictor products nor lagged predictors led to improved performance; this may reflect non-trivial coupling between cellular dynamics and the observed data.

2.4.4 Single-Cell Data

In the future it may become possible to obtain high-throughput measurements for single cell expression levels $\mathbf{X}^k$ non-destructively (e.g. by live cell imaging), producing truly longitudinal datasets. It is interesting to consider how such data may impact upon the performance of regression-based network inference. Under independent additive Gaussian measurement error $\mathbf{Y}(t) \sim \mathcal{N}(\mathbf{X}^k(t), \mathbf{M})$ uncertainty propagation for the single cell variance $\mathbb{V}(\dot{\mathbf{Y}} - \mathbf{f})$ over a time interval Δ, in analogy with Eqn. 2.16, is given by

$$\mathbf{M}\Delta^{-2} + (\mathbf{I}\Delta^{-1} + D\mathbf{F})\mathbf{M}(\mathbf{I}\Delta^{-1} + D\mathbf{F})' + \Delta^{-1}\mathbf{g}\mathbf{g}' + \dots \tag{2.18}$$

(see Section A.2.2). Thus a (single) longitudinal single cell dataset contains less information about the drift $\mathbf{f}$ than aggregate data (Eqn. 2.16) due to cellular stochasticity $\mathbf{g}$. However, multiple longitudinal datasets may jointly contain more information than a single aggregate dataset. To empirically test the utility of such data, we carried out network inference using 10 such longitudinal single-cell datasets on both the Cantone and Swat models, observed at even intervals with the same magnitude of measurement error as aggregate data. Results (Figs. 2.5(a), 2.5(b)) show a small improvement to the mean AUR scores, but reduction by a factor of about two in the variance of these scores (compared with the corresponding non-longitudinal data), implying that the network estimators may be converging to an incorrect network. Bias may occur when the cellular drift $\mathbf{f}$ is not well approximated by a linear function, as is the case for the Swat model. Consider the idealized scenario where $\mathbf{f} \equiv \mathbf{f}(\mathbf{X})$ is Markovian and it is possible to observe longitudinal, single cell expression levels. Under these apparently favourable circumstances even estimators obtained after a thorough exploration of state space may not offer good approximations. As a toy example consider the cellular drift

$$\mathbf{f} : [0, \pi]^2 \to \mathbb{R}, \quad \mathbf{f}(\mathbf{X}) = \begin{pmatrix} \sin(X_2) \\ \sin(X_1) \end{pmatrix} \tag{2.19}$$

which is not well approximated by a linear function over the state space $[0, \pi]^2$. In this case averaging leads to cancellation

$$\langle D\mathbf{f}(\mathbf{X}) \rangle_{\mathbf{X} \in [0,\pi]^2} = \left\langle \begin{pmatrix} 0 & \cos(X_2) \\ \cos(X_1) & 0 \end{pmatrix} \right\rangle_{\mathbf{X} \in [0,\pi]^2} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{2.20}$$

so that no interactions are inferred. Under such circumstances network inference is no longer possible using the naïve linear regression approach. This suggests that network inference rooted in non-linear models may be needed to fully exploit longitudinal single-cell data in the future. A related line of work addresses heterogeneity of the drift function in time by coupling DBNs with change point processes [Dondelinger *et al.*, 2012; Grzegorczyk and Husmeier, 2011; Kolar *et al.*, 2009; Lèbre *et al.*, , 2010]. A promising direction would be piecewise linear regression modelling for network inference applications, where the heterogeneity appears in the spatial domain.

2.4.5 Future Perspectives

We found that a simple linear model could approximately recover network structure from globally perturbed time-course data from the Cantone system. It is encouraging that inference based only on associations between variables, none of which were explicitly intervened upon, can in some cases be effective. Interventional designs should further enhance prospects for network inference. On the other hand, theoretical arguments, and the results we showed from the Swat system, emphasize that in some cases network structure may not be identifiable, even at the coarse level required for qualitative biological prediction. On balance, we believe that network inference can be useful in generating biological hypotheses and guiding further experiment. However, the concerns we raise motivate a need for caution in statistical analysis and interpretation of results. At the present time, we do not believe network inference should be treated as a routine analysis in bioinformatics applications, but rather as an open research area that may, in future, yield standard experimental and statistical protocols.

Some specific recommendations that arise from the results presented here are:

- *A default model.* Our results suggest that a reasonable default choice of model for typical applications uses the standard design matrix with no lagged predictors and a flat variance function, corresponding to the linear model

$$\dot{\mathbf{Y}}(t_j) \sim \mathcal{N}(\mathbf{A}\mathbf{Y}(t_{j-1}), \mathcal{D}(\sigma_1^2, \ldots, \sigma_P^2)). \tag{2.21}$$

  Coupled with the Bayesian variable selection scheme outlined in Section 2.2.4.1, this simple model produced empirically consistent network estimators for Cantone using evenly sampled global perturbation data. In Chapter 3 this model forms the basis for more sophisticated network inference procedures.

- *Diagnostics and validation.* Network inference as described in Section 2.2.1 does not enjoy general theoretical guarantees and the ability to successfully elucidate network structure depends on details of the specific system under study. Therefore careful empirical validation on a case-by-case basis is essential. This should include statistical assessment of model fit, robustness and predictive ability and where possible systematic validation using independent interventional data (though this may itself be challenging).

- *Experimental design.* We suggest sampling evenly in time as a default choice. Interventional designs may be helpful to effectively explore larger dynamical phase spaces. However, to control the burden of experimentally exploring multiple time points, molecular species, interventions, culture conditions and biological samples, adaptive designs that prune experiments based on informativeness for the specific biological setting may be helpful (e.g. [Xu *et al.*, 2010]).

In conclusion, linear statistical models for networks are closely related to models of cellular dynamics and can shed light on patterns of biochemical regulation. However, biological network inference remains profoundly challenging, and in some cases may not be possible even in principle. Nevertheless, studies aimed at elucidating networks from high-throughput data are now commonplace and play a prominent role in biology. For this reason there remains an urgent need in this application area for both (i) improved methodology and (ii) theoretical and empirical investigation of existing approaches. For (i) main challenges include resolving the problem of the dependence of statistical estimators upon a particular choice of time scale, accounting for measurement errors present in the predictor variables, addressing interpretation under the possibility of multicollinearity, and addressing the possibility of latent variables in this context. For (ii) main challenges include establishing mathematical conditions for consistency of the Bayes factors, establishing finite-sample properties for network estimators, and improving the realism of simulation benchmarking approaches. Furthermore, there remain many open questions in experimental design and analysis of designed experiments in this setting.

Chapter 3

Network Inference and Dynamical Prediction Using Chemical Kinetics

In Chapter 2 we saw that statistical estimation of networks is usually based on linear (or discrete) formulations. However biological networks N represent structural summaries of dynamical systems which are generally non-linear. In this Chapter we present methodology for network inference that is rooted in non-linear chemical kinetics. This is done by considering a dynamical system based on a chemical reaction graph G that summarises chemical reactions and associated parameters. Inference regarding G is carried out within a Bayesian framework that accounts for both model complexity and fit-to-data. Prediction of dynamical behaviour is achieved by averaging over both reaction graphs and associated parameters, allowing prediction even when the reaction graph itself is unknown or uncertain. We show results on data simulated from a recent mechanistic model of MAPK signalling and on phosphoproteomic data from cancer cell lines. Our results demonstrate that the use of non-linear kinetics within statistical network modelling can yield gains in estimation of biological networks as well as dynamical prediction.

3.1 Introduction

Statistical network models are typically rooted in linear or discrete descriptions of biological dynamics (see Chapter 2). The statistical and computational tractability of such formulations facilitates exploration of large spaces of networks. On the other hand, when the network topology is known, non-linear ordinary differential equations (ODEs) are widely used to model dynamics Chen *et al.* [2009]; Kholodenko [2006]; Leskovac [2003]; Steijaert *et al.* [2010]. The intermediate case where non-linear ODEs are employed to select between network models has received less attention.

For causal inference, linear formulations remain unsatisfactory for several reasons: (1) Variates may be highly correlated, often due to underlying dynamics. Flexibility inherent in the linear approach requires that modifications are made to the linear model in order to exclude non-causal but highly correlated variates [Cho and Fryzlewicz, 2012]. (2) Symmetry of the linear equivalence in general limits identification of underlying causal relationships [Pearl, 2009; Peters *et al.*, 2011]. (3) When the data generating model is non-linear, the linear model may produce inefficient or inconsistent estimation, attributing causal status to artefacts resulting from model misspecification [Heagerty and Kurland, 2001; Lv and Liu, 2010]. Indeed, we saw in Chapter 2 that such bias can prevent recovery of the correct network even in favourable asymptotic limits of large sample size and low noise.

Biochemical processes underlying biological networks are often highly non-linear and, in many settings, non-linear dynamical models of relevant biochemical processes are available (see Chapter 2). Where such models are available, it is natural to ask whether they may be exploited to facilitate network inference, since an appropriate non-linear formulation may have enhanced power to exclude non-causal variates. Note that due to the added complexity of non-linear formulations, it is not *a priori* obvious that they must outperform simpler models, under practical conditions of sample size and measurement noise. As we show below, such information can be valuable in guiding exploration of network topologies.

Kinetic formulations have been widely studied in the systems biology literature and, as discussed in Chapter 2, recently there has been much interest in statistical inference for such systems [e.g. Chen *et al.*, 2009; Xu *et al.*, 2010]. Our work is in a similar vein, but focuses on network inference *per se*.

While biochemical assays have become cheaper, it remains the case that experimental designs must often negotiate a trade off between more conditions (e.g. perturbations, biological samples, technical replicates) and temporal resolution. Methodologies which can exploit knowledge concerning relevant dynamical systems in the steady-state setting are therefore potentially valuable.

Here, we propose an approach that uses non-linear dynamical models to carry out both network inference and dynamical prediction. We proceed by considering a dynamical system $\boldsymbol{f}_G$ that depends on a chemical reaction graph G, characterising reactions in the system Craciun and Pantea [2008]. Letting $\boldsymbol{X}(t) \in \mathbb{R}_+^P$ denote a state vector describing system configuration at time t, we have $\dot{\boldsymbol{X}}(t) = \boldsymbol{f}_G(\boldsymbol{X}(t), \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ collects together all unknown parameters, such as rates of reaction. Given time-course data $\boldsymbol{y}$ consisting of noisy measurements of $\boldsymbol{X}$, we carry out inference within a Bayesian framework to obtain a posterior distribution over reaction graphs G,

$$p(G|\boldsymbol{y}) \propto p(G)p(\boldsymbol{y}|G) = p(G) \int p(\boldsymbol{y}|\boldsymbol{\theta}, G)p(\boldsymbol{\theta}|G)d\boldsymbol{\theta} \tag{3.1}$$

where the marginal likelihood $p(\boldsymbol{y}|G)$ captures how well the chemical reaction graph G describes data $\boldsymbol{y}$, taking into account model complexity. Recall that from Chapter 1 that a biological network $N \equiv N(G)$ is a coarse summary of the reaction graph G in which each species appears as a single node and directed edges indicate that the parent is involved in chemical reaction(s) which have the child as product. In contrast to linear or discrete approaches based on biochemical networks N, our likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}, G)$ depends on (richer) reaction graphs G with corresponding dynamical models $\boldsymbol{f}_G$. Importantly, we do not assume detailed knowledge of the dynamical system, but only the broad class to which dynamics and associated equilibria may belong. Indeed, the approach we describe does not require any kinetic parameters to be known *a priori*, nor knowledge of the reaction topology, and is in that sense directly comparable with conventional network inference methods. Its potential advantage stems from then rich yet constrained nature of the class of functional relationships that are considered. As recently discussed in Peters *et al.* [2011], non-linear functional forms can aid in identification of underlying causal relationships. Indeed non-linear formulations are able to confer asymmetries between nodes which may be sufficient to enable orientation of all edges [Peters *et al.*, 2011]. As a consequence, our proposal can in principle aid in causal network inference. We demonstrate empirically below that our approach outperforms linear methods with respect to causal network inference. Further, in contrast to linear models, in our approach the mechanistic roles of individual variables are respected. This facilitates analysis of interventional data and enhances scientific interpretability. Since prediction of dynamical behaviour (e.g. trajectories under intervention) in general depends on the reaction graph, in settings where the graph itself is unknown or uncertain, our methodology can aid in prediction. Empirical results in this Chapter demonstrate the use of our methods for dynamical prediction in this setting.

In a recent paper, [Xu *et al.*, 2010] demonstrated that statistical model selection based on four hand-crafted non-linear ODE models could be used to elucidate signalling mechanism. The work we present differs from [Xu *et al.*, 2010] in motivation and approach in that we carry out general network inference (over correspondingly large model spaces) and use automatically-generated rather than hand-crafted biochemical models.

The approach we propose is general and can be used in many settings where kinetic formulations are available to describe the dynamics, including gene regulation, metabolism and protein signalling. For definiteness we focus on protein signalling networks mediated by phosphorylation. As discussed in Section 1.1.3, protein phosphorylation is central to diverse biological processes and plays a key role in disease states including cancer [Weinberg, 2007]. Phosphorylation kinetics have been widely studied [Kholodenko, 2006]; here we employ ODEs based on Michaelis-Menten functionals (see Chapter 1, Section 1.3.4).

The remainder of the Chapter is organized as follows. First, we introduce the model and associated statistical formulation. Second, we discuss statistical inference and prediction using these models. Third, we show empirical results, comparing our proposal to existing approaches. We carry out *in silico* assessment using a mechanistic model of MAPK signalling [Xu *et al.*, 2010], under a range of data-generating regimes, and apply the method to phosphoproteomic time course data from human cancer cell lines. Finally, we discuss our findings, including directions for further work.

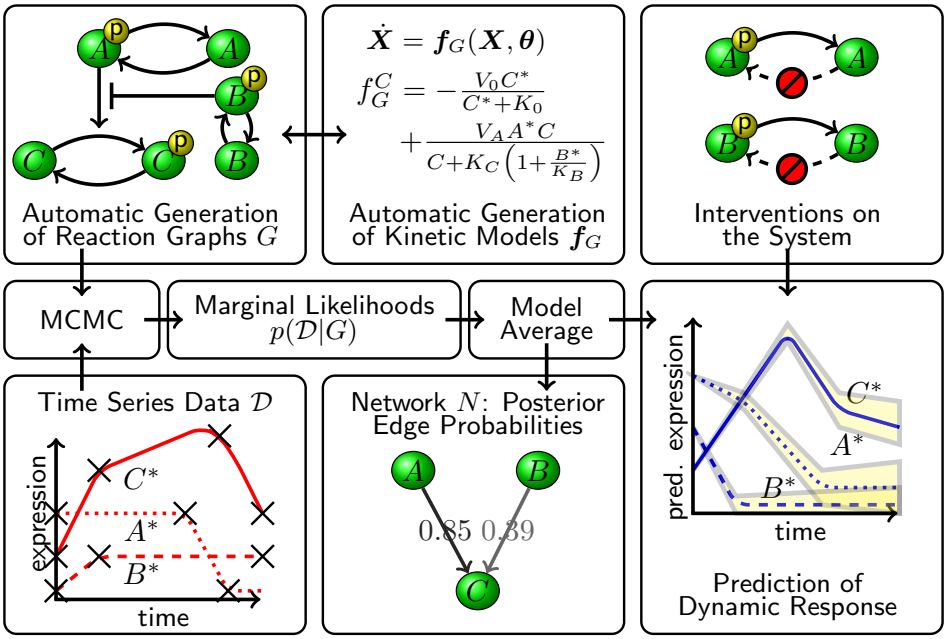


Figure 3.1: Chemical Model Averaging (CheMA). Chemical reaction graphs G summarize interplay that is described quantitatively by a kinetic model $\boldsymbol{f}_G$. Candidate graphs G are scored against observed time course data $\boldsymbol{y}$ in a Bayesian framework. Averaging over the space of reaction graphs G facilitates both network inference and dynamical prediction. A biological network N gives a coarse summary of the system; marginal posterior probabilities of edges in N quantify evidence in favour of causal relationships. Prediction of dynamic response to hypothetical drug regimens may be carried out even when the true reaction graph is unknown.

3.2 Methods

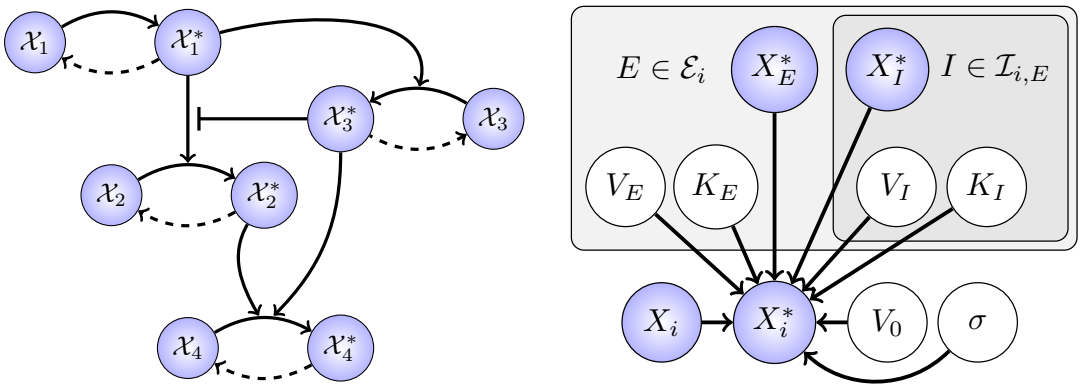
In this Section we describe the proposed methodology in the specific context of protein phosphorylation networks (Fig. 3.1). We begin by constraining chemical reaction graphs G to reflect known biochemistry in this setting and consider associated ODE models. We then discuss statistical inference and prediction.

3.2.1 Reaction Graphs for Protein Phosphorylation

We consider proteins $\mathcal{X}_1, \ldots, \mathcal{X}_P$. Each $\mathcal{X}_i$ can be phosphorylated to $\mathcal{X}_i^*$; the set of phosphorylated proteins is $\mathcal{X}^*$. Phosphorylation reactions $\mathcal{X}_i \to \mathcal{X}_i^*$ are catalysed by enzymes $\mathcal{X}_E^* : E \in \mathcal{E}_i$; the subscript indicates that each protein may have a specific set of enzymes (enzymes catalysing phosphorylation are known as kinases, we use both terms interchangeably). We consider the case in which the kinases themselves are phosphorylated proteins (if phosphorylation of $\mathcal{X}_i$ is not driven by an enzyme in $\mathcal{X}^*$, we set $\mathcal{E}_i = \varnothing$). For simplicity we do not consider multiple phosphorylation sites, other post-translational modifications such as ubiquitinylation, nor spatial effects. The ability of enzyme $\mathcal{X}_E : E \in \mathcal{E}_i$ to catalyse phosphorylation of $\mathcal{X}_i$ may be inhibited by phosphoproteins $\mathcal{X}_I : I \in \mathcal{I}_{i,E} \subset \mathcal{X}^*$; the double subscript indicates that inhibition is specific to both substrate $\mathcal{X}_i$ and enzyme E (see below). The chemical reaction graph G provides a visual representation of the sets $\mathcal{E}_i$ and $\mathcal{I}_{i,E}$; Fig. 3.1 contains an illustrative example (top left). A biological network $N(G)$ is formed by drawing exactly P vertices and edges (i, j) indicating that $\mathcal{X}_i^*$ is either an enzyme catalysing phosphorylation of $\mathcal{X}_j$, or an inhibitor of such an enzyme. That is, $(i, j) \in N \iff i \in \mathcal{E}_j \vee \exists E \cdot i \in \mathcal{I}_{j,E}$. For the example shown in Fig. 3.1, the corresponding biological network N is the directed graph $A \to C \leftarrow B$.

3.2.2 Phosphorylation Kinetics

The chemical reaction graph G can be decomposed into local reaction graphs G_i describing enzymes (and their inhibitors) for phosphorylation of protein $\mathcal{X}_i$. For simplicity of exposition, in what follows we



(a) (b)

Figure 3.2: Statistical models of enzyme kinetics. (a) An example of a chemical reaction graph for protein phosphorylation. (b) A statistical formulation (graphical model) for phosphorylation of species $\mathcal{X}_i$ is characterised by specifying the index set of kinases ($E \in \mathcal{E}_i$) and their inhibitors ($I \in \mathcal{I}_{i,E}$). [Bounding boxes are used to indicate multiplicity of variables, shaded nodes are observed with noise.]

consider inference concerning G_i; thus $\mathcal{X}_i$ plays the role of the substrate. We write $X_i, X_i^* \in \mathbb{R}_+$ for the concentrations of protein species $\mathcal{X}_i, \mathcal{X}_i^*$ respectively.

We use kinetic models $\boldsymbol{f}_G$ based on Michaelis-Menten functionals [Kholodenko, 2006; Leskovac, 2003; Steijaert *et al.*, 2010]. The rate of phosphorylation due to kinase E is given by $V_E X_E^* X_i^h / (X_i^h + K_E^h)$, which acknowledges variation of kinase concentration X_E^* and permits kinase-specific response profiles, parametrised by K_E and h, with rate constant V_E. In subsequent experiments the Hill coefficient h is taken equal to 1 (non-cooperative binding). We entertain competitive inhibition, where substrate and inhibitor I compete for the same binding site on the enzyme ($\mathcal{X}_E^* \mathcal{X}_I^* \rightleftharpoons \mathcal{X}_E^* \rightleftharpoons \mathcal{X}_E^* \mathcal{X}_i \rightarrow \mathcal{X}_E^* + \mathcal{X}_i^*$). When multiple inhibitors ($\mathcal{X}_I^*, \mathcal{X}_J^*$) are present, they are assumed to act exclusively, competing for the same binding site on the enzyme ($\mathcal{X}_E^* \mathcal{X}_I^* \rightleftharpoons \mathcal{X}_E^* \rightleftharpoons \mathcal{X}_E^* \mathcal{X}_J^*$), corresponding mathematically to a rescaling of the Michaelis-Menten parameter $K_E \mapsto K_E(1 + \sum_{I \in \mathcal{I}_{S,E}} X_I^*/K_I)$. We do not model phosphatase specificity; in particular, dephosphorylation is assumed to occur at a rate $V_0 X_i^*/(X_i^* + K_0)$, depending on a Michaelis-Menten parameter K_0 and taking a maximal value V_0. Combining these assumptions produces a kinetic model for phosphorylation of substrate $\mathcal{X}_i$, given by

$$f_{G,i}(\boldsymbol{X}, \boldsymbol{\theta}_i) = -\frac{V_0 X_i^*}{X_i^* + K_0} + \sum_{E \in \mathcal{E}_S} \frac{V_E X_E^* X_i}{X_i + K_E \left(1 + \sum_{I \in \mathcal{I}_{S,E}} \frac{X_I^*}{K_I}\right)} \tag{3.2}$$

where state vector $\boldsymbol{X}$ collects together concentrations $X_1, \ldots, X_P, X_1^*, \ldots, X_P^*$, parameter vector $\boldsymbol{\theta}_i$ contains the maximum rates $\boldsymbol{V}$ and Michaelis-Menten constants $\boldsymbol{K}$, and the (local) graph G_i specifies the sets $\mathcal{E}_i$ and $\mathcal{I}_{i,E}$ (Fig. 3.2(a)). Further information on the construction of Eqn. 3.2 can be found in Chapters 2-4 of Leskovac [2003]; see also Section 1.3.4. The complete dynamical system $\boldsymbol{f}_G$ is given by taking, for each species $i \in \mathcal{P}$, a model akin to Eqn. (3.2). In this way we are able to automate the generation of parametric ODE models for the system.

3.2.3 Statistical Formulation

Data $\boldsymbol{y}$ comprise observations $y_i(t_j)$ and $y_i^*(t_j)$ proportional to the concentrations X_i, X_i^* of unphosphorylated and phosphorylated forms, respectively, of species $\mathcal{X}_i, \mathcal{X}_i^*$ at discrete times t_j, $0 \leqslant j \leqslant n$. Data are scale-normalized to give unit mean for each protein. Observables are related to dynamics via an Euler approximation $z_i(t_j) = (y_i^*(t_j) - y_i^*(t_{j-1}))/(t_j - t_{j-1})$. Below we describe inference regarding the local reaction graph G_i for a single species $\mathcal{X}_i$; iterating over $i \in \mathcal{P}$ permits inference concerning the complete reaction graph G. The ODE model $f_{G,i}$ (Eqn. (3.2)) is formulated as a statistical model by constructing, conditional upon (unknown) Michaelis-Menten parameters $\boldsymbol{K}$, a design matrix $\boldsymbol{D}(\boldsymbol{K})$ with

rows

$$\left[-\frac{y_i^*}{y_i^* + K_0}, \ldots, \underbrace{\frac{y_E^* y_i}{y_i + K_E \left(1 + \sum_{I \in \mathcal{I}_{i,E}} \frac{y_I^*}{K_I}\right)}}_{E \in \mathcal{E}_i}, \ldots \right], \tag{3.3}$$

and then interpreting Eqn. (3.2) statistically as

$$\boldsymbol{z} \quad = \quad \boldsymbol{D}(\boldsymbol{K})\boldsymbol{V} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{3.4}$$

where $\boldsymbol{z} = [z_i(t_1), \ldots, z_i(t_n)]^T$. Here $\mathcal{N}$ denotes a normal distribution with uncorrelated covariance $\sigma^2 \boldsymbol{I}$ and, as above, $\boldsymbol{V}$ is the vector of maximum reaction rates (Fig. 3.2(b)). Note that Eqn. 3.4 follows the recommended statistical formulation (Eqn. 2.21) from Chapter 2 up to the specific form of the predictor variables.

3.2.4 Bayesian Inference

In the Bayesian setting, prior distributions over parameters are needed to complete the model specification. We use truncated normal priors $\mathcal{N}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ inherited from the untruncated distribution. Truncation ensures non-negativity of parameters, whilst normality facilitates partial conjugacy (below); additional information on truncated normals is provided in Section B.1. In order to elicit hyper-parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, we follow [Xu *et al.*, 2010] and assume all processes occur on observable time and concentration scales, that is $\boldsymbol{\mu_V}, \boldsymbol{\mu_K} \sim \mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)$ reflects that the data $\boldsymbol{y}$ are standardised *a priori*. For Michaelis-Menten parameter covariance $\boldsymbol{\Sigma_K}$ we assume independence of the components K_i, so that $p(\boldsymbol{K}|G_i) = \mathcal{N}_T(\boldsymbol{K}; \boldsymbol{\mu_K}, \nu \boldsymbol{I})$. For maximum reaction rate covariance $\boldsymbol{\Sigma_V}$ we take a unit information formulation of the truncated g-prior, so that $p(\boldsymbol{V}|\boldsymbol{K}, \sigma, G_i) = \mathcal{N}_T(\boldsymbol{V}; \boldsymbol{\mu_V}, n\sigma^2(\boldsymbol{D}'\boldsymbol{D})^{-1})$ [Zellner, 1986] and for the noise parameter we use a Jeffreys prior $p(\sigma|G_i) \propto 1/\sigma$. These latter choices render the formulation partially conjugate, with the conditional density $p(\boldsymbol{V}, \sigma|\boldsymbol{K}, G_i, \boldsymbol{y})$ given in closed form as

$$p(\boldsymbol{V}, \sigma|\boldsymbol{K}, G_i, \boldsymbol{y}) \quad = \quad \mathcal{N}_T(\boldsymbol{V}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{IG}(\sigma; a, b), \tag{3.5}$$

where $\boldsymbol{\mu} = \mathbf{1}/(n+1) + n/(n+1) \times (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{z}$, $\boldsymbol{\Sigma} = \sigma^2 n/(n+1) \times (\boldsymbol{D}'\boldsymbol{D})^{-1}$, $a = (n-1)/2$, $b = (1/2)(\mathbf{1}'\boldsymbol{D}'\boldsymbol{D}\mathbf{1}/n + \boldsymbol{z}'\boldsymbol{z} - n/(n+1) \times \boldsymbol{z}'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{z})$ and $\mathcal{IG}(\bullet; a, b)$ is an inverse gamma density with shape and scale parameters a, b respectively.

3.2.5 Marginal Likelihood

Write $\theta = (\boldsymbol{V}, \boldsymbol{K}, \sigma)$ for the vector of parameters associated with inference for variable i. (In what follows we suppress dependence of parameters on the variable i.) Partial conjugacy of the above formulation permits an efficient Metropolis-within-Gibbs Markov chain Monte Carlo (MCMC) sampling scheme for the parameter posterior distribution $p(\boldsymbol{\theta}|G_i, \boldsymbol{y})$. The conditional density $p(\boldsymbol{V}, \sigma|\boldsymbol{K}, G_i, \boldsymbol{y})$ is given in closed form as in Eqn. 3.5 above, while a Metropolis-Hastings acceptance step allows sampling from the remaining conditional $p(\boldsymbol{K}|\boldsymbol{V}, \sigma, G_i, \boldsymbol{y})$. To estimate marginal likelihoods from sampler output we use the method of [Chib and Jeliazkov, 2001], evaluating the identity

$$p(\boldsymbol{y}|G_i) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, G_i) p(\boldsymbol{\theta}|G_i)}{p(\boldsymbol{\theta}|\boldsymbol{y}, G_i)} \tag{3.6}$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ using a Monte Carlo estimate of the posterior ordinate $p(\boldsymbol{\theta}^*|\boldsymbol{y}, G_i)$. The point $\boldsymbol{\theta}^*$ is usually taken to be the posterior mode [Chib and Jeliazkov, 2001], however we found that in this application the posterior mean provided lower variance estimation, since in practice the mode is difficult to obtain. Below we describe the sampling scheme in detail:

3.2.5.1 Marginal likelihood from the Metropolis-within-Gibbs sampler

Partition the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_1 = \boldsymbol{K}$, $\boldsymbol{\theta}_2 = (\boldsymbol{V}, \sigma)$. As noted above, the conditional posterior density $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{y})$ is available in closed form, making it natural to implement a Gibbs sampler.

However the remaining conditional $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y})$ is not available analytically and a Metropolis-Hastings step must be used to facilitate sampling from this distribution [Roberts and Rosenthal, 2006].

Denote a Metropolis-Hastings proposal as $q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\boldsymbol{\theta}_2)$ so that the acceptance probability is

$$\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\boldsymbol{\theta}_2, \boldsymbol{y}) = \min\left\{1, \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_1', \boldsymbol{\theta}_2)p(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2)}{p(\boldsymbol{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \frac{q(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y})}{q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\boldsymbol{\theta}_2, \boldsymbol{y})}\right\}. \tag{3.7}$$

In practice the proposal density is taken to be $\mathcal{N}_T(\boldsymbol{\theta}_1, \lambda \boldsymbol{I})$ where λ is chosen to deliver an average acceptance probability of 23.4% [Roberts *et al.*, 1997]. The Metropolis-within-Gibbs scheme with I iterations is summarized in Algorithm 1.

Algorithm 1 A Metropolis-within-Gibbs scheme for sampling from the parameter posterior.

$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}) \leftarrow$ initial guess
for $i = 1$ **to** M **do**
  $\boldsymbol{\theta}_1' \sim q(\boldsymbol{\theta}_1^{(i-1)}, \boldsymbol{\theta}_1'|\boldsymbol{\theta}_2^{(i-1)}, \boldsymbol{y})$
  $r \sim U[0, 1]$
  if $r < \alpha(\boldsymbol{\theta}_1^{(i-1)}, \boldsymbol{\theta}_1'|\boldsymbol{\theta}_2^{(i-1)}, \boldsymbol{y})$ **then**
    $\boldsymbol{\theta}_1^{(i)} \leftarrow \boldsymbol{\theta}_1'$
  else
    $\boldsymbol{\theta}_1^{(i)} \leftarrow \boldsymbol{\theta}_1^{(i-1)}$
  end if
  $\boldsymbol{\theta}_2^{(i)} \sim p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)}, \boldsymbol{y})$
end for

Following Chib and Jeliazkov [2001] we construct the identity

$$p(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{y})p(\boldsymbol{\theta}_1|\boldsymbol{y})} \tag{3.8}$$

and seek an estimator $\hat{p}(\boldsymbol{\theta}_1|\boldsymbol{y})$ of the posterior ordinate $p(\boldsymbol{\theta}_1|\boldsymbol{y})$. Then an estimate for the marginal likelihood will be

$$\hat{p}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)}{p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \boldsymbol{y})\hat{p}(\boldsymbol{\theta}_1^*|\boldsymbol{y})}, \tag{3.9}$$

for some choice of $\boldsymbol{\theta}^*$. For minimizing estimator variance, Chib and Jeliazkov [2001] propose to take $\boldsymbol{\theta}^*$ to be the *maximum a posteriori* (MAP) estimate (or more conveniently the MAP estimator derived from the MCMC sample). In this application we found better performance to be achieved by taking $\boldsymbol{\theta}^*$ to be the arithmetic mean estimator; however in general the arithmetic mean may be unsuitable due to multi-modality or skew in the multidimensional likelihood.

An estimator is constructed based on the identity

$$p(\boldsymbol{\theta}_1^*|\boldsymbol{y}) = \frac{\mathbb{E}_{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y})}[\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2, \boldsymbol{y})q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2, \boldsymbol{y})]}{\mathbb{E}_{p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^*, \boldsymbol{y})q(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y})}[\alpha(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y})]}. \tag{3.10}$$

Estimation of the numerator is directly facilitated by the MCMC output, whereas estimation of the denominator requires an additional Monte Carlo integration, summarized in Algorithm 2. In practice the length J of this additional run is taken to be equal to the length M of the full run. For further details see Chib and Jeliazkov [2001]. The methodology, due to Chib and Jeliazkov [2001], has been demonstrated to perform well against state-of-the-art methods for estimation of marginal likelihood Friel and Wyse [2012].

3.2.5.2 Convergence diagnostics

We used standard diagnostics to assess convergence of the MCMC sampler, including both *within-run* and *between-run* diagnostics, using parallel runs from dispersed initial conditions [Cowles and Carlin, 1996]. In general the Metropolis-within-Gibbs sampler provided satisfactory convergence to stationary distributions. An example of within-run convergence for the cancer cell line data is shown in Figure 3.3.

