Communauté UNIVERSITÉ Grenoble Alpes

Turun yliopisto
University of Turku

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

préparée dans le cadre d'une cotutelle entre *la Communauté Université Grenoble Alpes* et *the University of Turku*

Spécialité : **ISCE/MODÈLES, MÉTHODES ET ALGORITHMES EN BIOLOGIE**

Arrêté ministériel : le 6 janvier 2005 - 7 août 2006

Présentée par

Sean ROBINSON

Thèse dirigée par **Laurent GUYON** et **Jaakko NEVALAINEN**

préparée au sein du **Laboratoire Biologie du Cancer et de l'Infection**

dans **l'École Doctorale Ingénierie Pour La Santé, La Cognition et l'Environnement (EDISCE)** et **the Doctoral Programme in Mathematics and Computer Sciences (MATTI)**

Applications en bioinformatique avec des modèles de Markov

Thèse soutenue publiquement le 1er juin 2018 devant le jury composé de :

Mervi EEROLA
Professor, University of Turku, Finland (Président)
Benno SCHWIKOWSKI
PhD, Institut Pasteur Paris, France (Examinateur)
Matti NYKTER
Professor, University of Tampere, Finland (Rapporteur)
Aki VEHTARI
Associate Professor, Aalto University, Finland (Rapporteur)
Laurent GUYON
PhD, CEA Grenoble, France (Directeur de thèse)
Jaakko NEVALAINEN
Professor, University of Tampere, Finland (Co-directeur de thèse)

Turun yliopisto
University of Turku

Communauté
UNIVERSITÉ Grenoble Alpes

APPLICATIONS IN BIOINFORMATICS WITH MARKOV MODELS

Sean Robinson

University of Turku

Faculty of Science and Engineering
Department of Mathematics and Statistics
Doctoral Programme in Mathematics and Computer Sciences (MATTI)

Université Grenoble Alpes

Le Laboratoire Biologie du Cancer et de l'Infection
l'École Doctorale Ingénierie Pour La Santé, La Cognition et l'Environnement (EDISCE)

Supervised by

Professor Jaakko Nevalainen
University of Tampere, Finland

Dr Laurent Guyon
CEA Grenoble, France

Reviewed by

Professor Matti Nykter
University of Tampere, Finland

Associate Professor Aki Vehtari
Aalto University, Finland

Opponent

Dr Benno Schwikowski
Institut Pasteur Paris, France

Custos

Professor Mervi Eerola
University of Turku, Finland

ISBN 978-951-29-7199-2 (PRINT)
ISBN 978-951-29-7200-5 (PDF)
Painosalama Oy - Turku, Finland 2018

Acknowledgements

Firstly I would like to thank my supervisors Jaakko Nevalainen and Laurent Guyon as well as Matthias Nees and Xavier Gidrol who were all involved in organising this compound international and interdisciplinary thesis. This was a particularly challenging project spanning two universities and research institutes, and I thank you all for your valuable contributions. I would also like to thank the opponent Benno Schwikowski for agreeing to take on the joint Finnish/French thesis defence requirements and both reviewers Matti Nykter and Aki Vehtari who also needed to go above and beyond the standard review process.

There have been many people inside and out of multiple research groups who have been incredibly welcoming and who have made my time in Finland and France immeasurably better. In particular I would like to thank Michael Courtney for an excellent final collaboration as well as Tero Aittokallio for kindly sharing his expertise. A big shout-out goes to Ilmari Ahonen for his friendship during the course of our concurrent PhD experiences. Finally I am truly grateful to Anni for all her love and support which has meant so much to me.

Funding for the first three years was jointly provided by VTT, Finland and CEA, France while finding for the final fourth year was provided by the MATTI doctoral school at the University of Turku and the Finnish Cultural Foundation, Varsinais-Suomi Regional Fund. Funding to participate in a number of international conferences as well as travel between Finland and France was provided by both labs at VTT and CEA, the Turku University Foundation, the MATTI doctoral school and the Institut français de Finlande.

Abstract

In this thesis we present four applications in bioinformatics with Markov models. That is, we extend the use of such models in the mathematical and statistical analysis of biological data. The data we consider are drawn from a broad range of areas. We consider applications at the genomic level with time series and network data as well as applications at the cellular level with microscopy image data of both cell culture and *in vivo* tissue.

Collections of objects such as genes, cells or pixels are of particular interest as a whole. We make use of associations within these collections, spatial, temporal or functional, and assume that closer objects are more strongly associated than those further apart. This allows for efficient inference within a Markov model framework and is encoded in terms of conditional independences between variables as represented by vertices and edges in an undirected graph.

Chapter 1 presents an overview of undirected graphical models in general and Markov models in particular. Chapter 2 presents inference for variables in hidden Markov random fields (MRFs) while Chapter 3 presents inference for the parameters of Gaussian MRFs. Chapter 4 outlines the four applications and how the Markov model framework is utilised in each case. For each application, the associated publication is also provided.

In Publication 1, hidden Markov models (HMMs) are used to achieve an alignment and classification of time series data. Publications 2 and 3 concern inference with hidden MRFs to obtain a segmentation of both digital image data and network data respectively. Spatial analysis with Gaussian MRFs is presented in Publication 4. We show that our particular use of Markov models in each of our applications enables us to achieve our aims.

Tiivistelmä

Tässä väitöskirjassa esitellään neljä Markovin mallien bioinformatiikan sovellusta. Väitöskirjassa laajennetaan Markovin mallien käyttöä monelta eri sovellusalalta kerätyn biologisen datan matemaattiseen ja tilastolliseen analyysiin. Malleja hyödynnetään genomitasolla aikasarjoissa ja verkkodatassa sekä solutasolla soluviljelyn ja *in vivo* -kudoksen mikroskooppikuva-aineistossa.

Kiinnostuksen kohteina tässä tutkimuksessa ovat geeneistä, soluista ja pikseleistä koostuvat objektien joukot. Työssä käytetään hyväksi näiden joukkojen sisällä vallitsevia spatiaalisia, temporaalisia tai funktionaalisia assosiaatiota olettaen, että toisiaan lähempänä olevat objektit ovat keskenään vahvemmin yhteyksissä kuin toisistaan kauempana olevat. Tämä oletus mahdollistaa tehokkaan päättelyn Markov-malliperheessä sekä ehdollisten riippumattomuuksien esittämisen satunnaismuuttujien välillä, joita kuvataan suuntaamattomissa graafeissa pisteillä ja viivoilla.

Luvussa 1 tarkastellaan yleisellä tasolla suuntaamattomia malleja, joista tarkemmin esitellään Markovin mallit. Luvussa 2 käsitellään muuttujia koskevaa päättelyä piilo-Markovin satunnaiskentissä (MRF) ja luvussa 3 keskitytään Gaussisten MRF -mallien parametreja koskevaan päättelyyn. Luvussa 4 esitellään neljä tutkimuksen sovellusta ja miten Markovin malleja on niissä laajennettu ja hyödynnetty. Sovelluksiin liittyvät osatyöt muodostavat väitöskirjan viimeisen luvun.

Osatyössä 1 Markovin piilomalleja (HMM) käytetään aikasarjadatan kohdennukseen ja luokitteluun. Osatöissä 2 ja 3 käsitellään päättelyä piilo-MRF-malleissa niiden hyödyntämiseksi digitaalisen kuva- ja verkkodatan segmentoimisessa. Osatyössä 4 käsitellään Gaussisten MRF -mallien käyttöä spatiaalisessa analyysissa. Huolellisesti rakennettujen Markovin mallien käyttö osoittautuu kaikissa sovelluksissa keskeiseksi osatutkimusten tavoitteiden saavuttamiseksi.

Résumé

Dans cette thèse nous présentons quatre applications en bioinformatique avec des modèles de Markov. Plus précisément, nous étendons ces modèles à l'analyse statistique et mathématique de données biologiques. Les données que nous étudions viennent de différentes sources. Nous considérons les applications au niveau génomique avec des séries temporelles et des données de réseau ainsi que les applications au niveau cellulaire avec des données d'images de microscopie de cultures cellulaires et de tissus *in vivo*.

Les ensembles d'objets tels que gènes, cellules ou pixels, présentent un intérêt dans leur intégralité. Nous utilisons des associations spatiales, temporelles ou fonctionnelles au sein de ces ensembles et nous supposons que les objets qui sont plus proches les uns des autres sont plus fortement liés que ceux qui sont plus éloignés. Cela permet une inférence efficace dans le cadre des modèles de Markov et est codé par des indépendances conditionnelles entre variables qui sont représentées par des sommets et arêtes dans un graphe non orienté.

Le chapitre 1 présente une vue d'ensemble des modèles graphiques non orientés en général et des modèles de Markov en particulier. Le chapitre 2 présente l'inférence des variables des champs aléatoires de Markov (MRFs) cachés tandis que le chapitre 3 présente l'inférence des paramètres des MRFs de Gauss. Le chapitre 4 expose les quatre applications traitées et comment le cadre des modèles de Markov est utilisé dans chaque cas. Pour chacune des applications, la publication associée est fournie.

Dans la publication 1, les modèles de Markov cachés (HMMs) sont utilisés pour atteindre un alignement et une classification des données de séries temporelles. Les publications 2 et 3 concernent l'inférence à la fois avec des MRFs cachés pour obtenir une segmentation des données d'images numériques et des données de réseau. Une analyse spatiale avec des MRFs de Gauss est présentée dans la publication 4. Nous montrons que notre utilisation particulière des modèles de Markov dans chaque publication nous a permis d'atteindre nos objectifs.

Publications

[1] Robinson, S., Glonek, G., Koch, I., Thomas, M. & Davies, C. (2015), 'Alignment of Time Course Gene Expression Data and the Classification of Developmentally Driven Genes with Hidden Markov Models', *BMC Bioinformatics* **16**(196), doi:10.1186/s12859-015-0634-9.

[2] Robinson, S., Guyon, L., Nevalainen, J., Toriseva, M., Åkerfelt, M. & Nees, M. (2015), 'Segmentation of Image Data from Complex Organotypic 3D Models of Cancer Tissues with Markov Random Fields', *PLOS One* **10**(12), doi:10.1371/journal.pone.0143798.

[3] Robinson, S., Nevalainen, J., Pinna, G., Campalans, A., Radicella, J. P. & Guyon, L. (2017), 'Incorporating Interaction Networks into the Determination of Functionally Related Hit Genes in Genomic Experiments with Markov Random Fields', *Bioinformatics* **33**(14), i170–i179, doi:10.1093/bioinformatics/btx244.

[4] Robinson, S. & Courtney, M. J. (2018), 'Spatial Quantification of the Synaptic Activity Phenotype Across Large Populations of Neurons with Markov Random Fields', *Bioinformatics* (Advance Access Publication), doi:10.1093/bioinformatics/bty322.