Algorithm 2 Computation of the Chib denominator.

for $j = 1$ **to** J **do**
  $\boldsymbol{\theta}_2^{(j)} \sim p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^*, \boldsymbol{y})$
  $\boldsymbol{\theta}_1^{(j)} \sim q(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(j)}, \boldsymbol{y})$
end for
$\hat{p}(\boldsymbol{\theta}_1^*|\boldsymbol{y}) \leftarrow \frac{1}{J}\sum_{j=1}^{J} \alpha(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1^{(j)}|\boldsymbol{\theta}_2^{(j)}, \boldsymbol{y})$

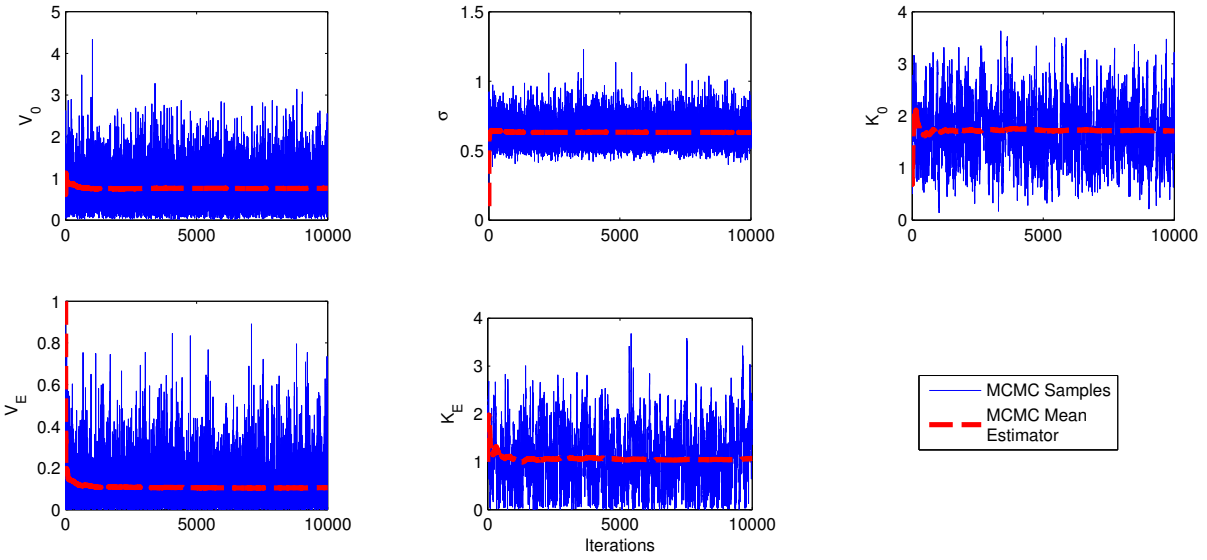


Figure 3.3: Within-run MCMC convergence diagnostics; cancer cell line data, typical trace plots.

3.2.6 Interventional Data

Interventions play a key role in experiments aimed at uncovering causal relationships [Hauser and Bühlmann, 2012; Pearl, 2009]. In interventional experiments, data are obtained under treatments that externally influence species in the chemical reaction graph G. Inhibitors of protein phosphorylation are now increasingly available; such inhibitors typically bind to the kinase domain of their target, preventing enzymatic activity (see Section 1.1.4). We consider such inhibitors in biological experiments below. Within our framework, we model inhibition by setting to zero those terms in the design matrix $\boldsymbol{D}$ corresponding to the inhibited enzyme $\mathcal{X}_E^*$ in the treated samples (this corresponds to *perfect certain* interventions in the terminology of [Eaton and Murphy, 2007]). This removes the causal influence of $\mathcal{X}_E^*$ for the inhibited samples in a way consistent with the candidate reaction graph.

3.2.7 Model Averaging

Following the recommendation of Chapter 2, evidence for a causal influence (either kinase or kinase inhibiting activity) of protein i on protein j is summarized by the marginal posterior probability of a directed edge (i, j) in the biological network N. This is obtained analogously to Eqn. 2.14 by averaging over all possible local chemical reaction graphs G_j, as

$$p((i,j) \in N|\boldsymbol{y}) = \frac{\sum_{G_j : i \in G_j} p(\boldsymbol{y}|G_j)p(G_j)}{\sum_{G_j} p(\boldsymbol{y}|G_j)p(G_j)}. \tag{3.11}$$

We refer to posterior probabilities computed in this way as *Chemical Model Averaging* (CheMA). Following work in structural inference for graphical models [Ellis and Wong, 2008; Friedman *et al.*, 2000] we bound graph in-degree; in particular, we bound $|\mathcal{E}_i| \leqslant d^1_{\max}$ and $|\mathcal{I}_{i,E}| \leqslant d^2_{\max}$. This allows explicit computation of Eqn. 3.11. In the same vein, model averaging is used to compute posterior mean predictive values (see Section B.5).

In all experiments we used a weakly-informative network prior $p(G)$. In high dimensions, Bayesian

variable selection requires multiplicity correction in order to avoid degeneracy [Scott and Berger, 2010]. Such correction is required to control the false discovery rate and is distinct from the penalty on model complexity provided by the marginal likelihood. In this Chapter, multiplicity correction is achieved with a prior probability distribution over reaction graphs G. In this context a uniform prior over reaction graphs G is inappropriate since as the number P of proteins increases, the prior weight on graphs with bounded parent set sizes $|G_i| \leqslant C$ vanishes super-exponentially quickly. Biological evidence in favour of restricted in-degree suggests using a prior which is better behaved in this limit. In the context of Bayesian variable selection, Scott and Berger [2010] and others argue for the use of a prior which is uniform over the number of predictors. Taking the above as heuristics, we used a non-informative network prior which is uniform over the number of kinases, and for a given kinase, uniform over the number of kinase inhibitors:

$$p(G) = \prod_{i=1}^{P} \binom{P}{|\mathcal{E}_i|}^{-1} \prod_{E \in \mathcal{E}_i} \binom{P}{|\mathcal{I}_{i,E}|}^{-1} \tag{3.12}$$

Under Eqn. 3.12 the prior weight on graphs G with bounded parent sets decreases slowly, at a rate $1/P$. Network priors which incorporate specific biological knowledge are also available in the literature [Mukherjee and Speed, 2008].

Since inference in our approach decomposes over proteins $i \in \mathcal{P}$ and for a given protein, over local models G_i, the computations were parallelised.

3.2.8 Prior Sensitivity and Reproducibility

We established suitable values for (i) the hyper-parameters $\boldsymbol{\mu_V}$, $\boldsymbol{\mu_K}$ and ν, (ii) maximum in-degree constraints $d^1_{\max}, d^2_{\max}$, and (iii) the number of Monte-Carlo iterations required for convergence. The suitability of (iv) the unit information g-prior, and (v) the Euler derivative approximation, are beyond the scope of this thesis. Below we describe how these values were elicited:

To investigate hyper-parameter sensitivity, we considered a fixed simulation regime (specifically we use the *in silico* model of MAPK signalling described in the following Section, with data-generating parameters $n = 100$ and $\sigma = 0.1$). For this regime we varied each of the three hyper-parameters one at a time with the other two held at the values $\boldsymbol{\mu_V} = \boldsymbol{\mu_K} = \mathbf{1}$, $\nu = 0.5$. We are not directly concerned with identification of dynamical parameters, rather we investigated whether network inference performance (quantified by AUPR and AUROC) was highly dependent on the precise values used for these hyper-parameters. Results are shown in Fig. 3.4. Both performance measures appear stable to changes in the hyper-parameters.

In addition to the hyper-parameters considered above, we also considered the influence of the maximum in-degree constraints c_1, c_2. Due to computational considerations, we did not carry out exhaustive exploration of hyper-parameter values on full networks. Instead, we constructed a smaller toy model, and explored sensitivity more fully using that model. The following model was used

$$\boldsymbol{X} \sim \mathcal{N}_T(\mathbf{1}_{10\times 1}, \boldsymbol{I}_{10\times 10}) \tag{3.13}$$

$$Z_1|\boldsymbol{X} \sim \mathcal{N}\left(f_{G,1}(\boldsymbol{X}, \boldsymbol{\theta}_1), \sigma^2 \boldsymbol{I}\right) \tag{3.14}$$

where we took

$$f_{G,1}(\boldsymbol{X}, \boldsymbol{\theta}_1) = -\frac{V_0 X_1^*}{X_1^* + K_0} + \frac{V_2 X_2^* X_1}{X_1 + K_1} + \frac{V_3 X_3^* X_1}{X_1 + K_3(1 + X_4^*/K_4)} \tag{3.15}$$

corresponding to two kinases X_2^* and X_3^*, the second of which is inhibited by X_4^*. All parameter values $\boldsymbol{\theta}_1$ were taken to be unity, in line with the observability hypothesis (see Section 3.2.4). For all experiments using the toy model we used $M = 10,000$ MCMC iterations (this was sufficient for convergence of posterior edge probabilities).

We first considered $\boldsymbol{\mu_V}$ and $\boldsymbol{\mu_K}$, along with the variance ν^2 of Michaelis-Menten parameters $\boldsymbol{K}$. Fixing $c_1 = 2$, $c_2 = 0$ we computed posterior edge probabilities (PEPs) whilst varying these hyper-parameters (Figs. 3.5(a-c)). In general we found that PEPs are stable, suggesting that results reported are not highly sensitive to the precise values used.

To investigate sensitivity to the in-degree constraint, we compared results obtained on the toy model using $(c_1 = 2, c_2 = 0)$ with $(c_1 = 2,\ c_2 = 1)$ and $(c_1 = 3,\ c_2 = 0)$ (with $\nu = 0.5$ in all cases). Results

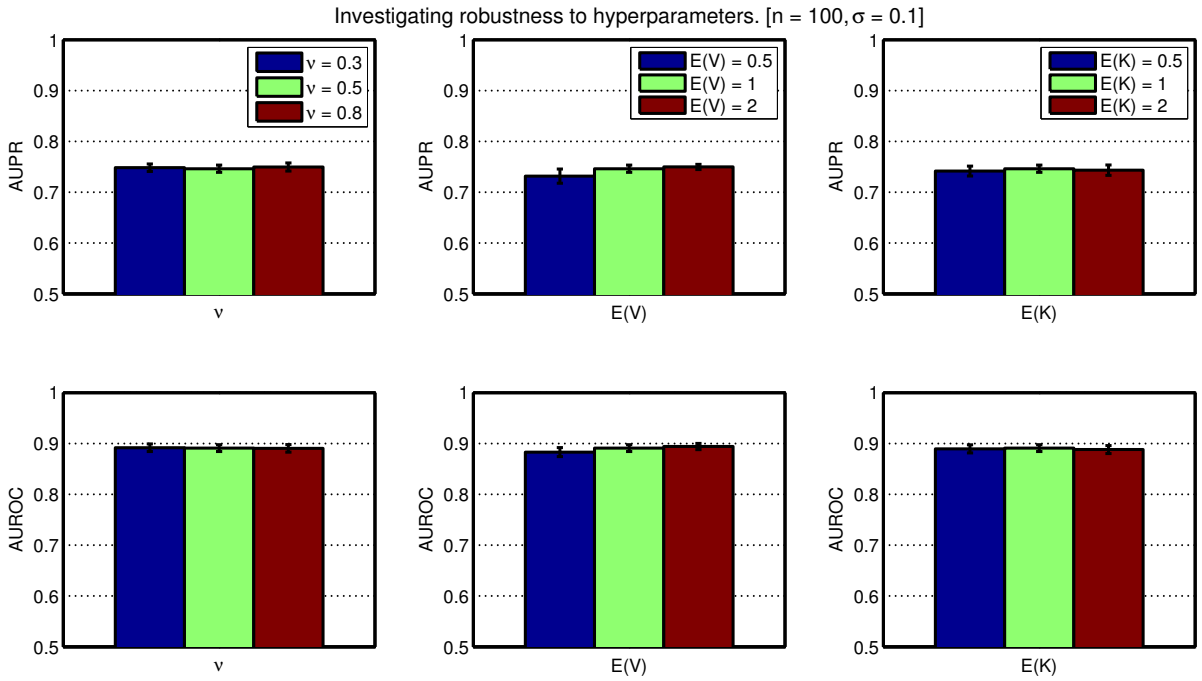


Figure 3.4: Sensitivity to hyper-parameter specification. Network inference performance (quantified by AUPR and AUROC) for various hyper-parameter values. [Here we present results over 5 independent datasets generated with $n = 100$, $\sigma = 0.1$. The 3 hyper-parameters were varied one at a time, with the remaining 2 hyper-parameters being set equal to the values used in Chapter 3, namely $\boldsymbol{\mu_V} = \boldsymbol{\mu_K} = \mathbf{1}$, $\nu = 0.5$.]

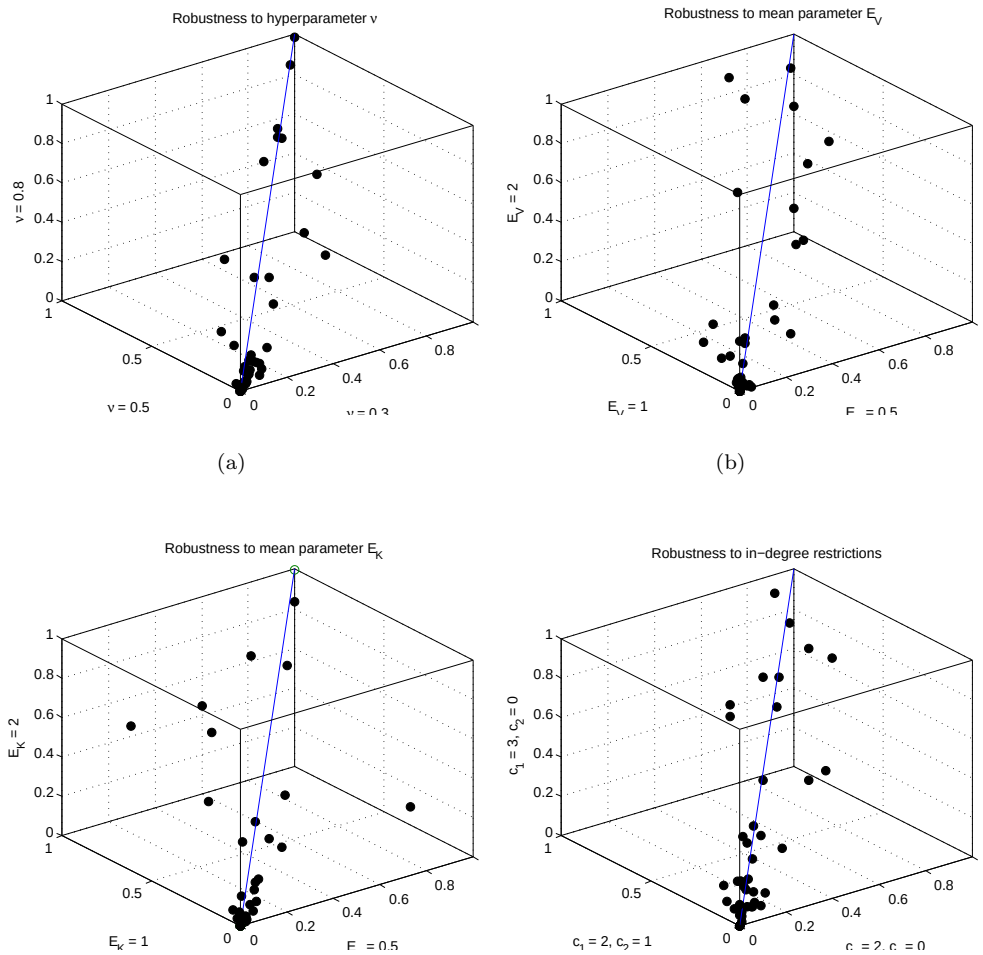
are shown in Fig. 3.5(b), comparing PEPs obtained under the three (c_1, c_2) regimes; we find good agreement between the three regimes, suggesting that the restriction is not highly influential in this setting. Results using $c_1 = 3$ suggest that models allowing 3 kinases to jointly influence a substrate are not needed in situations where the true number of kinases is $\leqslant 2$ (arguably a reasonable assumption for this thesis). Results for $c_1 = 2$, $c_2 = 1$ showed that the inhibitor X_4^* was difficult or impossible to identify from data (Fig. 3.5(b)). This suggests that time course data obtained experimentally may not contain enough information to identify such "second order" inhibitory effects, in line with previous reports that Michaelis-Menten parameters K_i (and hence inhibitory interactions) are only "weakly identifiable" from time course data Calderhead and Girolami [2011]. Note that in subsequent experiments we set $c_2 = 0$, implying that inhibitory effects are not considered.

3.3 Results

3.3.1 *In Silico* MAPK Pathway

Data were generated from a mechanistic model of the MAPK signalling pathway due to Xu *et al.* [Xu *et al.*, 2010]. The model is specified by a system of 25 ODEs of Michaelis-Menten type (described in Section B.2) and is outlined in Fig. 3.6. This archetypal protein signalling system provides an ideal test bed, since the causal graph is known and the model has been validated against experimental data [Xu *et al.*, 2010]. We considered performance under several regimes of sample size n and intrinsic noise σ (details of the simulation set-up appear in Section B.2.2).

For the estimation problem, we compared our approach to existing network inference methods which are compatible with time course data: (i) ℓ_1-penalized regression ("LASSO") , (ii) Time Series Network Identification ("TSNI" [Bansal *et al.*, 2006]; this is based on ℓ_2-penalized regression), (iii) dynamic Bayesian networks ("DBN" [Hill *et al.*, 2012a]); (iv) time-varying DBNs ("TVDBN" [Dondelinger *et al.*, 2010]) and (v) non-parametric (Gaussian process) regression with model averaging ("GP" [Äijö and Lähdesmäki, 2009]). Approaches (i-iii) are based on linear difference equations, (iv) relaxes the linear



(a) (b)

(c) (d)

Figure 3.5: Sensitivity to hyper-parameter specification, toy model. (a) prior variance $\mathrm{Var}(K) = \nu^2$, (b) prior mean $\boldsymbol{\mu_V}$ of $\boldsymbol{V}$, (c) prior mean $\boldsymbol{\mu_K}$ of $\boldsymbol{K}$, (d) in-degree restrictions c_1 and c_2,.

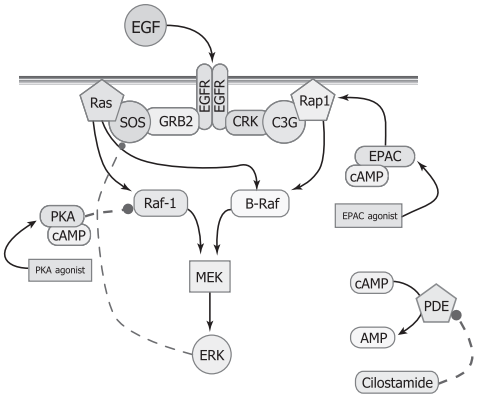
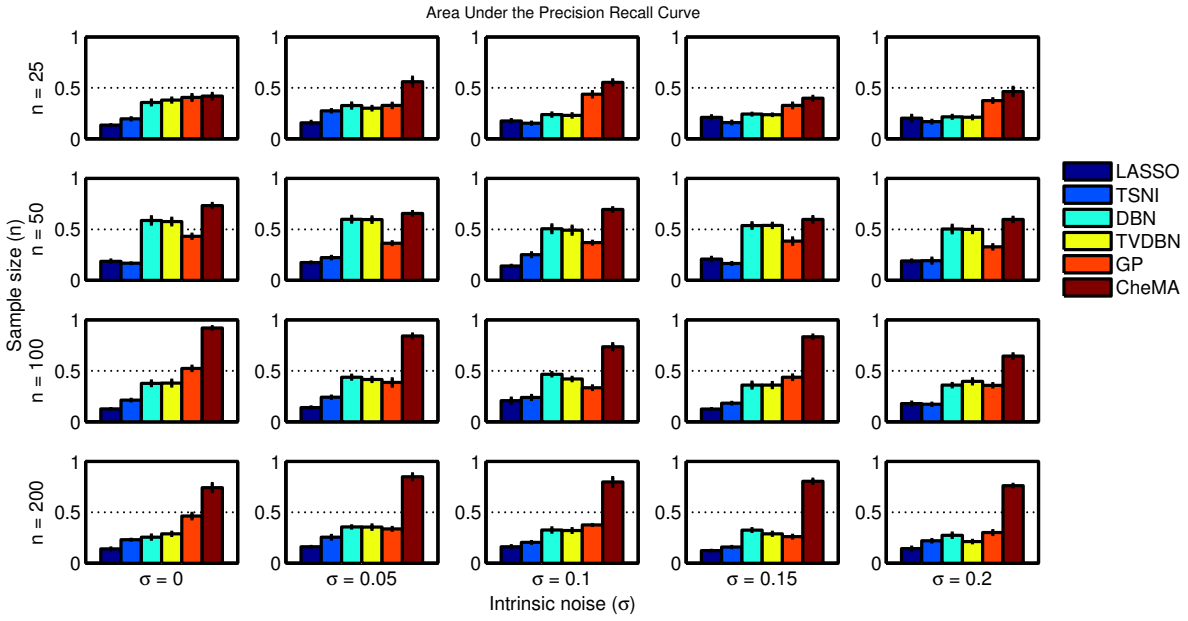
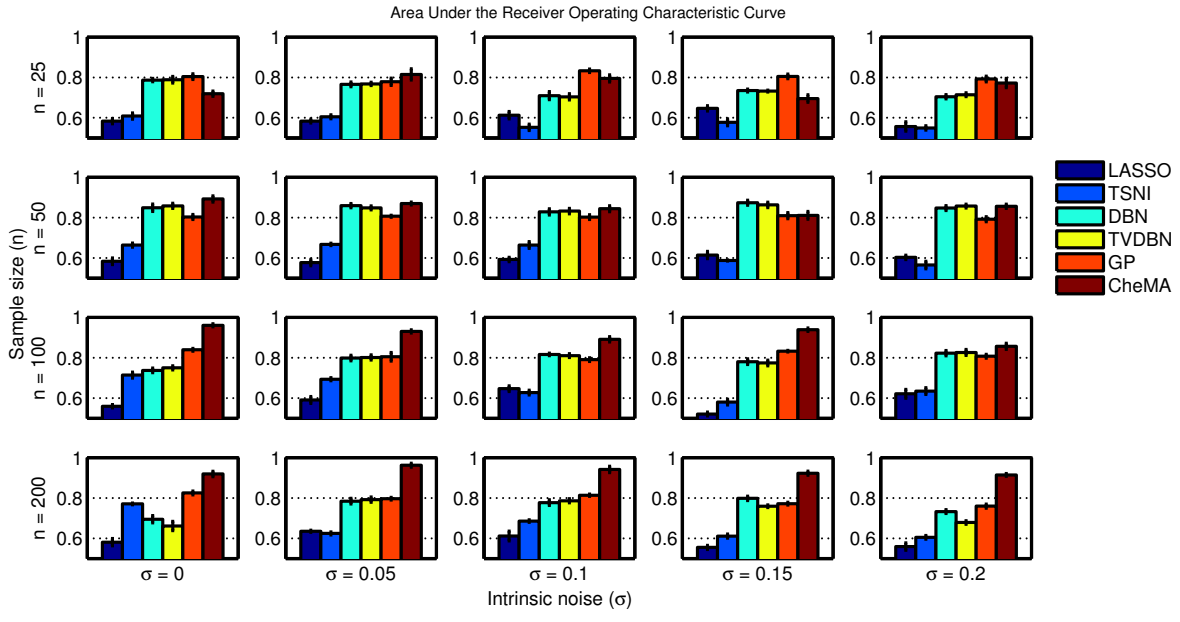


Figure 3.6: Model of the MAPK signalling pathway, reproduced from [Xu *et al.*, 2010].



(a)



(b)

Figure 3.7: Average area under the (a) PR and (b) ROC curves (AUPR,AUROC; with respect to the true causal graph). [Network inference methods: (i) LASSO, ℓ_1-penalized regression, (ii) TSNI, ℓ_2-penalized regression, (iii) DBN, dynamic Bayesian networks, (iv) TVDBN, time-varying DBNs, (v) GP, non-parametric regression, (vi) CheMA, the proposed approach. For each panel we averaged performance scores over 5 independent datasets. Sub-plots correspond to particular sample size n and noise level σ.]

assumption in a piecewise fashion, whereas (v) is a semi-parametric variable selection technique. We note that since TSNI cannot deal with multiple time courses we adapted it for use in this setting. Implementation details for all methods may be found in Section B.3.

To systematically assess estimation of network structure we computed the average area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) curves. Fig. 3.7(a) shows mean AUPR for all approaches, for 20 regimes of sample size n and noise σ. Our approach performs consistently well in all regimes, and outperforms (i-v) substantially at the larger sample sizes. It is interesting to note that the linear and piecewise linear DBNs (iii-iv) perform better at moderate sample sizes compared to higher sample sizes, possibly due to inconsistency arising from model mis-specification. AUROC results (Fig. 3.7(b)) showed a broadly similar pattern, but with CheMA offering gains only at larger sample sizes.

To investigate dynamical prediction in the setting where neither causal graph nor parameters are known, we generated data from an unseen intervention and assessed ability to predict the resulting dynamics (details of the simulation are included in Section B.5). The quality of a predicted trajectory was measured by the mean squared error (MSE) relative to the (held out) data points. To fix a length scale, both true and predicted trajectories were normalized by maximum protein expression in the test data. The network inference approaches (i-v) above have not yet been adapted for prediction in this setting. We therefore compared our chemical kinetic approach with the analogous linear formulation, which replaces Eqn. [3.2] by $f_{G,i}(\boldsymbol{X}, \boldsymbol{\theta}_i) = \beta_0 + \sum_{E \in \mathcal{E}_i} \beta_E X^*_E$ (see Section B.5.4 for details), along with the "benchmark" estimator which presumes protein concentrations do not change with time. Fig. 3.8(a) displays predictions for the effect of EPAC inhibition on the system. Here the chemical kinetic approach provides qualitatively correct prediction, whereas the linear approach rapidly diverges to infinity. This was likely due to error in the estimated eigenvalues being exaggerated geometrically at later times. We therefore focused only on short term prediction, specifically the first 25% of the time course, for which linear models may yet prove useful. Over all simulation regimes and experiments, we found that our approach produced significantly lower MSE than both the linear and benchmark models ($\text{MSE}_{\text{CheMA}} = 0.061$, $\text{MSE}_{\text{Lin.}} = 2.55$, $\text{MSE}_{\text{Bench.}} = 0.199$). Furthermore CheMA consistently produced lowest MSE at all fixed values of n and σ (Fig. 3.9; $p < 0.001$ binomial test).

3.3.2 *In Vitro* Signalling

Next, we considered experimental data obtained using reverse-phase protein arrays (see Section 1.2.2) from 15 human breast cancer cell lines, of which 10 were of HER2+ subtype [Neve *et al.*, 2006]. These data comprised observations for key phosphoproteins EGFR, Akt, MEK, GSK3ab, S6, 4EBP1 and their unphosphoryated counterparts. Data were acquired under pretreatment with inhibitors Lapatinib ("EGFRi"; an EGFR/HER2 inhibitor), GSK690693 ("Akti"; an Akt inhibitor) and without inhibition (DMSO) at 0.5,1,2,4 and 8 hours following serum stimulation, giving a total of $n = 15$ observations of each species in each cell line (see Section B.6.1 for full experimental protocol).

At present, inferred network topologies for the cell lines cannot be rigorously assessed since the true line-specific networks are not known. Inferred topologies partially concord with known signalling (Fig. B.3(b)), but the latter is based mainly on studies using wild type cells and may not reflect networks in genetically perturbed cancer lines. Therefore, for an unbiased test, we considered the problem of prediction of trajectories under an unseen intervention. We investigated whether the CheMA approach was able to outperform prediction based on "literature" signalling topology, focusing on the challenging regime where no prior topological information is made available to CheMA.

Training on DMSO and EGFRi (or AKTi) data, we assessed ability to predict the full dynamic response to Akt (or EGFR) inhibition. In this way, each held-out test set contained trajectories under a completely unseen intervention. Fig. 3.10 displays typical predictions for response to EGFRi. By considering all 15 cell lines, giving 30 held-out datasets, we found that in 19 out of 30 prediction problems our approach outperformed the literature predictor (Fig. 3.11). As for the simulated data, the linear model was not well-behaved for prediction and is not shown in Fig. 3.11. In the Akti test, of the 10 HER2+ cell lines 9 were better predicted by CheMA compared to literature prediction ($p = 0.01$, binomial test; $\text{MSE}_{\text{CheMA}} = 0.064$ vs $\text{MSE}_{\text{Lit.}} = 0.274$). Conversely 4 out of 5 HER2- lines were better predicted by literature ($\text{MSE}_{\text{Lit.}} = 0.145$ vs $\text{MSE}_{\text{CheMA}} = 0.240$), suggesting that signalling network topology in HER2+ lines may differ to the literature topology. This is a non-trivial finding, since *a priori* it is far from clear whether the training data, which involved only $P = 6$ species and $n = 10$ data points, contain sufficient information to predict the effect of an unseen intervention, even approximately.

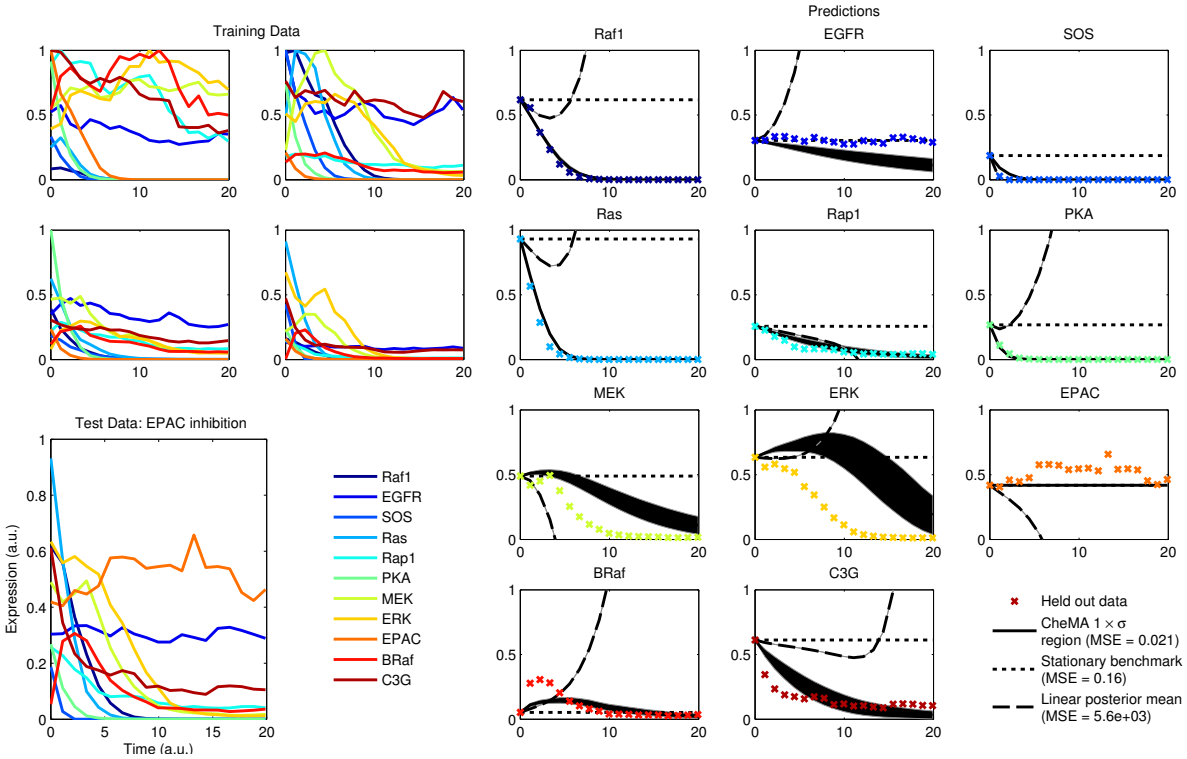


Figure 3.8: Predicting dynamical response to a novel intervention: Predicting the effect of EPAC inhibition under the data generating model of [Xu *et al.*, 2010]. [CheMA (solid) regions correspond to standard deviation of the posterior predictive distribution. Linear (dashed) replaces the non-linear chemical kinetic models with simple linear models. Benchmark (dotted) simply uses the initial data point as an estimate for all later data points. The true, test data are displayed as crosses. Here $n = 100$, $\sigma = 0.1$.]

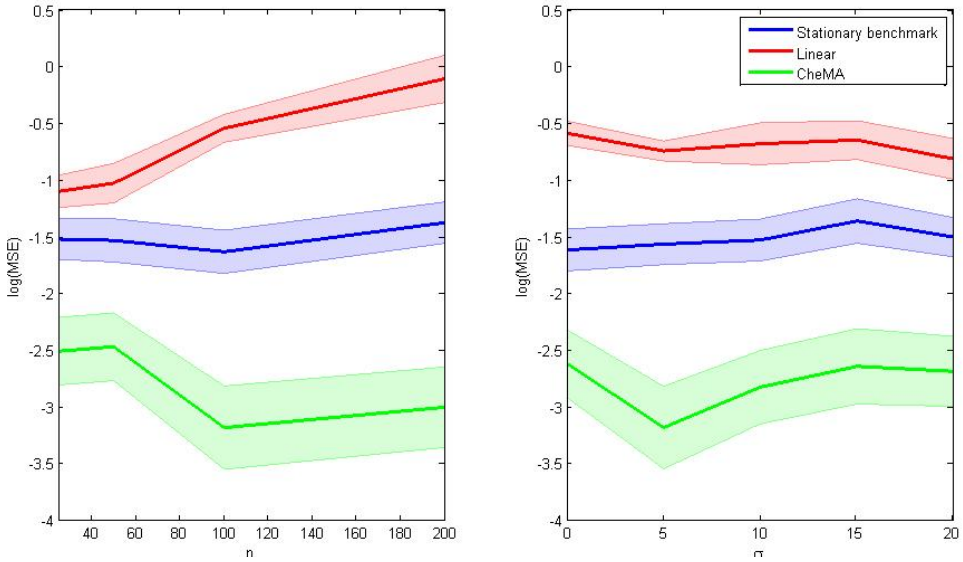


Figure 3.9: Assessment of predictive performance over varying sample size n and noise level σ; average normalized mean square error. [Shaded regions display standard error, computed over 15 independent datasets.]

However, in two of the failure cases (HCC 1569, HCC 1954; EGFRi test) CheMA produced extremely poor predictions ($\text{MSE}_{\text{CheMA}} > 1$), likely due to the extremely small training sample size.

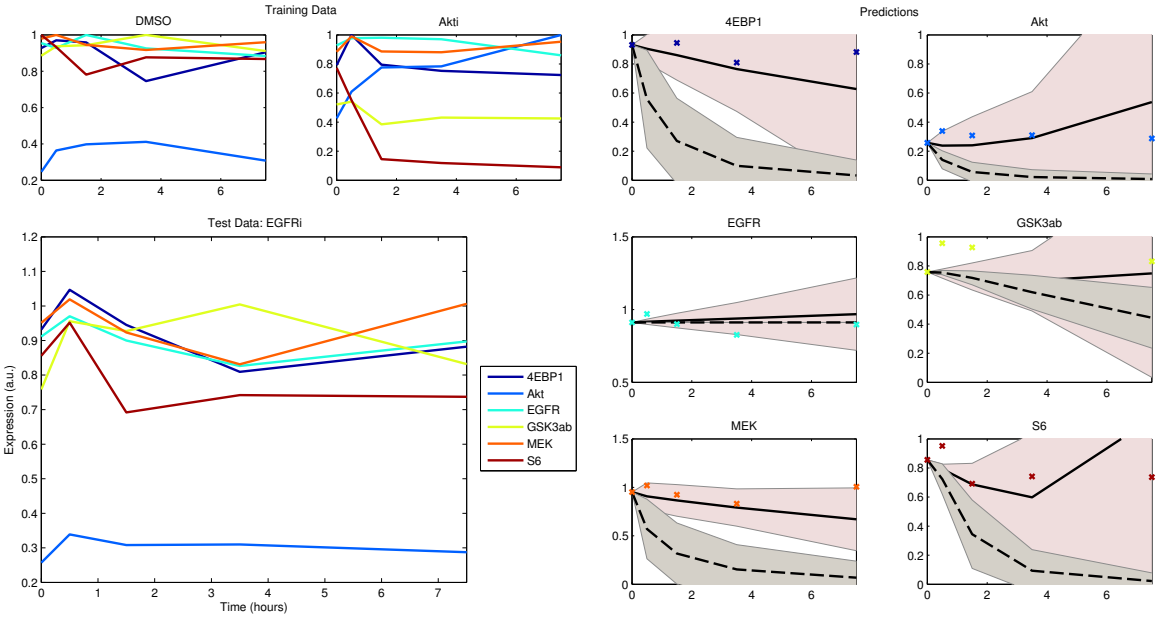


Figure 3.10: True and predicted *in vitro* trajectories for cell line HCC 70. [CheMA predictions (solid) with $1 \times \sigma$ confidence region (pink) and predictions based on the literature signalling topology (dashed) with $1 \times \sigma$ confidence region (gray).]

3.4 Discussion

We proposed an approach which integrates non-linear kinetics into network inference and dynamical prediction. Empirical results on simulated data demonstrated that the approach is capable of recovering causal network structure from time-course data and predicting the effect of unseen interventions. Whilst we restricted our investigation to protein signalling, the approach we propose is general and could be applied in other settings where automatic generation of kinetic equations is possible. In particular, extension to gene regulation is straightforward and indeed a Michaelis-Menten formulation could also be used in that setting [Hurley *et al.*, 2011].

At present dynamical predictions in systems biology require a known causal graph: a system of ODEs is usually specified conditional on such a graph and used for prediction [Nelander *et al.*, 2008]. However in many settings, including in disease biology, the causal graph cannot be assumed known. Then, the classical, known-graph approach cannot be used to predict dynamics. In contrast, our approach permits prediction of dynamical behaviour even when the reaction graph itself is unknown or uncertain. Unlike linear formulations [Maathuis *et al.*, 2010], our use of chemical kinetic models provides interpretable predictions. For example the dynamic behaviour of phosphoprotein concentrations obtained under our methodology are physically plausible (i.e. smooth, bounded and non-negative).

Network inference is naturally facilitated by interventional experiments, however adequate modelling of the effects of intervention is important to ameliorate statistical confounding [Eaton and Murphy, 2007; Hauser and Bühlmann, 2012; Pearl, 2009]. Within a chemical kinetic framework such factors may be naturally accounted for; for instance a *perfect* intervention simply corresponds to removal of the targeted species from the chemical model.

An important application in cancer is to predict the effect on signalling of a novel intervention, such as a drug treatment. In our approach this can be done without having to assume a single (potentially incorrect) reaction graph, for instance taken from literature or estimated from data [Nelander *et al.*, 2008]. Furthermore, by averaging predictions over reaction graphs, our approach should provide robustness in (typical) situations where it is unreasonable to expect to identify G precisely [Pearl, 2009].

Our approach differs from linear models, including conventional continuous (static or dynamic)

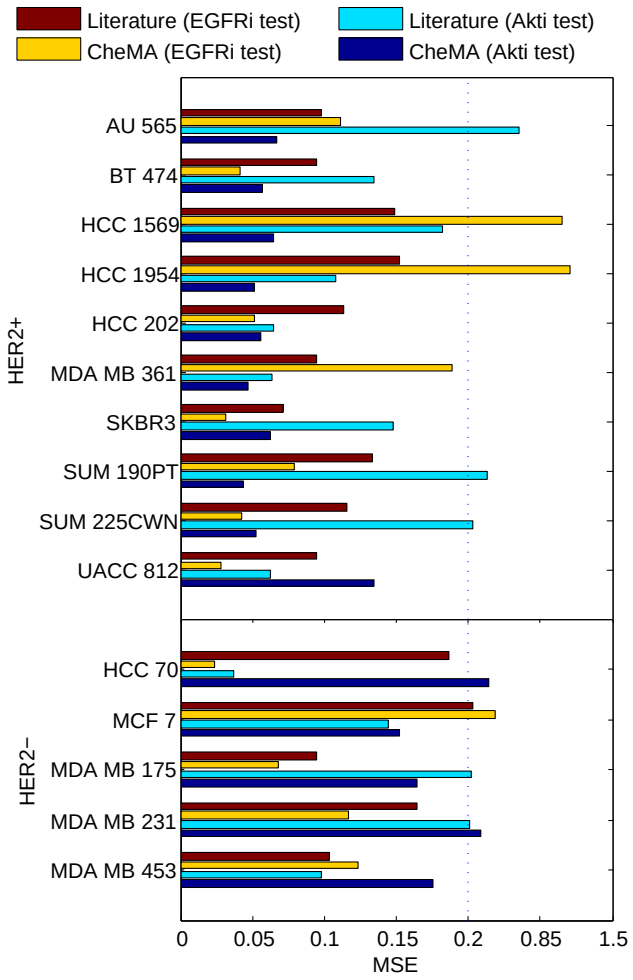


Figure 3.11: Predicting dynamical response to a novel intervention: Assessing prediction over a panel of 15 breast cancer cell lines. [Training data were time series under treatment with a single inhibitor; test data represented a second, held-out inhibitor. Normalized mean squared error was averaged over all protein species and all time points.]

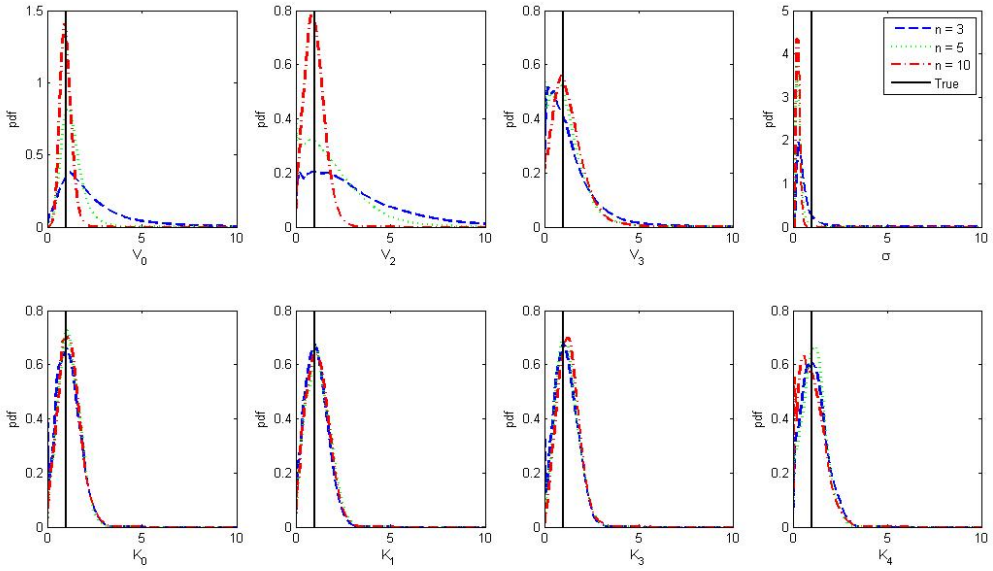


Figure 3.12: (Marginal) parameter posterior distributions for increasing sample size n. [For the Zellner g-prior, the $n = 3$ case is the closest well-defined analogue to a prior which we can plot.]

Bayesian networks, since the underlying non-linear models are not structurally symmetric. Peters *et al.* [2011] recently filled an important theoretical gap, demonstrating that within an *identifiable functional model class* (IFMOC) it is possible to consistently estimate causal relationships. This is an important step in thinking about causal inference using non-linear models and emphasises the limitations that arise from symmetry inherent in the the linear-additive-Gaussian model. However, in order to formally show that a given functional class constitutes an IFMOC, the theory at present requires strong assumptions, including noise-free observation, that do not hold in the systems we considered here. We demonstrated that basing the likelihood on a relevant non-linear dynamical system can lead to improved performance in both network inference and dynamical prediction, under practical conditions of sample size and noise. In this sense, our contribution complements the theoretical results of Peters *et al.* [2011].

At the lowest sample sizes, the chemical kinetic approach did not outperform linear methods in terms of AUROC, possibly due to the increased dimensionality of the statistical model. In order to better understand this small sample behaviour, we looked to see whether our approach was able to recover kinetic parameter values $\boldsymbol{\theta} = \{\boldsymbol{V}, \boldsymbol{K}\}$ in the case where the true graph G was known. Fig. 3.12 displays posterior probability distributions over parameters $\boldsymbol{\theta}$ for the toy model of Eqn. 3.15 (assuming known true graph G, else the parameters are not well defined) for varying sample size n. Results show that, whilst maximum reaction rates V_0, V_2, V_3 could be estimated from data, Michaelis-Menten parameters K_0, K_1, K_2, K_3, K_4 were much more difficult to infer, consistent with the *weak identifiability* reported by [Calderhead and Girolami, 2011]. Estimation for the noise parameter σ demonstrated bias toward lower values. In general, inference at the smaller sample size was much less successful. Nevertheless whilst individual parameters were not estimated precisely, the non-linear projection $\boldsymbol{f}_G(\boldsymbol{X}, \boldsymbol{\theta})$ was often identified from data.

Two ongoing challenges in Bayesian computation relevant to our work include inference of model parameters and computation of marginal likelihoods for model selection. The first has been tackled from many directions, including approximate Bayesian computation [Toni *et al.*, 2012], Gaussian processes [Calderhead *et al.*, 2009], MCMC [Wilkinson, 2006], particle filtering [Quach *et al.*, 2007], synthetic likelihoods [Wood, 2010] and tempering approaches [Campbell and Steele, 2012]. The second question is a comparatively under-developed area of statistical research, with candidate approaches including variational approximations [Rue *et al.*, 2009] and MCMC [Vyshemirsky and Girolami, 2008]. Here, we combined an MCMC scheme due to [Chib and Jeliazkov, 2001] with an Euler derivative approximation, although alternative approaches may offer advantages [Calderhead and Girolami, 2009]. The computational burden of our approach is higher than many methods, leading to greater run-times (see Section

B.3.6). By way of demonstration, inference for a 27 node network required over 12 hours computational time. In contrast, linear (or discrete) models offer improved scalability to high-dimensional systems by permitting closed form expression of model selection criteria. Thus, the approach proposed here can complement existing methodologies but is not at present applicable to high-dimensional problems with hundreds or thousands of nodes.

There are several directions in which this work can be extended. The statistical model proposed here does not explicitly distinguish between process and observation. An interesting direction for further research would be to integrate an explicit observational distribution. The automatic generation of kinetic equations clearly limits the extent to which in-depth knowledge about particular biochemical processes and dynamics may be incorporated. More generally, the simple form of kinetics used here will likely be sub-optimal in general, especially when the assumptions of the Michaelis-Menten approximation are violated [Leskovac, 2003]. We considered only the case in which the target of interventions is known. For interventions whose targets are unknown, the framework we propose could in principle be adapted to ask whether any of the observed nodes are likely targets, complementing via a non-linear model the work of [Bansal *et al.*, 2006].

3.5 Addendum: Steady-State Data

This Chapter focussed on inference using time course data, yet many excellent datasets are available which describe steady state protein expression levels. The Michaelis-Menten equations which formed the basis for CheMA exhibit unique equilibria, given by the Goldbeter-Koshland and related equations (Ex. 11). It is therefore natural to root a second version of CheMA in the Goldbeter-Koshland equations, suitable for the analysis of data obtained at steady state. Oates *et al.* [2012] described this application of CheMA to steady state data.

In the general case, the equilibrium probability distribution can be *unfaithful* to the equilibrium graph so that the *do*-calculus [Pearl, 2009] may not apply. Dash [2003] formulated a criterion, known as *equilibration-manipulation commutability* (EMC), which characterises causal faithfulness at equilibrium. Put simply, for causal reasoning based on the equilibrium graph to be valid, the *equilibrate* and the *do* operators must commute. The Goldbeter-Koshland solution of the Michaelis-Menten equations provide a unique equilibrium solution and hence it is possible to construct well defined structural equations for the equilibrium distribution.

Unlike DBNs, causal graphs defined on static data do not benefit from an identifiability result (see Ex. 13). However arguments similar to Aliferis *et al.* [2010]; Peters *et al.* [2011], described in Oates and Mukherjee [2012b] provide heuristic justification for identifiability in non-linear data-generating regimes. The approach of Oates *et al.* [2012], which relates the equilibrium solution of ODE models to structural equation models, has recently been formalised [Joris Mooij, personal correspondence]. Some of these causal aspects of the construction are discussed in Oates and Mukherjee [2012b].

The findings of Oates *et al.* [2012], on both simulated and real data, demonstrated that protein signalling network topology may be estimated more successfully under the CheMA approach than by conventional linear formulations, mirroring the conclusions of this Chapter. Further, we saw that apparently similar linear formulations can return very different recommendations for which variables ought to be included in the model, mirroring the findings of Chapter 2. Factors including model misspecification and missing variables may limit structural identifiability in general (see Chapter 2). Indeed, these studies we found that both CheMA and linear approaches performed poorly on proteomic data derived from luminal breast cancer cell lines at steady state. These findings reaffirm that the results of structural inference should be interpreted with caution and treated as hypotheses to be tested experimentally.

Chapter 4

Joint Estimation of Multiple Networks from Time Course Data

At this point we have highlighted many of the challenges associated with network inference and proposed a methodology, rooted in non-linear models of cellular chemistry, to alleviate some of these difficulties. However we have so far restricted attention to inference of single networks; in applications it is frequently the case that data are collected from multiple individuals whose networks may differ but are likely to share many features. In this Chapter we present a hierarchical Bayesian formulation for joint estimation of such networks. The formulation is general and can be applied to a number of specific graphical models, including those discussed in previous Chapters. Our methodology is accompanied by a computationally efficient, deterministic algorithm for exact inference. We show also how ancillary information, such as individual-specific genomic characteristics, can be incorporated into joint estimation. Application of the proposed method to simulated data demonstrates that joint estimation can improve ability to infer individual networks as well as differences between them. A real data study of protein signalling in breast cancer cell lines supports these conclusions. Finally, we describe approximations which are still more computationally efficient than the exact algorithm and demonstrate good empirical performance.

4.1 Introduction

In many applications, data are collected on multiple units (or individuals, we use both terms interchangeably) $j \in \mathcal{J}$ that may differ with respect to interplay between variables, such that corresponding networks N^j may be individual-specific. For example, in biology, units may correspond to different patients or cell lines and the networks themselves to gene regulatory or protein signalling networks. Interplay in such networks can depend on the genetic and epigenetic state of the individuals, such that even for a well-defined system, such as signalling downstream of a certain receptor class, or a sub-part of the transcriptional program, details may differ between even closely related samples [Csermely *et al.*, 2013; Ideker and Krogan, 2012]. For example, in yeast signalling, edges in the well-understood mitogen-activated protein kinase (MAPK) pathway can change depending on context [Zalatan *et al.*, 2012], whilst genetic networks have been shown to rewire following exposure to DNA-damaging agents [Bandyopadhyay *et al.*, 2010]. Furthermore, continuing reduction in the unit cost of biochemical assays has led to an increase in experimental designs that include panels of potentially heterogeneous individuals [Barretina *et al.*, 2012; Cao *et al.*, 2011; Maher, 2012; The Cancer Genome Atlas Network, 2012]. In such settings, given individual-specific data $\mathbf{y}^j$, there is scientific interest in individual-specific networks N^j and their similarities and differences.

Our work is motivated by questions concerning biological networks in cancer. Multiple studies have demonstrated the remarkable genomic heterogeneity of cancer [The 1000 Genomes Project Consortium, 2010; The Cancer Genome Atlas Network, 2012]. At the same time, the question of how such heterogeneity is manifested in terms of signalling or gene regulatory networks remains poorly understood. Recently, statistical approaches have been used to estimate cancer signalling networks from proteomic data [Bender *et al.*, 2010; Hill *et al.*, 2012a; Oates *et al.*, 2012]. Many open questions in cancer concern differences or similarities in such networks between different samples. This motivates studies in which data are obtained from multiple individuals (patient samples or cell lines). We discuss an example of

such an experimental design below. However, the case of multiple related individuals poses two key challenges for network inference:

- **Efficiency.** For related individuals whose networks are likely to have similarities, individual-level estimation (i.e. $\hat{N}^j = \hat{N}(\mathbf{y}^j)$) may be inefficient, since there is no sharing of information between individuals. Even in the favourable case of consistent network estimators that are well-behaved as the individual-specific sample size n_j grows large (e.g. Kalisch and Bühlmann [2007]), in practice small-to-moderate n_j's and the inherently high-dimensional nature of network inference render estimation challenging.

- **Data aggregation.** Aggregating data from multiple individuals and then performing network inference offers a way to obtain larger sample sizes. However, in settings where data from individuals are inhomogeneous (in the sense that the networks N^j may differ between individuals), inferences regarding network structure cannot in general be obtained from aggregated data (Simpson's paradox) and testing whether data aggregation is appropriate may be challenging [Pearl, 1998]. Estimating group structure using multivariate mixture models and related clustering approaches offers an alternative [Zhou *et al.*, 2009; Mukherjee and Hill, 2011; Rodríguez *et al.*, 2011; Vu *et al.*, 2012; Hill and Mukherjee, 2013], but remains challenging in the network setting.

In this Chapter we present a Bayesian approach to joint estimation of networks. Following Danaher *et al.* [2012], Penfold *et al.* [2012] and others (we discuss related work below), we focus on the case of networks N^j that are exchangeable in the sense that inference is invariant to permutation of individuals $j \in \mathcal{J}$. However, in general, the individual j's could have more complex, hierarchical relationships, for example with j's belonging to groups and sub-groups. We do not address estimation of networks with general hierarchical relationships, nor estimation of the hierarchy itself. The exchangeable case we consider corresponds to the simplest possible hierarchy in which each individual is dependent on a single latent graph (see Fig. 4.2). We note however that in settings where groups can be treated as approximately homogeneous, our approach can be trivially extended to give group-level estimates, by using j to index groups rather than individuals, with all data for group j modelled as dependent on graph N^j.

Following Werhli and Husmeier [2008] and others, we model data on all individuals $\{\boldsymbol{y}^j : j \in \mathcal{J}\}$ within a joint Bayesian framework. Regularisation of individual networks is achieved by introducing a latent network N to couple inference across all individuals. We report posterior marginal inclusion probabilities for every possible edge in each individual network N^j plus the latent network N. This provides a confidence measure for the inferred network topologies and may offer robustness in settings where posterior mass is not highly concentrated on a single model [Claassen and Heskes, 2012]. The high-level formulation we propose is general and could be applied to a wide range of graphical model formulations. That is, essentially any graphical model of interest could be embedded within our formulation to enable joint estimation. We present a detailed development for the time-course setting, focusing on directed graphical models called dynamic Bayesian networks (DBNs). These are directed acyclic graphs (DAGs) with explicit time-indices [Murphy, 2002].

The main contributions of this paper are:

- **Bayesian computation.** For the time-course setting, we put forward an efficient and deterministic algorithm. This is done by exploiting modularity of the DBN likelihood [Hill *et al.*, 2012a], analytic marginalization over continuous parameters, imposing a sparsity restriction on network topology and finally performing belief propagation on a graph whose vertices are themselves graphs (Fig. 4.2). In moderate-dimensional settings this allows exact joint estimation to be carried out in seconds to minutes (we discuss computational complexity below). To the best of our knowledge this is the first Bayesian approach for joint estimation that is sufficiently efficient to be suitable for interactive use.

- **Empirical investigation.** The availability of an efficient Bayesian algorithm enables, for the first time, a comprehensive empirical study of the statistical properties of joint estimators in the exchangeable network setting, including a wide range of simulation regimes and a study of protein signalling in a panel of breast cancer cell lines. We formulate joint estimators based on classical (non-joint) DBNs, including a recent variant suitable for interventional data [Spencer and Mukherjee, 2012]. We find that joint estimation outperforms the corresponding individual-level estimators. We also highlight a number of computationally favourable approximations to fully joint inference which perform well under a wide range of conditions.

Joint estimation of graphical models has recently been discussed in the penalised likelihood literature, with contributions including Chiquet *et al.* [2011]; Danaher *et al.* [2012]; Guo *et al.* [2011]; Hara and Washio [2012]; Mohan *et al.* [2013]; Yang *et al.* [2012]. These studies focus on the same exchangeable setting we consider here but differ from our work in that they use L_1 penalties, such as the fused graphical LASSO, to couple together inference of undirected Gaussian graphical models (GGMs). As such, the penalised likelihood methods are much more scalable to truly high dimensions.

Penalised methods derive computational efficiency from convexity of the objective function. However, for integration of diverse ancillary information the convexity requirement may become restrictive. In many applications ancillary information are available; for example, in gene regulation, the biological literature provides general information concerning gene-gene interplay, whilst patient-specific characteristics (e.g. genetic features) might also be available. In a breast cancer cell line panel we consider below, in addition to the time-course data (for each cell line) that is used to estimate networks, the mutational status of relevant genes is available for each of the cell lines. We discuss how ancillary information may be incorporated at both the "global" (all samples together) and "local" (individual) levels within our approach, with a demonstration in the cancer signalling example.

Further related work includes Werhli and Husmeier [2008], who propose a Bayesian approach to network inference based on multiple, steady-state datasets where in each dataset only a subset of the (shared) underlying network is identifiable. Dondelinger *et al.* [2012] extend the information sharing scheme from Werhli and Husmeier [2008] in the context of inference for time-varying networks. Hoff [2009] considers covariance estimation from a heterogeneous population, treating individual covariance matrices as samples from a matrix-valued probability distribution. Network priors have been discussed in the literature, including Imoto *et al.* [2003]; Mukherjee and Speed [2008]; Wei and Pan [2012]. Our work differs from these efforts by focusing on joint estimation; as we describe below, this leads to a different model structure and prior specification.

A recent paper by Penfold *et al.* [2012] considers Bayesian joint estimation for time-course data. Our work is in the same vein but differs in several respects. First, for the time-course setting, the exact algorithm we propose offers massive computational gains in comparison to the approach proposed by Penfold *et al.* [2012]. As we discuss in detail below the methodology of Penfold *et al.* [2012] is prohibitively computationally expensive for the applications we consider here. Second, the computational efficiency of our approach allows us to present a much more extensive study of joint estimation, using both simulated and real data, than has hitherto been possible. This adds to our understanding of the performance of hierarchical Bayesian formulations for joint estimation. Third, we allow for prior information regarding the network structure and ancillary information including individual-specific characteristics. Network priors and ancillary information can usefully constrain inference, not least in biological settings. For example in the cancer signalling example we consider below, much is known concerning relevant biochemistry (Fig. 4.1) and individual-specific information pertaining to e.g. mutation status and receptor expression is often available (nowadays also in the clinical setting).

The remainder of the Chapter is organized as follows. In Section 4.2 we lay out a hierarchical Bayesian formulation and in Section 4.3 we discuss computationally efficient joint inference. Empirical results are presented in Section 4.4, using both simulated (Section 4.4.2) and real (Section 4.4.3) breast cancer datasets. Finally we close with a discussion of our findings in Section 4.5.

4.2 Joint Network Inference

We carry out joint network inference using the hierarchical model shown in Fig. 4.2 that includes a prior network (N^0) as well as a latent network (N); each individual network (N^j; we use superscript notation when referring to a particular individual) is conceptually viewed as a variation upon the latter. Individual data $\boldsymbol{y}^j$ are then conditional upon individual networks. Estimates of the individual networks N^j are regularized by shrinkage towards the common latent network N which in turn may be constrained by an informative network prior. Since the latent network is itself estimated, this allows for adaptive regularization.

4.2.1 Hierarchical Model

Consider the space $\mathcal{N}$ of (directed) networks (not necessarily acyclic) on the vertex set $\mathcal{P} = \{1, \ldots, P\}$. A network $N \in \mathcal{N}$ decomposes over parent sets as $N = N_1 \times \cdots \times N_P$ where $N_p \subseteq \mathcal{P}$ are the network parents

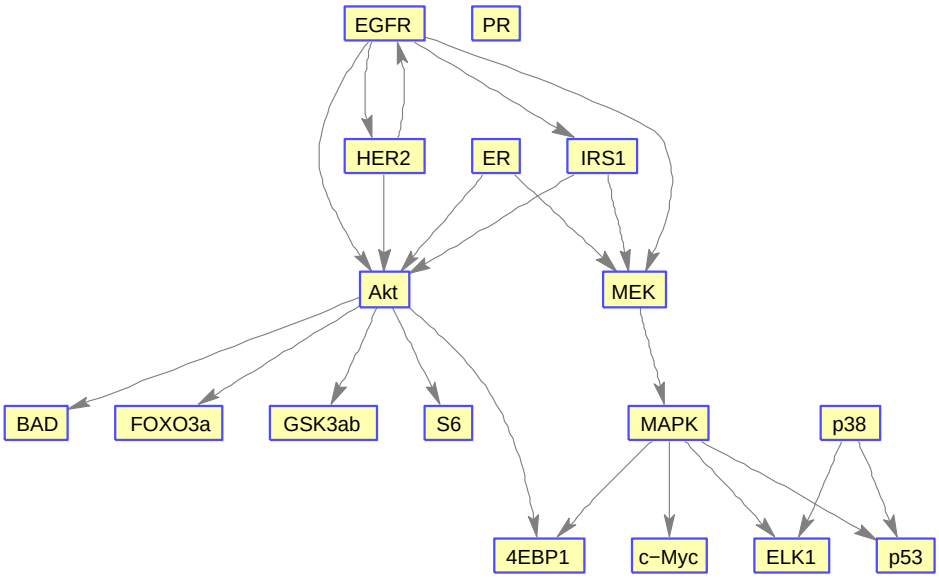


Figure 4.1: Epidermal growth factor receptor (EGFR) pathway for mammalian cells, characterized by extensive biochemistry. [Here edges represent high-level summaries of often complex molecular interactions that may involve latent chemical species.]

of $p \in \mathcal{P}$. Write $\mathcal{N}_p$ for the set of possible parent sets for variable p, such that formally $\mathcal{N} = \mathcal{N}_1 \times \cdots \times \mathcal{N}_P$. Write $\mathcal{J} = \{1, \ldots, J\}$ for the set of individuals in the population.

As shown in Fig. 4.2, each individual network N^j is conditional on a latent network N which in turn depends on a prior network N^0 (Section 4.2.2). As in any graphical model, data $\boldsymbol{y}^j$ is conditional on network N^j and parameters $\boldsymbol{\theta}^j$; A^j denotes any ancillary information available on individual j. In this Section we describe our general model and network priors, while in Section 4.3 we discuss the special case of inference for time-course data, giving full details of the likelihood for that case. The model is specified by

$$p(N|N^0, \eta) \propto \exp\left(-\eta d(N, N^0)\right) \tag{4.1}$$

$$p(N^1, \ldots, N^J|N, \boldsymbol{\lambda}, A^1, \ldots, A^J) \propto \exp\left(-\sum_{j \in \mathcal{J}} \lambda^j d^j(N^j, N; A^j)\right) \tag{4.2}$$

where the functionals $d^j, d : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$ and hyper-parameters $\eta, \lambda^1, \ldots, \lambda^J$ must be specified (Section 4.2.2). This formulation is borrowed from statistical mechanics, where d^j, d may be interpreted as energy terms, $\eta, \lambda^1, \ldots, \lambda^J$ as inverse temperature parameters and Eqns. 4.1,4.2 as Boltzmann (or Gibbs) distributions. Taken together with a suitable graphical model likelihood $p(\boldsymbol{y}^j|N^j, \boldsymbol{\theta}^j)$, we obtain the data-generating model. JNI performs inference jointly over $(N, N^1, \ldots, N^J)$, with information sharing occurring via the latent network N. The use of a latent network follows Guo *et al.* [2011]; Imoto *et al,* [2006]; Penfold *et al.* [2012]; Werhli and Husmeier [2008]. In some biological settings, it may be natural to think of the latent network as describing a "wild type" network however such an interpretation is not required. We refer to this general formulation as joint network inference (JNI).

4.2.2 Network Prior

Specifying a network prior (Eqn. 4.1) requires a penalty functional $d : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$ and a prior network $N^0 \in \mathcal{N}$, with the former capturing how close a candidate network $N \in \mathcal{N}$ is to the latter [Imoto *et*

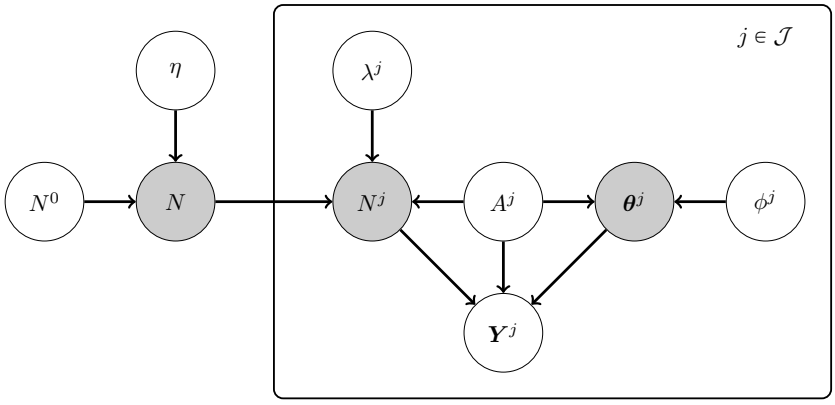


Figure 4.2: Joint network inference (JNI): A hierarchical model for analysis of multivariate data from a heterogeneous population. [Shaded nodes are unobserved. N^0 = prior network, N = latent network, N^j = individual j's network, $\boldsymbol{\theta}^j$ = parameters for individual j, $\boldsymbol{Y}^j$ = observables for individual j, A^j = ancillary information available on individual j, η, λ^j = inverse temperature hyper-parameters, ϕ^j = parameters defining a prior on $\boldsymbol{\theta}^j$.]

al., 2003; Mukherjee and Speed, 2008]. We discuss choice of N^0 below. Given N^0, a simple choice of penalty function d is the structural Hamming distance $d(N, N^0) = \text{SHD}(N, N^0) := \sum_{p \in \mathcal{P}} |N_p \Delta N_p^0|$. Here $A \Delta B$ denotes the symmetric difference of sets A and B and $|A|$ denotes cardinality of the set A. The hyper-parameter η controls the strength of the prior network N^0 (Eqn. 4.1). For brevity we follow Penfold *et al.* [2012] by restricting attention to SHD priors, however our formulation is general (see below) and compatible with other penalty functionals. For their work on joint estimation of inverse covariance matrices, Danaher *et al.* [2012]; Yang *et al.* [2012] employed the fused graphical LASSO (FGL) penalty, which may be interpreted as a real-valued extension of SHD (strictly speaking, there is no analogue of the latent network N here; FGL directly penalizes the difference between individual networks N^j, N^k). Another interesting extension due to Werhli and Husmeier [2008] distinguishes $N^0 \backslash N$ ("false prior positives") and $N \backslash N^0$ ("false prior negatives") by allocating a separate inverse temperature hyper-parameter for each case. Alternatively, one could employ a binomial prior as described in Dondelinger *et al.* [2012], which provides the same distinction, but allows for the hyper-parameters of the binomial to be integrated out.

Conditional on a latent network N, individual networks N^j are regularized in a similar way, as $d^j(N^j, N) = \text{SHD}(N^j, N)$. In their work on combining multiple data sources, Werhli and Husmeier [2008] allow the λ^j to vary over individuals (data sources) $j \in \mathcal{J}$, with λ^j reflecting the quality of dataset j. Likewise Penfold *et al.* [2012] learn the λ^j on an individual by individual basis. However, in both studies, hyper-parameter elicitation is non-trivial (see Section 4.2.4). To further limit scope, we consider only the special case where $\lambda^1 = \lambda^2 = \cdots = \lambda^J := \lambda$.

When ancillary information A^j is available regarding a specific individual network N^j, it is desirable to augment the prior specification in such a way as to condition upon A^j. In general such modification will be application specific. In Section 4.4.3.1 below we discuss the use of ancillary genetic and histological information in the context of protein signalling in breast cancer.

Although we focus on SHD priors, the inference procedures presented in this Chapter apply to the more general class of modular priors, which may be written in the form

$$d(N, N^0) = \sum_{p \in \mathcal{P}} d_p(N_p, N_p^0), \quad d^j(N^j, N; A^j) = \sum_{p \in \mathcal{P}} d_p^j(N_p^j, N_p; A^j) \tag{4.3}$$

for some functionals $d_p, d_p^j : \mathcal{N}_p \times \mathcal{N}_p \to \mathbb{R}$. Modularity here refers to a factorization over variables $p \in \mathcal{P}$, implying that only local information is available *a priori*. The SHD priors are clearly modular.