Contents

1 **Introduction** **1**

1.1 Graphical Models 1

1.2 Markov Models 3

1.3 Hammersley-Clifford Theorem 6

2 **Inference for Hidden MRFs** **7**

2.1 Energy Function 7

2.2 Graph Cuts 9

2.3 Alpha-Expansion Algorithm 13

3 **Inference for Gaussian MRFs** **15**

3.1 Precision Matrix 15

3.2 Conditional Autoregressive Models 17

3.3 Posterior Densities 20

4 **Presentation of Publications** **23**

Bibliography **31**

Publications

Publication 1 **37**

Publication 2 **51**

Publication 3 **85**

Publication 4 **101**

Chapter 1

Introduction

Markov models are a type of undirected graphical model, a collection of random variables whose conditional independence properties can be represented by an undirected graph. They are used for many different applications in a broad range of areas including the analysis of time series data, image data, network data and spatial data among many others (Rue & Held 2005, Koller & Friedman 2009, Blake et al. 2011, Banerjee et al. 2014).

Named after Russian mathematician Andrei Markov (1856–1922), such models are especially popular since the Markov structure allows for complex conditional independences to be modelled while still allowing for efficient inference. Moreover, the graphical representation of the conditional independence properties provides a valuable and intuitive visualisation of the model.

Below we define undirected graphical models in general before defining Markov models in particular. We then consider the Hammersley-Clifford Theorem, a fundamental result allowing for the factorisation of the model density. The content of this chapter is primarily based on the presentations of Whittaker (1990) and Lauritzen (1996).

1.1 Graphical Models

We use p to denote a probability density or probability mass function and use the argument to identify the random variable concerned. Hence for the random variables X and Z we write $p(x)$ to represent the marginal density function, $p(x, z)$ to represent the joint density function and $p(x|z)$ to represent the conditional density function. A random variable X takes values in its support Ω_X. For collections of random variables indexed by scalar subscripts, $\{X_1, X_2, \ldots\}$ we use set notation in the subscript, $X_{\mathcal{A}} = \{X_i \mid i \in \mathcal{A}\}$ along with standard set notation operations.

Definition 1.1 Conditionally Independent Random Variables.[†] *Random variables X and Y are conditionally independent given Z, written $X \perp\!\!\!\perp Y \mid Z$, if and only if*

$$p(x, y|z) = p(x|z)p(y|z)$$

for all $x \in \Omega_X$, $y \in \Omega_Y$ and $z \in \Omega_Z$ such that $p(z) > 0$.

Definition 1.2 Undirected Graph.[†] *An undirected graph $\mathcal{G}$ is a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$, written $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E}$ is a subset of the set $\mathcal{V} \times \mathcal{V}$ of pairs of distinct vertices such that $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$ for all $i, j \in \mathcal{V}$.*

Definition 1.3 Conditional Independence Graph.[†] *The conditional independence graph of the collection of random variables $X_\mathcal{V}$ is the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where*

$$(i, j) \notin \mathcal{E} \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_{\mathcal{V}\setminus\{i,j\}} \tag{1.1}$$

for all $i \neq j \in \mathcal{V}$.

We have defined a conditional independence graph in terms of the pairwise Markov property (1.1). Given a collection of random variables with a set of conditional independence properties, we are able to construct the corresponding conditional independence graph. Under certain conditions that are given below, we are able to deduce further conditional independence properties of the model directly from the corresponding graph.

Definition 1.4 Paths and Separation.[†] *In an undirected graph $\mathcal{G}$, a path between two vertices $i, j \in \mathcal{V}$ is a sequence $a_1, a_2, \ldots, a_K$ such that $a_1 = i$, $a_K = j$ and $(a_{k-1}, a_k) \in \mathcal{E}$ for $k = 2, 3, \ldots, K$. Two vertices $i, j \in \mathcal{V}$ are separated by the subset $\mathcal{C} \subseteq \mathcal{V}\setminus\{i, j\}$ if and only if all paths between i and j contain at least one member of $\mathcal{C}$.*

Theorem 1.1 Separation Theorem. *Let the collection of random variables $X_\mathcal{V}$ have density function $p(x_\mathcal{V}) > 0$ for all $x_\mathcal{V} \in \Omega_{X_\mathcal{V}}$ and conditional independence graph $\mathcal{G}$. For disjoint subsets of vertices $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{V}$, if in the conditional independence graph $\mathcal{G}$ each vertex in $\mathcal{A}$ is separated from each vertex in $\mathcal{B}$ by the subset $\mathcal{C}$, then*

$$X_\mathcal{A} \perp\!\!\!\perp X_\mathcal{B} \mid X_\mathcal{C}. \tag{1.2}$$

Proof of Theorem 1.1. A proof is given by Lauritzen (1996). □

[†] These definitions are taken directly from my Master of Philosophy thesis (Robinson 2012).

Theorem 1.1 is given in terms of the global Markov property (1.2). Hence for a collection of random variables with strictly positive density function, the corresponding conditional independence graph is equivalent to statements of the conditional independence properties of the model. That is, the conditional independence graph is a graphical visualisation of the model.

Lemma 1.2 Pairwise and Global Markov Properties. *If the collection of random variables $X_{\mathcal{V}}$ has density function $p(x_{\mathcal{V}}) > 0$ for all $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ then the pairwise* (1.1) *and global* (1.2) *Markov properties are equivalent.*

Proof of Lemma 1.2. The global Markov property clearly implies the pairwise Markov property while the pairwise Markov property implies the global Markov property if the collection of random variables has a strictly positive density function (Theorem 1.1). □

1.2 Markov Models

Definition 1.5 Neighbours and Cliques. *In an undirected graph $\mathcal{G}$ the neighbours of a vertex i are the members of the set $\nu(i) = \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\}$. A clique is a subset of vertices, $\mathcal{C} \subset \mathcal{V}$ such that for all $i \in \mathcal{C}$, if $j \in \mathcal{C}\backslash i$ then $j \in \nu(i)$. A clique $\mathcal{C}$ is maximal if there exists no other clique $\mathcal{C}^* \subseteq \mathcal{V}$ such that $\mathcal{C} \subset \mathcal{C}^*$.*

Each vertex is a clique and each pair of neighbouring vertices is also a clique. If a pair of neighbouring vertices have a mutual neighbour, then these three vertices are another clique. If all members of a clique don't have a mutual neighbour not already in the clique, then the clique is maximal.

Definition 1.6 Markov Model. *The collection of random variables $X_{\mathcal{V}}$ with conditional independence graph $\mathcal{G}$ is a Markov model if*

$$X_i \perp\!\!\!\perp X_{\mathcal{V}\backslash\{i,\nu(i)\}} \mid X_{\nu(i)} \tag{1.3}$$

for all $i \in \mathcal{V}$.

Definition 1.6 is given in terms of the local Markov property (1.3) which is the classic definition of a Markov chain. That is, each random variable is conditionally independent of all other random variables in the model given its 'neighbours' as represented in the conditional independence graph.

Theorem 1.3 Equivalency of the Markov Properties. *If the collection of random variables* $X_{\mathcal{V}}$ *has density function* $p(x_{\mathcal{V}}) > 0$ *for all* $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ *then the pairwise* (1.1)*, global* (1.2) *and local* (1.3) *Markov properties are equivalent.*

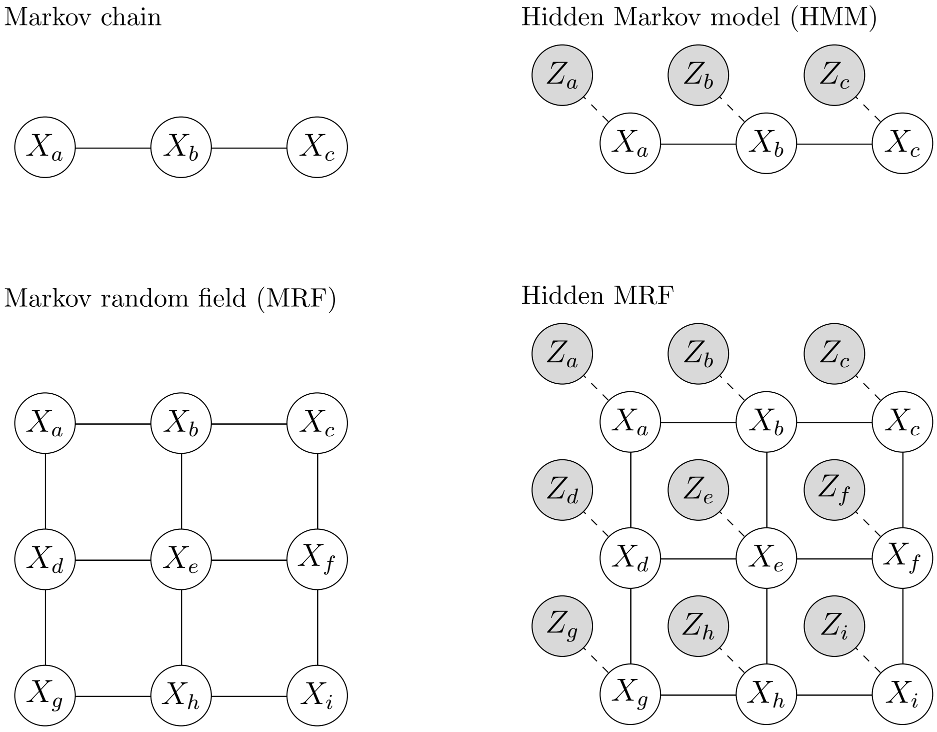
Proof of Theorem 1.3. The global Markov property clearly implies the local Markov property, which in turn implies the pairwise Markov property. From Lemma 1.2, all three properties are equivalent if the collection of random variables has a strictly positive density function. □

Hence a collection of random variables with strictly positive density function has a conditional independence graph for which Theorem 1.1 holds and hence is a Markov model. For all of the conditional independence graphs that we consider in the following we always assume that the corresponding collection of random variables has a strictly positive density function.

Definition 1.7 Trees. *An undirected graph* $\mathcal{G}$ *is a tree if there exists a unique path between each pair of vertices* $i, j \in \mathcal{V}$ *such that no vertex is contained in the path more than once.*

Figure 1.1 shows conditional independence graphs corresponding to a number of different Markov models. Note that the vertices are labelled by the corresponding random variables. The major differences between the models in Figure 1.1 are whether they are trees or not (above and below) and whether the Markov components are observed or 'hidden' (left and right). In the 'hidden' models it is taken that the random variables $Z_{\mathcal{V}}$ are observed while the random variables $X_{\mathcal{V}}$, constituting the Markov component of the model, are unobserved or missing. The defining difference between a Markov chain or hidden Markov model (HMM) as against a Markov random field (MRF) is that the latter is not a tree.

These differences determine the possible inference that can be achieved in the different models. It is the tree structure of both Markov chains and HMMs that allows for relatively straightforward inference of both the hidden variables and model parameters (Robinson 2012). In the following we consider inference for the variables in hidden MRFs (Chapter 2) and the model parameters of Gaussian MRFs (Chapter 3). Note that in all of these cases the conditional independence graph itself is assumed to be known. Furthermore the graphs are also assumed to be sparse, that is, there are few edges between the vertices compared to all those possible. This sparsity is implicit in the notion of a local neighbourhood and the local Markov property.



Markov chain

Hidden Markov model (HMM)

Markov random field (MRF)

Hidden MRF

Figure 1.1: Conditional independence graphs corresponding to different Markov models.

1.3 Hammersley-Clifford Theorem

Theorem 1.4 Hammersley-Clifford Theorem. *If the collection of random variables $X_{\mathcal{V}}$ has conditional independence graph $\mathcal{G}$ and density function $p(x_{\mathcal{V}}) > 0$ for all $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ then*

$$p(x_{\mathcal{V}}) \propto \prod_{\mathcal{C} \in \mathcal{M}} F_{\mathcal{C}}(x_{\mathcal{C}})$$

where $\mathcal{M}$ is the set of all maximal cliques in the graph and the factors $F_{\mathcal{C}}(x_{\mathcal{C}})$ are strictly positive.

Proof of Theorem 1.4. Details of a proof are given by Lauritzen (1996). □

The factors $F_{\mathcal{C}}(x_{\mathcal{C}})$ must be strictly positive but are otherwise arbitrary functions of their arguments. The Hammersley-Clifford Theorem is named after the unpublished paper by Hammersley & Clifford (1971).

Theorem 1.5 The Hammersley-Clifford Theorem and the Markov Properties. *If the collection of random variables $X_{\mathcal{V}}$ has density function $p(x_{\mathcal{V}}) > 0$ for all $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ then all three Markov properties and the Hammersley-Clifford Theorem are equivalent.*

Proof of Theorem 1.5. A proof is given by Lauritzen (1996). □

This equivalency and the fact that the Hammersley-Clifford Theorem was not originally published means that it appears in many different forms (Besag 1974, Lauritzen 1996, Cressie & Wikle 2015). We follow Lauritzen (1996) as this is the most convenient for us. It is the three Markov properties and the Hammersley-Clifford Theorem that allow for efficient inference so that Markov models can be gainfully used in practice.

Chapter 2

Inference for Hidden MRFs

Markov random fields (MRFs) have been extensively utilised in computer vision and in particular for digital image segmentation (Blake et al. 2011). Segmentation is the task of assigning a label to each pixel so that the image is partitioned into regions of pixels corresponding to mutually relevant features. Many different labelling problems can be approached as an inference problem in a suitably defined hidden MRF.

We consider inference for the random variables $X_{\mathcal{V}}$ in a hidden MRF. In general such inference in graphical models is NP-hard (Koller & Friedman 2009). However under certain conditions exact inference is computationally feasible for hidden MRFs while in other cases there exist efficient algorithms for approximate inference.

We first review the binary labelling problem as an inference problem for hidden MRFs. We present the energy minimisation framework and the use of graph cuts to find a computationally efficient and exact solution to the corresponding inference problem. Then we consider the multi-label problem and the established α-expansion algorithm for approximate energy minimisation.

2.1 Energy Function

For each $i \in \mathcal{V}$, let X_i be the random variable for the unobserved label we aim to infer and let Z_i be the random variable for the observed data. Let the collection of these random variables $\{X_{\mathcal{V}}, Z_{\mathcal{V}}\}$ be an hidden MRF with a conditional independence graph $\mathcal{G}$, an example of which is given in the bottom right of Figure 1.1. In this case, the model has a regular grid structure but this is not necessary in general, just that each Z_i has only X_i as a unique neighbour. We consider that the random variables are simply indexed by scalar values for generality and convenience.

The labelling task set as an inference problem is, given the hidden MRF $\{X_{\mathcal{V}}, Z_{\mathcal{V}}\}$ and having observed the data $Z_{\mathcal{V}} = z_{\mathcal{V}}$, find the maximum *a posteriori* labels

$$\begin{aligned}\hat{x}_{\mathcal{V}} &= \underset{x_{\mathcal{V}}}{\operatorname{argmax}}\ p(x_{\mathcal{V}}|z_{\mathcal{V}})\\ &= \underset{x_{\mathcal{V}}}{\operatorname{argmax}}\ \frac{p(x_{\mathcal{V}}, z_{\mathcal{V}})}{p(z_{\mathcal{V}})}\\ &= \underset{x_{\mathcal{V}}}{\operatorname{argmax}}\ p(x_{\mathcal{V}}, z_{\mathcal{V}}).\end{aligned}$$

For notational convenience let $\mathcal{G}$ be the conditional independence graph of only the random variables $X_{\mathcal{V}}$. That is, following our example graph in the bottom right of Figure 1.1, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the graph given in the bottom left of Figure 1.1. Recall we assume that all Markov models we consider have a strictly positive density function and hence we use the Hammersley-Clifford Theorem (Theorem 1.4) to write

$$p(x_{\mathcal{V}}, z_{\mathcal{V}}) \propto \prod_{i \in \mathcal{V}} F_i(x_i, z_i) \prod_{(i,j) \in \mathcal{E}} F_{(i,j)}(x_i, x_j)$$

where $F_i(x_i, z_i)$ and $F_{(i,j)}(x_i, x_j)$ are arbitrary, strictly positive functions of their arguments.

Definition 2.1 Energy Function. *Given the hidden MRF $\{X_{\mathcal{V}}, Z_{\mathcal{V}}\}$ with density function $p(x_{\mathcal{V}}, z_{\mathcal{V}}) > 0$ for all $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ and $z_{\mathcal{V}} \in \Omega_{Z_{\mathcal{V}}}$, and conditional independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ only corresponding to the random variables $X_{\mathcal{V}}$, the associated energy function is*

$$E(x_{\mathcal{V}}) = \sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \phi_{(i,j)}(x_i, x_j) \tag{2.1}$$

with energy potentials

$$\phi_i(x_i) = -\log(F_i(x_i, z_i)) \quad \textit{and} \quad \phi_{(i,j)}(x_i, x_j) = -\log(F_{(i,j)}(x_i, x_j)).$$

This is just rewriting the factorisation given by Theorem 1.4 for a hidden MRF although note that we have not explicitly included $z_{\mathcal{V}}$ in the energy function (2.1) as these terms are now accounted for in the subscripts of the energy potentials. Now we have that $p(x_{\mathcal{V}}, z_{\mathcal{V}}) \propto \exp\{-E(x_{\mathcal{V}})\}$ and so the maximum *a posteriori*/minimum energy labels are therefore

$$\hat{x}_{\mathcal{V}} = \underset{x_{\mathcal{V}}}{\operatorname{argmax}}\ p(x_{\mathcal{V}}|z_{\mathcal{V}}) = \underset{x_{\mathcal{V}}}{\operatorname{argmin}}\ E(x_{\mathcal{V}}).$$

The use of an energy function in computer vision goes back to Geman & Geman (1984), who established the connection with Gibbs distributions (Josiah Willard Gibbs (1839–1903) was an American mathematician and physicist). Geman & Geman (1984) also coined the term 'Gibbs sampling' in their approach to energy minimisation, which was superseded by the iterated conditional modes (ICM) algorithm in computer vision (Besag 1986).

The current use of energy minimisation in computer vision was brought about by the development of techniques based on graph cuts in the early 2000s, which allowed for much greater computational efficiency and improved performance as against ICM (Szeliski et al. 2008, Kappes et al. 2015). Note that in the hidden MRF given in Figure 1.1 the maximal cliques are pairs of vertices and for arbitrary graphs the maximal cliques may be larger than just pairs. However when the graph is still sparse a general approach is to just consider pairwise cliques in order to have an approximate energy function that can be optimised in practice (Boykov et al. 2001).

2.2 Graph Cuts

We first consider the binary label minimisation problem of (2.1), that is, where $x_i \in \{0,1\}$ for all $i \in \mathcal{V}$. Rewrite the energy potentials

$$\phi_i(x_i) = \begin{cases} \theta_{i;0} & \text{if } x_i = 0 \\ \theta_{i;1} & \text{if } x_i = 1 \end{cases} \qquad \text{and} \qquad \phi_{(i,j)}(x_i,x_j) = \begin{cases} \theta_{ij;00} & \text{if } x_i = x_j = 0 \\ \theta_{ij;01} & \text{if } x_i = 0, x_j = 1 \\ \theta_{ij;10} & \text{if } x_i = 1, x_j = 0 \\ \theta_{ij;11} & \text{if } x_i = x_j = 1 \end{cases}$$

for all $i \in \mathcal{V}$ and $(i,j) \in \mathcal{E}$. Hence we can rewrite the energy function (2.1) as

$$\begin{aligned} E(x_{\mathcal{V}}) = &\sum_{i\in\mathcal{V}} \Big(\theta_{i;1}x_i + \theta_{i;0}(1-x_i)\Big) \\ &+ \sum_{(i,j)\in\mathcal{E}} \Big(\theta_{ij;11}x_ix_j + \theta_{ij;10}x_i(1-x_j) + \theta_{ij;01}(1-x_i)x_j + \theta_{ij;00}(1-x_i)(1-x_j)\Big). \end{aligned} \tag{2.2}$$

The approach used to minimise the energy (2.2) is to recast the problem as one of finding the maximum flow in a suitably defined network. Then, provided the energy potentials satisfy certain conditions, the minimum energy solution exists and can be found using graph cuts. The following material on network flow is based on the presentation of Bondy & Murty (2008).

Definition 2.2 Directed Graph. *A directed graph $\overrightarrow{\mathcal{G}}$ is a set of vertices $\mathcal{V}$ and a set of edges $\overrightarrow{\mathcal{E}}$, written $\overrightarrow{\mathcal{G}} = (\mathcal{V}, \overrightarrow{\mathcal{E}})$, where $\overrightarrow{\mathcal{E}}$ is a subset of the set $\mathcal{V} \times \mathcal{V}$ of pairs of distinct vertices.*

A directed graph is an undirected graph (Definition 1.2) without the condition that if $(i,j) \in \overrightarrow{\mathcal{E}}$ then $(j,i) \in \overrightarrow{\mathcal{E}}$ for all $i,j \in \mathcal{V}$. That is, it may be the case that only one of the two possible edges between vertices i and j are in the edge set $\overrightarrow{\mathcal{E}}$. Hence there is a direction to each edge that allows for a different definition of neighbouring vertices.