4.2.3 Two Special Cases: INI and ANI

Up to inclusion of ancillary information, prior strength is fully determined, in this simplified setting, by the parameter pair (λ, η). Taking $\eta \to \infty$ requires that the latent network N is (almost surely) identical

to the prior network N^0; in the limit this corresponds to treating network inference for each individual separately, i.e. the estimator $\hat{N}^j = \hat{N}(\mathbf{y}^j)$. We call this approach *independent network inference* (INI). Conversely, taking $\lambda \to \infty$ requires that (almost surely) individual networks N^j do not deviate from the latent network N; this corresponds to assuming individuals have identical (unknown) network structure, but allowing parameter values $\boldsymbol{\theta}^j$ to vary between individuals, possibly becoming equal to zero. We call this approach *aggregated network inference* (ANI). Taking $\lambda, \eta \to \infty$ together corresponds to using only the prior. A further, cruder, approach would be to simply combine all data in order to estimate a single network and parameter set, an approach which Werhli and Husmeier [2008] call *monolithic*. We compare these approaches empirically in Section 4.4.

4.2.4 Network Prior Elicitation

Elicitation of hyper-parameters for network priors is an important and non-trivial issue. Hyper-parameters can be set using the data, but this poses a number of challenges, as reported in Dondelinger *et al.* [2012]; Penfold *et al.* [2012]; Werhli and Husmeier [2008]. In the context of sequential hierarchical network priors, Dondelinger *et al.* [2012] observed that when there is limited data available, hyper-parameters inferred from the data may be biased towards imposing too much agreement with the prior. Penfold *et al.* [2012] used an improper hyper-prior over the individual inverse temperature parameters λ^j, reporting that for most individuals posterior marginals did not differ greatly from the prior (possibly due to uninformative data). Similarly Werhli and Husmeier [2008] assigned improper flat prior distributions over the hyper-parameters, reporting that estimation was rather difficult. Due to such weak identifiability of hyper-parameters, we chose instead to specify the hyper-parameters λ, η in a subjective manner.

For subjective elicitation of network hyper-parameters, interpretable criteria are important. We present three criteria below which, for the special case of SHD which we consider, are simple to implement and can be used for expert elicitation. These heuristics seek to relate the hyper-parameters to more directly interpretable measures of the similarity and difference which they induce between prior, latent and individual networks.

Firstly, we note the following formula for the probability of maintaining edge status (present/absent) between the latent network N and an individual network N^j:

$$h_\lambda := p(i \notin N_p^j \Delta N_p) = \frac{e^{-\lambda \times 0}}{e^{-\lambda \times 0} + e^{-\lambda \times 1}} = \frac{1}{1 + e^{-\lambda}}. \tag{4.4}$$

This probability provides an interpretable way to consider the influence of λ. For example a prior confidence of $h_\lambda \approx 0.73$ that a given edge status in N is preserved in a particular individual N^j translates into a hyper-parameter $\lambda \approx 1$ (see Fig. 4.3). An analogous equation relates η and $h_\eta := p(i \notin N_p \Delta N_p^0)$, allowing prior strength to be set in terms of the probability that an edge status in the prior network N^0 is maintained in the latent network N.

A second, related approach is to consider the expected total SHD between an individual network N^j and the latent network N:

$$\mathbb{E}(\mathrm{SHD}(N^j, N)) = P^2(1 - h_\lambda) \tag{4.5}$$

This can be interpreted as the average number of edge changes needed to obtain N^j from N. An analogous equation holds for η and h_η.

Thirdly, in certain applications, the latent network N may not have a direct scientific interpretation, in which case the criteria presented above may be unintuitive. Then, hyper-parameters could be elicited by consideration of (a) similarity between individual networks N^j, N^k, and (b) concordance of individual networks N^j with the prior network N^0. Specifically, we suggest the following two-step procedure: (a) exploit the fact that (for an uniform prior on N) we have $s_1 := p(i \notin N_p^j \Delta N_p^k) = 1 - 2h_\lambda + 2h_\lambda^2$, which facilitates selection of h_λ via the formula $h_\lambda = (1 + \sqrt{2s_1 - 1})/2$. (b) elicit h_η using the observation that $s_2 := p(i \notin N_p^j \Delta N_p^0) = 1 - h_\lambda - h_\eta + 2h_\lambda h_\eta$, so that $h_\eta = (s_2 + h_\lambda - 1)/(2h_\lambda - 1)$. This two-step procedure uniquely determines a pair $(h_\eta, h_\lambda) \in [0.5, 1)^2$ and hence unique hyper-parameters $(\eta, \lambda) \in [0, \infty)^2$. One drawback of this approach is that λ is selected under an assumption of a uniform prior on N; that is, $\eta = 0$. The quality of this procedure will therefore depend on the actual informativeness η of the prior network N^0 on N selected in step (b). This approach to hyper-parameter selection has an analogous interpretation using expected total SHD.

The above heuristics may be useful in setting hyper-parameters in practice; we illustrate the use of

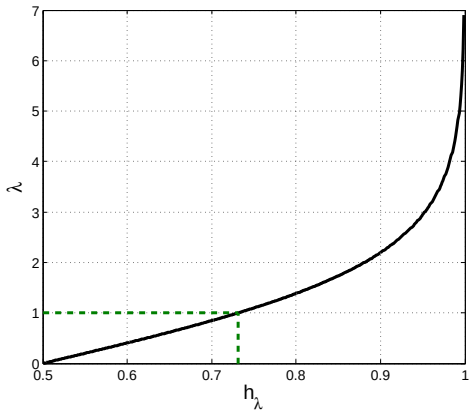


Figure 4.3: An heuristic for selecting hyper-parameter λ in the SHD prior. Here h_λ is the probability of a given edge changing status (present/absent) between the latent network N and an individual network N^j.

these procedures in empirical examples below. However, these heuristics are certainly no panacea and should be accompanied by checks of sensitivity to hyper-parameters, as we report below.

4.3 Joint Network Inference for Time-Course Data

The JNI model and network priors, as described above, are general. To apply the JNI framework in a particular context requires an appropriate likelihood at the individual level, that is, to specify the distribution $p(\boldsymbol{y}^j|N^j, \boldsymbol{\theta}^j)$ of data $\mathbf{y}^j$ conditional on a network N^j and parameters $\boldsymbol{\theta}^j$. In this Section we focus on time-course data, using DBNs to provide the likelihood.

4.3.1 Dynamic Bayesian Network Formulation

A DBN is a graphical model based on a DAG whose vertices have explicit time indices; see Murphy [2002] for details. Here, following Hill *et al.* [2012a] and others, we use stationary DBNs and permit only edges forwards in time (recall Ex. 12 in Chapter 1). Further assuming a modular network prior, structural inference for DBNs can be carried out efficiently, as described in detail in Hill *et al.* [2012a]. A novel contribution of this thesis is to extend these results to allow for efficient and exact *joint* estimation. In order to simplify notation, we define a data-dependent functional

$$\mathfrak{P}(\boldsymbol{X}) = p(\boldsymbol{X}(1)) \prod_{i=2}^{m} p(\boldsymbol{X}(i)|\boldsymbol{y}(i-1)) \tag{4.6}$$

which implicitly conditions upon observed history. Let $y_p^j(t)$ denote the observed value of variable p in individual j at time t. The above notation allows us to conveniently summarize the product

$$p(y_p^j(1)|N_p^j)p(y_p^j(2)|\boldsymbol{y}(1), N_p^j) \dots p(y_p^j(m)|\boldsymbol{y}(m-1), N_p^j). \tag{4.7}$$

as $\mathfrak{P}(\boldsymbol{y}_p^j|N_p^j)$. Thus, we have that, for DBNs, the full likelihood also satisfies modularity:

$$p(\boldsymbol{y}|N^1, \dots, N^J) \quad = \quad \prod_{j \in \mathcal{J}} \prod_{p \in \mathcal{P}} \mathfrak{P}(\boldsymbol{y}_p^j|N_p^j) \tag{4.8}$$

In other words, the parent sets N_p^j ($p \in \mathcal{P}$, $j \in \mathcal{J}$) are mutually orthogonal in the Fisher sense, so that inference for each may be performed separately.

For this Chapter, the local Bayesian score $\mathfrak{P}(\boldsymbol{y}_p^j|N_p^j)$ corresponds to the marginal likelihood for a linear autoregressive formulation described below, however we note that JNI is indifferent to how marginal likelihoods are obtained; in particular JNI is compatible with the CheMA approach of Chapter 3 (we return to this point in Chapter 5).

4.3.1.1 Linear Autoregressive Likelihood

We follow Aliferis *et al.* [2010]; Hill *et al.* [2012a]; Penfold *et al.* [2012] in formulating inference in DBNs as a regression problem. We entertain models for the response $y_p^j(t)$ as predicted by covariates $\boldsymbol{y}^j(t-1)$. In many cases multiple time series will be available. In this case the data vector $\boldsymbol{y}_p^j$ contains the concatenated time series. The DBN formulation gives rise to the following linear regression likelihood

$$\boldsymbol{y}_p^j = \boldsymbol{X}_0\boldsymbol{\alpha} + \boldsymbol{X}^j_{N_p^j}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.9}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \sigma^2\boldsymbol{I}_{n\times n})$. The matrix $\boldsymbol{X}_0 = [\mathbf{1}_{\{t=1\}}\ \mathbf{1}_{\{t>1\}}]_{n\times 2}$ contains a term for the initial time point in each experiment. The elements of $\boldsymbol{X}^j_{N_p^j}$ corresponding to initial observations $y_p^j(1)$ are simply set to zero. Parameters $\boldsymbol{\theta}_p^j = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma\}$ are specific to model N_p^j, variable p and cell line j. In the simplest case the model-specific component $\boldsymbol{X}^j_{N_p^j}$ of the design matrix consists of the raw predictors $\boldsymbol{y}^j_{N_p^j}(t-1)$ where $\boldsymbol{y}^j_A$ denotes the elements of the vector $\boldsymbol{y}^j(t-1)$ belonging to the set A, though more complex basis functions may be used. This Chapter restricts attention to this simple formulation of likelihood, however in principle our discussion applies to arbitrary likelihood functions, including those described in Chapters 2 and 3.

Experiment	1				2			
Time Point	1	2	3	4	1	2	3	4
Protein 1	0.5377	0.8622	-0.4336	2.7694	0.7254	-0.2050	1.4090	-1.2075
Protein 2	1.8339	0.3188	0.3426	-1.3499	-0.0631	-0.1241	1.4172	0.7172
Protein 3	-2.2588	-1.3077	3.5784	3.0349	0.7147	1.4897	0.6715	1.6302

Table 4.1: An example dataset for a single individual j, consisting of 3 variables, 2 time courses, each with 4 time points.

Example: We illustrate the linear autoregressive likelihood with a concrete example. Consider the datasets in Table 4.1, which is for a fixed individual j. For the particular model $N_1^j = \{2, 3\}$, i.e. the network parents of protein 1 are precisely proteins 2 and 3, the statistical model in Eqn. 4.9 translates as

$$\begin{bmatrix} 0.5377 \\ 0.8622 \\ -0.4336 \\ 2.7694 \\ 0.7254 \\ -0.2050 \\ 1.4090 \\ -1.2075 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0.3188 & -1.3077 \\ 0.3426 & 3.5784 \\ -1.3499 & 3.0349 \\ 0 & 0 \\ -0.1241 & 1.4897 \\ 1.4172 & 0.6715 \\ 0.7172 & 1.6302 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} \tag{4.10}$$

where the $e_i \sim \mathcal{N}(0, \sigma^2)$ are IID.

4.3.1.2 Modelling Interventions

Following Eaton and Murphy [2007]; Spencer and Mukherjee [2012] we model interventional data by modification to the DAG in line with a causal calculus [Pearl, 2009]. We mention briefly some of the key ideas and refer the interested reader to the references for full details. A *perfect intervention* corresponds to 100% removal of the target's activity with 100% specificity. In the context of protein phosphorylation, kinases may be intervened upon using agents such as monoclonal antibodies, small molecule inhibitors or even si-RNA [Lu *et al.*, 2011]. We make the simplifying assumptions that these interventions are perfect, and use the *perfect out fixed effects* (POFE) approach recommended by Spencer and Mukherjee [2012]. We refer the reader to Spencer and Mukherjee [2012] for an extended discussion of POFE. This changes the DAG structure to model the intervention and also estimates a fixed effect parameter to model the change under intervention in the log-transformed data.

Example: Assume that the protein data in Table 4.1 are already log-transformed. Then under the POFE approach, the information that experiment 2 was carried out in the presence of an inhibitor of

the activity of protein 3 would be incorporated into the statistical model of Eqn. 4.10 as follows:

$$\begin{bmatrix} 0.5377 \\ 0.8622 \\ -0.4336 \\ 2.7694 \\ 0.7254 \\ -0.2050 \\ 1.4090 \\ -1.2075 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0.3188 & -1.3077 & 0 \\ 0.3426 & 3.5784 & 0 \\ -1.3499 & 3.0349 & 0 \\ 0 & 0 & 1 \\ -0.1241 & 0 & 1 \\ 1.4172 & 0 & 1 \\ 0.7172 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} \tag{4.11}$$

4.3.1.3 Exact Marginal Likelihood

We employed a Jeffreys prior $p(\boldsymbol{\alpha}, \sigma | N_p^j, \phi^j) \propto 1/\sigma$ for $\sigma > 0$ over the common parameters. Prior to inference, the non-interventional components of the design matrix were orthogonalized using the transformation $(\boldsymbol{X}^j_{N_p^j})_{ik} \mapsto \sum_l (\boldsymbol{I}_n - \boldsymbol{P}_0)_{il} (\boldsymbol{X}^j_{N_p^j})_{lk}$, where $\boldsymbol{P}_0 = \boldsymbol{X}_0 (\boldsymbol{X}_0^T \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^T$ [Deltell *et al.*, 2012]. We then assumed a g-prior for regression coefficients [Zellner, 1986], given by $\boldsymbol{\beta} | \boldsymbol{\alpha}, \sigma, N_p^j, \phi^j \sim \mathcal{N}(\boldsymbol{0}_{b \times 1}, \phi^j \sigma^2 (\boldsymbol{X}^T_{N_p^j} \boldsymbol{X}_{N_p^j})^{-1})$ where $b = \dim(\boldsymbol{\beta})$. Using these priors for the DBNs with intervention as outlined above, the marginal likelihood can be obtained in closed-form:

$$\mathfrak{P}(\boldsymbol{y}_p^j | N_p^j, \phi^j) \propto \frac{1}{(\phi^j + 1)^{b/2}} \left(\boldsymbol{y}_p^{jT} \left(\boldsymbol{I}_{n \times n} - \boldsymbol{P}_0 - \frac{\phi^j}{\phi^j + 1} \boldsymbol{P}_{N_p^j} \right) \boldsymbol{y}_p^j \right)^{-\frac{n-a}{2}} \tag{4.12}$$

where $\boldsymbol{P}_{N_p^j} = \boldsymbol{X}_{N_p^j} (\boldsymbol{X}^T_{N_p^j} \boldsymbol{X}_{N_p^j})^{-1} \boldsymbol{X}^T_{N_p^j}$, $a = \dim(\boldsymbol{\alpha})$ and $b = \dim(\boldsymbol{\beta})$. Empirical investigations have previously demonstrated good results for network inference based on the above marginal likelihood [Hill *et al.*, 2012a; Spencer and Mukherjee, 2012].

4.3.1.4 Elicitation of the Zellner Parameter

We employed a marginal likelihood $\mathfrak{P}(\boldsymbol{y}_p^j | N_p^j, \phi^j)$ based on Bayesian linear regression using Zellner's g-prior [Zellner, 1986] as described above. The hyper-parameter $g = \phi^j$, which is related to the weight of the parameter prior $p(\boldsymbol{\beta} | \boldsymbol{\alpha}, \sigma)$ relative to the data $\boldsymbol{y}_p^j$, was selected using the conditional empirical Bayes procedure of George and Foster [2000], corresponding to

$$\hat{g}(N_p^j) = \arg\max_g \mathfrak{P}(\boldsymbol{y}_p^j | N_p^j, g). \tag{4.13}$$

In order to retain computational efficiency, we evaluated the argument over a finite set of eight candidate g values corresponding to prior weight of 0,10,20,30,40,50% and (100/n)% (the unit information prior). Alternative strategies for setting of the g hyper-parameter are discussed in Deltell *et al.* [2012]; Liang *et al.* [2008].

4.3.2 Computationally Efficient Joint Estimation

Previous studies have used MCMC to generate samples from the posterior distribution over networks [Penfold *et al.*, 2012; Werhli and Husmeier, 2008]. However, ensuring mixing has proven to be extremely challenging for joint estimation, with both studies reporting extremely slow convergence. Advances in MCMC and parallel computing may in the future ameliorate these issues [Lee *et al.*, 2010], but at present it remains the case that fast, interactive joint estimation is currently challenging or prohibitively demanding using MCMC. We therefore propose an exact approach, using an in-degree restriction coupled with prior modularity and a sum-product-type ("propagation") algorithm, to facilitate efficient estimation. For example, the DREAM4 problem ($P = 10$ variables, $J = 5$ individuals) considered by Penfold *et al.* [2012] was reported to require "several hours per node" for MCMC convergence; our approach solves the entire problem in ≈ 3 seconds. Our approach therefore complements MCMC-based inference, allowing fast, interactive investigation in moderate-dimensional settings.

Specifically, we use exact model averaging to marginalize over networks and report posterior marginal inclusion probabilities. We begin by computing and caching the local scores $\mathfrak{P}(\boldsymbol{y}_p^j | N_p^j)$ for all parent sets $N_p^j \in \mathcal{N}_p$, all variables $p \in \mathcal{P}$ and all individuals $j \in \mathcal{J}$; these could be obtained using essentially any

suitable likelihood. The posterior marginal probability for an edge (i, p) belonging to the latent network G is computed as Eqns. 4.14-4.19, where Eqn. 4.19 uses Lemma 1 in Appendix C.1 to interchange operators. This final step has important consequences for algorithmic complexity (see Section 4.3.3) and is a main advantage of these "propagation" algorithms [Pearl, 1982]. Note that, whilst this derivation can made without the explicit marginalization of Eqn. 4.15, the approach is quite general and may be used analogously to facilitate estimation of individual networks N^j (Eqns. 4.20-4.25) where again Lemma 1 justifies the exchange of operators.

$$
\begin{aligned}
p(i \in N_p|\boldsymbol{y}, N^0) &= \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} p(N_p|\boldsymbol{y}_p, N_p^0) && (4.14)\\
&= \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} \sum_{N_p^j \in \mathcal{N}_p : j \in \mathcal{J}} p(N_p^1, \ldots, N_p^J, N_p|\boldsymbol{y}_p, N_p^0) && (4.15)\\
&\propto \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} \sum_{N_p^j \in \mathcal{N}_p : j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|N_p^1, \ldots, N_p^J, N_p, N_p^0) p(N_p^1, \ldots, N_p^J, N_p|N_p^0) && (4.16)\\
&= \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} \sum_{N_p^j \in \mathcal{N}_p : j \in \mathcal{J}} p(N_p|N_p^0) \prod_{j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|N_p^j) p(N_p^j|N_p) && (4.17)\\
&= \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} p(N_p|N_p^0) \sum_{N_p^j \in \mathcal{N}_p : j \in \mathcal{J}} \prod_{j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|N_p^j) p(N_p^j|N_p) && (4.18)\\
&= \sum_{N_p \in \mathcal{N}_p} \mathbf{1}_{i \in N_p} p(N_p|N_p^0) \prod_{j \in \mathcal{J}} \sum_{N_p^j \in \mathcal{N}_p} \mathfrak{P}(\boldsymbol{y}_p|N_p^j) p(N_p^j|N_p) && (4.19)
\end{aligned}
$$

$$
\begin{aligned}
p(i \in N_p^j|\boldsymbol{y}, N^0) &= \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} p(N_p^j|\boldsymbol{y}_p, N_p^0) && (4.20)\\
&= \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} \sum_{N_p \in \mathcal{N}_p} \sum_{N_p^k \in \mathcal{N}_p : k \in \mathcal{J} \setminus \{j\}} p(N_p^1, \ldots, N_p^J, N_p|\boldsymbol{y}_p, N_p^0) && (4.21)\\
&\propto \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} \sum_{N_p \in \mathcal{N}_p} \sum_{N_p^k \in \mathcal{N}_p : k \in \mathcal{J} \setminus \{j\}} p(\boldsymbol{y}_p|N_p^1, \ldots, N_p^J, N_p, N_p^0) p(N_p^1, \ldots, N_p^J, N_p|N_p^0) && (4.22)\\
&= \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} \sum_{N_p \in \mathcal{N}_p} \sum_{N_p^k \in \mathcal{N}_p : k \in \mathcal{J} \setminus \{j\}} p(N_p|N_p^0) \prod_{l \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p^l|N_p^l) p(N_p^l|N_p) && (4.23)\\
&= \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} \sum_{N_p \in \mathcal{N}_p} p(N_p|N_p^0) \mathfrak{P}(\boldsymbol{y}_p^j|N_p^j) p(N_p^j|N_p) \sum_{N_p^k \in \mathcal{N}_p : k \in \mathcal{J} \setminus \{j\}} \prod_{l \in \mathcal{J} \setminus \{j\}} \mathfrak{P}(\boldsymbol{y}_p^l|N_p^l) p(N_p^l|N_p) && (4.24)\\
&= \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} \sum_{N_p \in \mathcal{N}_p} p(N_p|N_p^0) \mathfrak{P}(\boldsymbol{y}_p^j|N_p^j) p(N_p^j|N_p) \prod_{k \in \mathcal{J} \setminus \{j\}} \sum_{N_p^k \in \mathcal{N}_p} \mathfrak{P}(\boldsymbol{y}_p^k|N_p^k) p(N_p^k|N_p) && (4.25)
\end{aligned}
$$

4.3.3 Computational Complexity

Following Hill *et al.* [2012a] we reduced the space of parent sets $\mathcal{N}_p$ using an in-degree restriction of the form $|N_p^j| \leqslant d_{\max}$ for all $N_p^j \in \mathcal{N}_p$, $p \in \mathcal{P}$, $j \in \mathcal{J}$. Thus the cardinality of the space of parent sets $M = |\mathcal{N}_p| = \mathcal{O}(P^{d_{\max}})$ is polynomial in P, where it was previously exponential. This reduces summation over an exponential number of terms to a more manageable sum over polynomially many terms. Moreover, in the protein signalling example to follow, bounded in-degree is a reasonable biological assumption. Sensitivity to choice of $d_{\max}$ is discussed in Section 4.4.1.

Caching of selected probabilities is used to avoid redundant recalculation. Below we provide pseudo-code for computation of posterior marginal inclusion probabilities for edges in individual networks N^j:

for all $p \in \mathcal{P}$ **do**

Phase I:

Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall N_p \in \mathcal{N}_p]$

$\mathfrak{P}(\boldsymbol{y}_p^j|N_p) = \sum_{N_p^j \in \mathcal{N}_p} \mathfrak{P}(\boldsymbol{y}_p^j|N_p^j) p(N_p^j|N_p)$ $[\mathcal{O}(M)]$

Phase II:

Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall N_p^j \in \mathcal{N}_p]$

$p(N_p^j|\boldsymbol{y}_p, N_p^0) \propto \sum_{N_p \in \mathcal{N}_p} p(N_p|N_p^0) \mathfrak{P}(\boldsymbol{y}_p^j|N_p^j) p(N_p^j|N_p) \prod_{k \in \mathcal{J}\setminus\{j\}} \mathfrak{P}(\boldsymbol{y}_p^j|N_p)$ $[\mathcal{O}(MJ)]$

Phase III:

Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall i \in \mathcal{P}]$

$p(i \in N_p^j|\boldsymbol{y}, N^0) = \sum_{N_p^j \in \mathcal{N}_p} \mathbf{1}_{i \in N_p^j} p(N_p^j|\boldsymbol{y}, N_p^0)$ $[\mathcal{O}(M)]$

end for

Computational complexity of each operation is shown in parentheses. Pseudo-code for inference of the latent network N proceeds analogously.

The above pseudo-code consists of three phases of computation. Storage costs are dominated by Phases I and II, which each requiring the caching of $\mathcal{O}(PJM)$ real numbers. (Computational complexity of calculating marginal likelihoods $\mathfrak{P}(\boldsymbol{y}_p^j|G_p^j)$ will scale with sample size n; scaling exponents shown here assume $\mathcal{O}(n) = \mathcal{O}(1)$.) Phase II dominates computational effort, with total (serial) algorithmic complexity $\mathcal{O}(PJ^2M^2)$. However, within-phase computation is *embarrassingly parallel* in the sense that all calculations are independent (indicated by square parentheses notation in the pseudo-code).

4.4 Results

We tested our joint estimation procedure on both simulated and real proteomic time-course data. We compare our approach to the special cases of (i) inferring each network separately (INI); (ii) allowing parameters (but not networks) to change between individuals (ANI); (iii) the naive approach of aggregating all data (monolithic) and (iv) simple temporal correlations (absolute Pearson coefficient). For a fair comparison, all methods, with the exception of (iv), were implemented so as to take account of the interventional nature of the data. We note that it is not possible to directly compare our results with Danaher *et al.* [2012]; Guo *et al.* [2011]; Yang *et al.* [2012] since these methods do not apply to time-course data. The method of Penfold *et al.* [2012] applies to time-course data, but the computational demands of the approach precluded application in this setting. Specifically, in the simulated data example we report below, over 3000 rounds of inference were performed in total, on problems larger than DREAM4 ($P = 10$, $J = 5$). Using the approach of Penfold *et al.* [2012], these experiments would have required more than 10 years' computational time; in contrast our approach required less than 24 hours serial computation on a standard laptop.

4.4.1 Performance Metrics

The proposed methodology addresses three questions, some or all of which may be of scientific interest depending on application; (i) estimation of the latent network N, (ii) estimation of individual networks $N^1, \ldots, N^J$, and (iii) estimation of differences between individual networks. We quantify performance

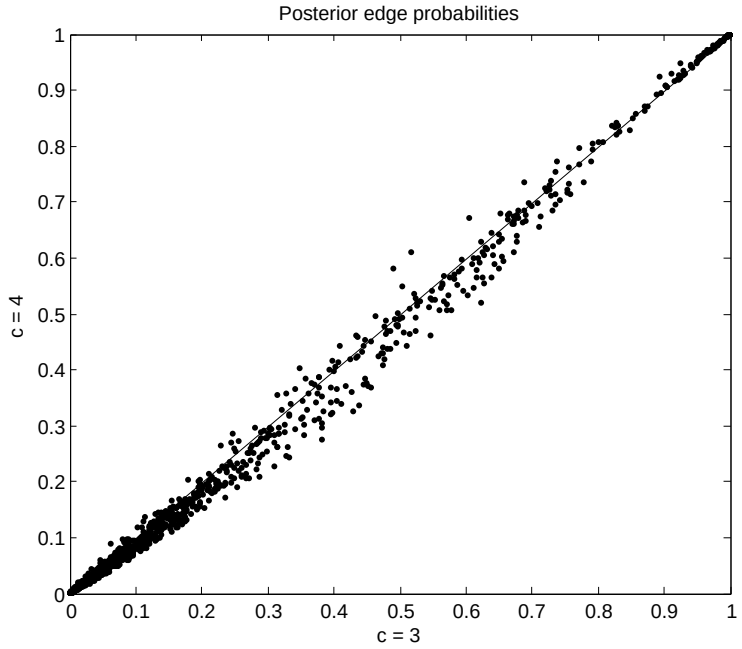


Figure 4.4: Insensitivity to the in-degree restriction. [Here we plot posterior edge probabilities obtained using in-degree restrictions $d_{\max} = 3$ and $d_{\max} = 4$.]

for tasks (i) and (ii) using the area under the receiver operating characteristic (ROC) curve (AUR). This metric, equivalent to the probability that a randomly chosen true edge is preferred by the inference scheme to a randomly chosen false edge, summarizes, across a range of thresholds, the ability to select edges in the data-generating network. AUR may be computed relative to the true latent network N, or relative to the true individual networks N^j, quantifying performance on tasks (i) and (ii) respectively. Both sets of results are presented below, in the latter case averaging AUR over all individual networks. For (iii), in order to assess ability to estimate individual heterogeneity, we computed AUR scores based on the statistics $F^j_{ip} = |p(i \in N^j_p|\boldsymbol{y}, N^0) - p(i \in N_p|\boldsymbol{y}, N^0)|$ which should be close to one if $i \in N^j_p \Delta N_p$, otherwise F^j_{ip} should be close to zero.

It is easy to show that inference for the latent network, under only the prior, attains mean AUR equal to h_η. Similarly, prior inference for the individual networks attains mean AUR equal to $1-h_\eta-h_\lambda+2h_\eta h_\lambda$. This provides a baseline for the proposed methodology at tasks (i) and (ii) and allows performance to be decomposed into AUR due to prior knowledge and AUR contributed through inference. Using a systematic variation of data-generating parameters, we defined 15 distinct data generating regimes. For all 15 regimes we considered 50 independent datasets; standard errors accompany average AUR scores. Results presented below use a computationally favourable in-degree restriction $d_{\max} = 3$. Note that when the maximum in-degree of any of the true networks exceeds the computational restriction $d_{\max}$, estimator consistency will not be guaranteed. In order to check robustness to $d_{\max}$, a subset of experiments were repeated using $d_{\max} = 4$, with close agreement observed (see Fig. 4.4).

4.4.2 Simulation Study

4.4.2.1 Data Generation

A latent network N on P vertices was drawn from the Erdös distribution with edge density ρ/P. (This Chapter restricts attention to Erdös random networks, but numerous other network models could be used; in particular there is evidence that certain bio-molecular networks are well described by a scale free network model.) In order to simulate heterogeneity, the individual networks N^j were obtained from N by maintaining the status (present/absent) of each edge independently with probability h_λ. A parameter β^j_{ip} for each parent $i \in N^j_p$ was independently drawn from the mixture normal distribution $0.5\mathcal{N}(-1, 0.1^2) + 0.5\mathcal{N}(1, 0.1^2)$ (the mixture distribution ensures that parameters are not vanishingly small, so that the structural inference problem is well-defined). Collecting together parameters produces

matrices $\boldsymbol{\beta}^j$, corresponding to networks N^j via $i \in N_p^j$ if and only if $\beta_{ip}^j \neq 0$. We also generate, for each individual j, intercept parameters $\boldsymbol{\alpha}^j \sim \mathcal{N}(\mathbf{0}_P, \boldsymbol{I}_{P\times P})$ representing baseline expression levels. Initial conditions were sampled as $\boldsymbol{y}^j(1) \sim \mathcal{N}(\mathbf{0}_P, \boldsymbol{I}_{P\times P})$. Data were then generated from the autoregressive model $\boldsymbol{y}^j(t) = \boldsymbol{\alpha}^j + \boldsymbol{y}^j(t-1)\boldsymbol{\beta}^j + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}_P, \sigma^2\boldsymbol{I}_{P\times P})$ are independent for $t = 2, \ldots, n$. In this way E such time courses were obtained; that is, from E distinct initial conditions, so the total number of data for individual j is $n_j = En$. In order to avoid issues of blow-up and to generate plausible datasets, the matrices $\boldsymbol{\beta}^j$ were normalized by their spectral radii prior to data generation.

In order to investigate the effect of using a prior network N^0, we do not simply want to set N^0 equal to the latent network N, since in practice this network is unknown. We therefore generated a prior network N^0 by correctly specifying each potential edge as either present or absent with probability h_η. In this way we mimic partial prior knowledge of the networks under study.

4.4.2.2 Alternative Data Generating Mechanisms

We augmented the above data-generating scheme to mimic interventional experiments. In this case, for each time course, a randomly chosen variable is marked as the target of an interventional treatment. Data are then generated according to the augmented likelihood described in Section 4.3.1.2 (fixed effects were taken to be zero). Furthermore, in order to investigate the impact of model misspecification, we also considered time series data generated from a computational model of protein signalling, based on non-linear ODEs [Xu *et al.*, 2010]. In order to extend this model, which is for a single cell type, to simulate a heterogeneous population, we randomly selected three protein species per individual and deleted their outgoing edges in the data-generating network.

4.4.2.3 Latent Network

Firstly we investigated ability to recover the latent network N. Initially all estimators are assigned approximately optimal hyper-parameter values $\eta = 1, \lambda = 4$ (for Xu *et al.* [2010], $\lambda = 3$) based on the heuristic of Eqn. 4.4; prior misspecification is investigated later in Section 4.4.2.6. We found little difference in the ability of JNI and ANI to recover the latent network structure across a wide range of regimes (Table 4.3). Since ANI enjoys favourable computational complexity, this estimator may be preferred for this task in practice. However, both approaches clearly outperformed monolithic inference, which was no better than inference based on the prior alone, demonstrating the importance of accounting for variation in parameter values. Correlations barely outperformed random sampling.

In practice, one could also estimate N using independent network inference (INI), via the *ad hoc* estimator $p(i \in N_p|\boldsymbol{y}, N^0) \approx \frac{1}{J}\sum_{j\in\mathcal{J}} p(i \in N_p^j|\boldsymbol{y}^j, N^0)$ which performs an unweighted average of J independent network inferences. However we found that INI offered no advantage over JNI and ANI, performing worse than both in 14 out of 15 regimes. We obtained qualitatively similar results for both alternative data-generating schemes (Tables 4.5,4.8).

4.4.2.4 Individual Networks

Secondly we investigated the ability to recover individual networks N^j. At this task, JNI outperformed INI in all 15 regimes (Table 4.2). This demonstrates a substantial increase in statistical power resulting from the hierarchical Bayesian approach. JNI also outperformed monolithic estimation and inference using temporal correlations in all 15 regimes, with the latter demonstrating substantial bias.

One may try to improve upon INI by firstly estimating the latent network N, and then taking this estimate as a prior network N^0 within a second round of INI. Informed by Section 4.4.2.3, we consider the approach whereby N is first estimated using ANI, referring to this two-step procedure as *empirical network inference* (ENI). We found that the performance of ENI consistently matched that of JNI over a wide range of regimes. Since ENI avoids all joint computation, this may provide a practical estimator of individual networks in higher dimensional settings. Similar results were observed using the alternative data-generating schemes, although JNI slightly outperformed ENI on the Xu *et al.* [2010] datasets (Tables 4.6,4.9).