Definition 2.3 Directed Neighbours. *In a directed graph $\overrightarrow{\mathcal{G}}$, the directed neighbours of a vertex i are the members of the sets $\overrightarrow{\mathcal{V}}(i) = \{j \in \mathcal{V} \mid (i,j) \in \overrightarrow{\mathcal{E}}\}$ and $\overleftarrow{\mathcal{V}}(i) = \{j \in \mathcal{V} \mid (j,i) \in \overrightarrow{\mathcal{E}}\}$.*

Definition 2.4 Network. *A network $\mathcal{N}$ is a directed graph $\overrightarrow{\mathcal{G}}$ with specified vertices $s, t \in \mathcal{V}$, known as the source and the sink respectively, and associated capacity function $c : \mathcal{V} \times \mathcal{V} \to [0, \infty)$ where $c(i, j) = 0$ if $(i, j) \notin \overrightarrow{\mathcal{E}}$, written $\mathcal{N} = (\mathcal{V}, \overrightarrow{\mathcal{E}}, c)$.*

A network is a particular type of directed graph where each edge has an associated capacity allowing for a flow between the specially designated source s and sink t vertices to be defined.

Definition 2.5 Flow in a Network. *A flow in a network $\mathcal{N}$ is a function $f : \mathcal{V} \times \mathcal{V} \to [0, \infty)$ such that $0 \leq f(i, j) \leq c(i, j)$ for all $(i, j) \in \overrightarrow{\mathcal{E}}$ and*

$$\sum_{j \in \overrightarrow{\mathcal{V}}(i)} f(i, j) = \sum_{j \in \overleftarrow{\mathcal{V}}(i)} f(j, i)$$

for all $i \in \mathcal{V} \backslash \{s, t\}$. The value of the flow is

$$val(f) = \sum_{j \in \overrightarrow{\mathcal{V}}(s)} f(s, j) - \sum_{j \in \overleftarrow{\mathcal{V}}(s)} f(j, s). \tag{2.3}$$

The flow on an edge is less than or equal to the capacity of that edge and for any network there exists a zero flow where $f(i, j) = 0$ for all $(i, j) \in \overrightarrow{\mathcal{E}}$. Flow is conserved for all vertices except the source s and sink t. That is, the flow into any vertex $i \in \mathcal{V} \backslash \{s, t\}$ (the sum of the flows on the edges $(j, i) \in \overrightarrow{\mathcal{E}}$) is equal to the flow out of i (the sum of the flows on the edges $(i, j) \in \overrightarrow{\mathcal{E}}$).

Lemma 2.1 Value of the Flow. *For a flow f in a network $\mathcal{N}$ the value of the flow is equivalently*

$$val(f) = \sum_{i \in \overleftarrow{\mathcal{V}}(t)} f(i, t) - \sum_{i \in \overrightarrow{\mathcal{V}}(t)} f(t, i). \tag{2.4}$$

Proof of Lemma 2.1. A proof is given by Bondy & Murty (2008). □

That is, the value of the flow is defined to be the flow out of the source minus the flow into the source (2.3), which is equal to the flow into the sink minus the flow out of the sink (2.4).

Definition 2.6 Graph Cut. *For a subset of vertices $\mathcal{S} \subset \mathcal{V}$ in a network $\mathcal{N}$ with $s \in \mathcal{S}$ and $t \notin \mathcal{S}$, a cut is the set of edges*

$$[\mathcal{S}] = \{(i, j) \in \overrightarrow{\mathcal{E}} \mid i \in \mathcal{S}, j \notin \mathcal{S}\}$$

and the value of the cut is

$$val[\mathcal{S}] = \sum_{(i,j) \in [S]} c(i, j).$$

A cut in a network $\mathcal{N}$ is the set of edges from the subset $\mathcal{S}$ to its complement that effectively 'cut' the network in two and hence 'graph cuts'.

Definition 2.7 Maximum Flow and Minimum Cut. *A cut $[\mathcal{S}]$ is a minimum cut if there exists no other cut $[\mathcal{S}^*]$ such that $val[\mathcal{S}^*] < val[\mathcal{S}]$. A flow f is a maximum flow if there exists no other flow f^* such that $val(f) < val(f^*)$.*

Theorem 2.2 Maximum Flow/Minimum Cut. *For a network $\mathcal{N}$, the value of the maximum flow is equal to the value of the minimum cut.*

Proof of Theorem 2.2. A proof is given by Bondy & Murty (2008). □

Given a network $\mathcal{N}$, the problem is to find a flow f and a cut $[\mathcal{S}]$ such that $val(f) = val[\mathcal{S}]$. That is, the value of the flow is maximised and the value of the cut is minimised. There exist classic algorithms for this task (Ford & Fulkerson 1956).

We will now consider how the maximum flow/minimum cut framework can be applied to the energy minimisation problem with a binary label energy function (2.2). First, we create a suitable network using the conditional independence graph of $X_{\mathcal{V}}$.

Algorithm 2.1 Network Creation Algorithm. *Given the collection of random variables $X_{\mathcal{V}}$ with conditional independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, create a corresponding network $\mathcal{N}^* = (\mathcal{V}^*, \overrightarrow{\mathcal{E}}^*, c^*)$ by defining:*

1. *The set of vertices $\mathcal{V}^* = \mathcal{V} \cup \{s, t\}$ where s is the source and t is the sink.*

2. *The edge set*

$$\overrightarrow{\mathcal{E}}^* = \Big\{(s,i), (i,s), (t,i), (i,t) \mid i \in \mathcal{V}\Big\} \cup \Big\{(i,j), (j,i) \mid (i,j) \in \mathcal{E}\Big\}.$$

3. *The capacity function*

$$c^*(i,j) = \omega_{ij} \geq 0$$

for all $(i,j) \in \overrightarrow{\mathcal{E}}^*$ with $c^*(i,s) = 0$ for all $i \in \mathcal{V}$ and $c^*(t,j) = 0$ for all $j \in \mathcal{V}$.

Consider that any cut $[\mathcal{S}]$ in the network $\mathcal{N}^*$ can be viewed as a binary labelling

$$x_i = \begin{cases} 0 & \text{if } i \in \mathcal{S} \\ 1 & \text{if } i \notin \mathcal{S} \end{cases} \tag{2.5}$$

for all $i \in \mathcal{V}$ and consider the energy function

$$E^*(x_{\mathcal{V}}) = \sum_{i \in \mathcal{V}} \Big(\omega_{si} x_i + \omega_{it}(1 - x_i)\Big) + \sum_{(i,j) \in \mathcal{E}} \Big(\omega_{ij}(1 - x_i)x_j + \omega_{ji} x_i (1 - x_j)\Big). \tag{2.6}$$

Theorem 2.3 Graph Cut Energy. *Given the collection of random variables $X_{\mathcal{V}}$ with conditional independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the value of a cut $[\mathcal{S}]$ in the corresponding network $\mathcal{N}^*$ is given by* (2.6) *when utilising the labelling* (2.5). *That is, $val[\mathcal{S}] = E^*(x_{\mathcal{V}})$.*

Proof of Theorem 2.3.

$$\begin{aligned}
val[\mathcal{S}] &= \sum_{(i,j)\in[S]} \omega_{ij} \\
&= \sum_{(i,j)\in\overrightarrow{\mathcal{E}}^*} \omega_{ij} I\{i \in \mathcal{S}, j \notin \mathcal{S}\} \\
&= \sum_{(s,i)\in\overrightarrow{\mathcal{E}}^*} \omega_{si} I\{i \notin \mathcal{S}\} + \sum_{(i,t)\in\overrightarrow{\mathcal{E}}^*} \omega_{it} I\{i \in \mathcal{S}\} + \sum_{(i,j)\in\overrightarrow{\mathcal{E}}^*} \omega_{ij} I\{i \in \mathcal{S}\backslash s, j \notin \mathcal{S} \cup t\} \\
&= \sum_{i\in\mathcal{V}} \Big(\omega_{si} I\{i \notin \mathcal{S}\} + \omega_{it} I\{i \in \mathcal{S}\}\Big) + \sum_{(i,j)\in\mathcal{E}} \Big(\omega_{ij} I\{i \in \mathcal{S}, j \notin \mathcal{S}\} + \omega_{ji} I\{i \notin \mathcal{S}, j \in \mathcal{S}\}\Big) \quad \square
\end{aligned}$$

Hence the energy (2.6) is minimised when we find the maximum flow/minimum cut in the network $\mathcal{N}^*$ and use the labelling (2.5). The original binary label energy function (2.2) is equivalent to the graph cut energy function (2.6) when

$$\theta_{i;1} = \omega_{si}, \quad \theta_{i;0} = \omega_{it}, \quad \theta_{ij;01} = \omega_{ij}, \quad \theta_{ij;10} = \omega_{ji} \quad \text{and} \quad \theta_{ij;11} = \theta_{ij;00} = 0. \tag{2.7}$$

Hence, under certain conditions on the potentials, we are able to reformulate the binary label energy minimisation problem into a network flow problem and find the minimum energy solution using graph cuts. Let us briefly consider that the potentials simply need to be submodular.

Definition 2.8 Submodular Potentials. *Let the collection of random variables $X_{\mathcal{V}}$ have conditional independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and energy function* (2.2). *The corresponding energy potentials are submodular if*

$$\theta_{ij;00} + \theta_{ij;11} \leq \theta_{ij;10} + \theta_{ij;01}$$

for all $(i, j) \in \mathcal{E}$.

For the binary label energy function (2.2), constants can be added and subtracted to the potentials so that the energy function remains unchanged, that is the function still gives the same output for the same input. This is known as a reparameterisation of the energy function and if the potentials are submodular, then the energy function (2.2) can be reparameterised so that the potentials are in the form (2.7) in a finite series of basic prescribed operations (Kolmogorov & Zabih 2004). There are standard energy minimisation algorithms based on graph cuts for the binary label, submodular potential problem (Boykov & Kolmogorov 2004). Below, we consider the case where there are more than 2 labels.