4.4.2.5 Feature Detection

Thirdly, we assessed ability to pinpoint sources of variation within the population. Interest is often directed toward individual-specific heterogeneity, or *features*. Informally, writing $N^j = N + \delta^j$, features

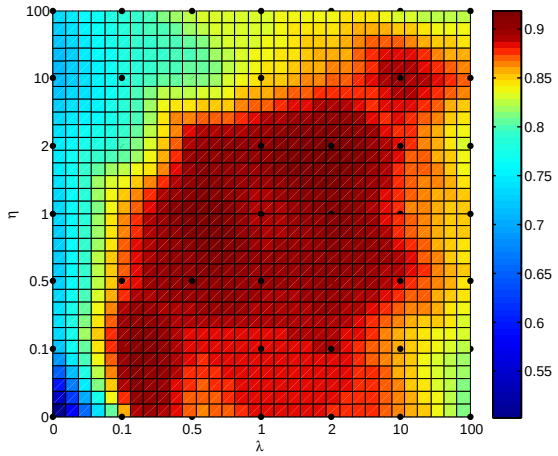


Figure 4.5: AUR as a function of hyper-parameters η, λ. Performance is insensitive to moderate variation in hyper-parameters. [Here we present AUR for inference of the latent network N when the data-generating hyper-parameters are $\lambda = \eta = 1$, but the result is typical for the other estimation problems and other data-generating regimes. A smooth interpolation is used to aid visualization.]

correspond to δ^j. JNI regularizes between individuals; it therefore ought to eliminate spurious differences, leaving only features which are strongly supported by data. Equivalently, since JNI offers improved estimation of the latent network N, the features $\delta^j = N^j - N$ ought also to be better estimated.

Feature detection may also be performed using INI or ENI, comparing an latent network estimator (see *ad hoc* estimator in Section 4.4.2.3) with individual networks. The performance of JNI was compared to the performance of INI and ENI (Table 4.4). We found that, whilst feature detection is much more challenging that previous tasks, JNI mostly outperformed both INI and ENI, with exceptions occurring whenever the underlying dataset was highly informative (in which case INI was often superior). This suggests that coherence of the JNI analysis aids in suppressing spurious features in the small sample setting. Alternative data-generating schemes produced qualitatively similar results, although JNI outperformed ENI on the Xu *et al.* [2010] datasets (Tables 4.7,4.10).

4.4.2.6 Robustness to Hyper-Parameter Misspecification

For the above investigation we used Eqn. 4.4 to elicit hyper-parameters λ, η. This was possible because the data-generating parameters h_λ, h_η were known by design; however in general this will not be the case. It is therefore important that estimator performance does not deteriorate heavily when alternative hyper-parameter values are employed. By fixing (h_λ, h_η) in the data generating process, we are able to investigate the robustness of JNI estimator to hyper-parameter misspecification. In particular, when finite values are ascribed to data-generating parameters (h_η, h_λ), ANI and INI may be interpreted as inference using misspecified prior distributions (see Section 4.2.3).

Fig. 4.5 displays how performance of the JNI estimator for latent networks depends on the choice of hyper-parameters $\lambda, \eta \in [0, \infty)$. Data were generated using $h_\lambda = h_\eta = 0.73$, corresponding to optimal hyper-parameters $\lambda = \eta = 1$. Inference was then performed using a range of misspecified prior distributions, with performance quantified by AUR. We notice that AUR remains close to that obtained for optimal λ, η over a fairly large interval, so that performance is not exquisitely dependent on prior elicitation.

Data Generating Regime								Estimator					
J	n	E	P	σ	ρ	h_η	h_λ	JNI	ANI	INI	Monolithic	Correl.	Prior
10	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0088	0.73 ± 0.011	0.87 ± 0.01	0.71 ± 0.012	0.55 ± 0.013	0.72
5	5	1	10	0.2	1	0.73	0.98	0.86 ± 0.0083	0.74 ± 0.01	0.85 ± 0.0092	0.75 ± 0.01	0.55 ± 0.015	0.72
20	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0057	0.74 ± 0.0098	0.88 ± 0.0074	0.68 ± 0.0089	0.59 ± 0.015	0.72
10	10	1	10	0.2	1	0.73	0.98	0.94 ± 0.0051	0.86 ± 0.0075	0.95 ± 0.0051	0.63 ± 0.012	0.56 ± 0.015	0.72
10	5	5	10	0.2	1	0.73	0.98	0.97 ± 0.0035	0.94 ± 0.0052	0.98 ± 0.0041	0.7 ± 0.011	0.6 ± 0.014	0.72
10	5	1	20	0.2	1	0.73	0.98	0.86 ± 0.0046	0.78 ± 0.0075	0.86 ± 0.0057	0.67 ± 0.0072	0.54 ± 0.0078	0.72
10	5	1	10	0.1	1	0.73	0.98	0.88 ± 0.009	0.75 ± 0.0094	0.88 ± 0.011	0.71 ± 0.012	0.53 ± 0.017	0.72
10	5	1	10	1	1	0.73	0.98	0.81 ± 0.0089	0.7 ± 0.0093	0.79 ± 0.013	0.72 ± 0.0084	0.51 ± 0.013	0.72
10	5	1	10	0.2	0.5	0.73	0.98	0.84 ± 0.012	0.67 ± 0.017	0.84 ± 0.013	0.69 ± 0.017	0.56 ± 0.016	0.72
10	5	1	10	0.2	2	0.73	0.98	0.88 ± 0.0068	0.73 ± 0.0087	0.84 ± 0.0089	0.7 ± 0.0099	0.54 ± 0.01	0.72
10	5	1	10	0.2	1	0.62	0.98	0.86 ± 0.0087	0.63 ± 0.012	0.86 ± 0.009	0.64 ± 0.013	0.53 ± 0.015	0.62
10	5	1	10	0.2	1	0.88	0.98	0.9 ± 0.0052	0.88 ± 0.0066	0.89 ± 0.0088	0.79 ± 0.0089	0.55 ± 0.011	0.87
10	5	1	10	0.2	1	0.73	0.73	0.57 ± 0.0041	0.56 ± 0.0043	0.56 ± 0.0044	0.54 ± 0.0037	0.52 ± 0.0036	0.61
10	5	1	10	0.2	1	0.73	1	0.9 ± 0.014	0.75 ± 0.019	0.9 ± 0.012	0.73 ± 0.014	0.56 ± 0.016	0.73
10	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0084	0.74 ± 0.012	0.89 ± 0.0095	0.71 ± 0.011	0.55 ± 0.013	0.72

Table 4.2: Assessment of estimators for inference of individual networks N^j; autoregressive dataset with interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient, "Prior" = estimation using only the prior network N^0.]

Data Generating Regime								Estimator					
J	n	E	P	σ	ρ	h_η	h_λ	JNI	ANI	INI	Monolithic	Correl.	Prior
10	5	1	10	0.2	1	0.73	0.98	0.93 ± 0.0097	0.93 ± 0.0094	0.77 ± 0.012	0.74 ± 0.015	0.54 ± 0.015	0.73
5	5	1	10	0.2	1	0.73	0.98	0.9 ± 0.009	0.9 ± 0.0091	0.78 ± 0.011	0.78 ± 0.011	0.54 ± 0.017	0.73
20	5	1	10	0.2	1	0.73	0.98	0.95 ± 0.006	0.95 ± 0.0059	0.79 ± 0.012	0.71 ± 0.011	0.56 ± 0.018	0.73
10	10	1	10	0.2	1	0.73	0.98	0.99 ± 0.0045	0.99 ± 0.0047	0.91 ± 0.0061	0.66 ± 0.014	0.53 ± 0.015	0.73
10	5	5	10	0.2	1	0.73	0.98	0.99 ± 0.0032	0.99 ± 0.0039	0.95 ± 0.0044	0.72 ± 0.012	0.56 ± 0.016	0.73
10	5	1	20	0.2	1	0.73	0.98	0.95 ± 0.0057	0.94 ± 0.006	0.86 ± 0.0091	0.73 ± 0.0095	0.51 ± 0.01	0.73
10	5	1	10	0.1	1	0.73	0.98	0.93 ± 0.0093	0.93 ± 0.0095	0.79 ± 0.011	0.75 ± 0.013	0.52 ± 0.018	0.73
10	5	1	10	1	1	0.73	0.98	0.85 ± 0.01	0.85 ± 0.01	0.74 ± 0.011	0.76 ± 0.01	0.51 ± 0.015	0.73
10	5	1	10	0.2	0.5	0.73	0.98	0.94 ± 0.012	0.94 ± 0.012	0.73 ± 0.023	0.74 ± 0.024	0.53 ± 0.024	0.73
10	5	1	10	0.2	2	0.73	0.98	0.91 ± 0.0073	0.91 ± 0.0075	0.76 ± 0.0094	0.71 ± 0.011	0.53 ± 0.011	0.73
10	5	1	10	0.2	1	0.62	0.98	0.91 ± 0.0094	0.91 ± 0.0092	0.66 ± 0.014	0.66 ± 0.015	0.51 ± 0.016	0.62
10	5	1	10	0.2	1	0.88	0.98	0.96 ± 0.0053	0.96 ± 0.0056	0.93 ± 0.0072	0.82 ± 0.01	0.52 ± 0.012	0.88
10	5	1	10	0.2	1	0.73	0.73	0.73 ± 0.011	0.72 ± 0.012	0.72 ± 0.012	0.65 ± 0.013	0.51 ± 0.012	0.73
10	5	1	10	0.2	1	0.73	1	0.92 ± 0.012	0.92 ± 0.011	0.76 ± 0.021	0.73 ± 0.014	0.54 ± 0.016	0.73
10	5	1	10	0.2	1	0.73	0.98	0.94 ± 0.0094	0.94 ± 0.0094	0.79 ± 0.014	0.75 ± 0.012	0.52 ± 0.016	0.73

Table 4.3: Assessment of estimators for inference of the latent network N; autoregressive dataset with interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient, "Prior" = estimation using only the prior network N^0.]

Data Generating Regime								Estimator				
J	n	E	P	σ	ρ	h_η	h_λ	JNI	INI	ENI	Monolithic	Correl.
10	5	1	10	0.2	1	0.73	0.98	0.58 ± 0.0086	0.54 ± 0.0095	0.55 ± 0.008	0.5 ± 0.0093	0.52 ± 0.0094
5	5	1	10	0.2	1	0.73	0.98	0.61 ± 0.014	0.54 ± 0.014	0.56 ± 0.011	0.49 ± 0.01	0.51 ± 0.015
20	5	1	10	0.2	1	0.73	0.98	0.54 ± 0.007	0.54 ± 0.0069	0.56 ± 0.0064	0.5 ± 0.0061	0.53 ± 0.0069
10	10	1	10	0.2	1	0.73	0.98	0.67 ± 0.011	0.7 ± 0.0093	0.67 ± 0.0076	0.49 ± 0.0087	0.53 ± 0.012
10	5	5	10	0.2	1	0.73	0.98	0.73 ± 0.012	0.79 ± 0.01	0.75 ± 0.0067	0.51 ± 0.009	0.56 ± 0.012
10	5	1	20	0.2	1	0.73	0.98	0.62 ± 0.0055	0.61 ± 0.006	0.59 ± 0.0049	0.5 ± 0.0062	0.54 ± 0.0069
10	5	1	10	0.1	1	0.73	0.98	0.57 ± 0.011	0.53 ± 0.0083	0.58 ± 0.0092	0.5 ± 0.0096	0.52 ± 0.012
10	5	1	10	1	1	0.73	0.98	0.54 ± 0.0097	0.52 ± 0.0091	0.52 ± 0.0095	0.49 ± 0.0074	0.51 ± 0.012
10	5	1	10	0.2	0.5	0.73	0.98	0.57 ± 0.0098	0.53 ± 0.0084	0.55 ± 0.0079	0.5 ± 0.0079	0.54 ± 0.0098
10	5	1	10	0.2	2	0.73	0.98	0.57 ± 0.0091	0.53 ± 0.0087	0.53 ± 0.0092	0.5 ± 0.0089	0.52 ± 0.0096
10	5	1	10	0.2	1	0.62	0.98	0.56 ± 0.0081	0.51 ± 0.0096	0.54 ± 0.0084	0.5 ± 0.0078	0.52 ± 0.011
10	5	1	10	0.2	1	0.88	0.98	0.59 ± 0.0093	0.6 ± 0.012	0.6 ± 0.0096	0.49 ± 0.0099	0.52 ± 0.011
10	5	1	10	0.2	1	0.73	0.73	0.53 ± 0.0029	0.51 ± 0.0025	0.51 ± 0.003	0.5 ± 0.003	0.51 ± 0.003
10	5	1	10	0.2	1	0.73	1	-	-	-	-	-
10	5	1	10	0.2	1	0.73	0.98	0.58 ± 0.0095	0.55 ± 0.01	0.57 ± 0.0062	0.51 ± 0.0093	0.52 ± 0.011

Table 4.4: Assessment of estimators for inference of the network features; Autoregressive dataset with interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "INI" = average J independent network inferences, "ENI" = a two step procedure described in Chapter 4, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

Data Generating Regime								Estimator					
J	n	E	P	σ	ρ	h_η	h_λ	JNI	ANI	INI	Monolithic	Correl.	Prior
10	5	1	10	0.2	1	0.73	0.98	0.95 ± 0.0077	0.95 ± 0.0074	0.79 ± 0.01	0.76 ± 0.012	0.54 ± 0.016	0.73
5	5	1	10	0.2	1	0.73	0.98	0.92 ± 0.0079	0.93 ± 0.0078	0.77 ± 0.011	0.76 ± 0.01	0.52 ± 0.015	0.73
20	5	1	10	0.2	1	0.73	0.98	0.95 ± 0.0083	0.96 ± 0.008	0.81 ± 0.011	0.76 ± 0.012	0.54 ± 0.014	0.73
10	10	1	10	0.2	1	0.73	0.98	0.99 ± 0.0026	0.99 ± 0.0026	0.93 ± 0.0059	0.68 ± 0.012	0.51 ± 0.018	0.73
10	5	5	10	0.2	1	0.73	0.98	0.99 ± 0.0031	0.99 ± 0.0033	0.96 ± 0.0041	0.7 ± 0.0098	0.51 ± 0.015	0.73
10	5	1	20	0.2	1	0.73	0.98	0.96 ± 0.0039	0.96 ± 0.0039	0.87 ± 0.008	0.75 ± 0.0089	0.53 ± 0.011	0.73
10	5	1	10	0.1	1	0.73	0.98	0.96 ± 0.0061	0.97 ± 0.0057	0.79 ± 0.012	0.74 ± 0.012	0.53 ± 0.014	0.73
10	5	1	10	1	1	0.73	0.98	0.87 ± 0.011	0.87 ± 0.011	0.76 ± 0.011	0.76 ± 0.014	0.53 ± 0.015	0.73
10	5	1	10	0.2	0.5	0.73	0.98	0.96 ± 0.01	0.96 ± 0.01	0.77 ± 0.015	0.74 ± 0.017	0.53 ± 0.023	0.73
10	5	1	10	0.2	2	0.73	0.98	0.92 ± 0.0074	0.92 ± 0.0075	0.79 ± 0.0078	0.74 ± 0.0091	0.53 ± 0.0089	0.73
10	5	1	10	0.2	1	0.62	0.98	0.95 ± 0.0078	0.95 ± 0.0077	0.69 ± 0.012	0.69 ± 0.013	0.49 ± 0.014	0.62
10	5	1	10	0.2	1	0.88	0.98	0.96 ± 0.0058	0.96 ± 0.0059	0.92 ± 0.0084	0.83 ± 0.0094	0.53 ± 0.018	0.88
10	5	1	10	0.2	1	0.73	0.73	0.75 ± 0.012	0.74 ± 0.012	0.74 ± 0.012	0.7 ± 0.014	0.51 ± 0.013	0.73
10	5	1	10	0.2	1	0.73	1	0.94 ± 0.0083	0.94 ± 0.0079	0.78 ± 0.013	0.77 ± 0.012	0.52 ± 0.014	0.73
10	5	1	10	0.2	1	0.73	0.98	0.97 ± 0.006	0.97 ± 0.0057	0.79 ± 0.011	0.77 ± 0.013	0.53 ± 0.015	0.73

Table 4.5: Assessment of estimators for inference of the latent network N; Autoregressive dataset without interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient, "Prior" = estimation using only the prior network N^0.]

Data Generating Regime								Estimator					
J	n	E	P	σ	ρ	h_η	h_λ	JNI	INI	ENI	Monolithic	Correl.	Prior
10	5	1	10	0.2	1	0.73	0.98	0.89 ± 0.0077	0.75 ± 0.0095	0.89 ± 0.0098	0.72 ± 0.01	0.56 ± 0.013	0.72
5	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0079	0.73 ± 0.01	0.89 ± 0.0083	0.73 ± 0.009	0.55 ± 0.013	0.72
20	5	1	10	0.2	1	0.73	0.98	0.89 ± 0.0079	0.77 ± 0.01	0.89 ± 0.0089	0.72 ± 0.01	0.55 ± 0.011	0.72
10	10	1	10	0.2	1	0.73	0.98	0.95 ± 0.0027	0.88 ± 0.0064	0.96 ± 0.0035	0.66 ± 0.0098	0.55 ± 0.015	0.72
10	5	5	10	0.2	1	0.73	0.98	0.98 ± 0.0031	0.95 ± 0.0047	0.99 ± 0.0028	0.68 ± 0.0079	0.56 ± 0.013	0.72
10	5	1	20	0.2	1	0.73	0.98	0.88 ± 0.0036	0.79 ± 0.0066	0.87 ± 0.0045	0.7 ± 0.0068	0.56 ± 0.0086	0.72
10	5	1	10	0.1	1	0.73	0.98	0.91 ± 0.0064	0.75 ± 0.011	0.92 ± 0.0075	0.71 ± 0.011	0.56 ± 0.013	0.72
10	5	1	10	1	1	0.73	0.98	0.83 ± 0.0093	0.72 ± 0.0093	0.8 ± 0.011	0.73 ± 0.012	0.54 ± 0.013	0.72
10	5	1	10	0.2	0.5	0.73	0.98	0.85 ± 0.009	0.7 ± 0.011	0.86 ± 0.0086	0.67 ± 0.012	0.58 ± 0.017	0.72
10	5	1	10	0.2	2	0.73	0.98	0.9 ± 0.0069	0.76 ± 0.0074	0.85 ± 0.01	0.72 ± 0.0082	0.54 ± 0.0087	0.72
10	5	1	10	0.2	1	0.62	0.98	0.89 ± 0.0075	0.66 ± 0.012	0.89 ± 0.0086	0.67 ± 0.011	0.52 ± 0.012	0.62
10	5	1	10	0.2	1	0.88	0.98	0.9 ± 0.0068	0.87 ± 0.0074	0.89 ± 0.01	0.79 ± 0.0083	0.56 ± 0.016	0.87
10	5	1	10	0.2	1	0.73	0.73	0.58 ± 0.0038	0.56 ± 0.004	0.57 ± 0.0042	0.55 ± 0.0041	0.52 ± 0.0036	0.61
10	5	1	10	0.2	1	0.73	1	0.93 ± 0.0085	0.77 ± 0.013	0.92 ± 0.0094	0.76 ± 0.012	0.55 ± 0.015	0.73
10	5	1	10	0.2	1	0.73	0.98	0.9 ± 0.0067	0.74 ± 0.0094	0.91 ± 0.008	0.73 ± 0.01	0.56 ± 0.013	0.72

Table 4.6: Assessment of estimators for inference of the individual networks N^j; Autoregressive dataset without interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "INI" = average J independent network inferences, "ENI" = a two step procedure described in Chapter 3, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient, "Prior" = estimation using only the prior network N^0.]

Data Generating Regime								Estimator				
J	n	E	P	σ	ρ	h_η	h_λ	JNI	INI	ENI	Monolithic	Correl.
10	5	1	10	0.2	1	0.73	0.98	0.58 ± 0.01	0.55 ± 0.012	0.57 ± 0.0091	0.5 ± 0.009	0.54 ± 0.01
5	5	1	10	0.2	1	0.73	0.98	0.63 ± 0.012	0.53 ± 0.012	0.58 ± 0.01	0.49 ± 0.017	0.55 ± 0.016
20	5	1	10	0.2	1	0.73	0.98	0.54 ± 0.0069	0.55 ± 0.0073	0.57 ± 0.0066	0.51 ± 0.007	0.52 ± 0.0078
10	10	1	10	0.2	1	0.73	0.98	0.69 ± 0.01	0.71 ± 0.011	0.67 ± 0.0088	0.51 ± 0.011	0.52 ± 0.011
10	5	5	10	0.2	1	0.73	0.98	0.79 ± 0.014	0.84 ± 0.011	0.77 ± 0.0082	0.51 ± 0.01	0.57 ± 0.013
10	5	1	20	0.2	1	0.73	0.98	0.64 ± 0.0059	0.6 ± 0.0067	0.6 ± 0.0055	0.5 ± 0.0052	0.54 ± 0.0059
10	5	1	10	0.1	1	0.73	0.98	0.58 ± 0.0092	0.55 ± 0.0099	0.58 ± 0.0081	0.49 ± 0.0095	0.52 ± 0.011
10	5	1	10	1	1	0.73	0.98	0.55 ± 0.0083	0.53 ± 0.0091	0.52 ± 0.01	0.5 ± 0.0095	0.51 ± 0.0073
10	5	1	10	0.2	0.5	0.73	0.98	0.57 ± 0.01	0.56 ± 0.012	0.56 ± 0.01	0.5 ± 0.0083	0.54 ± 0.011
10	5	1	10	0.2	2	0.73	0.98	0.55 ± 0.0078	0.53 ± 0.0083	0.55 ± 0.0087	0.5 ± 0.0095	0.5 ± 0.012
10	5	1	10	0.2	1	0.62	0.98	0.56 ± 0.0077	0.52 ± 0.0085	0.54 ± 0.0091	0.49 ± 0.0092	0.53 ± 0.009
10	5	1	10	0.2	1	0.88	0.98	0.58 ± 0.0084	0.57 ± 0.01	0.58 ± 0.01	0.48 ± 0.0095	0.53 ± 0.0097
10	5	1	10	0.2	1	0.73	0.73	0.53 ± 0.0027	0.5 ± 0.0024	0.51 ± 0.003	0.5 ± 0.0027	0.52 ± 0.0029
10	5	1	10	0.2	1	0.73	1	-	-	-	-	-
10	5	1	10	0.2	1	0.73	0.98	0.57 ± 0.0087	0.53 ± 0.0076	0.57 ± 0.0088	0.49 ± 0.0088	0.52 ± 0.01

Table 4.7: Assessment of estimators for inference of network features; Autoregressive dataset without interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "INI" = average J independent network inferences, "ENI" = a two step procedure described in Chapter 4, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

Data Generating Regime					Estimator				
J	n	E	σ	h_η	JNI	ANI	INI	Monolithic	Correl.
10	5	5	0.2	0.73	0.93 ± 0.0054	0.93 ± 0.0052	0.89 ± 0.006	0.78 ± 0.013	0.4 ± 0.0067
5	5	5	0.2	0.73	0.92 ± 0.0059	0.92 ± 0.0062	0.88 ± 0.0062	0.88 ± 0.0064	0.4 ± 0.0069
20	5	5	0.2	0.73	0.94 ± 0.0044	0.94 ± 0.0043	0.9 ± 0.0078	0.46 ± 0.0042	0.42 ± 0.007
10	10	5	0.2	0.73	0.95 ± 0.0051	0.95 ± 0.005	0.94 ± 0.0042	0.45 ± 0.0041	0.42 ± 0.0065
10	5	1	0.2	0.73	0.71 ± 0.011	0.69 ± 0.01	0.76 ± 0.011	0.88 ± 0.0074	0.41 ± 0.0074
10	5	10	0.2	0.73	0.95 ± 0.0036	0.95 ± 0.004	0.93 ± 0.0052	0.46 ± 0.0039	0.41 ± 0.0073
10	5	5	0.1	0.73	0.94 ± 0.0046	0.94 ± 0.0047	0.89 ± 0.0056	0.8 ± 0.011	0.4 ± 0.0059
10	5	5	1	0.73	0.86 ± 0.0083	0.85 ± 0.0088	0.83 ± 0.011	0.52 ± 0.0054	0.4 ± 0.0063
10	5	5	0.2	0.62	0.92 ± 0.0057	0.92 ± 0.006	0.83 ± 0.0085	0.77 ± 0.013	0.4 ± 0.0065
10	5	5	0.2	0.88	0.94 ± 0.0057	0.94 ± 0.0056	0.97 ± 0.0034	0.8 ± 0.011	0.41 ± 0.0062

Table 4.8: Assessment of estimators for inference of the latent network N; Xu *et al.* [2010] dataset. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

Data Generating Regime					Estimator				
J	n	E	σ	h_η	JNI	INI	ENI	Monolithic	Correl.
10	5	5	0.2	0.73	0.9 ± 0.0053	0.84 ± 0.0063	0.89 ± 0.0058	0.76 ± 0.011	0.43 ± 0.0025
5	5	5	0.2	0.73	0.9 ± 0.0051	0.84 ± 0.0076	0.88 ± 0.0071	0.85 ± 0.0066	0.43 ± 0.0036
20	5	5	0.2	0.73	0.91 ± 0.0047	0.84 ± 0.0076	0.88 ± 0.0056	0.45 ± 0.002	0.44 ± 0.0027
10	10	5	0.2	0.73	0.92 ± 0.0043	0.89 ± 0.0046	0.91 ± 0.0042	0.45 ± 0.0021	0.44 ± 0.0031
10	5	1	0.2	0.73	0.69 ± 0.0099	0.74 ± 0.011	0.57 ± 0.01	0.85 ± 0.0073	0.42 ± 0.0029
10	5	10	0.2	0.73	0.93 ± 0.0038	0.88 ± 0.0057	0.91 ± 0.0045	0.45 ± 0.002	0.44 ± 0.0029
10	5	5	0.1	0.73	0.91 ± 0.0048	0.84 ± 0.0069	0.9 ± 0.0052	0.78 ± 0.0093	0.43 ± 0.0029
10	5	5	1	0.73	0.83 ± 0.0078	0.78 ± 0.01	0.79 ± 0.009	0.52 ± 0.0048	0.42 ± 0.0031
10	5	5	0.2	0.62	0.9 ± 0.005	0.77 ± 0.0076	0.89 ± 0.0053	0.76 ± 0.012	0.43 ± 0.0032
10	5	5	0.2	0.88	0.91 ± 0.0057	0.92 ± 0.0044	0.88 ± 0.0064	0.79 ± 0.01	0.44 ± 0.003

Table 4.9: Assessment of estimators for inference of individual networks; Xu *et al.* [2010] dataset. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "INI" = average J independent network inferences, "ENI" = a two step procedure described in Chapter 4, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

Data Generating Regime					Estimator				
J	n	E	σ	h_η	JNI	INI	ENI	Monolithic	Correl.
10	5	5	0.2	0.73	0.66 ± 0.022	0.64 ± 0.022	0.61 ± 0.021	0.46 ± 0.021	0.63 ± 0.026
5	5	5	0.2	0.73	0.64 ± 0.019	0.66 ± 0.019	0.61 ± 0.019	0.42 ± 0.017	0.63 ± 0.023
20	5	5	0.2	0.73	0.68 ± 0.02	0.63 ± 0.018	0.61 ± 0.018	0.42 ± 0.016	0.59 ± 0.021
10	10	5	0.2	0.73	0.71 ± 0.019	0.69 ± 0.018	0.68 ± 0.016	0.44 ± 0.015	0.56 ± 0.022
10	5	1	0.2	0.73	0.5 ± 0.016	0.51 ± 0.022	0.51 ± 0.024	0.41 ± 0.015	0.59 ± 0.022
10	5	10	0.2	0.73	0.69 ± 0.023	0.7 ± 0.019	0.65 ± 0.019	0.42 ± 0.015	0.6 ± 0.023
10	5	5	0.1	0.73	0.68 ± 0.017	0.64 ± 0.019	0.6 ± 0.017	0.48 ± 0.019	0.65 ± 0.019
10	5	5	1	0.73	0.57 ± 0.018	0.55 ± 0.019	0.55 ± 0.023	0.49 ± 0.017	0.63 ± 0.021
10	5	5	0.2	0.62	0.64 ± 0.023	0.64 ± 0.02	0.59 ± 0.023	0.49 ± 0.025	0.63 ± 0.024
10	5	5	0.2	0.88	0.66 ± 0.023	0.69 ± 0.02	0.66 ± 0.018	0.5 ± 0.022	0.61 ± 0.022

Table 4.10: Assessment of estimators for inference of network features; Xu *et al.* [2010] dataset. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyper-parameters. "JNI" = joint network inference, "INI" = average J independent network inferences, "ENI" = a two step procedure described in Chapter 4, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

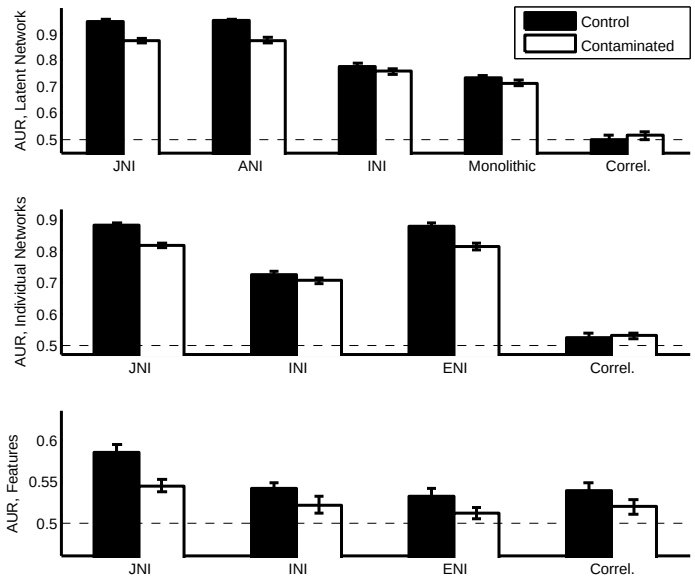


Figure 4.6: Investigating robustness to outliers and batch effects. [Here the autoregressive model, described in Chapter 4, was used to generate data. Mean AUR over 50 iterations and associated standard error are reported. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "ENI" = two-step estimation procedure described in Chapter 4, "Naive DBN" = as JNI but without integrating interventions, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient.]

4.4.2.7 Robustness to Outliers and Batch Effects

The biological datasets which motivate this study often contain outliers. At the same time, experimental design may lead to platform-specific batch effects. In order to probe estimator robustness, we generated data as previously described, with the addition of outliers and certain batch effects. Specifically, Gaussian noise from the contamination model $0.95\mathcal{N}(0, 0.1^2) + 0.05\mathcal{N}(0, 10^2)$ was added to all data prior to inference. At the same time, one individual's data were replaced entirely by Gaussian white noise to simulate a batch effect that could arise if preparation of a specific biological sample was incorrect. The relative decrease in performance at feature detection is reported in Fig. 4.6. We found that JNI remained the optimal estimator for all three estimation problems, in spite of these heavy violations to the modelling assumptions. However, the actual decrease in performance was more pronounced for JNI than for INI, suggesting that decoupled estimation (INI) may confer robustness to batch effects which affect single individuals.

4.4.3 Breast Cancer Data

In this Section we consider experimental data derived from human breast cancer cell lines. We focus on protein signalling networks, for which a substantial proportion of wild type network topology has been characterized by extensive biochemistry (EGFR pathway; Fig. 4.1). In this setting it is known that certain genetic aberrations which influence the network structure are relatively common across populations [Bachman *et al.*, 2004; Davies *et al.*, 2002]. Moreover, network heterogeneity is unlikely to be uniform across cell lines, since these genomic aberrations do not occur independently. This investigation serves three purposes: Firstly, the EGFR pathway allows us to validate a subset of the conclusions drawn from the simulation study, and thereby gain some confidence as to the applicability of those results to real data. Secondly, this study allows us to investigate the use of JNI in a realistic setting where ancillary information is available, in the form of mutational status and histological profiling. Finally, the tools developed in this Chapter allow us to shed light on the signalling heterogeneity across a panel of breast cancer cell lines.

Data were obtained using reverse-phase protein arrays [Hennessy *et al.*, 2010] from $J = 6$ breast cancer cell lines (AU565, BT474, HCC1954, MDAMB231, SKBR3 and SUM225CWN; experimental protocol is

described in brief in Section B.6.1). Data comprised observations for the $P = 17$ proteins shown in Fig. 4.1 (see also Table C.1). Specifically, $\boldsymbol{y}$ contains the logarithms of the measured concentrations. Data were acquired under treatment with an EGFR/HER2 inhibitor Lapatinib ("EGFRi"), an Akt inhibitor ("Akti"), EGFRi and Akti in combination, and without inhibition ("DMSO") at 0.5,1,2 and 4 hours following Serum stimulation, giving a total of $n_j = 16$ observations of each variable in each individual cell line.

4.4.3.1 Ancillary Information

In the context of cancer cell lines, ancillary information is available in the form of genetic aberrations (mutation statuses) and histological profiling, which may be integrated into a Bayesian prior. For instance, a loss-of-function mutation in the kinase domain of a protein corresponds to zero prior probability on edges emanating from that protein, since the protein is no longer functionally active as a kinase. Further, if the mutation also affects the ability of a protein to be phosphorylated, then incoming edges may also be assigned zero prior probability. As a second example, cell lines with ectopic expression of the receptor HER2 are known to depend heavily upon EGFR signalling. In this case the network prior would not penalize edges emanating from the EGFR receptor family. For our panel of breast cancer cell lines there is ancillary information available from published sources [Neve *et al.*, 2006] and on-line databases [Forbes *et al.*, 2011]. This data is reproduced in Table C.2. For our investigation we encoded ancillary information into a network prior following the general principles outlined in this paragraph. Full details are provided in Section C.3; we refer to these estimators as "JNI + A" etc.

For the EGFR pathway, extensive biochemistry on "wild type" cell lines (mainly fibroblasts) has produced a well-validated network representation N^0 (Fig. 4.1). However, individual breast cancer cell lines have received (comparatively) far less attention. It is therefore generally true that most of our prior knowledge on cell lines derives directly from assumed similarity with N^0. Whilst cancer signalling may differ with respect to wild type signalling, we expect the differences to be small in number. In light of these observations, we used our elicitation criteria from Section 4.2.4 to select hyper-parameters $\lambda = 4, \eta = 5$, corresponding variously to $h_\lambda = p(i \notin N_p^j \Delta N_p) = 0.982$, $h_\eta = p(i \notin N_p \Delta N_p^0) = 0.993$, $\mathbb{E}(\mathrm{SHD}(N^j, N)) = 5.2$, $\mathbb{E}(\mathrm{SHD}(N, N^0)) = 2.0$, $s_1 = p(i \notin N_p^j \Delta N_p^k) = 0.965$, $s_2 = p(i \notin N_p^j \Delta N_p^0) = 0.976$, $\mathbb{E}(\mathrm{SHD}(N^j, N^k)) = 10.1$ and $\mathbb{E}(\mathrm{SHD}(N^j, N^0)) = 6.9$.

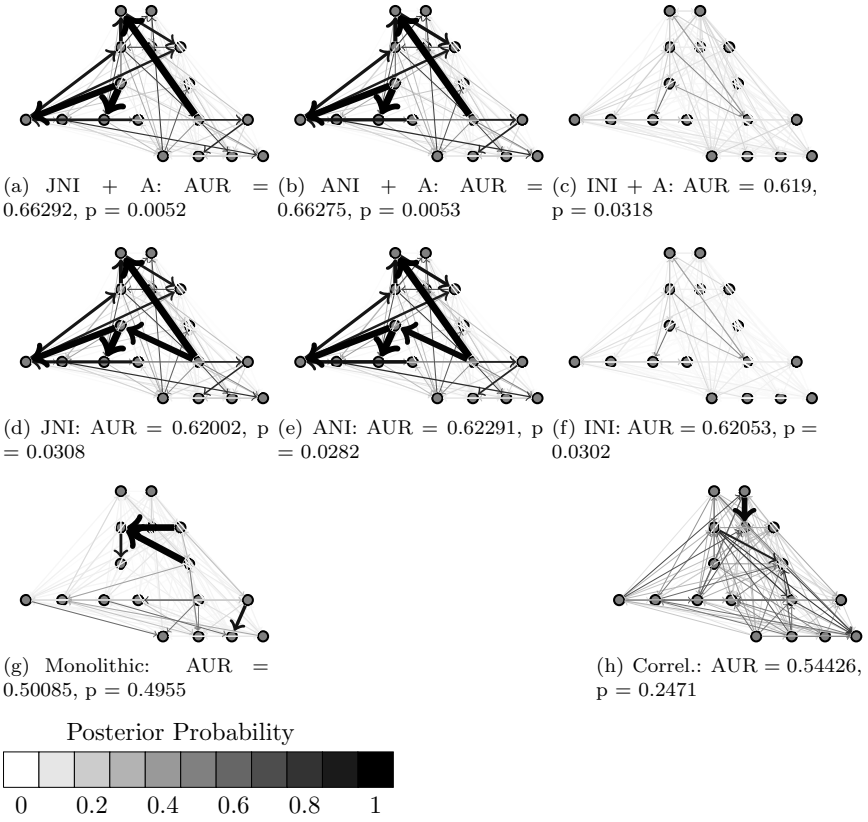
4.4.3.2 Validation of Estimators

In order to test estimator performance using real data, we firstly investigated inference for the latent network N, benchmarking estimates against the wild type network from literature (Fig. 4.1). For an unbiased assessment we used an empty prior network N^0. Inferred networks are displayed in Fig. 4.7(a). Results demonstrated good recovery of the literature network, with JNI attaining the highest AUR (0.66, $p < 0.01$, Fig. 4.7(b)). As in the simulation study, JNI outperformed INI, with ANI representing a good approximation to JNI. Interestingly, posterior inclusion probabilities for INI were consistently low, indicating that statistical power is sacrificed in independent inference. Inference was performed with and without using the ancillary data. We found that, for JNI and ANI, estimation improved as a result of including the ancillary data, demonstrating the potential strength of Bayesian estimation in this setting. Conversely, inclusion of ancillary data did not improve the performance of INI.

4.4.3.3 Inference for Cell Lines

We investigated inference for cell line specific networks (Fig. 4.8), taking the prior network N^0 from literature (Fig. 4.1). In order to assess correctness of the inferred heterogeneity, we exploited the fact that cell lines AU565 and SKBR3 derive from the same patient. We would therefore expect these two cell lines to be most similar at the network level. Reassuringly, under JNI and ENI, these cell line specific networks are the most similar, maximizing the Spearman correlation coefficient between corresponding posterior marginal inclusion probabilities over all ${}^6C_2 = 15$ possible pairs of cell lines. Both INI and correlation statistics fail to identify these two lines as the most similar.

A thorough assessment of the accuracy of these individual networks will require additional interventional experiments. However, Fig. 4.9 demonstrates that corresponding posterior edge probabilities are convincingly regularized under JNI and ENI compared with INI and correlation statistics; this is likely a necessary condition for successful network inference in this setting.



(a) JNI + A: AUR = 0.66292, p = 0.0052

(b) ANI + A: AUR = 0.66275, p = 0.0053

(c) INI + A: AUR = 0.619, p = 0.0318

(d) JNI: AUR = 0.62002, p = 0.0308

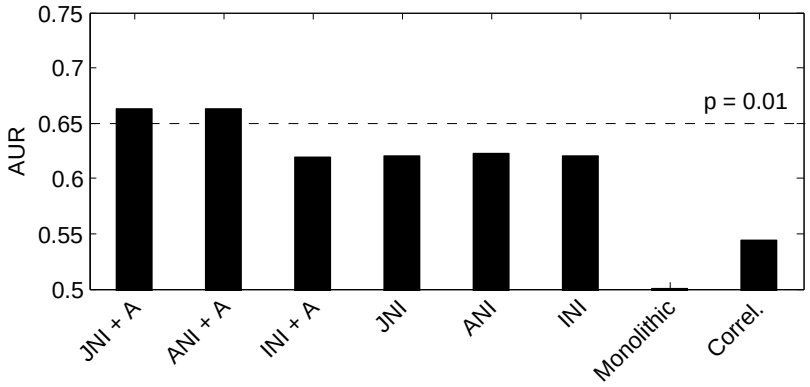
(e) ANI: AUR = 0.62291, p = 0.0282

(f) INI: AUR = 0.62053, p = 0.0302

(g) Monolithic: AUR = 0.50085, p = 0.4955

(h) Correl.: AUR = 0.54426, p = 0.2471

(a)



(b)

Figure 4.7: Breast cancer data; inference for the canonical EGFR pathway. [An empty prior N_0 was used. The literature network from Fig. 4.1 forms a benchmark for estimator assessment. For (a), the layout of vertices is congruent to Fig. 4.1, which may be used as a key. p-values were calculated by permutation test based on the AUR statistic, with 10,000 samples used to obtain an empirical null distribution. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient. "+A" indicates that ancillary information was integrated into inference.]

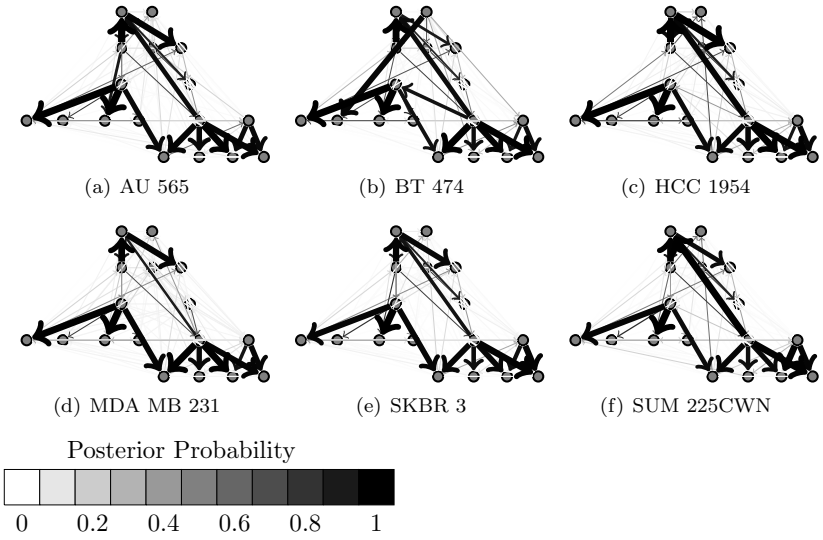


Figure 4.8: Breast cancer data; cell line specific networks inferred by JNI, using ancillary data. [Edge width and colour are proportional to posterior marginal inclusion probabilities. The layout of vertices is congruent to Fig. 4.1, which may be used as a key.]

4.5 Discussion

There are three distinct, though related, structure learning problems which may be addressed in the context of an heterogeneous population of individuals:

1. Recovering a shared or "wild type" network from the heterogeneous data.

2. Recovering networks for specific individuals.

3. Pinpointing network variation within the population.

Each problem may be of independent scientific interest and the joint approaches investigated here address all three problems simultaneously within a coherent framework. We considered simulated data, with and without model misspecification, as well as real data obtained from cancer cell lines. For all three problems we demonstrated that a joint analysis performs at least as well as independent or aggregate analyses.

Our analysis, based on exact Bayesian model averaging and a sum-product type ("propagation") algorithm, was massively faster then the sampling-based schemes of Penfold *et al.* [2012]; Werhli and Husmeier [2008]. Moreover, our estimators are deterministic, so that difficulties pertaining to MCMC convergence were avoided. Indeed, attaining convergence on joint models of this kind appears to be challenging [Werhli and Husmeier, 2008]. The proposed methodology is scalable, with an embarrassingly parallel algorithm provided in Section 4.3.3. Furthermore, we described approximations to a joint analysis which enjoy further reduced computational complexity whilst providing almost equal estimator performance across a wide range of data-generating regimes.

Whilst we considered the simplest form of regularization, based on prior modularity, there is potential to integrate richer knowledge into inference. One possibility would be hierarchical regularization that allows entire pathways to be either active or inactive. However, in this setting it would be important to revisit hyper-parameter elicitation; the procedures which we have described are specific to SHD priors. In particular we restricted the joint model to have equal inverse temperatures $\lambda^1 = \cdots = \lambda^J := \lambda$. Relaxing this assumption may improve robustness to batch effects which target single individuals, since then weak informativeness ($\lambda^j \approx 0$) may be learned from data. It would also be interesting to distinguish between $N \backslash N^j$ ("loss of function") and $N^j \backslash N$ ("gain of function") features. However, as we have seen, hyper-parameter elicitation in these hierarchical models requires a degree of care. In this work we did not explore information sharing through parameter values $\boldsymbol{\theta}^j$, yet this may yield more powerful estimators of network structure in settings where individuals' parameters $\boldsymbol{\theta}^j, \boldsymbol{\theta}^k$ are not independent.

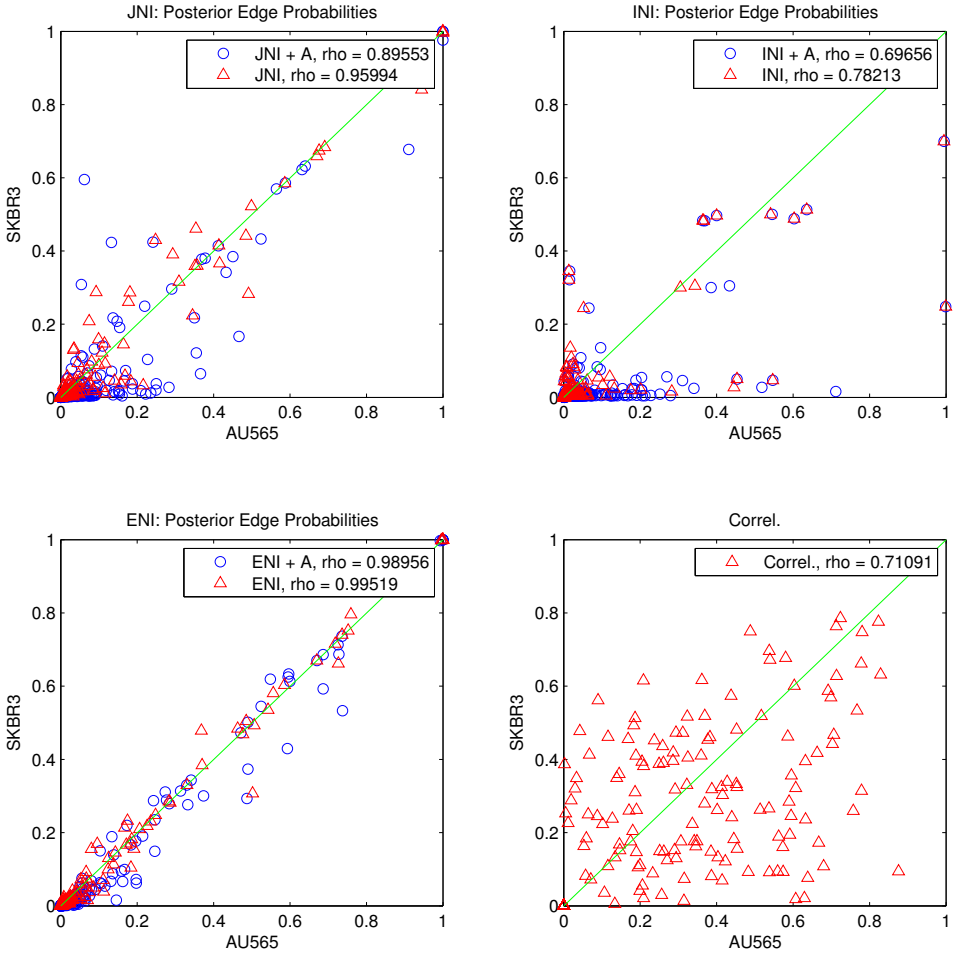


Figure 4.9: Two breast cancer cell lines from the same patient. Joint inference (JNI) and approximate joint inference (ENI) improve the Spearman correlation coefficient ("rho") between posterior marginal inclusion probabilities for AU565 and SKBR3 compared to independent inference (INI) and inference based on absolute Pearson temporal correlation coefficients (Correl). ["+A" refers to the integration of ancillary information into inference, as described in the Main Text.]

The JNI model could be formulated as a penalized (log-)likelihood

$$\log(p(\boldsymbol{y}|N^1,\ldots,N^J)) - \sum_{j\in\mathcal{J}} \lambda^j d^j(N^j,N) - \eta d(N,N^0;A^j). \tag{4.26}$$

The frequentist approaches described by Danaher *et al.* [2012]; Guo *et al.* [2011]; Mohan *et al.* [2013]; Yang *et al.* [2012] enjoy favourable computational complexity (esp. Danaher *et al.* [2012] who provide an example with $P = 22,283$ variables and $J = 187$ individuals). However, in small to moderate dimensional settings, the Bayesian methods proposed here are complementary in several respects: (i) Bayesian approaches provide a confidence measure for inferred topology, dealing with non-identifiable and multi-modal problems; (ii) no convexity assumptions are required on the form of the penalty functions d, d^j in the Bayesian setting, which may assist with integration of ancillary information; (iii) the above penalized likelihood methods do not apply directly to time course data (but could be extended to do so).

These experiments employed a promising formulation of likelihood under intervention due to Spencer and Mukherjee [2012]. There are a number of interesting extensions which may be considered in future work: (i) In high dimensions, Bayesian variable selection requires multiplicity correction in order to avoid degeneracy [Scott and Berger, 2010]. Such correction is required to control the false discovery rate and is independent to the penalty on model complexity provided by the marginal likelihood. In this moderate-dimensional work, in order to simplify the presentation, we did not employ a multiplicity correction; this should be an avenue for future development. (ii) Inference was based upon a local score borrowed from Bayesian linear regression. We chose to employ the g-prior due to Zellner [1986], where following George and Foster [2000] we used (conditional) empirical Bayes to select the g hyper-parameter. Others have suggested setting $g = n$ [unit information prior; Smith *et al.*, 2001], whilst Deltell *et al.* [2012] and Liang *et al.* [2008] propose prior distributions over g with attractive theoretical properties. Our empirical investigation suggested that the choice of hyper-parameter elicitation is influential, but a thorough comparison of linear model specifications is beyond the scope of this thesis. (iii) As discussed in Chapter 2, linear autoregressive formulations may be inadequate in realistic settings; in particular, samples which are obtained unevenly in time can be problematic. Recent advances which incorporate mechanistic detail into the likelihood, such as those in Chapter 3 may prove advantageous. Since the JNI approach decouples the marginal likelihood and model averaging computations, it may be applied directly to the output of more sophisticated models. (iv) In the case of linear models, Barbieri and Berger [2004] showed that the median probability model (i.e. model averaging) provides superior predictive performance over the maximum *a posteriori* (MAP) model in the Bayesian setting. However we are unaware of an analogous result for causal inference.

Techniques for modelling heterogeneous data are clearly widely applicable. The methodology presented here may be applicable in other disciplines. For example, our approach is suited to meta-analyses of network analyses [Weile *et al.*, 2012], integration of multiple data sources [Kato *et al.*, 2005; Wei and Pan, 2012; Werhli and Husmeier, 2008] or data arising from context dependent networks [Baumbach *et al.*, 2009]. The ideas discussed here share many connections with time-heterogeneous DBNs which, for brevity, we did not discuss in this thesis [Dondelinger *et al.*, 2010, 2012; Grzegorczyk and Husmeier, 2011; Song *et al.*, 2009].

4.6 Addendum: Structured Populations

This Chapter focussed on unstructured populations of individuals; in particular our model was *exchangeable* in the sense that each individual was treated equally within the joint likelihood. However in many applications individuals will not satisfy such an exchangeability assumption. For example the cancer cell line panel considered in this Chapter may be better described by firstly stratifying cell lines according to cancer subtype and then constructing a JNI model within each subtype. Here the population as a whole will not be exchangeable, since cell lines within the same subtype are more closely related than cell lines in different subtypes (the population is *structured*). For brevity we did not discuss structured populations in this thesis. Oates *et al.* [2013c] extends the JNI framework to encompass populations admitting a (known) tree structure. In this more general setting, efficient computation is once again facilitated by belief propagation. However hyper-parameter elicitation requires greater care since each branch of the tree, in principle, requires its own inverse temperature hyper-parameter. Oates *et al.* [2013c] circumvent hyper-parameter elicitation by introducing a novel, non-parametric regularisation between adjacent networks in the tree. The reader is referred to Oates *et al.* [2013c] for full details.

Chapter 5

Outlook

In this thesis we have discussed some statistical challenges in the area of cellular signalling systems, proposed specific solutions and explicated the methodology in the context of emerging high-throughput phosphoproteomic data obtained from reverse phase protein arrays (RPPAs). In particular we focussed on the data-driven characterisation of context-specific biological networks. It is worth stressing that, despite enjoying a recent surge in popularity, research into network inference remains underdeveloped and care should be taken when drawing scientific conclusions from such analyses.

This thesis contributed three novel Chapters to the emerging literature.

- In Chapter 2 we surveyed the existing statistical literature for approaches to network inference from time course data, exploring the connection between simplified statistical models and complex data-generating processes at the single-cell level. We highlighted issues surrounding the effectiveness of such inference procedures on realistic data and proposed a specification of the linear model which could served as a default for subsequent work.

- Chapter 3 extended this statistical specification to integrate non-linear chemical kinetics into network inference. We found that such models (i) facilitate improved reconstruction of biological network topology *in silico* and (ii) have the potential to boost predictive power in situations where it is not possible to reliably extract network topology from literature. Using RPPA data obtained on breast cancer cell lines we validated (ii), finding that response of HER2+ cell lines to small molecule inhibitors EGFRi and Akti were better predicted by the proposed CheMA approach than by fitting a model derived from literature to data.

- Chapter 4 generalised the modelling approach to integrate multiple data obtained from a heterogeneous population. Here we found that a hierarchical model was able to increase statistical efficiency in recovering network structure across the population at little additional computational cost (in moderate dimensional settings). Using RPPA data, we found that the proposed JNI approach successfully regularised networks across a panel of breast cancer cell lines and demonstrated ability to recover "wild type" signalling topology from a heterogeneous biological sample.

A technical commentary was reserved for the relevant Chapters, but below we provide a high level summary of the many opportunities for further statistical research in this field:

- *Modelling.* This thesis restricted attention to Gaussian DBNs and ODE models of cellular chemistry. Yet it is known that cellular dynamics are frequently non-Gaussian [Paulsson, 2005] and are not well described by the well-mixed assumption of mass action chemistry [Ando *et al.*, 2010; Konopka *et al.*, 2006]. More general continuous time Markov processes, with spatial effects being explicitly modelled, offers one solution. However inference for such systems is extremely challenging, with even direct simulation requiring sophisticated computational techniques [Vigelius *et al.*, 2010].

- *Observational.* Statistical inference for "real world" continuous time stochastic processes is necessarily based on data obtained discretely in time, typically with measurement noise. Moreover, the microscopic molecular systems arising in cell biology often necessitate indirect approaches to experimental data collection, involving aggregation of possibly asynchronous cell populations in

order to generate sufficient biological material, or destructive sampling, where "time course" data are in reality non-longitudinal. For many experimental platforms it is possible to simultaneously measure the abundance of many ($P >> 1$) species, yet the total number n of measurements is often limited due to cost and/or labour, leading to a paucity of data ($P > n$). All of these difficulties suggest avenues for the development of novel statistical methodologies.

- *Emerging biotechnologies.* We considered data derived from RPPA, a relatively new high-throughput technology [Hennessy *et al.*, 2010]. Several other experimental platforms are emerging and it will be important to investigate adequate statistical representations for the resulting data. For instance RNA-seq data contain integer counts of the number of each RNA in a biological sample; the discrete nature of this data call for novel statistical analyses.

- *Causal reasoning and latent variables.* We focused on the simplest possible case of fully observed, low-dimensional systems. There is a rich literature in high-dimensional variable selection and related graphical models [Bühlmann and van de Geer, 2011; Friedman *et al.*, 2008; Hans *et al.*, 2007; Maathuis *et al.*, 2010; Meinshausen and Bühlmann, 2006] which applies equally to the regression models described here. Many of the issues raised in this thesis remain relevant in the high-dimensional setting. Indeed, in practice even high-dimensional observations are likely to be incomplete, since it is not currently possible to measure all relevant chemical species. Therefore, inferred relationships between variables may be indirect. This may be acceptable for the purpose of predicting the outcome of biochemical interventions (e.g. inhibiting gene or protein nodes), but requires greater care in experimental design in order to ensure causal sufficiency (Section 1.4.2). Latent variable approaches are available [Beal *et al.*, 2005], but model selection can be challenging and remains an open area of research [Colombo *et al.*, 2012; Gao *et al.*, 2008; Knowles and Ghahramani, 2011]. Further work is required to better understand these issues in the context of inference for biological networks.

- *Causal reasoning and interventions.* There now exist a number of graphical formulations for causal reasoning, including Bayesian networks [Pearl, 2009], structural equation models [Peters, 2012] and chain event graphs [Smith and Anderson, 2008]. Each methodology is able to facilitate the principled analysis of data obtained under intervention, or to predict the effect of an intervention from observational data. Physical limitations mean that an idealised "step function" representation of interventions, such as receptor stimulation or administration of an inhibitor, is far from reality. In the latter case, treatment with a small molecule inhibitor of Akt kinase activity, for example, requires several hours to diffuse through out the cellular population. As a consequence, during pretreatment the cellular population has several hours to adjust to the new environment by altering gene expression profiles. Accounting for such non-local interventions and (real-valued) experimental data is potentially challenging, since intervention may change, in principle, the empirical form of any conditional distribution in the system. Here JNI and related approaches may prove useful, allowing for modest changes in global structural to result from intervention.

- *Statistical model selection.* It is well known that the *high-dimensional* regime, where the number P of predictors is allowed to grow with the number n of samples, poses challenges to classical statistical methodology, which focusses on consistent estimation as $n \to \infty$ with P fixed. Recent developments in high-dimensional statistics have led to the development of consistent estimators under restrictions of sparsity [Meinshausen and Bühlmann, 2006]. These procedures have been extended to settings where variables are highly correlated, such as Gaussian graphical models [Loh and Wainwright, 2012], and in settings where latent variables may be present [Colombo *et al.*, 2012]. However such techniques do not integrate specific knowledge about the data-generating system; in this respect these estimators may be sub-optimal against model-based inference. More generally, consistent estimation is currently far removed from the reality of inference in systems biology, with enormous model misspecification and often highly multi-modal likelihood surfaces which must be explored.

- *Bayesian computation.* The likelihood resulting from mechanistic models of cellular dynamics is typically intractable, in the sense that the normalising constant is not easily computed. This would preclude model selection using information criteria or Bayes factors, so it is an important problem to estimate the normalising constant for a given model. Unfortunately such estimation is extremely

challenging [Vyshemirsky and Girolami, 2008], though increasingly sophisticated approaches are being developed to address this bottleneck [Calderhead and Girolami, 2009].

- *Data integration.* It is typically the case that ancillary data are available on the biological samples under investigation. For example, in the case of breast cancer cell lines, data are routinely collected on gene expression profiles and expression of receptor proteins (Example 4). It has been shown that such data are highly predictive [Heiser *et al.*, 2011] of cellular dynamics, yet it is far from clear how to optimally exploit such data in inference. More generally, sample heterogeneity presents a challenge to statistical efficiency, since although cell lines from the same lineage share more commonalities than cell lines from different lineages, within diseases such as breast cancer there remains remarkable heterogeneity.

- *Application.* Whilst network inference offers much promise for our understanding of complex multivariate systems, it remains the case that estimation from real data is extremely challenging. Initial applications of network inference have focussed on exploiting the inferred network in order to constrain experimental design, and in this sense are scientifically valid. A smaller number of papers have drawn conclusions directly from the results of network inference algorithms. Chapter 2 of this thesis attempts to demonstrate significant shortcomings with the latter strategy. An important area for research will be to undertake a systematic analysis of network inference algorithms on large corpora of validation data, following e.g. [Maathuis *et al.*, 2010; Marbach *et al.*, 2012; Prill *et al.*, 2010].

- *Translation.* Whilst inference for biological systems is still an emerging research field, it may soon become important to consider issues surrounding the clinical translation of these statistical techniques. At the time of writing, there appears to be much potential for network analyses to contribute to research in oncology; if network heterogeneity proves to explain some of the heterogeneity observed in cancer, there will be important translational questions to address in the clinical application of network inference techniques. For example, if patient-specific signalling network topology is demonstrated to be predictive of response to a particular therapy, then an important question becomes how to achieve adequate network reconstruction in minimal time with minimal experimental and computational cost. At the moment this prospect seems several years away, but theoretical progress on understanding the properties of statistical network estimators could be achieved as of present.

- *Immediate extensions.* The work performed in this thesis has some immediate extensions: (i) The CheMA and JNI algorithms could easily be adapted to perform inference and prediction in other biological networks. In particular, CheMA easily extends to facilitate inference and prediction in gene regulation, where Michaelis-Menten formulations are widely used to model dynamics [Kholodenko, 2006]. (ii) We rooted JNI in a linear formulation of likelihood which allowed for rapid, interactive exploration of data. However JNI may apply directly to more sophisticated formulations of likelihood, including CheMA. Thus an immediate extension of this work could be to explore inference for protein signalling networks under a hierarchical model rooted in non-linear chemical kinetics.

Clearly there remain major challenges for statistical applications in cellular biology. At present, the extent to which multivariate statistical techniques will contribute in this field is unclear. Yet within oncology, there is reason to believe a systems approach to cellular dynamics will be essential to understand the complex multivariate properties of the disease. The fundamental limitations on experimental drug screens discussed in the Introduction, which result from combinatoric arguments, preclude a brute-force experimental approach to fundamental biological discovery. Coupled with the rapid emergence of large multivariate datasets, it seems likely that multivariate statistical tools will be essential both in the short-to-medium term and further into the future.

Appendix A

Supplemental Material for Chapter 2

A.1 Dynamical Systems

A.1.1 Model 1: Cantone *et al.*

$$x_1 = [Cbf1],\ x_2 = [Gal4],\ x_3 = [Swi5],\ x_4 = [Gal80],\ x_5 = [Ash1]$$

$$\begin{aligned}
\frac{dx_1}{dt} &= \alpha_1 + v_1\left(\frac{x_3^{h_1}(t-\tau)}{(k_1^{h_1} + x_3^{h_1}(t-\tau))\left(1+\frac{x_5^{h_2}}{k_2^{h_2}}\right)}\right) - d_1x_1 \\
\frac{dx_2}{dt} &= \alpha_2 + v_2\left(\frac{x_1^{h_3}}{k_3^{h_3} + x_1^{h_3}}\right) - d_2x_2 \\
\frac{dx_3}{dt} &= \alpha_3 + v_3\left(\frac{x_2^{h_4}}{k_4^{h_4} + x_2^{h_4}(1+\frac{x_4^4}{\gamma^4})}\right) - d_3x_3 \\
\frac{dx_4}{dt} &= \alpha_4 + v_4\left(\frac{x_3^{h_5}}{k_5^{h_5} + x_3^{h_5}}\right) - d_4x_4 \\
\frac{dx_5}{dt} &= \alpha_5 + v_5\left(\frac{x_3^{h_6}}{k_6^{h_6} + x_3^{h_6}}\right) - d_5x_5
\end{aligned}$$

Parameter values were taken from the "switch-on" experimental conditions; see Cantone *et al.* [2009].

A.1.2 Model 2: Swat *et al.*

$$\begin{aligned}
x_1 = [pRB],\ x_2 = [E2F1],\ x_3 = [CycD_i],\ x_4 = [CycD_a],\ x_5 = [AP-1] \\
x_6 = [pRB_p],\ x_7 = [pRB_{pp}],\ x_8 = [CycE_i],\ x_9 = [CycE_a]
\end{aligned}$$

$$
\begin{aligned}
\frac{dx_1}{dt} &= k_1 \frac{x_2}{k_{m1}+x_2}\frac{J_{11}}{J_{11}+x_1}\frac{J_{61}}{J_{61}+x_6} - k_{16}x_1x_4 + k_{61}x_6 - \phi_1 x_1 \\
\frac{dx_2}{dt} &= k_p + k_2 \frac{a^2+x_2^2}{k_{m2}^2+x_2^2}\frac{J_{12}}{J_{12}+x_1}\frac{J_{62}}{J_{62}+x_6} - \phi_2 x_2 \\
\frac{dx_3}{dt} &= k_3 x_5 + k_{23}x_2 \frac{J_{13}}{J_{13}+x_1}\frac{J_{63}}{J_{63}+x_6} + k_{43}x_4 - k_{34}x_3\frac{x_4}{k_{m4}+x_4} - \phi_3 x_3 \\
\frac{dx_4}{dt} &= k_{34}x_3\frac{x_4}{k_{m4}+x_4} - k_{43}x_4 - \phi_4 x_4 \\
\frac{dx_5}{dt} &= F_m + k_{25}x_2\frac{J_{15}}{J_{15}+x_1}\frac{J_{65}}{J_{65}+x_6} - \phi_5 x_5 \\
\frac{dx_6}{dt} &= k_{16}x_1x_4 - k_{61}x_6 - k_{67}x_6x_9 + k_{76}x_7 - \phi_6 x_6 \\
\frac{dx_7}{dt} &= k_{67}x_6x_9 - k_{76}x_7 - \phi_7 x_7 \\
\frac{dx_8}{dt} &= k_{28}x_2\frac{J_{18}}{J_{18}+x_6}\frac{J_{68}}{J_{68}+x_7} + k_{98}x_9 - k_{89}x_8\frac{x_9}{k_m 9 + x_9} - \phi_8 x_8 \\
\frac{dx_9}{dt} &= k_{89}x_8\frac{x_9}{k_{m9}+x_9} - k_{98}x_9 - \phi_9 x_9
\end{aligned}
$$

Parameter values were taken as in Swat *et al.* [2004].

A.2 Derivations

A.2.1 Deriving a Model in the Large Sample Limit

Suppose the true large sample process obeys $d\mathbf{X}^\infty/dt = \mathbf{F}(\mathbf{X}^\infty)$. Then a Taylor expansion of $\mathbf{X}^\infty$ about t gives (dropping the superscript ∞)

$$\mathbf{X}(t+\Delta) = \mathbf{X}(t) + \Delta\mathbf{F}(\mathbf{X}(t)) + \dots \tag{A.1}$$

so that when we account for measurement error $\mathbf{Y} = \mathbf{X} + \mathbf{w}$ we have

$$\mathbf{Y}(t+\Delta) - \mathbf{w}(t+\Delta) = \mathbf{Y}(t) - \mathbf{w}(t) + \Delta\mathbf{F}(\mathbf{Y}(t) - \mathbf{w}(t)) + \dots. \tag{A.2}$$

A Taylor expansion of $\mathbf{F}$ about $\mathbf{Y}(t)$ and a rearrangement gives

$$\frac{\mathbf{Y}(t+\Delta) - \mathbf{Y}(t)}{\Delta} - \mathbf{F}(\mathbf{Y}(t)) = \frac{\mathbf{w}(t+\Delta)}{\Delta} - \left[\frac{\mathbf{I}}{\Delta} + (D\mathbf{F})(\mathbf{Y}(t))\right]\mathbf{w}(t) + \dots \tag{A.3}$$

so that the variance

$$\mathbb{V}\left(\frac{\mathbf{Y}(t+\Delta) - \mathbf{Y}(t)}{\Delta} - \mathbf{F}(\mathbf{Y}(t))\right) = \frac{\mathbf{M}}{\Delta^2} + \left(\frac{\mathbf{I}}{\Delta} + D\mathbf{F}\right)\mathbf{M}\left(\frac{\mathbf{I}}{\Delta} + D\mathbf{F}\right)' + \dots. \tag{A.4}$$

A.2.2 Deriving a Model for Longitudinal Single Cell Measurements

An Euler-Maruyama approximation for single cell expression $\mathbf{X}$ gives

$$\mathbf{X}(t+\Delta) = \mathbf{X}(t) + \Delta\mathbf{f}(\mathbf{X}(t)) + \mathbf{g}(\mathbf{X}(t))\Delta\mathbf{B} + \dots \tag{A.5}$$

so that when we account for measurement error $\mathbf{Y} = \mathbf{X} + \mathbf{w}$ we have

$$\mathbf{Y}(t+\Delta) - \mathbf{w}(t+\Delta) = \mathbf{Y}(t) - \mathbf{w}(t) + \Delta\mathbf{f}(\mathbf{Y}(t) - \mathbf{w}(t)) + \mathbf{g}(\mathbf{Y}(t) - \mathbf{w}(t))\Delta\mathbf{B} + \dots. \tag{A.6}$$

Taking a diffusion $\mathbf{g}(\mathbf{X}) = \sigma\mathcal{D}(\mathbf{X})$, a Taylor expansion of $\mathbf{f}$ about $\mathbf{Y}(t)$ and a rearrangement gives

$$\begin{aligned}\frac{\mathbf{Y}(t+\Delta)-\mathbf{Y}(t)}{\Delta} - \mathbf{f}(\mathbf{Y}(t)) \\ = \frac{\mathbf{w}(t+\Delta)}{\Delta} - \quad \left[\tfrac{\mathbf{I}}{\Delta} + (D\mathbf{f})(\mathbf{Y}(t))\right]\mathbf{w}(t) + \sigma[\mathcal{D}(\mathbf{Y}(t)) - \mathcal{D}(\mathbf{w}(t))]\Delta\mathbf{B} + \dots\end{aligned} \tag{A.7}$$

so that

$$\mathbb{V}\left(\frac{\mathbf{Y}(t+\Delta)-\mathbf{Y}(t)}{\Delta} - \mathbf{f}(\mathbf{Y}(t))\right) = \frac{\mathbf{M}}{\Delta^2} + \left[\frac{\mathbf{I}}{\Delta} + D\mathbf{f}\right]\mathbf{M}\left[\frac{\mathbf{I}}{\Delta} + D\mathbf{f}\right]' + \frac{\mathbf{g}\mathbf{g}'}{\Delta} + \dots. \tag{A.8}$$

Notice that this variance is larger than the corresponding variance in Eqn. A.4, showing that a single cell data set contains less information than the corresponding dataset for an averaged process.

A.2.3 Approximating $h_{\mathbf{true}}$ for Cantone

From Eqn. A.4 a natural choice of variance function h_{true} is

$$h_{\text{true}}(\Delta)\mathcal{D}(\sigma_1^2, \dots, \sigma_P^2) \approx \mathbf{M}\Delta^{-2} + (\mathbf{I}\Delta^{-1} + D\mathbf{F})\mathbf{M}(\mathbf{I}\Delta^{-1} + D\mathbf{F})' \tag{A.9}$$

where the large sample process obeys $d\mathbf{X}^{\infty}/dt = \mathbf{F}$. This can be made precise under a given matrix norm:

$$h_{\text{true}}(\Delta) \approx \frac{\left\|\mathbf{M}\Delta^{-2} + (\mathbf{I}\Delta^{-1} + D\mathbf{F})\mathbf{M}(\mathbf{I}\Delta^{-1} + D\mathbf{F})'\right\|}{\|\mathcal{D}(\sigma_1^2, \dots, \sigma_P^2)\|} \tag{A.10}$$

Under the (strong) assumption that $\mathbf{F} \equiv \mathbf{F}(\mathbf{X}^{\infty})$ we have that $D\mathbf{F}|_{\mathbf{x}^{\infty}=\mathbf{0}} = D\mathbf{f}|_{\mathbf{x}^{k}=\mathbf{0}}$ since $\mathbf{X}^{\infty} = \mathbf{0}$ if and only if almost all the $\mathbf{X}^k = \mathbf{0}$. So it suffices to find the Jacobian of the single cell drift $\mathbf{f}$.

We seek an approximation to h_{true} for Cantone *et al*, so for simplicity ignore the delay term in the regulation of Cbf1 by Swi5. Then

$$D\mathbf{f}|_{\mathbf{x}=\mathbf{0}} = \begin{bmatrix} -d_1 & 0 & v_1/k_1 & 0 & 0 \\ v_2/k_3 & -d_2 & 0 & 0 & 0 \\ 0 & 0 & -d_3 & 0 & 0 \\ 0 & 0 & v_k/k_5 & -d_4 & 0 \\ 0 & 0 & v_5/k_6 & 0 & -d_5 \end{bmatrix}. \tag{A.11}$$

Substituting Eqn. A.11 for $D\mathbf{F}$ in Eqn. A.10 provides an approximation to h_{true}.

Appendix B

Supplemental Material for Chapter 3

B.1 Truncated Gaussian Distributions

We used truncated normal distributions as priors for kinetics parameters, as described in Chapter 3. Here, we define truncated normal distributions and discuss how we sampled from them.

B.1.1 Definition

A random variable $\boldsymbol{Y} \in \mathbb{R}^n$ has a truncated multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (taken from the untruncated distribution), denoted $\boldsymbol{Y} \sim \mathcal{N}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\boldsymbol{Y}$ has probability density function

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) \propto \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right)\mathbb{I}(\boldsymbol{y} \geqslant \boldsymbol{0}). \tag{B.1}$$

The notation $\boldsymbol{y} \geqslant \boldsymbol{0}$ is taken to mean that $y_i \geqslant 0$ for all $i = 1, \ldots, n$. The density $p_{\boldsymbol{Y}}$ is related to the standard normal probability density ϕ via $p_{\boldsymbol{Y}}(\boldsymbol{y}) = C^{-1}\phi(\boldsymbol{y})\mathbb{I}(\boldsymbol{y} \geqslant \boldsymbol{0})$, so evaluation of $p_{\boldsymbol{Y}}$ requires

$$C = \int_{\boldsymbol{y}\geqslant\boldsymbol{0}} \phi(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{y} = \int_{\boldsymbol{z}\leqslant\boldsymbol{0}} \phi(\boldsymbol{z}; -\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{z} := \Phi(\boldsymbol{0}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{B.2}$$

where Φ is the normal cumulative distribution function.

B.1.2 Sampling

In general, sampling efficiently from truncated multivariate normal distributions is challenging. For example a rejection sampler based on an unconditioned normal density becomes inefficient when the measure of the target density's support is small. One approach is to construct a Gibbs sampler based on Eqn. B.1 (see Rodriguez-Yam *et al.* [2002, 2004]) but this is considerable effort for obtaining random samples for our purposes. However if the target distribution is non-degenerate (i.e. $\boldsymbol{\Sigma}$ is positive definite) then there exists a bijective mapping onto a product of standard truncated normal densities, which we exploit for sampling. Specifically, if $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then we can write $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{A}$ arises from the Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Positive definiteness ensures that the Cholesky decomposition exists and is unique. Moreover $\boldsymbol{A}$ is invertible, being lower triangular with strictly positive diagonal entries. Since $\boldsymbol{Y} \geqslant \boldsymbol{0}$ if and only if $\boldsymbol{Z} \geqslant -\boldsymbol{A}^{-1}\boldsymbol{\mu}$, we have the basis for efficient sampling (Algorithm 3). In the case that the target distribution approximates a point mass (this arises from conditioning on a rare event in the tails of a normal distribution), the algorithm uses numerical regularization.

B.2 ODE model of MAPK signalling for simulation

B.2.1 Dynamical system

The *in silico* model used for our investigation was published by Xu *et al.* [2010], with the ODE formulation $\dot{\boldsymbol{X}} = \boldsymbol{f}_G(\boldsymbol{X}; \boldsymbol{\theta})$ reproduced in Fig. B.1. Parameter values $\boldsymbol{\theta}$ were chosen an in order to ensure signalling

Algorithm 3 Efficient sampling from the (non-degenerate) truncated multivariate normal $\boldsymbol{Y} \sim \mathcal{N}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with numerical regularization. Here U is the uniform distribution, p is the dimension of $\boldsymbol{Y}$ and ϵ is taken to be machine precision.

$\boldsymbol{A} \leftarrow \text{Cholesky}(\boldsymbol{\Sigma})$
$\boldsymbol{b} \leftarrow -\boldsymbol{A}^{-1}\boldsymbol{\mu}$
for $i = 1$ **to** p **do**
  $u \sim U[\Phi(b_i), 1]$
  if $u > 1 - \epsilon$ **then**
    $z_i \leftarrow b_i$
  else
    $z_i \leftarrow \Phi^{-1}(u)$
  end if
end for
$\boldsymbol{y} \leftarrow \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{z}$

was identifiable in principle from the dynamics.

B.2.2 Simulation regimes

In order to accurately assess the impact of sample size upon performance, it is important that the amount of information in the simulated data increases with n. Given that the informative range of the dynamics is determined by the choice of parameters (approximately $0 \leqslant t \leqslant 20$), adding noise to deterministic data will not satisfy the above requirement, since additional data will merely replicate existing information. We therefore introduced intrinsic stochasticity into the data generating process, interpreting the Xu *et al.* model as the drift in a stochastic differential equation:

$$\boldsymbol{X}(0) = \boldsymbol{x}_0 \tag{B.3}$$
$$d\boldsymbol{X} = \boldsymbol{f}_G(\boldsymbol{X}, \boldsymbol{\theta})dt + \sigma d\boldsymbol{B} \tag{B.4}$$

where σ controls the magnitude of the stochastic fluctuations. Initial state $\boldsymbol{x}_0$ was drawn from the truncated standard normal distribution.

To generate time courses, we simulated solutions $\boldsymbol{X}(t)$ of this SDE for times $0 \leqslant t \leqslant 20$ and then selected $[n/4]$ evenly spaced samples; four such time courses constituted a dataset. Data regimes were characterized by total observation sample size $n \in \{25, 50, 100, 200\}$ and noise magnitude $\sigma \in \{0, 0.05, 0.1, 0.15, 0.2\}$. A time course with 100 evenly spaced samples is shown in Fig. B.2. Simulated datasets differ in both the initial state $\boldsymbol{x}_0$ and the realization of the Brownian motion $\boldsymbol{B}$.

B.2.3 Details of assessment

Of the 25 state variables, 3 denote drug compounds; these were not considered for the purpose of network inference. The remaining 22 variables denote the active and inactive forms of 11 signalling proteins; Raf1, EGFR, SOS, Ras, Rap1, PKA, MEK, ERK, EPAC, BRaf and C3G. Network inference was therefore performed for these 11 proteins, in each of the experimental regimes, using each available method. Disregarding self-edges made a total of $(2^{10})^{10} \approx 10^{29}$ possible networks.

B.3 Implementation

All of the methods used in Chapter 3 have a number of user-set parameters or configurations. We used default configurations for each method, as described below. Experiments involving LASSO, DBN and TVDBN (below) were carried out by Frank Dondelinger at the Netherlands Cancer Institute, Amsterdam.

B.3.1 LASSO

We used the R package `glmnet` Friedman *et al.* [2000] to train an l1-regularised linear model (known as LASSO, for Least Absolute Shrinkage and Selection Operator) on the input data. The optimal setting of the regularisation parameter λ was determined for each dataset separately using cross-validation. For

$$\dot{unboundEGFR} = -p_5 \cdot EGF \cdot unboundEGFR + p_6 \cdot boundEGFR$$

$$\dot{removed\text{Raf-1}} = \frac{p_{23} \cdot PKA \cdot \text{Raf-1}}{p_{24} + \text{Raf-1}}$$

$$\dot{removedSOS} = \frac{p_1 \cdot ERKPP \cdot inactiveSOS}{p_2 + inactiveSOS} + \frac{p_1 \cdot ERKPP \cdot activeSOS}{p_2 + activeSOS}$$

$$\dot{inactiveSOS} = -\frac{p_3 \cdot boundEGFR \cdot inactiveSOS}{p_4 + inactiveSOS} + \frac{p_8 \cdot activeSOS}{p_7 + activeSOS} - \frac{p_1 \cdot ERKPP \cdot inactiveSOS}{p_2 + inactiveSOS}$$

$$\dot{inactiveRas} = -\frac{p_9 \cdot activeSOS \cdot inactiveRas}{p_{10} + inactiveRas} + \frac{p_{11} \cdot Gap \cdot activeRas}{p_{12} + activeRas}$$

$$\dot{inactiveRap1} = -\frac{p_{37} \cdot EPAC \cdot inactiveRap1}{p_{38} + inactiveRap1} + \frac{p_{39} \cdot Gap \cdot activeRap1}{p_{40} + activeRap1} - \frac{p_{50} \cdot activeC3G \cdot inactiveRap1}{p_{51} + inactiveRap1}$$

$$\dot{inactivePKA} = -\frac{p_{25} \cdot PKAA \cdot inactivePKA}{p_{26} + inactivePKA} - \frac{p_{27} \cdot Cilostamide \cdot inactivePKA}{p_{28} + inactivePKA} + \frac{p_{30} \cdot PKA}{p_{29} + PKA}$$

$$\dot{inactiveEPAC} = -\frac{p_{31} \cdot EPACA \cdot inactiveEPAC}{p_{32} + inactiveEPAC} - \frac{p_{33} \cdot Cilostamide \cdot inactiveEPAC}{p_{34} + inactiveEPAC} + \frac{p_{36} \cdot EPAC}{p_{35} + EPAC}$$

$$\dot{\text{Raf-1}PP} = \frac{p_{13} \cdot activeRas \cdot \text{Raf-1}}{p_{14} + \text{Raf-1}} - \frac{p_{16} \cdot \text{Raf-1}PP}{p_{15} + \text{Raf-1}PP}$$

$$\dot{\text{Raf-1}} = -\frac{p_{13} \cdot activeRas \cdot \text{Raf-1}}{p_{14} + \text{Raf-1}} + \frac{p_{16} \cdot \text{Raf-1}PP}{p_{15} + \text{Raf-1}PP}$$

$$\dot{boundEGFR} = p_5 \cdot EGF \cdot unboundEGFR - p_6 \cdot boundEGFR$$

$$\dot{activeSOS} = \frac{p_3 \cdot boundEGFR \cdot inactiveSOS}{p_4 + inactiveSOS} - \frac{p_8 \cdot activeSOS}{p_7 + activeSOS} - \frac{p_1 \cdot ERKPP \cdot activeSOS}{p_2 + activeSOS}$$

$$\dot{activeRas} = \frac{p_9 \cdot activeSOS \cdot inactiveRas}{p_{10} + inactiveRas} - \frac{p_{11} \cdot Gap \cdot activeRas}{p_{12} + activeRas}$$

$$\dot{activeRap1} = \frac{p_{37} \cdot EPAC \cdot inactiveRap1}{p_{38} + inactiveRap1} - \frac{p_{39} \cdot Gap \cdot activeRap1}{p_{40} + activeRap1} + \frac{p_{50} \cdot activeC3G \cdot inactiveRap1}{p_{51} + inactiveRap1}$$

$$\dot{PKA} = \frac{p_{25} \cdot PKAA \cdot inactivePKA}{p_{26} + inactivePKA} + \frac{p_{27} \cdot Cilostamide \cdot inactivePKA}{p_{28} + inactivePKA} - \frac{p_{30} \cdot PKA}{p_{29} + PKA}$$

$$\dot{MEKPP} = \frac{p_{17} \cdot \text{Raf-1}PP \cdot MEK}{p_{18} + MEK} - \frac{p_{20} \cdot MEKPP}{p_{19} + MEKPP} + \frac{p_{45} \cdot \text{B-Raf}PP \cdot MEK}{p_{46} + MEK}$$

$$\dot{MEK} = -\frac{p_{17} \cdot \text{Raf-1}PP \cdot MEK}{p_{18} + MEK} + \frac{p_{20} \cdot MEKPP}{p_{19} + MEKPP} - \frac{p_{45} \cdot \text{B-Raf}PP \cdot MEK}{p_{46} + MEK}$$

$$\dot{ERKPP} = \frac{p_{21} \cdot MEKPP \cdot ERK}{p_{22} + ERK} - \frac{p_{55} \cdot ERKPP}{p_{54} + ERKPP}$$

$$\dot{ERK} = -\frac{p_{21} \cdot MEKPP \cdot ERK}{p_{22} + ERK} + \frac{p_{55} \cdot ERKPP}{p_{54} + ERKPP}$$

$$\dot{EPAC} = \frac{p_{31} \cdot EPACA \cdot inactiveEPAC}{p_{32} + inactiveEPAC} + \frac{p_{33} \cdot Cilostamide \cdot inactiveEPAC}{p_{34} + inactiveEPAC} - \frac{p_{36} \cdot EPAC}{p_{35} + EPAC}$$

$$\dot{EGF} = -p_5 \cdot EGF \cdot unboundEGFR + p_6 \cdot boundEGFR$$

$$\dot{\text{B-Raf}PP} = \frac{p_{41} \cdot activeRap1 \cdot \text{B-Raf}}{p_{42} + \text{B-Raf}} - \frac{p_{44} \cdot \text{B-Raf}PP}{p_{43} + \text{B-Raf}PP} + \frac{p_{52} \cdot activeRas \cdot \text{B-Raf}}{p_{53} + \text{B-Raf}}$$

$$\dot{\text{B-Raf}} = -\frac{p_{41} \cdot activeRap1 \cdot \text{B-Raf}}{p_{42} + \text{B-Raf}} + \frac{p_{44} \cdot \text{B-Raf}PP}{p_{43} + \text{B-Raf}PP} - \frac{p_{52} \cdot activeRas \cdot \text{B-Raf}}{p_{53} + \text{B-Raf}}$$

$$\dot{activeC3G} = \frac{p_{47} \cdot boundEGFR \cdot inactiveC3G}{p_{48} + inactiveC3G} - p_{49} \cdot activeC3G$$

$$\dot{inactiveC3G} = -\frac{p_{47} \cdot boundEGFR \cdot inactiveC3G}{p_{48} + inactiveC3G} + p_{49} \cdot activeC3G$$

Figure B.1: *In silico* ODE model of the EGFR/ERK signalling pathway due to Xu *et al.* [2010].

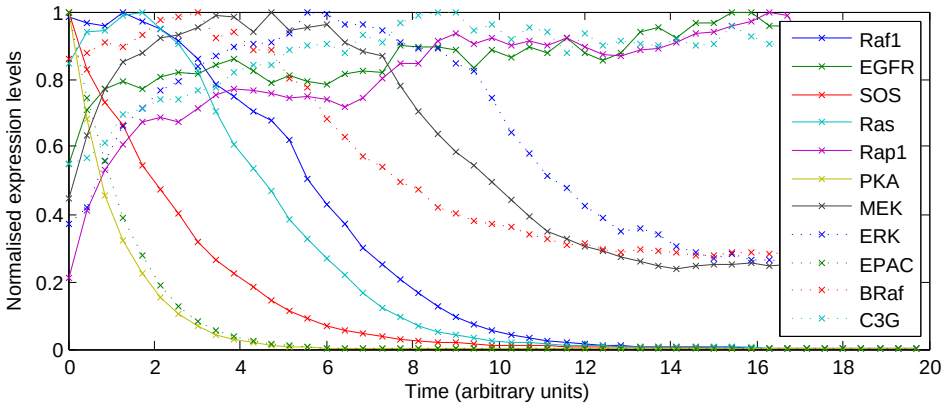


Figure B.2: Typical simulated time course from the ODE model of Xu *et al.* [2010]. [Initial conditions were drawn from a truncated standard Gaussian; four such time courses constitute a dataset. Here 100 evenly spaced samples are shown with intrinsic noise of magnitude $\sigma = 0.05$. Species expression is normalized to unit maximum to improve presentation.]

each node i in the network, we learn a regression model for observations $Y_i^*(t)$ with respect to the remaining nodes $Y_j^*(t-1)$ $(j \neq i)$ at time $t-1$. LASSO automatically sets the regression coefficients of some nodes to zero. We used the absolute values of the regression coefficients to give an indication of the strength of each edge in the network. We used the default settings of the `glmnet`, and the input data for each regression were standardised to mean 0 and variance 1.

B.3.2 TSNI

Time Series Network Inference (TSNI) Bansal and di Bernardo [2007] was run according to the recommended settings provided at `http://dibernardo.tigem.it/wiki/index.php/Time_Series_Network_Identification_TSNI-integral`. Since TSNI only accepts single time series, the resulting weighted adjacency matrices corresponding to separate time courses were subsequently averaged to obtain a single network estimate.

B.3.3 DBN

To learn dynamic Bayesian networks (DBNs) from the data, we used the model described in Hill *et al.* [2012a], which also corresponds to the model in Dondelinger *et al.* [2010] when one imposes the restriction of not allowing change-points. For obtaining the results in this Chapter, we therefore used the R software package `EDISON` that implements the model in Dondelinger *et al.* [2010] and samples from it via reversible-jump MCMC. We fixed the change-point settings so that no change-points would be inferred during the network inference. The sampled networks were evaluated based on the marginal posterior probability of each edge. We used the default settings of the software package, except for the maximum number of iterations, which was set to 1e6. The data was standardised to mean 0 and variance 1. Note that alternative implementations of linear DBNs may enjoy computational advantages Hill *et al.* [2012a].

B.3.4 TVDBN

For inferring time-varying DBNs, we again used the R software package `EDISON` that implements the model in Dondelinger *et al.* [2010]. In this case, change-points were allowed to be inferred during the reversible-jump MCMC, which potentially allows for modelling non-linear effects. The sampled networks were evaluated based on the marginal posterior probability of each edge. We used the default settings of the software package, except for the maximum number of iterations, which was set to 1e6. The data was standardised to mean 0 and variance 1. We also used information sharing with a soft coupling of nodes, as described in Dondelinger *et al.* [2010], to regularise the number of changes at each change-point.

Method:	CheMA	LASSO	TSNI	DBN	TVDBN	GP
Time (secs):	2×10^4	1	1	4×10^3	4×10^3	3×10^2

Table B.1: Computational times (approximate) for inference of the Xu *et al.* network. [Implementational details for the various methods are contained in Section B.3. Note that certain methods may enjoy more favourable computational implementations, e.g. Hill *et al.* [2012a] for linear DBNs.]

B.3.5 GP

GP [Äijö and Lähdesmäki, 2009] was run in MATLAB R2012b using code generously supplied by Tarmo Äijö. On noise-free data ($\sigma = 0$) this code could encounter numerical loss of positive-definiteness so, when required, covariance matrices were regularized using Tikhonov regularization prior to Cholesky decomposition. GP was then run using the following settings; optimization iterations = 50, no delay terms, zero order model = used, maximum in-degree = 2, prior covariance = $0.01 \times \mathbf{I}$, prior mean = $\mathbf{0}$.

B.3.6 Computational times

Table B.1 contains the approximate computational time requirements of the competing methodologies. It may be seen that the chemical kinetic approach is considerably more demanding compared with competing approaches, requiring at least 5 times more computation. Note that these time requirements are empirical and implementation-dependent; a formal time complexity analysis of the algorithms is beyond the scope of this Chapter.

For illustration of computation for larger networks, we ran CheMA using data obtained on breast cancer cell line AU565 (see Section B.6.1) based on 27 phosphoproteins (network not shown, since its interpretation and assessment is beyond the scope of this thesis). This required over 12 hours of computational time. This illustrates that in principle CheMA could be used for larger networks. However, there were fewer samples ($n = 24$) than protein species in the dataset, and only 2 targeted interventions, so caution would need to be exercised in interpreting the results.

B.4 *In silico* results

B.5 Prediction of signalling response

For the prediction problem we are given training data $\boldsymbol{y}$ and an initial condition $\boldsymbol{x}_0$, from which the goal is to predict the entire time course $\boldsymbol{x}(t)$. Below we describe how these data were generated and how training data were used. The quality of a prediction was assessed by mean square error (MSE) with respect to the test data. All protein species were normalized by their maximum value in the training data $\boldsymbol{y}$. The network inference algorithms used in Section B.4 have not been modified for prediction; we therefore considered simple stationary and linear benchmark predictors (described in Chapter 3).

B.5.1 Data generation

Training data $\mathcal{D}$ were generated as described in Section B.2. For test data, one randomly chosen protein X_i was selected as the target of an intervention. One time course $\boldsymbol{x}(t)$ was generated under this intervention by forcing terms X_i^* corresponding to the target(s) of intervention to equal zero in the drift $\boldsymbol{f}_G$ of Eqn. B.4.

B.5.2 Stationary benchmark

The benchmark mean square error was computed by predicting $\boldsymbol{x}(t) = \boldsymbol{x}_0$ for all t.

B.5.3 CheMA

Our approach returns samples from the joint posterior distribution $p(G, \boldsymbol{\theta}|\mathcal{D})$ over reaction graphs G and parameters $\boldsymbol{\theta}$. In order to facilitate prediction of $\boldsymbol{x}(t)$, we perform model-averaging as described in Algorithm 4. For the experiments reported in Chapter 3 we used $I = 1,000$ samples to construct an averaged prediction. Note that, since we do not model genetic variation, prediction is conditional upon

the noisy measurements of unphosphorylated protein expression in $\boldsymbol{x}$; linear interpolation of noisy data is used to approximate unphosphorylated protein concentrations at any given time.

Algorithm 4 CheMA prediction

for $i = 1$ **to** I **do**
  $G^{(i)} \sim p(G|\mathcal{D})$
  $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|G^{(i)}, \mathcal{D})$
  Numerically solve the ODE $\dot{\boldsymbol{X}} = \boldsymbol{f}_{G^{(i)}}(\boldsymbol{X}, \boldsymbol{\theta}^{(i)})$ from the initial condition $\boldsymbol{X}(0) = \boldsymbol{x}_0$. Denote the solution by $\boldsymbol{X}^{(i)}$.
end for
Predict $\boldsymbol{x}(t) \approx \frac{1}{I}\sum_{i=1}^{I} \boldsymbol{X}^{(i)}(t)$.

B.5.4 Linear kinetics

For an unbiased assessment of the importance of non-linearity in inference, the same approach to prediction was employed based on the linear model $f_{G,i}(\boldsymbol{X}, \boldsymbol{\theta}) = \beta_{0,i} + \sum_{E \in \mathcal{E}_i} \beta_{E,i} X_E^*$ where, following Hill *et al.* [2012a], the parameters $\boldsymbol{\beta}_i$ and σ_i for a given target i are assigned (untruncated) Zellner prior distributions with zero mean. Models G involving kinase inhibition were excluded from inference (inhibitory effects are accommodated by allowing coefficients to become negative). We believe this to be the closest (reasonable) linear approximation to the chemical kinetic framework described above.

B.6 *In vitro* results

B.6.1 Experimental Data

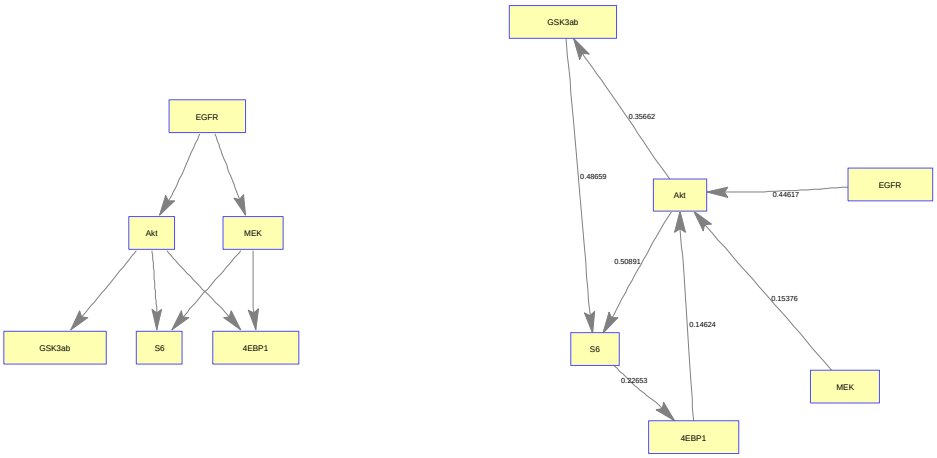
This thesis exploits experimental data derived from breast cancer cell lines; in particular we consider time course RPPA data on protein phosphorylation. Below we reproduce the experimental protocol as described by collaborators (Gray Lab, Knight Cancer Center, OHSU, Portland, OR, USA).

B.6.1.1 RPPA Data

Time course data on 15 breast cancer cell lines were obtained. For selected species, total protein and/or phosphoprotein levels were obtained. Cells were plated into 10 cm^2 dishes at a density of $1 - 2 \times 10^6$ cells. After 24 hours, cells were treated with 250 nM Lapatinib or 250 nM AKTi (GSK690693). For treatment with both inhibitors, 250 nM of Lapatinib and 250 nM of AKTi were used. DMSO served as a control. Cells were grown in full serum and harvested in RPPA lysis buffer at 30 min, 1h, 2h, 4h, 8h, 24h, 48h, and 72h post-treatment. Cell lysates were quantitated, diluted, arrayed, and probed as described previously Tibes *et al.* [2006]. Imaging and quantitation of signal intensity was done as described previously Tibes *et al.* [2006]. Data will be made available as Korkola *et al.* [2013].

B.6.2 *In Vitro results*

From literature we obtained a canonical protein signalling network (Fig. B.3(a)). Many of the networks inferred by CheMA shared topology with the literature network (Fig. B.3(b)). However it is not possible to validate inferred line-specific topology without extensive biochemistry. We therefore focused on the predictive power of CheMA, comparing this to the predictive power afforded by the literature network coupled with kinetic equations as described in Chapter 3. RPPA experimental protocol is described in Section B.6.1. Pre-treatment allowed for protein phosphorylation levels to respond to kinase inhibition treatment. In this way, the initial time point contains considerable information concerning the effect of treatment. The particular protein species analyzed were 4EBP1(pT37), AKT(pS473), EGFR(pY1173), GSK3ab(pS21), MEK1/2(pS217), S6(pS240).



(a) (b)

Figure B.3: (a) Protein signalling network derived from literature. (b) Inferred topology for cell line HCC 70. [Edge weights correspond to posterior probabilities. Only the most probable edges are displayed.]

Appendix C

Supplemental Material for Chapter 4

C.1 Propagation and the Sum-Product Lemma

Efficient inference in Bayesian networks relies on the *belief propagation* algorithm [Pearl, 1982] and related extensions [Kschischang *et al.*, 2001]. The *sum-product* lemma, which forms the basis for several exact inference procedures in graphical models, can be expressed in its most basic form as follows:

Lemma 1 (The sum-product lemma)**.** *For a finite set of functionals $f_i : \mathbb{X}_i \to \mathbb{R}$ on finite domains $\mathbb{X}_i$ indexed by $1 \leqslant i \leqslant I$ we have*

$$\sum_{x_1 \in \mathbb{X}_1, \ldots, x_I \in \mathbb{X}_I} \prod_{i=1}^{I} f_i(x_i) = \prod_{i=1}^{I} \sum_{x_i \in \mathbb{X}_i} f_i(x_i). \tag{C.1}$$

The proof is straight forward (induction on I) and can be found in e.g. Kschischang et al. [2001].

The sum-product lemma is typically used to reduce algorithmic complexity, replacing the $\mathcal{O}(|\mathbb{X}_1| \times \cdots \times |\mathbb{X}_I| \times I)$ expression on the left hand side by the $\mathcal{O}(|\mathbb{X}_1| + \cdots + |\mathbb{X}_I|)$ expression on the right hand side.

C.2 Additional Simulation Protocol

The ODE model of ERK signalling proposed by Xu *et al.* [2010] is based on a 25-dimensional state vector $\boldsymbol{x}$ and 56 parameters $p_1, \ldots, p_{56}$ (reproduced here for convenience in Fig. B.1). The ODEs define a protein signalling network via $i \in N_p$ if and only if variable i appears on the right hand side of the ODE describing the rate of change of variable p. Network heterogeneity was simulated as described in Chapter 4. In order to simulate parameter heterogeneity between individuals, we independently sampled parameters $p_i^j \sim \mathcal{N}(0, 1/4)$ and then made sure these were positive by taking the absolute value $p_i^j \mapsto |p_i^j|$. Initial conditions for each time course experiment were independently sampled as $(\boldsymbol{x}_0)_i \sim U(0, 1)$.

C.3 RPPA Data and Ancillary Information

The RPPA data described previously (Sec. B.6.1) was analysed. Specifically, Chapter 4 considered the phosphoforms listed in Table C.1.

In this Section we briefly describe how ancillary data were accounted for during analysis. Histological data on cell lines were obtained from previously published literature [Neve *et al.*, 2006]. Mutational status of known cancer genes were extracted from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [Forbes *et al.*, 2011]. Additional background on signalling processes may be found in Section 1.1.2.

Receptor tyrosine kinases (RTKs) are the high-affinity cell surface receptors for many external stimuli. Our data contain three RTKs, namely EGFR, HER2 (both from the ErbB family) and IRS1. Upon ligand binding, EGFR forms homo- or hetero-dimers with ErbB family members, resulting in phosphorylation and activation of intracellular kinase domains. Similarly, ligand binding of IRS1 leads to phosphorylation

Protein	Phosphoform
4EBP1	T37
AKT	S473
BAD	S112
c-Myc	T58
EGFR	Y1173
ELK1	S383
ER	–
FOXO3a	S318
GSK3ab	S21
HER2	–
IRS1	S307
MAPK	T202
MEK12	S217
p38	T180
p53	–
PR	–
S6	S240

Table C.1: RPPA data; measured proteins. ["–" denotes total protein. For example, phosphoform "S473" denotes phosphorylation on serine residue 473.]

and activation. Once activated, RTKs are able to initiate a cascade of phosphorylation events culminating in a ligand-specific cellular response. HER2 is frequently over expressed in breast cancer ("HER2+"), leading to a HER2 dependence which renders the ErbB signalling pathway an attractive therapeutic target. In our data 5 out of 6 cell lines were HER2+ (Table C.2). We employed a network prior over $N^j|N$ which did not penalize edges emanating from either of the ErbB family members EGFR or HER2, for HER2+ cell lines.

PI3Ks are key intra-cellular components mediating RTK signalling. Class IA PI3Ks are heterodimers, composed of a catalytic subunit (p110) and an adaptor/regulatory subunit (p85), such that the PI3Ks are activated by RTKs' interaction with p85. There are three variants of the p110 catalytic subunit (p110α, p110β, p110δ), expressed by genes PIK3CA, PIK3CB and PIK3CD respectively. PIK3CA is located in the chromosome 3q26, a region that is frequently amplified in several human cancers [Velculescu *et al.*, 2004], including breast cancers [Bachman *et al.*, 2004]. The resulting gain-of-function for this PI3K can lead to constitutive activation of the entire PI3K/Akt pathway [Samuels *et al.*, 2004]. In our data 2 out of 6 cell lines harbor activating mutations in the PIK3CA gene. The network prior for $N^j|N$ did not permit edges entering Akt in PIK3CA mutant lines, reflecting the fact that Akt is no longer regulated at the receptor level.

Our panel also contained BRAF, KRAS and TP53 mutations which may affect the kinase activity of their protein products to a lesser extent (Table C.2). However, since this would alter the parameter prior $p(\boldsymbol{\theta}^j|N^j)$, we chose not to integrate this ancillary information in this Chapter (though this may merit further research). The remaining mutations were considered not to impact heavily upon the interaction between observed species.

Cell Line	Ba/Lu	HER2	ER	PR	Gene	A.A. Mut.	Info.
AU 565	L	+	-	-	CDH1	H1047R	Kinase domain
					TP53	R175H	DNA binding domain; no effect [Petitjean *et al.*, 2007]
BT 474	L	+	+	+	PIK3CA	K111N	p85-binding domain
					TP53	E285K	DNA binding domain; no effect [Petitjean *et al.*, 2007]
HCC 1954	B	+	-	-	PIK3CA	H1047R	Kinase domain
					TP53	Y163C	DNA binding domain; no effect [Petitjean *et al.*, 2007]
MDA MB 231	B	−	-	-	BRAF	G464V	G-loop domain; double kinase activity [Davies *et al.*, 2002]
					CDKN2A	gene deletion	
					CDKN2a(p14)	gene deletion	
					KRAS	G13D	Increase kinase activity [Hollestelle *et al.*, 2007]
					NF2	E231*	
					TP53	R280K	DNA binding domain; no effect [Petitjean *et al.*, 2007]
SKBR 3	L	+	-	-			
SUM 225CWN	B	+	-	-			

Table C.2: Breast cancer cell lines; ancillary information. ["+/−" indicates receptor over/under expression; "B/L" indicates basal/luminal gene expression profile; "A.A. mutation" gives the precise location of mutated amino acids; "Domain" indicates local protein structure. Data sourced from Forbes *et al.* [2011]; Neve *et al.* [2006].]

Bibliography

Äijö, T., Lähdesmäki, H. (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics* **25**(22):2937-44.

Aliferis *et al.* (2010) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification, Part I: Algorithms and Empirical Evaluation. *J. Mach. Learn. Res.* **11**:171-234.

Alon, U. (2007) An Introduction to Systems Biology: Design Principles of Biological Circuits. *Mathematical and Computational Biology Series* **10**, Chapman & Hall/CRC, London.

Altay, G., Emmert-Streib, F. (2010) Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics* **26**(14):1738-1744.

Ando, T., Skolnick, J. (2010) Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proc. Natl. Acad. Sci.* **107**(43):18457-18462.

Arya *et al.* (2012) Recapitulating tumour microenvironment in chitosangelatin three-dimensional scaffolds: an improved in vitro tumour model. *J. R. Soc. Interface* **9**(77):3288-3302..

ATCC (2013) http://www.lgcstandards-atcc.org/, accessed 13/02/2013.

Avraham, R., Yarden, Y. (2011) Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Bio.* **12**(2):104-117.

Bachman *et al.* (2004) The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol. Ther.* **3**(8):772-775.

Bahia *et al.* (2002) Karyotypic variation between independently cultured strains of the cell line MCF-7 identified by multicolour fluorescence in situ hybridization. *Int. J. Oncol.* **20**:489-494.

Baker, M. (2010) Cellular imaging: Taking a long, hard look. *Nature* **466**(26):11371140.

Bandyopadhyay *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science Signaling* **330**(6009):1385.

Bansal, M., di Bernardo, D. (2007) Inference of gene networks from temporal gene expression profiles. *IET Syst. Biol.* **5**:306-312.

Bansal, M., Della Gatta, G., di Bernardo, D. (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* **22**(7):815-822.

Bansal *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Sys. Bio.* **3**(78).

Barbieri, M.M., Berger, J.O. (2004) Optimal predictive model selection. *Ann. Stat.* **32**(3):870-897.

Barclay, L.M., Hutton, J.L., Smith, J.Q. (2013) Refining a Bayesian Network with a Chain Event Graph. *Int. J. Approx. Reason.*, to appear.

Barretina *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**(7391):603-607.

Baumbach *et al.* (2009) Reliable transfer of gene regulatory networks between taxonomically related organisms. *BMC Bioinformatics* **3**:8.

Beal *et al.* (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* **21**(3):349-356.

Beasley *et al.* (2012) A Phase I Multi-Institutional Study of Systemic Sorafenib in Conjunction with Regional Melphalan for In-Transit Melanoma of the Extremity. *Ann. Surg. Oncol.* **19**(12):3896.

Bender *et al* (2010) Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics* **26**(ECCB 2010):i596-i602.

Bolstad, A., Van Veen, B., Nowak, R. (2011) Causal Network Inference via Group Sparse Regularization. *IEEE T. Signal Proces.* **59**(6):2628-2641.

Bonneau, R. (2008) Learning biological networks: from modules to dynamics. *Nat. Chem. Bio.* **4**:658-664.

Boyd *et al.* (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.

Breastcancer.org (2012) Genetics. Breastcancer.org, accessed 30-03-2013.
Bühlmann, P., van de Geer, S.A. (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Berlin Heidelberg.
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.
Burdall *et al.* (2003) Breast cancer cell lines: friend or foe? *Breast. Cancer. Res.* **5**:89-95.
Burnham, K.P., Anderson, D.R. (2002) Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach. Springer, New York.
Cabrera *et al.* (2006) Identity tests: determination of cell line cross-contamination. *Cytotechnology* **51**(2):4550.
Calderhead, B., Girolami, M. (2009) Estimating bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data An.* **53**:4028-4045.
Calderhead, B., Girolami, M., Lawrence, N. (2009) Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes. *Adv. Neur. In.* **21**:217-224.
Calderhead, B., Girolami, M. (2011) Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *J. Roy. Soc. Interface* **1**(6):821-835.
Camacho, D.M., Collins, J.J. (2009) Systems biology strikes gold. *Cell* **137**(1):24-6.
Campbell, D., Steele, R.J. (2012) Smooth functional tempering for nonlinear differential equation models. *Stat. Comput.* **22**(2):429-443.
Cancer Research UK (2013) CancerStats. www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/Allcancerscombined/, accessed 17/04/2013.
Cantone *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* **137**(1):172-181.
Cao *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**:956-963.
Carlson *et al.* (2009) Breast cancer: Clinical practice guidelines in oncology. *J. Natl. Compr. Canc. Netw.* **7**(2):122-192.
Cizkova *et al.* (2012) PIK3CA mutation impact on survival in breast cancer patients and in ER, PR and ERBB2-based subgroups. *Breast Cancer Res.* **14**:R28.
Challenger, J. D., McKane, A. J., Pahle, J. (2012) Multi-compartment linear noise approximation. *J. Stat. Mech.-Theory E.* **2012**(11):P11010.
Chandrasekaran *et al.* (2012) Latent variable graphical model selection via convex optimization. *Ann. Stat.* **40**(4):1935-1967.
Chen *et al.* (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.* **5**:239.
Chib, S., Jeliazkov, I. (2001) Marginal Likelihood from the Metropolis-Hastings Output. *J. Am. Stat. Assoc.* **96**(453):270-281.
Chiquet, J., Grandvalet, Y., Ambroise, C. (2011). Inferring multiple graphical structures. *Statistics and Computing* **21**(4):537-553.
Cho, H., Fryzlewicz, P. (2012) High dimensional variable selection via tilting. *J. R. Statist. Soc. B* **74**(3):593-622.
Choudhary, C., Mann, M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11**:427-439.
Claassen, T., Heskes, T. (2012) A Bayesian Approach to Constraint Based Causal Inference. *Proceedings of the 28th Conference on Uncertainty and Artificial Intelligence, Santa Catalina CA USA*.
ClinicalTrials.gov (2013a) IRESSA (Gefitinib) in Breast Cancer Patients. NCT00632723.
ClinicalTrials.gov (2013b) Cabozantinib in Women With Metastatic Hormone-Receptor-Positive Breast Cancer. NCT01441947.
ClinicalTrials.gov (2013c) A Phase 1/2 Study Of HKI-272 In Combination With Herceptin In Subjects With Advanced Breast Cancer. NCT00398567.
Colombo *et al.* (2012) Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* **40**:294-321.
Cowles, M.K., Carlin, B.P. (1996) Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *J. Am. Stat. Assoc.* **91**(434):883-904.
Craciun, G., Pantea, C. (2008) Identifiability of chemical reaction networks. *J. Math. Chem.* **44**:244-259.
Crick, F. (1970) Central dogma of molecular biology. *Nature* **227**(5258):561-563.

Crowther, D.J. (2002) Applications of microarrays in the pharmaceutical industry. *Curr. Opin. Pharmacol.* **2**(5):551-554.
Csermely *et al.* (2013) Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol. Therap.* **138**:333-408.
Curtis C *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403):346-352.
Danaher, P., Wang, P., Witten, D.M. (2012) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B*, to appear.
Dash, D. (2003) Caveats for Causal Reasoning with Equilibrium Models. PhD thesis, Intelligent Systems Program, University of Pittsburgh.
Dattner, I., Klaassen, C.A.J. (2013) Estimation in Systems of Ordinary Differential Equations Linear in the Parameters. arXiv:1305.4126.
Davies *et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature* **417**(6892):949-54.
Davies *et al.* (2012) Phase I study of the combination of sorafenib and temsirolimus in patients with metastatic melanoma. *Clin. Cancer Res.* **18**(4):1120-1128.
Dawid, A.P. (1980) Conditional independence for statistical operations. *Ann Statist* **8**:598-617.
Deltell, A. *et al.* (2012) Criteria for Bayesian Model Choice with Application to Variable Selection. *Ann. Stat.* **40**(3):1550-1577.
Dondelinger, F., Lebre, S., Husmeier, D. (2010) Heterogeneous continuous dynamic Bayesian networks with flexible structure and inter-time segment information sharing. *Proceedings of the 27th International Conference on Machine Learning*, 303-310.
Dondelinger, F., Lebre, S., Husmeier, D. (2012) Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Mach. Learn.* **90**(2):191-230.
Dondelinger *et al.* (2013) ODE parameter inference using adaptive gradient matching with Gaussian processes. *Sixteenth International Conference on Artificial Intelligence and Statistics; AISTATS 2013.*
Du *et al.* (2008) Bead-based profiling of tyrosine kinase phosphorylation identifies SRC as a potential target for glioblastoma therapy. *Nat. Biotechnol.* **27**:77-83.
Eaton, D., Murphy, K. (2007) Exact Bayesian structure learning from uncertain interventions, *AI & Statistics* **2**:107-114.
Ellis, B, Wong, W.H. (2008) Learning causal Bayesian network structures from experimental data, *J. Am. Stat. Assoc.* **103**(482):778-789.
Elowitz *et al.* (2002) Stochastic gene expression in a single cell. *Science* **297**(5584):1129-1131.
Engvall, E., Perlmann, P. (1971). Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **8**:871-874.
Eroles *et al.* (2012) Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treat. Rev.* **38**(6):698-707.
Fawcett, T. (2005) An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**:861-874.
Finkenstädt *et al.* (2013) Quantifying Intrinsic and Extrinsic Noise in Gene Transcription Using the Linear Noise Approximation: An Application to Single Cell Data, in preparation.
Forbes *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* **39**(Suppl 1):D945-D950.
Flaherty *et al.* (2010) Inhibition of mutated, activated BRAF in metastatic melanoma. *New Engl. J. Med.* **363**(9):809-819.
Frese, K.K., Tuveson, D.A. (2007) Maximizing mouse cancer models. *Nat. Rev. Cancer* **7**:654-658.
Friedman *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comp. Bio.* **7**:601-620.
Friedman, J., Hastie, T., Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3):432-441.
Friedman, J, Koller, D. (2003) Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach. Learn.* **50**(1):95-125.
Friel, N., Wyse, J. (2012) Estimating the statistical evidence – a review. arXiv:1111.1957.
Fuchs, C. (2013) Inference for Diffusion Processes with Applications in Life Sciences. Springer, Heidelberg.
Gao *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* **24**(16):i70-i75.
George, E.I., Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika* **87**(4):731-747.
Gillespie, D.T. (2009) The deterministic limit of stochastic chemical kinetics. *J. Phys. Chem. B* **113**:1640-

1644.
Girolami, M., Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B* **73**(2):123214.
Goldbeter, A., Koshland, D.E. (1981) An amplified sensitivity arising from covalent modification in biological systems. *Proc. Natl. Acad. Sci. USA* **78**(11):6840-6844.
Golightly, A., Wilkinson, D. (2011) Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1**:807-820.
Grzegorczyk, M., Husmeier, D. (2011) Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics* **27**(5):693-699.
Gujral *et al.* (2012) Profiling phospho-signaling networks in breast cancer using reverse-phase protein arrays. *Oncogene* **378**:1-7.
Guo *et al.* (2011) Joint estimation of multiple graphical models. *Biometrika* **98**:115.
Hache, H., Lehrach, H., Herwig, R. (2009) Reverse Engineering of Gene Regulatory Networks: A Comparative Study. *EURASIP Journal on Bioinformatics and Systems Biology* (2009):8.
Hanahan, D., Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**(5):646-74.
Hans, C., Dobra, A., West, M. (2007) Shotgun stochastic search for "large p" regression. *J. Am. Stat. Assoc.* **102**(478):507-516.
Hara, S., Washio, T. (2012) Learning a common substructure of multiple graphical Gaussian models. *Neural Networks* **38**:2338.
Harsha, H.C., Pandey, A. (2010). Phosphoproteomics in cancer. *Mol. Oncol.* **4**:482-495.
Hauser, A., Bühlmann, P. (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13**:2409-2464.
Heagerty, P.J., Kurland, B.F. (2001) Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models. *Biometrika* **88**(4):973-985.
Hecker, M., Lambeck, S., Toepfer, S., *et al.* (2009) Gene regulatory network inference: Data integration in dynamic models - A review. *Biosystems* **96**(1):86-103.
Heiser *et al.* (2011) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. USA* **109**(8):2724-2729.
Hennessy, B.T. *et al.* (2010) A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Nonmicrodissected Human Breast Cancer. *Clin. Proteomics* **6**:129-151.
Herzenberg *et al.* (2002) The history and future of the fluorescence activated cell sorter and fow cytometry: A view from Stanford. *Clin. Chem.* **48**:1819-1827.
Hill *et al.* (2012a) Bayesian Inference of Signaling Network Topology in a Cancer Cell Line. *Bioinformatics* **28**(21):2804-2810.
Hill, S.M. (2012a) Sparse Graphical Models for Cancer Signalling. PhD thesis, University of Warwick, UK.
Hill *et al.* (2012b) Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. *BMC Bioinformatics* **13**:94.
Hill, S.M., Mukherjee, S. (2013) Network-based clustering with mixtures of L1-penalized Gaussian graphical models: an empirical investigation. arXiv:1301.2194.
Hoff, P.D. (2009) A hierarchical eigenmodel for pooled covariance estimation. *J. Roy. Stat. Soc. B* **71**(5):971-992.
Hollestelle *et al.* (2007) Phosphatidylinositol-3-OH Kinase or RAS Pathway Mutations in Human Breast Cancer Cell Lines. *Mol. Cancer. Res.* **5**(2):195:201.
Honkela *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. USA* **107**(17):77937798.
Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **4**(11):682-690.
Hsiao *et al.* (2012) 384 hanging drop arrays give excellent Z-factors and allow versatile formation of co-culture spheroids. *Biotechnol. Bioeng.* **109**(5):12931304.
Hsieh *et al.* (1997) Multidimensional chromatography coupled with mass spectrometry for target-based screening. *Mol. Divers.* **2**(4):189-196.
Hu *et al.* (2007) Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**:1986-1994.
Hurley *et al.* (2011) Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic Acids Res.* **39**(21):1-22.

Hurn, A., Jeisman, J., Lindsay, K. (2007) Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *J. Financ. Economet.* **5**(3):390-455.
Iadevaia *et al.* (2010) Identification of optimal drug combinations targeting cellular networks: Integrating phosphoproteomics and computational network analysis. *Can. Res.* **70**:6704-6714.
Ideker, T., Lauffenburger, D. (2003) Building with a scaffold: emerging strategies for high to low level cellular modelling. *Trends Biotechnol.* **21**(6):255-262.
Ideker, T., Krogan, N.J. (2012) Differential network biology. *Mol. Syst. Biol.* **8**(1).
Imoto *et al.* (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc. IEEE Computer Society Bioinformatics Conference (CSB'03)*, 104-113.
Imoto *et al.* (2006) Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Stat. Method.* **3**(1):116.
Iwasaki, Y., Simon, H.A. (1994) Causality and model abstraction. *Artif. Intel.* **67**(1):143194.
Kalisch, M., Bühlmann, P. (2007) Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *J. Mach. Learn. Res.* **8**:613-636.
Kalogeropoulos *et al.* (2010) Inference for stochastic volatility models using time change transformations. *Ann. Stat.* **38**(2):784:807.
Kato, T., Tsuda, K., Asai, K. (2005) Selective integration of multiple biological data for supervised network inference. *Bioinformatics* **21**(10):2488-2495.
Kholodenko, B.N. (2006) Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Bio.* **7**(3):165-176.
Kim, S.Y., Imoto, S., Miyano, S. (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* **4**(3):228-235.
Knowles, D.A., Ghahramani, Z. (2011) Nonparametric Bayesian Sparse Factor Models with Application to Gene Expression Modelling. *Ann. Appl. Stat.* **5**(2B):1534-1552.
Kolar, M., Song, L., Xing, E.P. (2009) Sparsistent learning of varying-coefficient models with structural changes. *Adv. Neur. In.* **22**:10061014.
Komorowski *et al.* (2009) Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**:343.
Komorowski *et al.* (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl. Acad. Sci. USA* **108**(21):8645-8650.
Konopka *et al.* (2006) Crowding and confinement effects on protein diffusion in vivo. *J. Bacteriol.* **188**(17):6115-6123.
Korkola *et al.* (2012) PIK3CA hotspot mutations predict synergistic response between lapatinib and AKT inhibitors in HER2 positive breast cancer cells. In submission.
Kou, S., Sunney, X., Jun, L. (2005) Bayesian analysis of single-molecule experimental data (with discussion). *J. R. Statist. Soc. C* **54**:469-506.
Kschischang, F.R., Frey, B.J., Loeliger, H.-A. (2001) Factor Graphs and the Sum-Product Algorithm. *IEEE T. Inform. Theory* **47**(2):498-519.
Lèbre *et al.* (2010) Statistical inference of the time-varying structure of gene- regulation networks. *BMC Syst. Biol.* **4**:130.
Lashkari *et al.* (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* **94**(24):13057-13062.
Lauritzen, S.L. (1996) Graphical models. Oxford University Press.
Lavezzari *et al.* (2012) Monitoring phosphoproteomic response to targeted kinase inhibitors using reverse-phase protein microarrays. *Kinase Inhibitors*, Humana Press p.203-215.
Lee, A. *et al.* (2010) On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods. *J. Comp. Graph. Stat.* **19**(4):769-789.
Lee, W.P., Tzou, W.S. (2009) Computational methods for discovering gene networks from expression data. *Brief. Bioinform.* **10**(4):408-423.
Lee *et al.* (2012) Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks. *Cell* **149**(4):780-794.
Leskovac, V. (2003) Comprehensive enzyme kinetics. Kluwer Academic / Plenum Publishers, New York.
Li, C-W., Chen, B-S. (2010) Identifying Functional Mechanisms of Gene and Protein Regulatory Networks in Response to a Broader Range of Environmental Stresses. *Comp. Funct. Genom.*:408705.
Li, Z., Li, P., Krishnan, A. (2011) Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* **27**(19):2686-2691.

Li, H., Petzold, L. (2008) Efficient Parallel Simulation for Stochastic Simulation of Biochemical Systems on the Graphics Processing Unit. *Int. J. High Perform. C.* **24**(2):107-116.
Liang *et al.* (2008) Mixtures of g Priors for Bayesian Variable Selection. *J. Am. Stat. Assoc.* **103**(481):410-423.
Liu *et al.* (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178**(3):1763-1776.
Loh, P., Wainwright, M.J. (2012) High-dimension regression with noisy and missing data: Provable guarantees with non-convexity. *Ann. Stat.* **40**(3):1637-1664.
Lu *et al.* (2011) Kinome siRNA-phosphoproteomic screen identifies networks regulating Akt signaling. *Oncogene* **30**:4567-4577.
Lv, J., Liu, J.S. (2010) Model Selection Principles in Misspecified Models. arXiv:1005.5483v1.
Maathuis *et al.* (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7**:247-248.
Maher, B. (2012) ENCODE: The human encyclopaedia. *Nature* **489**(7414):46-48.
Malinowsky *et al.* (2012) Common Protein Biomarkers Assessed by Reverse Phase Protein Arrays Show Considerable Intratumoral Heterogeneity in Breast Cancer Tissues. *PloS one* **7**(7):e40285.
Marbach *et al.* (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods, *J. Comput. Biol.* **16**(2):229-239.
Marbach *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**:796-804.
Markowetz, F., Spang, R. (2007) Inferring cellular networks - A review. *BMC Bioinformatics* **8**(Suppl. 6): S5.
Masters, J.R.W. (2000) Human cancer cell lines: fact and fantasy. *Nat. Rev. Mol. Biol.* **1**:233-236.
Masters *et al.* (2001) Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc. Natl. Acad. Sci. USA* **98**:8012-8017.
Mauro, M.J. (2006) Defining and managing imatinib resistance. *ASH Education Program Book* **1**:219-225.
McAdams, H.H., Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94**:814-819.
McKeage, K., Perry, C.M. (2002) Trastuzumab: A Review of its Use in the Treatment of Metastatic Breast Cancer Overexpressing HER2. *Drugs* **62**(1):209-243.
M. D. Anderson RPPA Core Facility (2013) Standard Antibody List. http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/antibody-lists—protocols/corestdablist-1-16-13.pdf. Accessed 14-02-13.
Meinshausen, N., Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **34**(3):1436-1462.
Michaelis, L., Menten, M.L. (1913) Die kinetik der invertinwirkung. *Biochem Z* **49**:333-369.
Minty, J.J., Varedi, K.S.M., Nina, L.X. (2009) Network benchmarking: a happy marriage between systems and synthetic biology. *Chemistry and Biology* **16**(3):239-41.
Moen *et al.* (2007) Imatinib: A Review of its Use in Chronic Myeloid Leukaemia. *Drugs* **67**(2):299-320.
Mohan *et al.* (2013) Node-based learning of multiple Gaussian graphical models. arXiv:1303.5145.
Morrissey *et al.* (2010) On reverse engineering of gene interaction networks using time course data with repeated measurements. *Bioinformatics* **26**(18):2305-2312.
Mueller, C., Liotta, L.A., Espina, V. (2010) Reverse phase protein microarrays advance to use in clinical trials. *Mol. Oncol.* **4**:461-481.
Mugler *et al.* (2011) Statistical method for revealing form-function relations in biological networks. *Proc. Natl. Acad. Sci. USA* **108**(2):446-451.
Mukherjee, S., Speed, T.P. (2008) Network inference using informative priors. *Proc. Natl. Acad. Sci. USA* **105**(38):14313-14318.
Mukherjee, S., Hill, S.M. (2011) Network clustering: probing biological heterogeneity by sparse graphical models. *Bioinformatics* **27**(7):994-1000.
Murphy, K. (2002) Dynamic Bayesian Networks: Representation, Inference and Learning. PhD Thesis, University of California, Berkeley.
Nam, D., Yoon, S.H., Kim, J.F. (2007) Ensemble learning of genetic networks from time-series expression data. *Bioinformatics* **23**(23):3225-3231.
Nelander *et al.* (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol. Syst. Biol.* **4**(1):216.
Nelson-Rees, W.A., Daniels, D.W., Flandermeyer, R.R. (1981) Cross-contamination of cells in culture.

Science **212**:446-452.
Neve *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**(6):515-527.
Nita-Lazar, A., Saito-Benz, H., White, F.M. (2008) Quantitative phosphoproteomics by mass spectrometry: Past, present, and future. *Proteomics* **8**:4433-4443.
Nodelman, U., Shelton, C. R., Koller, D. (2002) Continuous time Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 378-387.
Novartis Pharma AG. (2006) Gleevec (imatinib mesylate) tablets prescribing information. East Hanover, NJ, USA.
Oates *et al.* (2012) Network Inference Using Steady State Data and Goldbeter-Koshland Kinetics. *Bioinformatics* **28**(18):2342-2348.
Oates C.J., Mukherjee, S. (2012b) Causal Variable Selection Using Equilibrium Relations from Nonlinear Dynamics. *Workshop on Causal Structure Learning, Uncertainty in Artificial Intelligence (UAI'12)*. Santa Catalina, CA, USA.
Oates *et al.* (2012c) Structure Learning Trees. In submission.
Øksendal, B. (1998) Stochastic Differential Equations, An Introduction with Applications (5th ed.). Springer.
Opgen-Rhein, R., Strimmer, K. (2007) Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* **8** (Suppl. 2):S3.
Osborne, C.K., Hobbs, K., Trent, J.M. (1987) Biological differences among MCF-7 human breast cancer cell lines from different laboratories. *Breast Cancer Res. Treat.* **9**:111-121.
Pahlajani, C.D., Atzberger, P.J., Khammash, M. (2011) Stochastic reduction method for biological chemical kinetics using time-scale separation. *J. The. Bio.* **272**(1):96-112.
Papaspiliopoulos *et al.* (2012) Nonparametric estimation of diffusions: a differential equations approach. *Biometrika* **99**(3):511-531.
Paweletz *et al.* (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**:1981-1989.
Pearl, J. (1998) Why there is no statistical test for confounding, why many think there is, and why they are almost right. Technical report, Department of Statistics, UCLA, UC Los Angeles.
Pearl, J. (2009) Causality: models, reasoning and inference (Second Edition). Cambridge: MIT press.
Pearl, J. (1982) Reverend Bayes on inference engines: A distributed hierarchical approach. *Proceedings of the Second National Conference on Artificial Intelligence. AAAI-82: Pittsburgh, PA. Menlo Park, California: AAAI Press.*, 133136.
Penfold *et al.* (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics* **28**(12):i233-i241.
Perez, O.D., Nolan, G.P. (2002). Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat. Biotech.* **20**:155-162.
Peters *et al.* (2011) Identifiability of Causal Graphs using Functional Models. *Proc. 27th Ann. Conf. Uncertainty in Artificial Intelligence (UAI-11)*, 589-598.
Peters, J., Bühlmann, P. (2012) Identifiability of Gaussian structural equation models with same error variances. arXiv:1205.2536.
Peters, J. (2012) Restricted Structural Equation Models for Causal Inference. PhD thesis, ETH Zurich.
Petitjean, A. *et al.* (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum Mutat.* **28**(6):622-9.
Pierobon *et al.* (2012) Development and Clinical Implementation of Reverse Phase Protein Microarrays for Protein Network Activation Mapping: Personalized Cancer Therapy. *Systems Biology in Cancer Research and Drug Discovery* Springer Netherlands, p.309-323.
Prill *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS one* **5**:e9202.
Paulsson, J. (2005) Models of stochastic gene expression, *Phys. Life Rev.* **2**(2):157-175.
Quach M, Brunel N, D'Alché-Buc F (2007) Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference. *Bioinformatics* 23(23):3209-3216.
Queen, C.M., Smith, J.Q. (1993) Multiregression Dynamic Models. *J. Roy. Statist. Soc. B* **55**(4):849-870.
Quintás-Cardama *et al.* Reverse phase protein array profiling reveals distinct proteomic signatures associated with chronic myeloid leukemia progression and with chronic phase in the CD34-positive com-

partment. *Cancer* **118**(21):5283-5292.
Ramaswamy *et al.* (2005) Application of protein lysate microarrays to molecular marker verification and quantification. *Proteome Science* **3**(1):9.
Richardson, T., Spirtes, P. (2002) Ancestral graph Markov models. *Ann. Stat.* **30**:962-1030.
Roberts, G.O., Gelman, A., Gilks, W.R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**(1):110-120.
Roberts, G.O., Rosenthal, J.S. (2006) Harris Recurrence of Metropolis-within-Gibbs and Trans-Dimensional Markov Chains. *Ann. Appl. Probab.* **16**(4):2123-2139.
Roberts, P., Der, C.J. (2007) Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **26**:3291-3310.
Rodríguez, A., Lenkoski, A., Dobra, A. (2011) Sparse covariance estimation in heterogeneous samples. *Electron. J. Statist.* **5**:981-1014.
Rodriguez-Yam, G., Davis, R.A., Scharf, L.L. (2002) A Bayesian model and Gibbs sampler for hyperspectral imaging. *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, p.105-109.
Rodriguez-Yam, G., Davis, R.A., Scharf, L.L. (2004) Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression, unpublished manuscript.
Rogers, S., Khanin, R., Girolami, M. (2007) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics* **8**(S2).
Rue, H., Martino, S., Chopin, N. (2009) Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *J. R. Statist. Soc. B* **71**(2):319-392.
Sachs *et al* (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**:523529.
Samuels *et al.* (2004) High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**(5670):554.
Sanft *et al.* Legitimacy of the stochastic Michaelis-Menten approximation. *Systems Biology, IET* **5**(1):58-69.
Sayikli, C., Bagci, E.Z. (2011) Limitations of Using Mass Action Kinetics Method in Modeling Biochemical Systems: Illustration for a Second Order Reaction. *Computational Science and Its Applications-ICCSA 2011*, Springer Berlin Heidelberg, p.521-526.
Schena *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235):467470.
Schnell S, Mendoza C (1997) Closed Form Solution for Time-dependent Enzyme Kinetics. *J. Theor. Biol.* **187**:207-212.
Schoeberl *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* **20**(4):370-375.
Scott, J.G., Berger, J.O. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Stat.* **38**(5):2587-2619.
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.
Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* **432**(7019):862-865.
Silverman, E. (2012) Are Drug Pipelines Really Improving? *Forbes* 15-09-12.
Smith *et al.* (2001) Nonparametric regression using linear combinations of basis functions. *Stat. Comp.* **11**(4):313-322.
Smith, J.Q., Anderson, P.E. (2008) Conditional independence and chain event graphs. *Artificial Intelligence* **172**(1):42-68.
Sokol, A., Hansen, N.R. (2013) Causal interpretation of stochastic differential equations. arXiv:1304.0217.
Song, L., Kolar, M., Xing, E.P. (2009) Time-Varying Dynamic Bayesian Networks. *Adv. Neur. In.* **22**:1732-1740.
Sotiriou, C., Pusztai, L. (2009) Gene-Expression Signatures in Breast Cancer. *N. Engl. J. Med.* **360**:790-800.
Souza *et al.* (2010) Three-dimensional tissue culture based on magnetic cell levitation. *Nat. Nanotechnol.* **5**:291-296.
Spencer, S., Hill, S.M., Mukherjee, S. (2012) Dynamic Bayesian networks for interventional data. *CRiSM Working Paper Series, The University of Warwick, UK* **12**:24.
Spirtes *et al.* (2000) Causation, Prediction, and Search, 2nd ed. MIT Press, Cambridge, MA.
Spurrier, B., Ramalingam, S., Nishizuka, S. (2008) Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat. Protocols* **3**:1796-1808.

Stathopoulos, V., Girolami, M. (2013) Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Phil. Trans. R. Soc. A* **371**:1984.

Steijaert *et al.* (2010) Computing the Stochastic Dynamics of Phosphorylation Networks. *J. Comput. Biol.* **17**(2):189-199.

Stoevesandt *et al.* (2009) Protein microarrays: high-throughput tools for proteomics. *Expert Rev. Proteomic.* **6**(2):145-157.

Sudhakar, A. (2009) History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.* **1**(2):1-4.

Swain, P.S., Elowitz, M.B., Siggia, E.D. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA* **99**(12):795-800.

Swat, M., Kel, A., Herzel, H. (2004) Bifurcation analysis of the regulatory modules of the mammalian G_1/S transition. *Bioinformatics* **20**(10):1506-1511.

Telesca *et al.* (2011) Modeling Protein Expression and Protein Signaling Pathways. *J. Am. Stat. Assoc.* **107**(500):1372-1384.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* **467**(7319):1061-1073.

The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418):61-70.

Tibes *et al.* (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**(10):2512-2521.

Toni *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**(31):187-202.

Toni *et al.* (2012) Elucidating the in vivo phosphorylation dynamics of the ERK MAP kinase using quantitative proteomics data and Bayesian model selection. *Molecular BioSystems* **8**:1921-1929.

Van den Bulcke, T., Van Leemput, K., Naudts, B., *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* **7**(1):43.

van Kampen, N.G. (1976) The expansion of the master equation. *Adv. Chem. Phys.* **34**:245-309.

van Kampen, N.G. (2007) Stochastic Processes in Physics and Chemistry (3rd ed.), North Holland.

van 't Veer *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871):530-536.

Velculescu, V.E. *et al.* (2004) High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**(5670):554.

Vigelius *et al.* (2010) Accelerating Reaction-Diffusion Simulations with General-Purpose Graphics Processing Units. *Bioinformatics* **27**(2):288-290.

Vogelstein, B., Kinzler, K. (2004) Cancer genes and the pathways they control. *Nat. Med.* **10**(8):789-799.

Voortman *et al.* (2010) Learning Why Things Change: The Difference-Based Causality Learner. *Proceedings of the Twenty-Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI).*

Vu, D., Hunter, D.R., Schweinberger, M. (2012) Model-based clustering of large networks. *Ann. Appl. Stat.*, to appear.

Vyshemirsky V, Girolami M (2008) Bayesian ranking of biochemical system models. *Bioinformatics* **24**(6):833-839.

Wainwright, M.J., Jordan, M.I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1-2**:1-305.

Wallace, E., Gillespie, D.T., Sanft, K.R., Petzold, L.R. (2012) A New Perspective on the Linear Noise Approximation. *IET Syst. Biol.*, to appear.

Wei, P., Pan, W. (2012) Bayesian Joint Modeling of Multiple Gene Networks and Diverse Genomic Data to Identofy Target Genes of a Transcription Factor. *Ann. Appl. Stat.* **6**(1):334-355.

Weile *et al.* (2012) Bayesian integration of networks without gold standards. *Bioinformatics* **28**(11):1495-1500.

Weinberg R (2007) The Biology of Cancer. Garland Science, New York.

Werhli, A.V., Grzegorczyk, M., Husmeier, D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* **22**(20):2523-2531.

Werhli, A.V., Husmeier, D. (2008) Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of Bioinformatics and Computational Biology* **6**(3):543-572.

Wetterskog *et al.* (2013) Identification of novel determinants of resistance to lapatinib in ERBB2-amplified cancers. *Oncogene*, to appear.

Wilkinson, D.J. (2006) Stochastic Modelling for Systems Biology. Chapman and Hall/CRC.

Wilkinson, D.J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* **10**(2):122-133.

Wood SN (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**(26):1102-1104.

Xu *et al.* (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species, *Science Signaling* **3**(113):ra20.

Yang, S. *et al.* (2012) Fused Multiple Graphical Lasso. arXiv:1209.2139.

York *et al.* (2012) Network analysis of reverse phase protein expression data: Characterizing protein signatures in acute myeloid leukemia cytogenetic categories t (8; 21) and inv (16). *Proteomics* **12**(13):2084-2093.

Zalatan *et al.* (2012) Conformational Control of the Ste5 Scaffold Protein Insulates Against MAP Kinase Misactivation. *Science* **337**(6099):1218-1222.

Zellner, A. (1986) On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions, *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti, eds. P. K. Goel and A. Zellner*, p.233-243.

Zhang, H., Pelech, S. (2012) Using Protein Microarrays to Study Phosphorylation-mediated Signal Transduction. *Seminars in Cell & Developmental Biology*, Academic Press.

Zhang, J. (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172**:18731896.

Zhou, H., Pan, W., Shen, X. (2009) Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* **3**:1473-1496.

Zou, C., Feng, J. (2009) Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics* **10**(12):122.