2.3 Alpha-Expansion Algorithm

Now consider the case where we have more than 2 labels, that is $x_i \in \mathcal{L}$ for all $i \in \mathcal{V}$ where $\mathcal{L}$ is a set of more than 2 elements. Consider the original energy function (2.1) and rewrite

$$\phi_i(x_i = l) = \theta_{i;l} \quad \text{and} \quad \phi_{(i,j)}(x_i = l, x_j = k) = \theta_{ij;lk}$$

for all $i \in \mathcal{V}$, $(i,j) \in \mathcal{E}$ and $l, k \in \mathcal{L}$ so that

$$E(x_\mathcal{V}) = \sum_{i\in\mathcal{V}}\sum_{l\in\mathcal{L}} \theta_{i;l} I\{x_i = l\} + \sum_{(i,j)\in\mathcal{E}}\sum_{l\in\mathcal{L}}\sum_{k\in\mathcal{L}} \theta_{ij;lk} I\{x_i = l, x_j = k\}. \tag{2.8}$$

Clearly the multi-label energy function (2.8) is equal to the binary label energy function (2.2) if $\mathcal{L} = \{0,1\}$. Minimising the multi-label energy function (2.8) is NP-hard (Boykov et al. 2001). However, we can approximate the minimum energy solution using an algorithm that is based on graph cuts (Boykov et al. 2001, Boykov & Kolmogorov 2004, Kolmogorov & Zabih 2004). The main idea is to recast the multi-label problem in terms of a binary label problem by considering an initial labelling and then a binary choice for each variable: either change label or stay the same. In order to properly determine the problem, 'changing label' needs to be defined and one option is to only allow a change to a specified label, $\alpha \in \mathcal{L}$. The following material is based on the presentation of Koller & Friedman (2009).

Let the initial labelling be $x^*_\mathcal{V}$. We construct a binary label problem with dummy variables $Y_\mathcal{V}$ such that $y_i \in \{0,1\}$ for all $i \in \mathcal{V}$. Consider that from the initial labelling $x^*_\mathcal{V}$, the dummy variables $Y_\mathcal{V}$ and a specified label $\alpha \in \mathcal{L}$, we obtain a new labelling $\hat{x}_\mathcal{V}$ given by

$$\hat{x}_i = \begin{cases} x^*_i & \text{if } y_i = 0 \\ \alpha & \text{if } y_i = 1 \end{cases}$$

for all $i \in \mathcal{V}$. Hence each variable either keeps its initial label x^*_i or changes to α depending on the value of y_i. By substituting $\hat{x}_\mathcal{V}$ into (2.8) we obtain the equivalent energy function

$$\tilde{E}(y_\mathcal{V}) = \sum_{i\in\mathcal{V}} \tilde{\phi}_i(y_i) + \sum_{(i,j)\in\mathcal{E}} \tilde{\phi}_{(i,j)}(y_i, y_j) \tag{2.9}$$

where

$$\tilde{\phi}_i(y_i) = \begin{cases} \theta_{i;x^*_i} & \text{if } y_i = 0 \\ \theta_{i;\alpha} & \text{if } y_i = 1 \end{cases} \quad \text{and} \quad \tilde{\phi}_{(i,j)}(y_i, y_j) = \begin{cases} \theta_{ij;x^*_i x^*_j} & \text{if } y_i = y_j = 0 \\ \theta_{ij;x^*_i \alpha} & \text{if } y_i = 0, y_j = 1 \\ \theta_{ij;\alpha x^*_j} & \text{if } y_i = 1, y_j = 0 \\ \theta_{ij;\alpha\alpha} & \text{if } y_i = y_j = 1. \end{cases}$$

That is, we have a binary label energy function (2.9) for the binary label problem for $Y_{\mathcal{V}}$. By construction, $\tilde{E}(y_{\mathcal{V}}) = E(\hat{x}_{\mathcal{V}})$ and hence finding the optimal labels $\hat{y}_{\mathcal{V}}$ is equivalent to finding the optimal labels $\hat{x}_{\mathcal{V}}$ in the restricted space of only allowing each variable to keep its initial label or change to α. This is a constrained energy minimisation problem where we are only (potentially) increasing the number of vertices with the label α and hence known as ‘alpha-expansion’. In practice, all of the possible labels are considered iteratively and multiple times.

Algorithm 2.2 Alpha-Expansion Algorithm. *Given the collection of random variables $X_{\mathcal{V}}$ with conditional independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, energy function* (2.8)*, initial labelling $x^*_{\mathcal{V}}$ and specified label $\alpha \in \mathcal{L}$,*

1. *Construct the binary label problem for the dummy variables $Y_{\mathcal{V}}$ as outlined above.*
2. *Solve the binary label problem for $Y_{\mathcal{V}}$ using graph cuts as outlined in Section 2.2.*
3. *Update the labelling and repeat for another label until some convergence criteria is satisfied.*

Recall that in order to use graph cuts for the binary label problem, the energy potentials must be submodular. The potentials for the above binary label energy (2.9) are submodular when

$$\theta_{ij;\alpha\alpha} + \theta_{ij;x_i^* x_j^*} \leq \theta_{ij;x_i^*\alpha} + \theta_{ij;\alpha x_j^*}$$

for all $(i, j) \in \mathcal{E}$. This condition is satisfied when the energy potentials are a metric.

Definition 2.9 Semi-metric and Metric. *A function $D : \mathcal{V} \times \mathcal{V} \to [0, \infty)$ is a semi-metric if*

$$D(i, j) = 0 \Leftrightarrow i = j \quad \textit{and} \quad D(i, j) = D(j, i)$$

for all $i, j \in \mathcal{V}$, and is a metric if in addition

$$D(i, k) \leq D(i, j) + D(j, k)$$

for all $i, j, k \in \mathcal{V}$.

However, even when the energy potentials are only a semi-metric, the α-expansion algorithm may still be gainfully used in practice (Boykov et al. 2001).

Chapter 3

Inference for Gaussian MRFs

Gaussian Markov random fields (MRFs) are multivariate Gaussian distributions parameterised by a mean vector and a correlation matrix where there is a one-to-one relationship between the graphical representation of the model and the inverse correlation (precision) matrix (Rue & Held 2005). They are used in particular as the spatial random effects component within larger hierarchical models for the analysis of spatial data (Banerjee et al. 2014, Cressie & Wikle 2015).

We are interested in inference for the parameters of a Gaussian MRF with a spatial interpretation through the precision matrix. This inference is considered in a Bayesian framework where the model parameters are considered as random variables and the posterior density of the parameters is proportional to the likelihood of the model multiplied by a prior density of the parameters.

We first consider multivariate Gaussian distributions, their precision matrices and how this leads to a natural definition of Gaussian MRFs. Then we consider how Gaussian MRFs can be defined in terms of full conditional distributions which allows for the construction of the precision matrix of the whole model. We then outline finding the posterior densities for the model parameters which can be used for sampling and inference in a Bayesian framework.

3.1 Precision Matrix

In the following we consider vectors rather than sets of random variables but still use set notation in the subscripts, $X_{\mathcal{V}} = [X_1, X_2, \ldots, X_n]^T$ when the meaning is unambiguous. For a matrix A we write the individual elements of the matrix as A_{ij} and use standard vector and matrix notation.

Definition 3.1 Symmetric Matrix. *A square matrix A is symmetric if and only if $A^T = A$.*

Definition 3.2 Positive Definite Matrix. *A square matrix A is positive definite if and only if $v^T Av > 0$ for all vectors $v \neq 0$.*

Lemma 3.1 Inverse of a Symmetric Positive Definite Matrix. *If A is symmetric positive definite then A^{-1} is symmetric positive definite.*

Proof of Lemma 3.1. Symmetry is clear and for any vector $u \neq 0$ take $v = A^{-1}u$. Since A is invertible, $v \neq 0$ and hence $v^T Av > 0$ implies $u^T A^{-1} u > 0$. □

Definition 3.3 Multivariate Gaussian Distribution. *The collection of random variables $X_{\mathcal{V}} = [X_1, X_2, \ldots, X_n]^T$ has an n-variate Gaussian distribution with mean vector $\mu \in \mathbb{R}^n$, symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ and support $\Omega_{X_{\mathcal{V}}} = \mathbb{R}^n$, written $X_{\mathcal{V}} \sim N(\mu, \Sigma)$, if and only if its density function has the form*

$$p(x_{\mathcal{V}}) \propto |\Sigma|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}(x_{\mathcal{V}} - \mu)^T \Sigma^{-1} (x_{\mathcal{V}} - \mu)\Big\}.$$

Gaussian distributions are named after German mathematician Carl Friedrich Gauss (1777–1855). Consider the matrix $Q = \Sigma^{-1}$ known as the precision matrix and note that since Σ is symmetric positive definite, Q is symmetric positive definite (Lemma 3.1). There is a direct connection between the Markov properties of the model and the precision matrix Q. The content of this section is based on the presentation of Rue & Held (2005).

Theorem 3.2 Conditional Independence and the Precision Matrix. *Let the collection of random variables $X_{\mathcal{V}}$ have a multivariate Gaussian distribution with precision matrix $Q = \Sigma^{-1}$. Then*

$$Q_{ij} = 0 \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j | X_{\mathcal{V} \setminus \{i,j\}}$$

for all $i \neq j \in \mathcal{V}$.

Proof of Theorem 3.2. A proof is given by Rue & Held (2005). □

That is, for any multivariate Gaussian model, we have a connection between the form of the precision matrix and the conditional independence properties of the model. The covariance matrix Σ gives information about the marginal dependence structure of the model whereas the precision matrix Q gives information about the conditional independence structure of the model. This leads to a natural definition of a Gaussian MRF through the precision matrix.

Definition 3.4 Gaussian MRF. *The collection of random variables $X_{\mathcal{V}}$ is a Gaussian MRF with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean vector $\mu \in \mathbb{R}^n$, symmetric positive definite precision matrix $Q \in \mathbb{R}^{n \times n}$ and support $\Omega_{X_{\mathcal{V}}} = \mathbb{R}^n$ if and only if its density function has the form*

$$p(x_{\mathcal{V}}) \propto |Q|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}(x_{\mathcal{V}} - \mu)^T Q (x_{\mathcal{V}} - \mu)\Big\}$$

and

$$Q_{ij} = 0 \quad \Leftrightarrow \quad (i,j) \notin \mathcal{E}$$

for all $i \neq j \in \mathcal{V}$.

It is clear that the three Markov properties (pairwise (1.1), global (1.2) and local (1.3)) are all equivalent for Gaussian MRFs since $p(x_{\mathcal{V}}) > 0$. Hence a Gaussian MRF is defined through the precision matrix Q which is one-to-one with a graphical representation of the model. The non-zero pattern of Q determines the conditional independence graph $\mathcal{G}$. That is, if $Q_{ij} = 0$ then there is no edge between vertices i and j in the conditional independence graph and hence X_i and X_j are conditionally independent given $X_{\mathcal{V}\setminus\{i,j\}}$. So we can simply read from the elements of Q whether two variables in the model are conditionally independent.

Now Q is generally taken as a sparse matrix (most of its elements are zero) while Σ may be very dense (most elements are non-zero). The Markov structure of a Gaussian MRF allows for efficient inference due to a sparse precision matrix, which corresponds to a sparse conditional independence graph (Chapter 1). Although it is possible to define reasonable dependence structures through a covariance function, there are no guarantees on the sparsity of the inverse covariance matrix and it is non-trivial to construct reasonable dependence structures ensuring sparsity (Rue & Held 2005). So rather than considering covariance matrices, we consider conditional independencies and the construction of reasonable and useful precision matrices.

3.2 Conditional Autoregressive Models

To utilise a Gaussian MRF in practice, instead of specifying the full covariance matrix for every pair of variables, we consider the 'local' behaviour for each variable and the full conditional distributions, as introduced by Besag (1974). That is, rather than defining a joint density $p(x_{\mathcal{V}})$, we consider the much simpler task of defining the conditional densities $p(x_i|x_{\mathcal{V}\setminus i})$ for all $i \in \mathcal{V}$.

Definition 3.5 Conditional Autoregressive Model. *The collection of random variables $X_{\mathcal{V}}$ is a conditional autoregressive (CAR) model with parameters $\omega_{ij} \in \mathbb{R}$ and $\sigma_i^2 > 0$ if and only if the collection of full conditionals distributions are*

$$X_i | X_{\mathcal{V}\setminus i} \sim N\Big(\sum_{j \in \mathcal{V}\setminus i} \omega_{ij} X_j, \sigma_i^2 \Big) \tag{3.1}$$

for all $i \in \mathcal{V}$.

That is, each random variable has a conditional Gaussian distribution given all the other variables with mean, a weighted sum of the other variables and a given variance. We now consider finding the joint distribution of the CAR model.

Lemma 3.3 Brook's Lemma. *Consider the collection of random variables $X_{\mathcal{V}}$ with density function $p(x_{\mathcal{V}}) > 0$ for all $x_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$. For any $x_{\mathcal{V}}, x'_{\mathcal{V}} \in \Omega_{X_{\mathcal{V}}}$ we have*

$$\begin{aligned} \frac{p(x_{\mathcal{V}})}{p(x'_{\mathcal{V}})} &= \prod_{i=1}^{n} \frac{p(x_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}{p(x'_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)} && (3.2) \\ &= \prod_{i=1}^{n} \frac{p(x_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n)}{p(x'_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n)}. && (3.3) \end{aligned}$$

Proof of Lemma 3.3. A proof is given by Rue & Held (2005). □

Lemma 3.3 (Brook 1964) allows for the construction of the joint density of a CAR model from the full conditionals. However, there are conditions on the full conditionals in order that the density is proper. It is not the case that any arbitrary set of full conditionals will determine a proper distribution (Gelfand & Vounatsou 2003).

Theorem 3.4 Joint Density of a CAR Model. *If the collection of random variables $X_{\mathcal{V}}$ has full conditional distributions*

$$X_i | X_{\mathcal{V}\setminus i} \sim N\Big(\phi \sum_{j \in \mathcal{V}\setminus i} \frac{e_{ij}}{\partial_i} X_j, \frac{\sigma^2}{\partial_i} \Big) \tag{3.4}$$

for all $i \in \mathcal{V}$ where $e_{ij} \geq 0$, $e_{ij} = e_{ji}$, $e_{ii} = 0$, $\partial_i = \sum_{j \in \mathcal{V}} e_{ij}$ and with parameters $\phi \in (-1, 1)$ and $\sigma^2 > 0$, the joint distribution of the model is

$$X_{\mathcal{V}} \sim N(0, \sigma^2 (D(I - \phi W))^{-1}) \tag{3.5}$$

where $D = diag(\partial_i)$ and W is the symmetric matrix with entries e_{ij}.

Proof of Theorem 3.4. First consider that for a zero-mean multivariate Gaussian distribution we have

$$\begin{aligned}\log(p(x_{\mathcal{V}})) &\propto -\frac{1}{2}x_{\mathcal{V}}^T Q x_{\mathcal{V}}\\ &= -\frac{1}{2}\Big(\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}} Q_{ij}x_i x_j\Big)\\ &= -\frac{1}{2}\Big(\sum_{i\in\mathcal{V}} Q_{ii}x_i^2 + \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}\setminus i} Q_{ij}x_i x_j\Big). \qquad (3.6)\end{aligned}$$

Consider the full conditionals as given as (3.1). We have the density functions

$$p(x_i|x_{\mathcal{V}\setminus i}) \propto \frac{1}{\sigma_i}\exp\Big\{-\frac{1}{2\sigma_i^2}(x_i - \sum_{j\in\mathcal{V}\setminus i}\omega_{ij}x_j)^2\Big\}$$

for all $i \in \mathcal{V}$. Using (3.2) we can write

$$\begin{aligned}\frac{p(x_{\mathcal{V}})}{p(0)} &= \prod_{i=1}^{n}\frac{\frac{1}{\sigma_i}\exp\Big\{-\frac{1}{2\sigma_i^2}(x_i - \sum_{j=1}^{i-1}\omega_{ij}x_j)^2\Big\}}{\frac{1}{\sigma_i}\exp\Big\{-\frac{1}{2\sigma_i^2}(\sum_{j=1}^{i-1}\omega_{ij}x_j)^2\Big\}}\\ &= \prod_{i=1}^{n}\exp\Big\{-\frac{1}{2\sigma_i^2}\Big(x_i - \sum_{j=1}^{i-1}\omega_{ij}x_j\Big)^2 + \frac{1}{2\sigma_i^2}\Big(\sum_{j=1}^{i-1}\omega_{ij}x_j\Big)^2\Big\}\\ &= \prod_{i=1}^{n}\exp\Big\{-\frac{1}{2\sigma_i^2}\Big(x_i^2 - 2\sum_{j=1}^{i-1}\omega_{ij}x_i x_j\Big)\Big\}.\end{aligned}$$

Hence

$$\log\Big(\frac{p(x_{\mathcal{V}})}{p(0)}\Big) = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}x_i^2 - \sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{\omega_{ij}}{\sigma_i^2}x_i x_j. \qquad (3.7)$$

Now we use (3.3) to obtain the corresponding expression

$$\log\Big(\frac{p(x_{\mathcal{V}})}{p(0)}\Big) = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}x_i^2 - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{\omega_{ij}}{\sigma_i^2}x_i x_j. \qquad (3.8)$$

Since (3.7) and (3.8) are equal,

$$\log(p(x_{\mathcal{V}})) \propto -\frac{1}{2}\Big(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}x_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}\frac{\omega_{ij}}{\sigma_i^2}x_i x_j\Big).$$

Comparing this to the form of the zero-mean multivariate Gaussian distribution (3.6), we have $Q_{ii} = \frac{1}{\sigma_i^2}$ and $Q_{ij} = \frac{\omega_{ij}}{\sigma_i^2}$ with $\frac{\omega_{ij}}{\sigma_i^2} = \frac{\omega_{ji}}{\sigma_j^2}$. Finally we write $\omega_{ij} = \phi\frac{e_{ij}}{\partial_i}$ and $\sigma_i^2 = \frac{\sigma^2}{\partial_i}$. □

Hence the CAR model (3.5) is a Gaussian MRF with $Q = \frac{1}{\sigma^2}D(I - \phi W)$. There is a one-to-one connection between the adjacency matrix W and the corresponding conditional independence graph since $e_{ij} = 0$ if and only if $Q_{ij} = 0$ and $e_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$. That is, the sum over $j \in \mathcal{V}\backslash i$ in (3.4) may as well be over $j \in \nu(i)$ due to the zero pattern of W. Note that in general the adjacency matrix W is taken as a sparse matrix and hence Q is sparse. Recall that this is the point of constructing the precision matrix in this way and is necessary for efficient inference.

We additionally have interpretations for the parameters when considering $X_{\mathcal{V}}$ as spatially embedded variables through the adjacency matrix W. If $\phi = 0$ then the variables $X_{\mathcal{V}}$ have a multivariate Gaussian distribution with a diagonal covariance matrix and hence the marginal distributions are independent. A value of $\phi > 0$ results in spatial autocorrelation between neighbouring variables while a value of $\phi < 0$ results in spatial anti-autocorrelation (easiest to see in the full conditional distributions (3.4)). Hence ϕ is known as the spatial autocorrelation parameter while σ^2 is a general variance term.

3.3 Posterior Densities

We consider inference for the parameters ϕ and σ^2 of the CAR model (3.5). Although CAR models are most often used as a spatial random effects component within larger hierarchical models in the analysis of spatial data (Banerjee et al. 2014, Cressie & Wikle 2015), inference for the parameters can also be considered when the models are used directly (Bell & Broemeling 2000, De Oliveira 2012, Ren & Sun 2013). This section is primarily based on the work of Bell & Broemeling (2000).

Theorem 3.5 Bayes' Theorem. *For random variable Y with associated parameters θ and taking the parameters themselves as random variables we have*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta).$$

Proof of Theorem 3.5. This is the definition of conditional probability density, $p(x|z) = \frac{p(x,z)}{p(z)}$. □

That is, the posterior density of the parameters is proportional to the likelihood of the model and a prior density of the parameters. This is the cornerstone of 'Bayesian inference' named after English mathematician Thomas Bayes (1701–1761). Rather than point estimates of model parameters, treating the parameters as random variables results in a posterior distribution that if not in an analytical form can usually be numerically sampled from.

Consider the CAR model with joint density given by (3.5) and parameters ϕ and σ^2. We use the re-parameterisation $\sigma^2 = \tau^{-1} > 0$ so that the model density can be written

$$p(x_{\mathcal{V}}|\phi,\tau) \propto |\tau B(\phi)|^{\frac{1}{2}} \exp\Big\{-\frac{\tau}{2}x_{\mathcal{V}}^T B(\phi) x_{\mathcal{V}}\Big\}$$

where $B(\phi) = D(I - \phi W)$. We set priors $p(\phi) \propto$ constant and $p(\tau) \propto \tau^{-1}$. We assume *a priori* that ϕ and τ are independent, that is $p(\phi,\tau) = p(\phi)p(\tau) = \tau^{-1}$. Hence using Bayes' Theorem (Theorem 3.5) the joint posterior density is

$$\begin{aligned} p(\phi,\tau|X_{\mathcal{V}} = x_{\mathcal{V}}) &\propto p(x_{\mathcal{V}}|\phi,\tau)p(\phi,\tau) \\ &\propto |\tau B(\phi)|^{\frac{1}{2}} \exp\Big\{-\frac{\tau}{2}x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}}\Big\}\tau^{-1} \\ &= |B(\phi)|^{\frac{1}{2}}\tau^{\frac{n}{2}-1} \exp\Big\{-\frac{\tau}{2}x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}}\Big\}. \end{aligned} \tag{3.9}$$

Definition 3.6 Gamma Distribution. *The random variable Y has a Gamma distribution with parameters $\alpha > 0$, $\beta > 0$ and support $\Omega_Y = (0,\infty)$, written $Y \sim \Gamma(\alpha,\beta)$, if and only if its density function is*

$$p(y) = \frac{1}{G(\alpha)}\beta^\alpha y^{\alpha-1}\exp\{-y\beta\} \tag{3.10}$$

where $G(z) = \int_0^\infty x^{z-1}\exp\{-x\}dx$.

Since (3.10) is a properly scaled density function we have the identity

$$\int_0^\infty \frac{1}{G(\alpha)}\beta^\alpha y^{\alpha-1}\exp\{-y\beta\}dy = 1. \tag{3.11}$$

Now for fixed ϕ it can be seen from (3.9) that τ has a Gamma distribution. That is,

$$\tau|\{\phi, X_{\mathcal{V}} = x_{\mathcal{V}}\} \sim \Gamma\Big(\frac{n}{2}, \frac{1}{2}x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}}\Big).$$

In order to find the posterior density of ϕ we integrate $p(\phi,\tau|X_{\mathcal{V}} = x_{\mathcal{V}})$ with respect to τ using the identity (3.11) to obtain

$$\begin{aligned} p(\phi|X_{\mathcal{V}} = x_{\mathcal{V}}) &= \int_0^\infty p(\phi,\tau|X_{\mathcal{V}} = x_{\mathcal{V}})d\tau \\ &= |B(\phi)|^{\frac{1}{2}}\int_0^\infty \tau^{\frac{n}{2}-1}\exp\Big\{-\frac{\tau}{2}x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}}\Big\}d\tau \\ &= |B(\phi)|^{\frac{1}{2}}\Big(\frac{1}{2}x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}}\Big)^{-\frac{n}{2}} G\Big(\frac{n}{2}\Big) \\ &\propto |B(\phi)|^{\frac{1}{2}}(x_{\mathcal{V}}^T B(\phi)x_{\mathcal{V}})^{-\frac{n}{2}}. \end{aligned} \tag{3.12}$$

Hence in order to sample from the joint posterior $p(\phi,\tau|X_{\mathcal{V}} = x_{\mathcal{V}}) = p(\tau|\phi, X_{\mathcal{V}} = x_{\mathcal{V}})p(\phi|X_{\mathcal{V}} = x_{\mathcal{V}})$ we first sample from $p(\phi|X_{\mathcal{V}} = x_{\mathcal{V}})$ and then sample from $p(\tau|\phi, X_{\mathcal{V}} = x_{\mathcal{V}})$ for fixed ϕ.

We consider some final details about sampling using the posterior density function (3.12). Firstly, in applications with large n, the calculation of the determinant of $B(\phi)$ can easily become computationally infeasible while the exponent $-\frac{n}{2}$ will result in computational zeros. For the later problem we consider log space (scaling and then converting back as necessary) and write

$$\log(p(\phi|X_{\mathcal{V}} = x_{\mathcal{V}})) = \log(|B(\phi)|^{\frac{1}{2}}) - \frac{n}{2}\log\left(x_{\mathcal{V}}^T B(\phi) x_{\mathcal{V}}\right).$$

Hence the exponent is no longer a problem but a potential problem with the calculation of the determinant remains.

Definition 3.7 Cholesky Decomposition. *Every symmetric positive definite matrix A can be written as*

$$A = LL^T$$

where L is a unique lower triangular matrix with positive diagonal entries.

We use this decomposition, named after French mathematician André-Louis Cholesky (1875–1918), to rewrite

$$\begin{aligned}\log(|A|^{\frac{1}{2}}) &= \log(|LL^T|^{\frac{1}{2}})\\ &= \log((|L||L|)^{\frac{1}{2}})\\ &= \log(|L|)\\ &= \log\left(\prod_{i=1}^{n} L_{ii}\right)\\ &= \sum_{i=1}^{n}\log(L_{ii}).\end{aligned}$$

Now $B(\phi) = \tau^{-1}Q$ is symmetric positive definite and so by making use of the Cholesky decomposition the calculation of $\log(|B(\phi)|^{\frac{1}{2}})$ is greatly simplified. Although finding the Cholesky decomposition may be difficult in general, since Q is sparse it is possible to use numerical methods for sparse matrices to efficiently find the Cholesky decomposition and hence use CAR models in practice (Rue & Held 2005).

Chapter 4

Presentation of Publications

We make use of Markov models in four bioinformatics applications each of which has a corresponding publication presented below. Firstly, Table 4.1 sets out the different objects of interest and what the elements of the Markov model represent in each case. This ranges from vertices representing genes (Publications 1 and 3) to cells (Publication 4) along with pixels in digital images of cell culture models (Publication 2). The edges may represent either spatial or temporal relations between the variables as well as functional interactions that are neither spatial nor temporal.

Table 4.1: The objects and relations represented in the graphical model for each application.

	Vertices	Edges
Publication 1	Genes	Temporal relations (gene expression in time)
Publication 2	Pixels	Spatial relations
Publication 3	Genes	Functional relations
Publication 4	Neurons	Spatial relations

Figure 4.1 shows a representative conditional independence graph for each application, a description of which is summarised in Table 4.2. Note that in this discussion we focus on the Markov components $X_{\mathcal{V}}$ within each model. The most important differences between the models as seen in Figure 4.1 are whether the models are trees or not and whether the variables $X_{\mathcal{V}}$ are ‘hidden’ or not (compare to Figure 1.1). Recall that these difference define the particular type of Markov model and determine the possibilities for inference. We briefly consider general descriptions of the models and how they relate to each application.

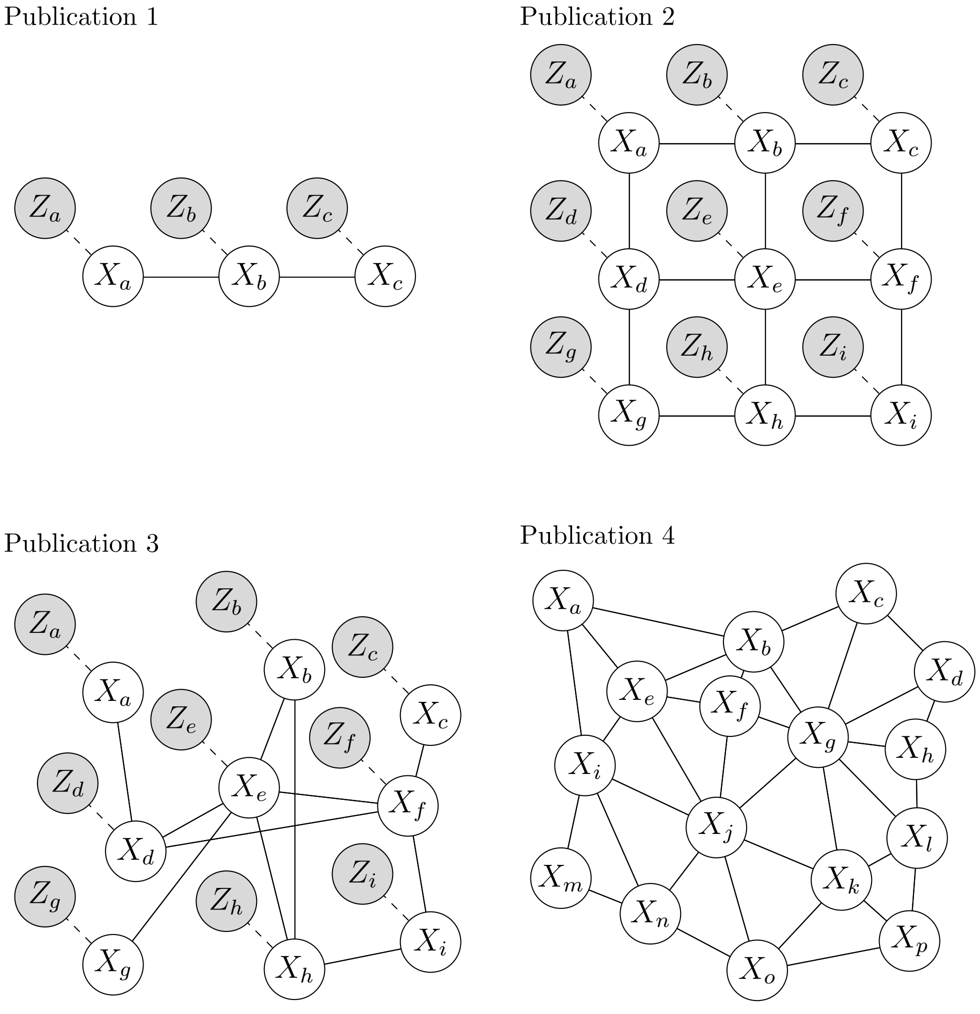
Table 4.2: The graphical model structure corresponding to each application.

	Non-tree	Variables $X_{\mathcal{V}}$ are 'hidden'	Markov Model	Description
Publication 1		X	HMM	Regular chain
Publication 2	X	X	Hidden MRF	Regular grid
Publication 3	X	X	Hidden MRF	Scale free
Publication 4	X		Gaussian MRF	Triangulation

By a regular chain for Publication 1 we mean a classic Markov chain where all vertices corresponding to the variables $X_{\mathcal{V}}$ have the same degree (except the beginning and end vertices). This application concerned time course gene expression data and we model each instance of the gene expression in time using a hidden Markov model (HMM). In Publication 2 we are now considering a non-tree Markov random field (MRF). By a regular grid we again mean that all vertices corresponding to the variables $X_{\mathcal{V}}$ have the same degree (except the border vertices). This was an image analysis application concerning microscopy image data of cell culture models. In this case the vertices represent pixels in digital images and the modelling is carried out for each pixel and its four immediate neighbours, above, below, left and right in the digital image.

The graph for Publication 3 is also a hidden MRF but rather than a regular grid appears to be scale free. That is, in general terms, there are a few vertices with many edges and many vertices with a few edges. This isn't completely apparent in Figure 4.1 with only nine vertices constituting the Markov component of the model but we can still see the contrast to the regular grid case where every vertex has the same degree. Here the application was to do with genomic network data and the edges represent interactions between genes. While some genes may interact with most other genes, the majority of genes interact with only a few others.

The final graph corresponding to Publication 4 is also an MRF, however in this case the variables $X_{\mathcal{V}}$ are not 'hidden'. In this application the vertices represent neurons and connections between vertices represent spatial interactions between neurons modelled by their relative spatial positions as obtained from image data. The neurons are not distributed within the image in the same way as the actual pixels, that is, they are not laid out in rows and columns. Hence each vertex may have a different number of edges but unlike the scale free case in Publication 3, the degree in this application is bounded in practice (see Publication 4).


Publication 1

Publication 2

Publication 3

Publication 4

Figure 4.1: Representative conditional independence graphs corresponding to each application.

Table 4.3: The inference carried out in each application.

	Infer 'hidden' $X_{\mathcal{V}}$	Infer parameters	Framework	Reference
Publication 1	X	X	Maximum likelihood	(Robinson 2012)
Publication 2	X		Maximum *a posteriori*	Chapter 2
Publication 3	X		Maximum *a posteriori*	Chapter 2
Publication 4		X	Bayesian	Chapter 3

We are now able to consider how each of the Markov models are utilised, what kind of inference is carried out and for what aims based on the type of Markov model and what the vertices and edges represent in each application. As noted, the most important information in Table 4.2 is given in the first two columns corresponding to the particular type of Markov model. This determines the inference that can be achieved in each application, which is set out in Table 4.3 and discussed below. Note that final results for each application are summarised in Table 4.4 while the comparison and evaluation of our use of Markov models against the diverse literature associated to each application is covered within the publications themselves.

Our first application concerns time course gene expression data from grapevine plants undergoing berry development at multiple vineyards. The time series are modelled as HMMs where $Z_{\mathcal{V}}$ are the observed gene expression measurements and $X_{\mathcal{V}}$ are hidden underlying states (note that in Publication 1 the notation for hidden states is S and observed variables are X, W and C). An alignment between expression profiles from different vineyards is achieved by parameterising a mapping to a common state sequence for each gene. Exact inference of the underlying 'hidden' states is computationally feasible for HMMs owing to their tree structure while parameter inference is carried out with the EM algorithm (Robinson 2012).

Both Publications 2 and 3 concern inference of the 'hidden' Markov component of hidden MRFs. For Publication 2 this results in a segmentation of digital image data corresponding to different biological structures. The structures of interest are represented as contiguous objects within the digital images and hence using a MRF, where neighbouring pixels are induced to have the same segmentation label, is inherently appropriate for this application. Moreover, the smoothness of the segmentation output is particularly complimentary for the subsequent investigation of physical contact between the different cell types. The inference for Publication 2 is carried out using energy minimisation and requires the material from Chapter 2 concerning graph cuts and the α-expansion algorithm (Algorithm 2.2).

In Publication 3, genomic networks are similarly segmented by genes being labelled as a 'hit' or not, that is, whether they are identified as valuable for further subsequent analysis. Each gene is assumed to have similar behaviour to the others it interacts with in the genomic experiments considered as represented by edges in the genomic network. We utilise an MRF to incorporate the additional network information in the identification of network based gene hits in the same way that adjacent pixels in a digital image are induced to have the same segmentation label. Hence Publication 3 also requires the material from Chapter 2 and we again use graph cuts and the α-expansion algorithm (Algorithm 2.2) for energy minimisation.

Publication 4 concerns modelling the spatial component of neuronal synaptic activity in image data. In this case the variables $X_{\mathcal{V}}$ are observed directly rather than 'hidden' and it is the spatial autocorrelation parameter within the Gaussian MRF (CAR) model that is of interest. Similar to the consideration of interacting genes, we model the mutual synaptic interaction of individual neurons through their neighbours as defined by relative spatial distance. Publication 4 requires the theory reviewed in Chapter 3 and sampling from the posterior distribution of the parameters of a CAR model in a Bayesian framework.

Table 4.4 summarises the results for each application. Hence we have shown that Markov models can be used to generate new knowledge in diverse applications relating to multiple domains of biological research. The Markov structure allows for a graphical and intuitive model of conditional independencies within a standard and general framework, making such models very popular in many different areas. Furthermore, as we have shown, there are many associated computational schemes and algorithms available in order to efficiently utilise Markov models in practice.

Table 4.4: The results for each application.

Publication 1	We propose a novel alignment method for time series gene expression data based on hidden Markov models (HMMs). The modelling extensions we employ account for specific features of the data, in particular the sparsity of the time series commonly encountered in such gene expression time course experiments. Moreover we show that the alignment obtained under the model is robust against subsets of genes that do not align. We subsequently use a measure of alignment based on the inferred Viterbi paths to classify genes as likely to be developmentally driven or not and additionally validate the classification accuracy. This results in a set of ∼1200 genes with no current annotated function for which we have found new evidence that they are likely to be controlled in a developmental manner.
Publication 2	We extend the established use of hidden Markov random fields (MRFs) for digital image segmentation of 3D fluorescence image data of cell culture models containing both multicellular tumour and cancer associated fibroblast (CAF) structures. Using both Gaussian mixture distributions and local entropy filtering allows for a segmentation with multiple labels to be achieved in an automatic and principled way. The segmentation output subsequently allows for the investigation of physical contact between tumour cells and CAFs on the surface of tumour organoids, which has important implications for tumour biology (Åkerfelt et al. 2015). We demonstrate how each step of our proposed method improves segmentation performance. Additionally we show our MRF based method is more generally applicable to other types of image data, not just fluorescence microscopy.

Continued on the next page

Table 4.4 – *Continued from the previous page*

Publication 3	As an analogue to segmentation of image data, we use a hidden MRF based method to label ‘hit’ genes across genomic networks together with genomic screening data. The particular advance here concerns multiple ‘hit’ labels being easily achieved within the MRF framework. This allows for applications with multivariate measurements for each gene as well as further discrimination of the labelling. Such an approach had not been considered up until now and we demonstrate its advantages by finding additional pathway enrichment in previously analysed data concerning lymphoma. We show that our MRF based method is widely applicable to genomic screening data in general, including from RNA interference experiments. Additionally we compared to other competing methods using an independent simulation experiment and obtained the best performance results.
Publication 4	We use a Gaussian MRF (CAR) model to quantify the synaptic activity phenotype of large populations of neurons. Previous analysis was limited to spike binning in an artificial square grid defined across the field of view whereas our model is based directly on the relative spatial positions of the individual neuronal somata. Spatial autocorrelation of the neuronal activity is then quantified through a Bayesian posterior density of the model parameters. Our proposed approach is inherently suitable for the complexity and resolution of the image data and allows for the analysis of upwards of thousands of neurons over hundreds of time points. We analyse data specifically developed for this study as well as from the literature. The quantification results both support and conform to previous analysis as well as demonstrate that our method has potential for screening applications.

Following my Master of Philosophy thesis (Robinson 2012), the underlying research framework of the use of Markov models was driven by me while biologically relevant applications were developed within multiple research groups. Table 4.5 lists the published author contributions. For each application I was primarily responsible for:

- Methodological development, implementation and analysis of results
- Literature review, evaluation and presentation of results
- Production of figures, movies and code for publication
- Writing the manuscript and organising the contributions of co-authors
- Submission of the manuscript and responding to reviewers

Table 4.5: The author contributions for each application.

Publication 1	Conceived and designed the time course microarray experiments: MT CD. Performed the experiments: MT CD. Developed the alignment methodology: SR GG IK. Implemented the methodology: SR. Analysed the data: SR GG IK. Contributed to the analysis and drafting of the paper: IK MT. Wrote the paper: SR GG CD.
Publication 2	Conceived and designed the experiments: MÅ MN. Performed the experiments: MT MÅ. Wrote the paper: SR MÅ MN. Manually segmented the image data: MT MÅ. Developed the segmentation methodology: SR LG JN. Implemented the methodology and analysed the data: SR. Contributed to the analysis and drafting of the paper: LG JN MT.
Publication 3	Developed and implemented the methodology; wrote the paper: SR. Proposed the project; supervised development of the methodology and writing of the paper: LG. Contributed to the analysis and drafting of the paper: JN GP JPR. Conceived, designed and performed the RNAi screen: GP AC JPR.
Publication 4	Conceived, designed and performed the experiments: MJC. Developed and implemented the modelling methodology; analysed the data: SR. Contributed to the development and analysis: MJC. Wrote the paper: SR MJC.

Bibliography

Åkerfelt, M., Byramoglu, N., Robinson, S., Toriseva, M., Schukov, H.-P., Härmä, V., Virtanen, J., Kaakkinen, M., Eklund, L., Kannala, J., Heikkilä, J. & Nees, M. (2015), 'Tracking Morphology and Dynamics of Tumor-Stroma Interactions in Three Dimensional Co-Cultures by Automated Image Analysis', *Oncotarget* **6**(30), 30035–30056.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition, CRC Press.

Bell, B. S. & Broemeling, L. D. (2000), 'A Bayesian Analysis for Spatial Processes with Application to Disease Mapping', *Statistics in Medicine* **19**(7), 957–974.

Besag, J. (1974), 'Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion)', *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 192–236.

Besag, J. (1986), 'On the Statistical Analysis of Dirty Pictures (with Discussion)', *Journal of the Royal Statistical Society. Series B (Methodological)* **48**(3), 259–302.

Blake, A., Kohli, P. & Rother, C., eds (2011), *Markov Random Fields for Vision and Image Processing*, MIT Press.

Bondy, J. A. & Murty, U. S. (2008), *Graph Theory*, Springer.

Boykov, Y. & Kolmogorov, V. (2004), 'An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(9), 1124–1137.

Boykov, Y., Veksler, O. & Zabih, R. (2001), 'Fast Approximate Energy Minimization via Graph Cuts', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(11), 1222–1239.

Brook, D. (1964), ‘On the Distinction Between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems’, *Biometrika* **51**(3/4), 481–483.

Cressie, N. & Wikle, C. K. (2015), *Statistics for Spatio-Temporal Data*, John Wiley & Sons.

De Oliveira, V. (2012), ‘Bayesian Analysis of Conditional Autoregressive Models’, *Annals of the Institute of Statistical Mathematics* **64**(1), 107–133.

Ford, L. R. & Fulkerson, D. R. (1956), ‘Maximal Flow Through a Network’, *Canadian Journal of Mathematics* **8**(3), 399–404.

Gelfand, A. E. & Vounatsou, P. (2003), ‘Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis’, *Biostatistics* **4**(1), 11–15.

Geman, S. & Geman, D. (1984), ‘Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Hammersley, J. M. & Clifford, P. (1971), ‘Markov Fields on Finite Graphs and Lattices’, Unpublished.

Kappes, J. H., Andres, B., Hamprecht, F. A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B. X., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., & Rother, C. (2015), ‘A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems’, *International Journal of Computer Vision* **115**(2), 155–184.

Koller, D. & Friedman, N. (2009), *Probabilistic Graphical Models: Principles and Techniques*, MIT Press.

Kolmogorov, V. & Zabih, R. (2004), ‘What Energy Functions Can Be Minimized via Graph Cuts?’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(2), 147–159.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford University Press.

Ren, C. & Sun, D. (2013), ‘Objective Bayesian Analysis for CAR Models’, *Annals of the Institute of Statistical Mathematics* **65**(3), 457–472.

Robinson, S. (2012), ‘Alignment of Time Course Microarray Data with Hidden Markov Models’, Master of Philosophy Thesis, The University of Adelaide.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, CRC Press.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M. & Rother, C. (2008), 'A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(6), 1068–1080.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons.