Statistical Modelling of Lattice Data

with Applications in Flow Cytometry

By

Kevin Christopher Brosnan

UNIVERSITY of LIMERICK

OllSCOIL LUIMNIGH

Department of Mathematics & Statistics

UNIVERSITY OF LIMERICK

A thesis submitted for the award of Ph.D.

Supervised by: Dr. Kevin Hayes & Dr. Norma Bargary

Submitted to the University of Limerick, October 2018
Declaration

This thesis is presented in fulfilment of the requirements for the award of Doctorate of Philosophy. It is entirely my own work, completed without collaboration with others except my supervisors, Dr. Norma Bargary and Dr. Kevin Hayes. Where use has been made of the work of other people it has been fully acknowledged and referenced accordingly.

Signature: __________________________
Kevin Christopher Brosnan
October 2018
Acknowledgements

I would like to thank my supervisors Dr. Norma Bargary and Dr. Kevin Hayes for their patient guidance, invaluable support and useful critiques for the duration of my postgraduate studies. This work would not have been possible without them.

To the MACSI group, I thank you all for making my time as a postgraduate student extremely enjoyable. To Dr. Sinéad Burke, Peg Hanrahan and Dr. Joanna Jordan for your reliability and hard-working nature which allowed me to focus solely on my research.

My parents, Mark and Linda, and my brothers, Shane and Steven, who have shown encouragement, support and trust in me throughout my years as a postgraduate and undergraduate student. When times got tough I could always rely on them to see the silver lining and get me back on track.

The financial support I have received from the Irish Research Council through the Government of Ireland Postgraduate Scholarship scheme [GOIPG/2014/19] has made this research project possible.
Abstract

Statisticians are among the most in demand problem solvers in today’s data driven world, with their expertise often employed to address emerging issues in other scientific fields by adopting, modifying and/or developing statistical methods. This application and redeployment of statistical methods allows statisticians to provide invaluable and actionable insights to scientific experts in the chosen field. This thesis provides statistical methodology for the gating of flow cytometry data.

Flow cytometry is a technology that simultaneously measures and analyses multiple physical and chemical characteristics of single cells as they flow in a fluid stream through a beam of laser light. This technology has become an emerging state-of-the-art device in microbiology and dairy science, and is also used extensively in medical diagnostics. Unfortunately, the lack of a robust statistical analysis toolbox for flow cytometry has restricted the deployment of this world-leading sensor technology.

Gating, the identification of homogeneous cell populations, of these complex data sets are performed using expert opinion or naive clustering algorithms rather than employing a specialised statistical framework. Gating is the equivalent of clustering however the data from flow cytometry lies on a structured lattice grid. The focus of this thesis is to provide a methodology for the gating of flow cytometry data which respects the underlying structure of the data.

This research provides statistical methods for the analysis of binary lattice data, which are based on the statistical properties of the observed lattice. A Bernoulli outcome governs the binary value taken at each node. From this the energy distribution, a measure of similarity of nodes across the lattice, for each node can be specified, allowing the expected value and variance of the complete lattice energy to be calculated. For small lattices the complete probability distribution is provided, while the energy function for larger lattices is known for specific values. The novel probabilistic and statistical quantities lead to an algorithm for gating FCM data, which incorporates an improved Markov chain Monte Carlo and a hierarchical approach, that overcomes many of the limitations facing the deployment of this world leading sensor technology.
# Table of Contents

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Flow Cytometry</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Gating</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Thesis Outline</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1 Chapter 2</td>
<td>7</td>
</tr>
<tr>
<td>1.3.2 Chapter 3</td>
<td>7</td>
</tr>
<tr>
<td>1.3.3 Chapter 4</td>
<td>7</td>
</tr>
<tr>
<td>1.3.4 Chapter 5</td>
<td>8</td>
</tr>
<tr>
<td>1.3.5 Chapter 6</td>
<td>8</td>
</tr>
<tr>
<td>1.3.6 Chapter 7</td>
<td>8</td>
</tr>
<tr>
<td>2 Flow Cytometry</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Data Collection</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1 Data Measurement</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Data Structure</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Sequential Gating</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Software Approaches</td>
<td>15</td>
</tr>
<tr>
<td>2.5 FlowCAP-I</td>
<td>16</td>
</tr>
<tr>
<td>2.6 FlowClust</td>
<td>16</td>
</tr>
<tr>
<td>2.7 Summary</td>
<td>19</td>
</tr>
<tr>
<td>3 Bernoulli Random Variables</td>
<td>20</td>
</tr>
<tr>
<td>3.1 Configurations on a Lattice</td>
<td>20</td>
</tr>
<tr>
<td>3.2 The Generalised Binominal Distribution</td>
<td>23</td>
</tr>
<tr>
<td>3.3 Properties of the Lattice Energy</td>
<td>24</td>
</tr>
<tr>
<td>3.4 Densities of the Lattice Energy</td>
<td>27</td>
</tr>
<tr>
<td>3.4.1 Node Energy Probability Mass Function</td>
<td>27</td>
</tr>
<tr>
<td>3.4.2 Lattice Energy Probability Mass Function</td>
<td>28</td>
</tr>
<tr>
<td>3.4.3 Minimum &amp; Maximum Energy</td>
<td>30</td>
</tr>
<tr>
<td>3.5 Properties of the Estimands</td>
<td>31</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>33</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

4 Adapted MCMC 35  
4.1 Classical Energy Minimisation .......................... 35  
4.2 Metropolis-Hasting Algorithm .......................... 38  
4.3 Simulated Annealing ................................. 40  
4.4 Adapted MCMC ................................... 43  
4.5 Probability Map .................................. 47  
4.6 Application to Binary Images ......................... 47  
4.7 Summary ...................................... 50  

5 Hierarchical Approach 52  
5.1 Hierarchical Methods ................................. 52  
5.2 Pruning ....................................... 54  
5.3 Pruned Hierarchical Method ......................... 56  
5.4 Application to Binary Images ......................... 62  
5.5 Summary ...................................... 64  

6 Application to FCM 66  
6.1 Gating ....................................... 66  
6.2 Application to Rituximab Data ......................... 67  
6.3 Application to GvHD Data ........................... 69  
6.4 Summary ...................................... 72  

7 Summary & Future Work 75  
7.1 Summary ...................................... 75  
7.2 Further Work ................................... 76  
  7.2.1 Neighbourhood Systems ......................... 76  
  7.2.2 Non-Binary Observables ......................... 77  
  7.2.3 Implementation ................................ 78  

A Bernoulli Algebra 79  
A.1 The Generalised Binomial Distribution ............ 79  
A.2 Expected Value and Variance ....................... 80  
A.3 Covariance ................................... 80  
A.4 Node Energy Function ............................ 81  
A.5 Overlapping Neighbourhood Covariances ............ 82  
A.6 Properties of Lattice Energy ....................... 84  
A.7 Probability of Node Energy ......................... 86  
A.8 Probability of Lattice Energy ...................... 89
# TABLE OF CONTENTS

## B  FCM Data File Standard

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1 Header Section</td>
<td>95</td>
</tr>
<tr>
<td>B.2 Text Section (Metadata)</td>
<td>95</td>
</tr>
<tr>
<td>B.3 Data Section</td>
<td>95</td>
</tr>
<tr>
<td>B.4 Analysis Section</td>
<td>96</td>
</tr>
</tbody>
</table>

## Bibliography

<table>
<thead>
<tr>
<th></th>
<th>97</th>
</tr>
</thead>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 (a)</td>
<td>The original rituximab flow cytometry data for forward light-scatter (FSC) and sideward light-scatter (SSC). (b) The subjective manual gating used in industry for two expert cytometrists. (c) The approach of Lo et al. (2008) to the data. (d) A zoomed in version of the rituximab data whereby the lattice structure is evident. (e) The probability map associated with the methodology of chapter 5. (f) The clusters identified via the methodology of chapter 5.</td>
</tr>
<tr>
<td>2.1</td>
<td>A simple flow cytometer (Wulff et al., 2006).</td>
</tr>
<tr>
<td>2.2</td>
<td>Schematic illustration of the FCM data analysis framework (Bashashati and Brinkman, 2009).</td>
</tr>
<tr>
<td>2.3</td>
<td>Pictorial view of the scattering of the laser beam (Diamantina Institute, University of Queensland, 2015).</td>
</tr>
<tr>
<td>2.4</td>
<td>Optics in a flow cytometer (AVR Optics, 2017).</td>
</tr>
<tr>
<td>2.5</td>
<td>Gating of the rituximab data via the approach of Lo et al. (2008).</td>
</tr>
<tr>
<td>2.6</td>
<td>Gating of the GvHD data via the t-mixtures approach.</td>
</tr>
<tr>
<td>3.1 (a)</td>
<td>Corner node; (b) Edge node; (c) Inner node neighbourhoods, where red is the node of interest and blue is its neighbourhood.</td>
</tr>
<tr>
<td>3.2</td>
<td>Covariance structures are only required when the neighbourhood systems overlap. (a) shows non-overlapping neighbourhood. (b) shows two overlapping neighbourhoods. Red is the node of interest, blue its non-overlapping neighbourhood and green overlapping neighbours.</td>
</tr>
<tr>
<td>3.3</td>
<td>Overlap between neighbouring neighbourhoods. Y, Z, P and Q represent Bernoulli nodes, X are the neighbours of Y not in the neighbourhood of Z and W are the neighbours of Z not in the neighbourhood of Y.</td>
</tr>
<tr>
<td>3.4</td>
<td>The probability mass function of the energy of a 2 × 2 lattice for ( \theta = {0.25, 0.5, 0.75} ).</td>
</tr>
<tr>
<td>3.5</td>
<td>The probability mass function of the energy of a 3 × 3 lattice for ( \theta = {0.25, 0.5, 0.75} ).</td>
</tr>
<tr>
<td>3.6</td>
<td>Probabilities of observing minimum energy state for ( \theta ). The red curve is for a 2 × 2 lattice, while the purple curve is 2^6 × 2^6. The remaining lines represent the dyadic structures between these two dimensions.</td>
</tr>
</tbody>
</table>
3.7 Heatmap of the energy function on a $3 \times 3$ lattice for $\theta = \{0, 0.02, \ldots, 1\}$. The red line is the average theoretical energy, the quartiles are given in green, while the purple and blue are the 90% and 95% bounds.

3.8 A comparison on a $3 \times 3$ lattice between the theoretical variance function (red line) and the variance from Bernoulli simulations (boxplots).

4.1 Proposed Metropolis algorithm movement steps. White circles indicate initial particle positions, while the grey circle is the initial location of the selected node. The blue square signifies the potential movement range of the grey node, with the black arrow and red circle showing a potential move.

4.2 Examples of initial configurations of a lattice: (a) shows a lattice with all nodes in agreement; (b) shows a lattice with one disagreeing node; (c) shows a lattice with alternating nodes.

4.3 Sample binary images each comprised of $128 \times 128$ pixels.

4.4 The noisy letter A with the solutions from a simulated annealing approach and the accelerated method proposed in this chapter.

4.5 The noisy letter G with the solutions from a simulated annealing approach and the accelerated method proposed in this chapter.

4.6 A set of example $4 \times 4$ lattices: (a) and (b) two lattices which are organised; (c) a lattice with one outlying node; (d) is a disorganised lattice.

5.1 The pyramid structure.

5.2 Hierarchical approach to image segmentation.

5.3 A decision tree model for identifying survivors on board the titanic using the individual’s sex, class and age. (a) shows the complete decision tree model, which is overfitting the training data. (b) shows the pruned decision tree model, which produces less decision nodes.

5.4 Hierarchical approach to identifying clusters within an image.

5.5 Probabilities of maximum energy state for $\theta$ across varying lattice sizes. The red curve is for a $2 \times 2$ lattice, while the purple curve is $2^n \times 2^n$, the remaining lines represent the dyadic structures between these two dimensions.

5.6 A set of example $4 \times 4$ lattices: (a) and (b) two lattices which are organised; (c) a lattice with one outlying node; (d) is a disorganised lattice.

5.7 The letter A and letter G images, along with the noisy images.

5.8 The letter A and letter G binary images solutions for the accelerated simulated annealing (ASA) and the hierarchical pruning algorithm (HPA) approaches.
6.1 (a)-(c) The initial gating stage based on FSC and SSC variables. (a) The t-mixture approach to gating. (b) The probability map generated from the proposed methodology. (c) The gates identified from the probability map, blue and grey points denote gates and unclassified points respectively. (d)-(f) The second gating stage based on 7 AAD v Anti-BrdU FITC. (d) The t-mixture approach to gating. (e) The probability map whereby red, orange, yellow and white regions denote high to low active probabilities. (f) The gates identified from the probability map denoted by blue points, while grey points denote unclassified points. 68

6.2 GvHD Control: (a)-(c) Gating based on CD4 FITC and CD8β PE variables. (a) The t-mixture approach to gating. (b) The hierarchical model probability map showing high, medium and low active probabilities in red, orange and yellow respectively. (c) The clustering solution from the proposed methodology. (d)-(f) Gating solutions for CD4 FITC against CD3 PerCP. (d) The t-mixture approach to gating. (e) The probability map showing regions of probabilities where red, orange and yellow denote high, medium and low probabilities. (f) The sub-populations identified from the probability map. 70

6.3 GvHD Positive: (a)-(c) Gating based on CD4 FITC and CD8β PE variables. (a) The t-mixture approach to gating. (b) The probability map of the proposed methodology. (c) The clustering solution from the proposed methodology. (d)-(f) Gating solutions for CD4 FITC against CD3 PerCP. (d) The t-mixture approach to gating. (e) The probability map showing regions of probabilities where red, orange and yellow denote high, medium and low probabilities. (f) The sub-populations identified by the proposed methodology. 71

7.1 Lattice grid showing the first (red), second (red & blue) and third (red, blue & green) order neighbourhoods of a node (black) for the von Neumann and Moore systems. 76

A.1 Configurations of corner nodes with associated energy and the observation probability. Red is an active node observed with probability θ, while blue is an inactive node which occurs with probability 1 − θ. 87

A.2 Configurations of edge nodes with the associated energy and the observation probability. Red is an active node, while blue is an inactive node. 87

A.3 Configurations of inner nodes with the associated energy and the probability of observing such configurations. Red indicates an active node, while blue indicates an inactive node. 88
A.4 Configurations of $2 \times 2$ lattice with the associated energy and the probability of observing such configurations. Red indicates an active node, while blue indicates an inactive node. The circled numbers denotes the number of each type of configuration. 89

A.5 Configurations of $3 \times 3$ lattice with the energy and the probability of observing such configurations. Red indicates an active node, while blue is an inactive node. The circled numbers denotes the number of each configuration. 91
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Method comparison between the original image, the simulated annealing procedure and the proposed adaptation for the letter A.</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Method comparison between the original image, the simulated annealing procedure and the proposed adaptation for the letter G.</td>
<td>50</td>
</tr>
<tr>
<td>5.1</td>
<td>The letter A: chapter 4 approach versus chapter 5 approach.</td>
<td>63</td>
</tr>
<tr>
<td>5.2</td>
<td>The letter G: chapter 4 approach versus chapter 5 approach.</td>
<td>64</td>
</tr>
<tr>
<td>A.1</td>
<td>The number of each of the lemma structures from section 3.1 required for the variance of given sized lattices.</td>
<td>86</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Flow Cytometry

Flow cytometry (FCM) is an emerging state-of-the-art sensor technology used in medical diagnostics for the diagnosis and monitoring of leukaemia, in dairy sciences for ensuring stringent quality control, and for analysing the abundance of species in marine biology. Although the instrumentation and underlying technologies supporting FCM have advanced significantly in recent years (De Rosa et al., 2003; Perfetto et al., 2004), the development of automated data analysis tools remain elusive, the absence of which has been cited as a reason for the restricted deployment of the sensor technology (Lizard, 2007). The conventional method for analysing FCM data has relied heavily on manual, expert-driven approaches rather than an automated statistical framework (Eudey, 1996). Over the last two decades calls for the development of such a statistical framework have been made by the FCM community (Bagwell, 2004; Bashashati and Brinkman, 2009; Braylan, 2004; De Rosa et al., 2003; Eudey, 1996; Lugli et al., 2010), facilitating reproducible and standardised analyses to be conducted while reducing the considerable time-investment currently required for manual FCM analysis. This thesis will provide a statistical methodology for FCM analysis, which: exploits the structure of the measured data; removes subjectivity from the analysis of gating (clustering) of FCM data; is reproducible across experts in the field; and reduces the considerable time investment currently required for FCM analysis.
CHAPTER 1. INTRODUCTION

Figure 1.1: (a) The original rituximab flow cytometry data for forward light-scatter (FSC) and sideward light-scatter (SSC). (b) The subjective manual gating used in industry for two expert cytometrists. (c) The approach of Lo et al. (2008) to the data. (d) A zoomed in version of the rituximab data whereby the lattice structure is evident. (e) The probability map associated with the methodology of chapter 5. (f) The clusters identified via the methodology of chapter 5.
1.2 Gating

Gating is a filtering operation applied to FCM data to identify subpopulations that exist within the collected cell samples. In a sense, gating is similar to data clustering where groups of cells which appear to be similar are identified as belonging to the same gate or cluster. This allows FCM experts to identify unique regions within the data and analyse regions of interest in greater detail. The gating procedure often involves the manual drawing of gates to specify regions of interest in a 2D graphical representation of a pair of FCM variables (Bagwell, 2004; Finak et al., 2016; Parks, 1997; Suni et al., 2003). This manual expert-driven approach is highly subjective across laboratories and across colleagues within individual laboratories (Maecker et al., 2012; Mair et al., 2016). In order to reduce the subjectivity, and hence variability associated with manual gating, an objective statistical methodology with appropriate software implementations is required (Maecker et al., 2012; Mair et al., 2016).

An example of FCM data, the rituximab data (Gasparetto et al., 2004), is shown in figure 1.1(a). The expert-driven approach to identifying gates within FCM data, shown in figure 1.1(b), is completed by drawing a polygon enclosing the region of interest using a mouse on a computer screen. The polygons drawn are inherently biased by the expert cytometrist, and inconsistent across experts (and potentially across a single expert over time), which further exemplifies the need for a novel statistical approach to gating.

The statistical approach currently used by academics is presented in Lo et al. (2008). The aim of the method, known as flowClust, is to use model-based clustering to identify the gates or clusters, where the data in each cluster is
represented by a t-distribution (Andrews and McNicholas, 2012). More details on flowClust are provided in chapter 2. The one cluster solution of flowClust on the rituximab data is shown in figure 1.1(c). In comparison to manual gating the flowClust encloses points which are a considerable distance from the dense population.

Figure 1.1(d) illustrates the intrinsic lattice structure of FCM data, a feature imposed by the use of an analogue-to-digital converter in the technology. Thus the FCM data measurements are integer valued on a dyadic length scale, that is the data lie on the interval $[0, 2^n]$ where $n$, an integer, represents the power of the converter. Model-based clustering assumes that the data measurements are continuous, which is not the case in FCM. Consequently, the treatment of the FCM data by Lo et al. (2008) is inappropriate when the underlying structure of the data is explored. The t-mixtures approach also restricts the size, shape and orientation of the identified clusters, however biological knowledge provides no basis for such a constraint.

This research presents a novel approach for gating FCM data. The proposed statistical methodology utilises the underlying lattice structure of FCM data by considering Bernoulli outcomes at each node of the lattice, detailed in chapter 3. A classical method for analysing lattice data is the Ising model (Ising, 1925), originating in physics for the exploration of spontaneous magnetisation of ferromagnetic materials, which identifies particles in spin agreement by minimising an energy function. The statistical properties of the energy function, and the underlying Bernoulli lattice structure, provide the basis for an algorithmic approach to identify gates within FCM data. Utilising the probabilistic quantities of the energy allows gates to be identified (see
CHAPTER 1. INTRODUCTION

figure 1.1(f)) and a probability map to be estimated (see figure 1.1(e)). The probability map provides a quantification of energy across the entire lattice, which indirectly identifies the required gates. The new methodology provided in chapter 5 respects the underlying data structure and is not restrictive on the size, shape or orientation of identified gates.

In general, the exact definition of a cluster or gate is not universally agreed. This is not surprising as the area in which the methodology is applied often provides the context. An analyst tasked with market segmentation has a definition of identified clusters which is completely different to that of an analyst interested in identifying if sub-types of breast cancer exist based on a set of recorded features from patients with breast cancer (James et al., 2013). The use of cluster analysis for a number of tasks including exploratory analysis, dimension reduction or the identification of underlying structure makes it even more difficult to define a true cluster.

Hennig (2015) remarks:

“If researchers want to find true or real clusters, they have to specify what kind of truth they are interested in and what should constitute a “real” cluster. An appropriate clustering method can be found by connecting the characteristics of the clustering method to what is desired according to the researchers’ cluster concept.”

In line with these views, it is important to define the cluster characteristics required for FCM gates in an effort to provide a methodology which is statistically robust and relevant to the objectives of FCM analysis. The novel methodology presented in this thesis adheres to the following characteristics
for identified gates:

- homogeneity within gates and heterogeneity between gates;

- represent the true gates at the global level;

- represent the local perturbations in the data;

- and no assumption on the underlying gate distribution.

1.3 Thesis Outline

Specifying FCM data to lie on a regular lattice grid, respecting the recorded measurements, allows the energy function of Ising (1925) to be used to identify regions of similarity and dissimilarity within FCM data. The statistical properties of the energy function (provided in chapter 3) allow an informed decision on the lattice structure to be made based on the lattice dimension and the number of active nodes within the lattice. Using an adapted simulated annealing approach (detailed in chapter 4) and a hierarchical procedure (specified in chapter 5) provides a method which identifies regions of similarity (gates) at varying levels of detail. Combining these solutions into a single representation of the gates observed in the underlying data will result in the identification of both local and global features within FCM data. By utilising the statistical and probabilistic quantities of the energy function of the lattice, there is no requirement to specify the number of clusters within the image a priori and no assumption is made on the distribution of the gates. This allows flexibility in terms of size, shape and orientation.
1.3.1 Chapter 2

Chapter 2 provides detail on the data collection process in a flow cytometer. The use of an analogue-to-digital converter for data conversion results in the data lying on a discrete lattice grid in two dimensions, a structural layer that has remained unused in modelling approaches to date. A critical analysis of existing commercial and open-source software solutions for automated gating of FCM data follows. The chapter concludes with a summary of FlowCAP-I, a challenge set to identify the best algorithms for gating FCM data.

1.3.2 Chapter 3

Chapter 3 explores Bernoulli random variables defined on a discrete lattice. A neighbourhood system is defined which governs the interaction between nodes on the lattice. Expressing this neighbourhood as a sum of Bernoulli outcomes, resulting in a Binomial random variable, the expected value of the energy and its variance across the lattice is calculated. Finally, the complete energy distribution for smaller lattices is provided, with important quantities specified for larger lattices. The methodology outlined in this chapter is utilised throughout the remainder of the thesis.

1.3.3 Chapter 4

To date Markov chain Monte Carlo (MCMC) procedures have been used to identify edges and clusters within lattices. Chapter 4 adapts a MCMC approach for binary lattices. Utilising the methodology outlined in chapter 3, an improved simulated annealing approach is proposed. The proposed accelerated simulated annealing is demonstrated on two binary images, resulting in optimal solutions requiring fewer annealing iterations.
1.3.4 Chapter 5

Hierarchical analysis is a technique often employed to explore global and local features in data. Similar to the procedure employed in wavelet multi-resolution analysis, chapter 5 provides an iterative procedure for identifying regions within binary lattices. Utilising the methodology developed in chapter 3 a pruning approach is developed, which incorporates the accelerated simulated annealing approach of chapter 4. This method improves the accuracy of the resulting images, while reducing the time required to reach optimal solutions.

1.3.5 Chapter 6

This chapter presents an analysis of FCM data utilising the novel methodology of chapter 5. A comparison between the model-based clustering approach of Lo et al. (2008) and the hierarchical analysis procedure provided in chapter 5 is carried out on two flow cytometry data sets.

1.3.6 Chapter 7

Chapter 7 outlines some areas of possible future research.
2 Flow Cytometry

This chapter provides an overview of the scientific background of FCM. The measurement technology is presented, along with the structure of the resulting data. Existing commercial software for gating and the open-source flowClust (Lo et al., 2008) approach are critically examined. The chapter concludes with a specification of the problem for which this thesis provides a novel solution.

2.1 Introduction

The general structure of a flow cytometer is illustrated in figure 2.1. The sample of cells to be analysed enter the flow chamber together. On arrival, a faster flowing suspension fluid forces the cells into a narrow core region, resulting in a flow of single cells through the laser beam(s) and the collection of data. The single line flow of the cells allows properties of all individual cells in the sample to be measured by the technology.

Figure 2.1: A simple flow cytometer (Wulff et al., 2006).
A flow cytometer is similar to a microscope that yields very detailed output. A microscope provides an image of the overall cell, however a flow cytometer measures multiple physical and chemical characteristics of single cells as they flow in a stream of fluid through a beam of laser light (Wulff et al., 2006). Additionally, the number of cells analysed per second in a FCM scanner far exceeds that of a microscope, with some scanners analysing more than 10000 cells per second. Features such as cell size, granularity and shape are easily measured using a flow cytometer. In addition, chemical (e.g. size of the nucleus) and biological (e.g. classifying cells as living or dead) features of cells can be identified if the cells have been treated with a fluorescent dye prior to the experiment. This results in the creation of large multivariate data sets, in need of advanced statistical methods from which meaningful, usable and robust conclusions can be drawn.

![FCM Data Analysis Framework](image)

Figure 2.2: Schematic illustration of the FCM data analysis framework (Bashashati and Brinkman, 2009).

While the last two decades have seen the continued improvement of the underlying FCM technology, the time and challenges posed by the analysis of the data recorded accounts for the largest level of uncertainty in
gating solutions (Bashashati and Brinkman, 2009). In an effort to overcome subjective gating procedures as discussed in section 1.2, while reducing the substantial time required, the cytometry community have set out a framework encompassing all elements of the analysis scheme, from data pre-processing to expert interpretation. Figure 2.2 provides the flowchart associated with this framework, as detailed in Bashashati and Brinkman (2009). Quality Assessment, Normalisation and Outlier Removal can be incorporated into a single phase which identifies, investigates and removes cells which have been influenced by external factors such as instrumental parameters, instrumental variability or cell debris. The Gating and Cluster Labelling phases are also often conducted together, resulting in each cell being assigned a label of a particular gate. Each of the identified gates are then explored to identify the features of the subpopulations, this stage is referred to as Feature Extraction. Finally, the Interpretation of the gating solution is dependent on the overall objective of the study. Of particular focus in this research are the gating and cluster labelling components of the established FCM framework, however the proposed statistical methodology also addresses the normalisation and outlier removal stages detailed in the analysis framework.

2.2 Data Collection

2.2.1 Data Measurement

A flow cytometer records two scatter signals from a biological sample, specifically forward-scatter (FSC) and side-scatter (SSC). FSC is the light scattered in the same direction as the laser beam, while SSC is the light scattered at up to a $90^\circ$ angle to the laser beam. FSC recordings represent the size of the cell being analysed while SSC measurements can be interpreted
as a measure of the density of the cell. Moreover, a combination of these
can identify the initial population of interest for further analysis. Figure 2.3
depicts the scattering of the laser beam when it excites a cell, with a particular
focus on the scatter associated with FSC and SSC.

Figure 2.3: Pictorial view of the scattering of the laser beam (Diamantina
Institute, University of Queensland, 2015).

In addition to the two scatter signals, a FCM scanner can measure multiple
staining fluorescents simultaneously. The staining of particles with one or more
fluorescent dyes is used to identify additional biological features of the cells
being analysed, such as DNA content, metabolic activity and specific inter-
cellular markers. The stained cells will emit light of a particular wavelength
when excited by the laser if the cell contains the feature(s) of interest. Each
fluorescent being measured must be collected by a single optical detector called
a photomultiplier tube (PMT) and thus the number of fluorescents being
analysed is limited to the number of PMTs available in the flow cytometer.
Figure 2.4 shows how this can be achieved in a flow cytometer recording both
scatter signals and up to four fluorescents using optical filters and mirrors to
decompose the emitted light into the required wavelengths.
The recording of the voltage pulse commences as the cell enters the beam of laser light and continues to record until the cell has completely traversed the diameter of the beam. An obstruction bar placed in line with the laser on the opposite side of the fluid stream (see figure 2.1) ensures that when no particles are in line with the laser no measurements are taken.

Figure 2.4: Optics in a flow cytometer (AVR Optics, 2017).

2.2.2 Data Structure

The data collected from a flow cytometer adheres to a standardised format and is saved with a .fcs extension as outlined by the International Society for Advancement of Cytometry (Spidlen et al., 2010). A synopsis of the detailed document outlining this policy is provided in appendix B. The advantages of a strict policy on the collection of FCM data allows the development of analysis software deployable with any flow cytometer. The structured form of FCM data allows for general algorithms to be applied as a first step.
Of particular interest in this research project is the use of an analogue-to-digital converter (ADC) to process the individual wavelength intensities in FCM scanners, resulting in integer valued output data. The range of discrete integer values observed in the data is governed by the resolution of the ADC, where a higher resolution allows for considerably more unique integer values to be assigned to varying wavelength intensities (Shapiro, 2004). Historically, commercial FCM detectors have been restricted to 10-bit ADCs. Recently this restriction has been overcome due to improvements in the underlying technology and modern scanners can now provide up to 20-bit ADCs. Regardless, this underlying data capture protocol can be expressed as a lattice of dimension $2^n \times 2^n$, where the integer value at each node can be treated as an intensity and $n$ is the accuracy of the ADC. While several approaches for the automation of FCM gating have been proposed, none to date have exploited this underlying data structure.

### 2.3 Sequential Gating

While FCM can record multiple variables of interest on each cell population analysed, the identification of gates has remained restricted to pairs of variables due to the interpretability of gates in two dimensions. The process, known as sequential gating (Hahne et al., 2006), is initiated by identifying a single cluster on the FSC and SSC variables, which filter residue cells from the sample. The population identified is carried forward for further analysis, while the remaining cells are deemed to be cell debris and not relevant to the analysis. The next stage consists of examining the fluorescent markers of the cells identified in the initial stage. The gating procedure iterates through the recorded variables until the expert identifies the subpopulations which address the research question.
2.4 Software Approaches

The existing methods available for the statistical analysis of flow cytometry data include both commercial and open-source offerings. FlowJo™ (Tree Star, Ashland, OR), FCS Express™ (De Nova Software, Los Angeles, CA), WinList (Verity Software House, Topsham, Maine) and Kaluza™ (Beckman-Coulter, Brea, CA) are the leading commercial software solutions for the visualisation, gating and reporting of FCM data globally. Each offers scientists a software solution which visualises, analyses and reports on any FCM data collected. They provide point-and-click software solutions to FCM labs worldwide to improve the quality and efficiency of the analysis of this highly-complex data. Although these software solutions are well established as industry standards, the gating of cytometry data remains a manual process where gates are defined using a mouse on a computer screen. As a result these analysis tools are based on experience rather than a unified, robust statistical framework. This allows the analysis of a single FCM experiment to differ substantially between two cytometry experts.

The most substantial effort to develop open source statistical tools for FCM analysis has been completed by Bioconductor, based on the R programming language (R Core Team, 2017), in their comprehensive suite of flow cytometry packages (Huber et al., 2015). Bioconductor have produced packages for all stages of FCM analysis (as shown in figure 2.2), ranging from data input, to the production of analysis reports. However, few of the statistical methodologies in the gating and cluster labelling stages go beyond simple clustering approaches such as k-means (MacQueen, 1967), an issue which led to the formulation of the Flow Cytometry: Critical Assessment of Population Identification Methods
CHAPTER 2. FLOW CYTOMETRY

project (FlowCAP), (Aghaeepour et al., 2016, 2013), a group tasked with advancing FCM analysis. A substantial review of the computational analysis tools used for FCM can be found in Robinson et al. (2012).

2.5 FlowCAP-I

The FlowCAP project in 2009 was of particular importance to the development of advanced statistical solutions for FCM analysis. The FlowCAP project set about making computational algorithms an essential component of standard FCM data analysis. To address the expert-driven approach of FCM analysis, the FlowCAP project initiated four challenges focusing on the development and evaluation of automated gating solutions. The FlowCAP-I challenge focused on determining if automated algorithms could reproduce the results achieved from manual expert analysis, however it should be noted that comparing to manual experts may not be adequate due to inherent bias. The challenge provided 5 FCM experiments to competitors and ranked the competing automated solutions based on F-measure (Van Rijsbergen, 1979) and runtime. One of the competing algorithms in FlowCAP-I was flowClust (Lo et al., 2008) which has seen considerable usage in flow cytometry research with 228 paper citations and 118 package citations. As such flowClust will be discussed in the next section, and used as a comparison to the proposed methods of this research.

2.6 FlowClust

Model-based clustering has been shown to be an excellent method for identifying subpopulations across a variety of disciplines (Fraley and Raftery,
The standard approach is to assume that each cluster is represented by a single Gaussian distribution, thus the data can be represented via a mixture of Gaussian distributions (McLachlan and Basford, 1988). Lo et al. (2008) propose that a mixture of t-distributions is more appropriate to model FCM data than Gaussian mixtures due to the presence of outliers. In addition, the authors note that the gates identified within FCM do not conform to elliptical clusters as is the restriction within Gaussian model-based clustering. A Box-Cox transformation (Box and Cox, 1964) is also applied to the data to bring skewed data back to symmetry. The method is applied to the rituximab (Gasparetto et al., 2004) and a Graft-versus-Host-Disease (GvHD) experiment (Brinkman et al., 2007) in the paper. The analysis will be reproduced and critically reviewed in this section.

Figure 2.5: Gating of the rituximab data via the approach of Lo et al. (2008).

The rituximab data consists of 1545 cells measured for FSC and SSC, along with two fluorescent markers. Prior to analysis 36 observations which recorded maximum intensities in either FSC or SSC were removed and categorised as doublets (that is cells which have been stuck together in the streaming process). Figure 2.5(a) presents the initial gating solution of the t-mixture model, restricted to identifying a single cluster. The cells contained within
the initial gate were then gated on the remaining FCM variables, 7 AAD and Anti-BrdU FITC. Figure 2.5(b) and figure 2.5(c) present the two and three cluster solutions. The three gate solution of \textit{flowClust} is chosen as it minimises the Bayesian information criterion (BIC), (Schwarz, 1978).

![Gating of the GvHD data via the $t$-mixtures approach.](image)

Figure 2.6: Gating of the GvHD data via the $t$-mixtures approach.

The GvHD data contains a sample from two individual patients each measured across four fluorescent markers with in excess of 6800 cells per sample. The GvHD data used in this paper has already undergone the initial gating procedure on the FSC and SSC variables, and as such is only subject to the fluorescent marker gating. The control patient included in the GvHD data contains 6809 cells, as initial gating is already completed all cells are included in fluorescent gating. The gating solution of CD4 FITC against CD3 PerCP with 4, 5 and 6 gates are shown in figures 2.6(a)-(c), where the five gate solution minimises the BIC.

In the results shown above for the rituximab and GvHD data there appears to be significant overlapping of the identified gates due to the underlying modelling assumptions of \textit{flowClust}. The statistically best gating solution, as chosen by BIC, of the rituximab data is the three gate solution, however
the overlapping gates exhibit a shared region which contain a large proportion of the total cell population of both gates. The optimal five gate solution of the GvHD contains gates which exhibit the same issue. The suggested flexibility of the \( t \)-mixtures and the deployment of a Box-Cox transformation to correct for skewed data appears to be inadequate even when applied to well studied FCM data sets. This analysis further strengthens the requirement for a robust statistical model which incorporates the fundamental structure of the data and is not restrictive on the size, shape or orientation of resulting gates.

2.7 Summary

A detailed explanation of the scientific background of FCM is provided within this chapter. An examination of the data collection process provides evidence of the underlying lattice structure of FCM data. The \( t \)-mixture approach of Lo et al. (2008) known as flowClust is currently considered the statistical “gold standard”, however the treatment of the data by the method is inappropriate. The lack of a statistical framework which incorporates the underlying data structure, while removing the subjectivity of manual gating will continue to restrict the deployment of this world-leading sensor technology. The remainder of this thesis will focus on developing a statistically robust, unified framework which respects the inherent structure of the recorded data through the specification of regular lattice grids. Chapter 3 provides the novel statistical and probabilistic quantities required for such a framework to be developed.
3 Bernoulli Random Variables

This thesis exploits the underlying lattice structure of FCM data (discussed in chapter 2) to develop a novel statistical gating solution. This chapter develops useful mathematical results of Bernoulli random variables observed on a finite dimensional lattice. A local neighbourhood is specified whereby nodes within a confined region are conditionally dependent on each other. Let energy be a measure of similarity of nodes (such that low energy means largely similar nodes). The energy function of Ising (1925), the simplest energy function for lattice data, is defined on the lattice which acts as a measure of similarity among neighbouring nodes. However, any function which provides a measure of similarity and dissimilarity across a lattice grid can be utilised as the energy function. A Bernoulli random variable is specified at each node which allows the distribution of the energy at each node to be found. In turn, the complete distribution of the energy across the lattice can be calculated by considering a finite set of possible configurations of nodes on the lattice. The new statistical and probabilistic properties defined in this chapter form the basis for a novel gating procedure of FCM data in chapter 4 and chapter 5. Appendix A provides proofs of the mathematical results outlined in this chapter.

3.1 Configurations on a Lattice

Let $L$ be an $N \times M$ regular lattice grid with nodes at all locations $(i, j)$ for $1 \leq i \leq N$ and $1 \leq j \leq M$. With respect to FCM data the dimension of
the lattice is restricted to $2^n \times 2^m$, however all of the results in this chapter can be applied to lattices of any dimension. Let $y = (y_{ij})_{(i,j) \in L}$ represent the observed data at each node. Assume that the node quantities $y_{ij}$ are restricted to binary values, where a value of 0 signifies an inactive node and a value of 1 denotes an active node. Let each of the nodes $L_{ij} \in \{0, 1\}$ form a random field. The lattice $L$ with neighbourhood system $\eta_L$ is a Markov random field if and only if

$$\Pr\{L_{ij} = \ell_{ij}|L_{i'j'} = \ell_{i'j'}, (i', j') \in L, (i', j') \neq (i, j)\} = \Pr\{L_{ij} = \ell_{ij}|L_{i'j'} = \ell_{i'j'}, (i', j') \in \eta_{ij}(h)\};$$

that is the probability of $L_{ij} = \ell_{ij}$ is conditional only on its neighbouring nodes as defined by the specified neighbourhood system.

At each node $(i, j)$ of $L$ let the neighbourhood system of range $h$, which connects $(i, j)$ to nodes within its locality, be of the simplest form, $h = 1$. This is known as the von Neumann neighbourhood of range one (von Neumann and Burks, 1996). Mathematically the local neighbourhood system is

$$\eta_{ij} := \{(i', j') : 0 < \sqrt{(i - i')^2 + (j - j')^2} \leq 1\}, \quad (3.1)$$

where $i' \in \{1, \ldots, N\}$ and $j' \in \{1, \ldots, M\}$. The collection of all local neighbourhoods $\eta_{ij}$ across all nodes of the lattice $L$ is given by

$$\eta_L = \{\eta_{ij} : (i, j) \in L\}. \quad (3.2)$$

Assuming a ‘free-edge’ boundary condition\(^1\) on $L$ the neighbourhood system has four neighbours for interior nodes, two neighbours for corner nodes and three neighbours for the remaining exterior nodes, as illustrated in figure 3.1.

---

\(^1\) A ‘free-edge’ boundary condition is used in contrast to a periodic boundary condition which wraps the lattice grid into a torus like structure.
Definition 3.1. [Node Energy.] At each node \((i,j)\) the energy contributed is dependent only on its neighbours as specified in equation (3.1). The node contribution to the energy is given by

\[
\xi_{ij} = -(2\ell_{ij} - 1) \sum_{\eta_{ij}} (2\ell_{\eta_{ij}} - 1),
\]

where a node achieves a minimum energy state when it agrees with all of its neighbours and maximum energy if it disagrees with each of its neighbours.

Definition 3.2. [Lattice Energy.] Consider the energy across the lattice based on the neighbourhood system defined in equation (3.2). The energy of the complete lattice \(L\) is defined as the sum of the node energies across the lattice such that

\[
\xi_L = \sum_{(i,j)} \xi_{ij} = - \sum_{(i,j)} (2\ell_{ij} - 1) \sum_{\eta_{ij}} (2\ell_{\eta_{ij}} - 1),
\]

where minimum energy is achieved when all nodes of the lattice are in agreement and maximum energy is achieved where each node disagrees with each of its neighbours.
3.2 The Generalised Binomial Distribution

**Definition 3.3.** [The Generalised Binomial Distribution.] Consider the independent random variables $Y \sim \text{Bern}(\pi)$ and $X \sim \text{Binom}(\theta, n)$, where $n$ is the number of trials and $\pi$ and $\theta$ are probabilities of observing an active node. Let the random variable $G$ be the product of $Y$ and $X$, such that the probability mass function of $G = YX$ is given by

$$
\Pr(G = g) = \begin{cases} 
(1 - \theta)^n + (1 - \pi) \sum_{i=1}^{n} \binom{n}{i} \theta^i (1 - \theta)^{n-i}, & g = 0 \\
\pi \binom{n}{g} \theta^g (1 - \theta)^{n-g}, & g > 0.
\end{cases}
$$

The distribution of a random variable $G$ is referred to as the generalised binomial distribution, denoted $G \sim \text{GenBinom}(\pi, \theta, n)$. The expected value of $G$ is $E[G] = n\pi\theta$, while its variance is $\text{Var}[G] = n\pi\theta(1 - \theta + n\theta - n\pi\theta)$. □

**Definition 3.4.** [The Symmetric Generalised Binomial Distribution.] The random variable $G \sim \text{GenBinom}(\theta, \theta, n)$ is a symmetric generalised binomial distribution with probability mass function given by

$$
\Pr(G = g) = \begin{cases} 
(1 - \theta)^n + (1 - \theta) \sum_{i=1}^{n} \binom{n}{i} \theta^i (1 - \theta)^{n-i}, & g = 0 \\
\binom{n}{g} \theta^{g+1} (1 - \theta)^{n-g}, & g > 0,
\end{cases}
$$

with $E[G] = n\theta^2$ and $\text{Var}[G] = n\theta^2(1 + n\theta)(1 - \theta)$. □

**Lemma 3.1.** [Covariance between Generalised Binomials.] Consider the independent random variables $Y \sim \text{Bern}(\pi_Y)$, $Z \sim \text{Bern}(\pi_Z)$ and $X \sim \text{Binom}(\theta, n_X)$. Let $G_1 = YX \sim \text{GenBinom}(\pi_Y, \theta, n_X)$ and let $G_2 = YZ \sim \text{GenBinom}(\pi_Z, \pi_Y, 1)$. The covariance between $G_1$ and $G_2$ is

$$
\text{Cov}[G_1, G_2] = n_X \theta \pi_Y \pi_Z (1 - \pi_Y),
$$

where $G_1$ and $G_2$ are both generalised binomial distributions. □
3.3 Properties of the Lattice Energy

Let the random variables $X_{ij} \sim \text{Binom}(\theta, n_X)$ and $Y_{ij} \sim \text{Bern}(\theta)$ represent the number of active neighbours and the activity of the node of interest respectively, where $\theta$ is the probability of a node being active and $n_X$ is the number of neighbours a node. Redefine the node energy stated in definition (3.1) as

$$
\xi_{ij} = -(2Y_{ij} - 1)(2X_{ij} - n_X) = -4Y_{ij}X_{ij} + 2n_XY_{ij} + 2X_{ij} - n_X. \quad (3.3)
$$

Using the linearity property of expectation the expected value is written as

$$
E \left[ \sum_{(i,j)} \xi_{ij} \right] = \sum_{(i,j)} E [\xi_{ij}] .
$$

The expected energy at a node can be calculated by taking the expectation of equation (3.3) with the assumption that $Y$ and $X$ are independent. Thus

$$
E [\xi_{ij}] = -4E[Y_{ij}]E[X_{ij}] + 2n_XE[Y_{ij}] + 2E[X_{ij}] - n_X,
$$

$$
= -n_X (2\theta - 1)^2.
$$

The expected value of the node energy is dependent on the number of nodes within its neighbourhood. Corner nodes have two neighbours while edge and inner nodes have three and four neighbours respectively. Let $\#C$, $\#E$ and $\#I$ be the number of corner, edge and inner nodes respectively. Only lattices where $N, M \geq 2$ are considered (i.e. the smallest considered is a $2 \times 2$ lattice) and for these lattices$^2$, $\#C = 4$, $\#E = 2N + 2M - 8$ and $\#I = NM - 2N - 2M + 4$. Thus the expected energy of the lattice can be expressed as

$$
E[\xi_L] = -\left[ 2\#C + 3\#E + 4\#I \right] (2\theta - 1)^2,
$$

$$
= -\left( 4NM - 2N - 2M \right) (2\theta - 1)^2. \quad (3.4)
$$

$^2$Vector based lattices are not of interest in this research, however a similar approach can be used to calculate the expected energy on such a lattice.
The variance of the energy across the complete lattice is the sum of the
covariances between each of the node energies. The defined neighbourhood
system on the lattice forces many of the covariance terms to zero as each node
is conditional only on those nodes within its local neighbourhood, a visual
example is given in figure 3.2. As such the variance on the lattice can be
expressed as

$$\text{Var} \left[ \sum_{(i,j)} \xi_{ij} \right] = \sum_{(i,j)} \text{Var}[\xi_{ij}] + \sum_{(i,j)} \sum_{(l,m) \neq (i,j)} \text{Cov}[\xi_{ij}, \xi_{lm}].$$

The variance of the energy on individual nodes of the lattice can be expressed
in terms of $Y_{ij}$ and $X_{ij}$ as defined above. Taking the variance of equation (3.3)
and using definition (3.4) the node variance is

$$\text{Var}[\xi_{ij}] = 16\text{Var}[Y_{ij}X_{ij}] + 4n_{X}^{2}\text{Var}[Y_{ij}] + 4\text{Var}[X_{ij}],$$

$$= 4n_{X}\theta(1 - \theta)(n_{X}(2\theta - 1)^{2} + 1).$$

Figure 3.2: Covariance structures are only required when the neighbourhood
systems overlap. (a) shows non-overlapping neighbourhood. (b) shows two
overlapping neighbourhoods. Red is the node of interest, blue its non-
overlapping neighbourhood and green overlapping neighbours.

The covariance between node energies can be formulated in terms of three
unique lattice structures which account for the overlap between neighbouring
neighbourhoods (figure 3.3). The following three lemmas provide the covari-
ances required to calculate the variance of the lattice energy.
Figure 3.3: Overlap between neighbouring neighbourhoods. \( Y, Z, P \) and \( Q \) represent Bernoulli nodes, \( X \) are the neighbours of \( Y \) not in the neighbourhood of \( Z \) and \( W \) are the neighbours of \( Z \) not in the neighbourhood of \( Y \).

**Lemma 3.2.** Let \( Y \sim Z \sim \text{Bern}(\theta) \), \( X \sim \text{Binom}(\theta, n_X) \) and \( W \sim \text{Binom}(\theta, n_W) \), structured as shown in figure 3.3(a). Define \( \xi_Y \) and \( \xi_Z \) to be the energy of nodes \( Y \) and \( Z \). The covariance is given by

\[
\text{Cov}[\xi_Y, \xi_Z] = 4\theta(1 - \theta)((1 + n_X + n_W)(2\theta - 1)^2 + 1)
\]

by the results outlined in lemma (3.1). \( \square \)

**Lemma 3.3.** Let \( Y \sim Z \sim P \sim \text{Bern}(\theta) \), \( X \sim \text{Binom}(\theta, n_X) \) and \( W \sim \text{Binom}(\theta, n_W) \), depicted in figure 3.3(b). Define \( \xi_Y \) and \( \xi_Z \) to be the energy of nodes \( Y \) and \( Z \) respectively. The covariance between \( \xi_Y \) and \( \xi_Z \) is given by

\[
\text{Cov}[\xi_Y, \xi_Z] = 4\theta(1 - \theta)(2\theta - 1)^2
\]

by the results outlined in lemma (3.1). \( \square \)

**Lemma 3.4.** Let \( Y \sim Z \sim P \sim Q \sim \text{Bern}(\theta) \), \( X \sim \text{Binom}(\theta, n_X) \) and \( W \sim \text{Binom}(\theta, n_W) \), constructed as shown in figure 3.3(c). Define \( \xi_Y \) and \( \xi_Z \) to be the energy of nodes \( Y \) and \( Z \) respectively. The covariance between \( \xi_Y \) and \( \xi_Z \) is given by

\[
\text{Cov}[\xi_Y, \xi_Z] = 8\theta(1 - \theta)(2\theta - 1)^2
\]

by the results outlined in lemma (3.1). \( \square \)
The covariance terms can be calculated by counting the number of each of the structures specified in lemma (3.2), lemma (3.3) and lemma (3.4), further details can be found in appendix A (page 82). Thus, the variance of the energy on a complete lattice is

\[
\text{Var}[\xi_L] = 16\theta(1-\theta)[(14NM-13N-13M+8)(2\theta-1)^2+(2NM-N-M)] \quad (3.5)
\]

### 3.4 Densities of the Lattice Energy

The probability density functions of the energy at individual nodes and complete lattices, provides a metric by which the probability of agreement or disagreement of the nodes on a lattice can be quantified. This measure of agreement allows regions of similarity within the lattice to be identified.

#### 3.4.1 Node Energy Probability Mass Function

The probability distribution of the node energy can be calculated directly from equation (3.3). The probability distribution is dependent on \( \theta \), the probability that a node is active, and \( n \), the number of neighbours of the node of interest. The number of neighbours of a node varies depending on its position within the lattice. Thus the probability density of the energy for an inner node of the lattice is given by

\[
\Pr(\xi = \epsilon; \theta, n = 4) = \begin{cases} 
\theta^5 + (1-\theta)^5, & \epsilon = -4 \\
4\theta(1-\theta)(\theta^3 + (1-\theta)^3), & \epsilon = -2 \\
6\theta^2(1-\theta)^2, & \epsilon = 0 \\
4\theta^2(1-\theta)^2, & \epsilon = +2 \\
\theta(1-\theta)(\theta^3 + (1-\theta)^3), & \epsilon = +4.
\end{cases} \quad (3.6)
\]
In a similar manner each edge node has a set of 3 neighbours. The density of the energy at an edge node is specified as

\[
\begin{align*}
\Pr(\xi = \epsilon; \theta, n = 3) &= \begin{cases} 
\theta^4 + (1 - \theta)^4, & \epsilon = -3 \\
3\theta(1 - \theta)(\theta^2 + (1 - \theta)^2), & \epsilon = -1 \\
6\theta^2(1 - \theta)^2, & \epsilon = +1 \\
\theta(1 - \theta)(\theta^2 + (1 - \theta)^2), & \epsilon = +3.
\end{cases}
\end{align*}
\] (3.7)

The final nodes to classify are the four corner nodes which exist in all lattices considered in this thesis, where the dimension is at least \(2 \times 2\). For each corner node there are 2 neighbours with

\[
\begin{align*}
\Pr(\xi = \epsilon; \theta, n = 2) &= \begin{cases} 
\theta^3 + (1 - \theta)^3, & \epsilon = -2 \\
2\theta(1 - \theta), & \epsilon = 0 \\
\theta(1 - \theta), & \epsilon = +2.
\end{cases}
\end{align*}
\] (3.8)

giving the probability density function on the energy of these nodes. Further detail on the node energy probability densities is provided in appendix A.7 (page 86).

### 3.4.2 Lattice Energy Probability Mass Function

#### The \(2 \times 2\) Lattice

The energy function for the \(2 \times 2\) lattice is

\[
\begin{align*}
\Pr(\xi = \epsilon_L) &= \begin{cases} 
\theta^4 + (1 - \theta)^4, & \epsilon = -8 \\
4\theta(1 - \theta)(\theta^2 - \theta + 1), & \epsilon = 0 \\
2\theta^2(1 - \theta)^2, & \epsilon = +8.
\end{cases}
\end{align*}
\] (3.9)

The probability function for the energy across a \(2 \times 2\) lattice is found by considering each of the 16 possible lattice configurations (see appendix A.8
(page 89)). For each lattice configuration the energy can be calculated and the probability of observing the lattice is the product of the appropriate Bernoulli random variables.

Figure 3.4: The probability mass function of the energy of a $2 \times 2$ lattice for $\theta = \{0.25, 0.5, 0.75\}$.

Figure 3.5: The probability mass function of the energy of a $3 \times 3$ lattice for $\theta = \{0.25, 0.5, 0.75\}$. 
The $3 \times 3$ Lattice

The distribution of the energy on a $3 \times 3$ lattice is given by

$$
\Pr(\xi_L = \epsilon) = \begin{cases} 
\theta^9 + (1 - \theta)^9, & \epsilon = -24 \\
4\theta(1 - \theta)(\theta^7 + (1 - \theta)^7), & \epsilon = -16 \\
4\theta(1 - \theta)(\theta^5 + (1 - \theta)^5), & \epsilon = -12 \\
\theta(1 - \theta)(-23\theta^6 + 69\theta^5 - 81\theta^4 + 47\theta^3 - 11\theta^2 - \theta + 1), & \epsilon = -8 \\
12\theta^2(1 - \theta)^2(3\theta^2 - 3\theta + 1), & \epsilon = -4 \\
2\theta^2(1 - \theta)^2(4\theta^4 - 8\theta^3 + 16\theta^2 - 12\theta + 5), & \epsilon = 0 \\
12\theta^3(1 - \theta)^3, & \epsilon = +4 \\
\theta^3(1 - \theta)^3(17\theta^2 - 17\theta + 10), & \epsilon = +8 \\
4\theta^3(1 - \theta)^3, & \epsilon = +12 \\
4\theta^4(1 - \theta)^4, & \epsilon = +16 \\
\theta^4(1 - \theta)^4, & \epsilon = +24,
\end{cases}
$$

where $\theta$ is the probability of a node being active. The specification of the probability distribution on the $3 \times 3$ lattice is constructed by calculating the energy contributed by each of the 512 configurations (see appendix A.8 (page 89)).

### 3.4.3 Minimum & Maximum Energy

For each lattice the proportion of active nodes $\theta$ and the dimension of the lattice is known. While the complete probability density function for the energy can be found for larger lattices, the number of configurations to examine is huge. However, from the probability mass functions for the $2 \times 2$ and the $3 \times 3$ lattice it can be seen that the probability of observing a minimum energy state of a
lattice (see appendix A.8 (page 86)) is given by

\[ \Pr(\min \xi_L) = \theta^\alpha + (1 - \theta)^\alpha, \]  

(3.10)

where \( \alpha = N \times M \) is the number of nodes on the lattice. The minimum energy state probabilities for varying lattice sizes are shown in figure 3.6. In a similar manner, the lattice being in a state of maximum energy is observed with probability

\[ \Pr(\max \xi_L) = \theta \left\lfloor \frac{N M}{2} \right\rfloor (1 - \theta) \left\lceil \frac{N M}{2} \right\rceil + \theta \left\lceil \frac{N M}{2} \right\rceil (1 - \theta) \left\lfloor \frac{N M}{2} \right\rfloor. \]  

(3.11)

Quantifying the minimum and maximum energy states allows for the identification of regions of nodes approaching complete agreement (i.e. a gate) or complete disagreement (i.e. isolated nodes).

Figure 3.6: Probabilities of observing minimum energy state for \( \theta \). The red curve is for a 2 \( \times \) 2 lattice, while the purple curve is 2\(^6\) \( \times \) 2\(^6\). The remaining lines represent the dyadic structures between these two dimensions.

### 3.5 Properties of the Estimands

The quantification of the uncertainty, the statistical properties of the energy function and the probabilistic quantities is of particular importance for use as an information criterion. Figure 3.7 shows the energy function and
exact percentiles at \{2.5\%, 5\%, 25\%, 75\%, 95\%, 97.5\%\} levels. The heatmap represents the energy of a simulated $3 \times 3$ lattice, where the scale is from grey to blue, and white denotes an unobserved value. The theoretical average of the energy function tracks the shape of the heatmap intensity accurately. The percentiles provide more information, relevant to the observed energy, towards the centre of the distribution which would suggest that the variance is minimised close to 0 and 1, as expected.

Figure 3.7: Heatmap of the energy function on a $3 \times 3$ lattice for $\theta = \{0, 0.02, \ldots, 1\}$. The red line is the average theoretical energy, the quartiles are given in green, while the purple and blue are the 90\% and 95\% bounds.

Figure 3.8 shows the variance of the energy of a $3 \times 3$ lattice. It shows that the theoretical variance of the energy (red line) is maximised at approximately $\theta = \{0.19, 0.81\}$. As the dimension of the lattice tends to infinity the maximum of the variance occurs at $\theta = \left\{\frac{1}{2} - \sqrt{\frac{3}{28}}, \frac{1}{2} + \sqrt{\frac{3}{28}}\right\} \approx \{0.17267, 0.82732\}$, which is calculated by taking the limit of the derivative of the variance. The boxplots show the estimates of the variance for $3 \times 3$ lattices for 10000 simulations, each with 1000 runs for each value of $\theta = \{0, 0.01, \ldots, 1\}$.
3.6 Summary

This chapter has focused on developing novel statistical methodology for data lying on a regular lattice grid, a feature exhibited by FCM recordings. The specification of a local neighbourhood dependence similar to those seen in the study of Markov random fields is utilised. The novelty of this methodology is the specification of each node as a Bernoulli random variable. This statistical property of the nodes allows the expected value, variance and in some cases the density of node and lattice energy to be quantified. While these properties inform the agreement or disagreement of nodes within the lattice, ultimately an algorithm which relies on these quantities is required for gating FCM data. To achieve minimum energy across a lattice Markov chain Monte Carlo procedures, specifically simulated annealing, are currently employed. Utilising the node energy probability densities given in this chapter an adapted
simulated annealing procedure is developed in chapter 4. In addition, an innovative hierarchical approach is presented in chapter 5, which is based on statistical and probabilistic quantities derived in this chapter. The adapted simulated annealing approach is embedded in the hierarchical methodology, and this unified framework is applied to gating of FCM data in chapter 6.
4 Adapted MCMC

The minimisation of the energy function of Ising (1925) detailed in chapter 3 is often carried out using Markov chain Monte Carlo methods, in particular a simulated annealing approach. For lattices of large dimensions, as would arise with FCM data, standard simulated annealing approaches are inefficient due to the arbitrary choice of updated configurations, which results in low acceptance rates. This chapter proposes an adapted simulated annealing approach which utilises the probabilistic properties of the node energies outlined in chapter 3. An example of the novel adaptation is provided on binary images, structures equivalent to binary lattices, reaching optimal solutions while reducing the computational time and simulated annealing iterations required. The proposed adaptation provides a computational advantage over standard simulated annealing approaches for FCM data, while also providing a probability map which quantifies the agreement between nodes across the lattice structure.

4.1 Classical Energy Minimisation

Early and direct methods for identifying regions within lattice data were manual classification based approaches (Grasselli, 1969; Narasimhan, 1969; Schowengerdt, 2012), largely due to the lack of computing power. The development of probabilistic methods, and the availability of computing resources, moved beyond knowledge based approaches allowing for the dependence on
expert knowledge to be reduced and a probability density to be applied to model a lattice. The two main areas of research within probabilistic approaches over the last decade have been established within the Markov random field framework and the belief networks\(^1\) framework.

Pearl (1986) introduced the concept of belief networks which allow the prior knowledge of an expert to be incorporated into the computational architecture of a network. Pearl (1988) provides a complete discussion on the area of belief networks, with a focus on their use in intelligent systems, in particular their use in artificial intelligence.

The probabilistic approaches within the Markov random field framework are of particular importance within this research. The seminal papers of Besag (1974), Geman and Geman (1984) and subsequently Besag (1986) introduced the concept of a Markov random field for the identification of features within lattices. Much of the work to date relies heavily on the three aforementioned papers, as will the proposed method outlined in the following section. While a significant body of research has focused on adapting this early work, none have exploited the Bernoulli outcomes evident in a binary lattice.

The attractiveness of the Markov random field approach is the local dependence of neighbouring nodes (see section 3.1). The value of a node at a particular location is more likely to be similar to those nodes within a confined region in comparison with nodes a large distance away. The Hammersley-Clifford theorem (Hammersley and Clifford, 1971) provides a probabilistic

\(^1\)Probabilistic reasoning networks have not followed a strict naming policy and as such belief networks in this context encompasses Bayesian, belief, independence and causal networks.
density on the Markov random field through the specification of the Gibbs distribution

\[
\Pr(\Omega = \omega) = \frac{1}{Z} e^{-\xi(\omega)},
\]

where \(\omega\) is the configuration of the lattice, \(\xi(\cdot)\) is the energy function and \(Z\) is the normalising constant. The normalising constant is specified as

\[
Z = \sum_{\omega \in \Omega} e^{-\xi(\omega)}
\]

which requires the infeasible sum over all possible lattice configurations. The distribution allows the lattice to be represented by a probability map. The local dependence and the probability map provide a framework within which images, binary and non-binary, can be analysed. The Ising model (Ising, 1925), briefly discussed in section 1.2, is one of the simplest Markov random fields used in the setting of ferro-magnetism. The Ising model provides the basis of the models discussed throughout the rest of this chapter, with a proposed adaptation using the underlying Bernoulli probabilities on the lattice grid.

The estimation of a Markov random field results in the energy function being non-convex, making the identification of a global minimum computationally challenging. Problems of this form are categorised as combinatorial optimisation problems. For an \(N \times M\) binary image, that has only two possible labels at each node, the number of possible configurations is \(2^{NM}\). Even for a small \(8 \times 8\) image there are \(1.84 \times 10^{19}\) possible configurations, making an exploration of the complete space computationally infeasible. As such, a number of MCMC procedures have been developed to fit Markov random fields (Kato and Zerubia, 2012; Winkler, 2012). The Metropolis-Hasting algorithm (Metropolis et al., 1953), one of the first proposed solutions, is utilised throughout this chapter.
4.2 Metropolis-Hasting Algorithm

In its original formulation the Metropolis-Hasting algorithm examines $N$ particles contained within a finite sized square. The potential energy of particles within the square can be calculated if the particle positions are known and the potential between particles is quantifiable. The potential energy of the system is

$$\xi = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1 \atop j \neq i}^{N} V(\delta_{ij}),$$

where $V$ denotes the potential between molecules and $\delta_{ij}$ the distance between two nodes. As $\delta_{ij}$ increases the function $V(\cdot)$ increases, therefore the optimisation process focuses on the minimisation of the complete energy of the space. The computation of the following integral

$$\bar{F} = \frac{\int F \exp \left( -\frac{\xi}{\kappa T} \right) d^{2N}p d^{2N}q dF}{\int \exp \left( -\frac{\xi}{\kappa T} \right) d^{2N}p d^{2N}q dF}$$

is required to calculate the equilibrium state where $d^{2N}p d^{2N}q$ is a 4$N$-dimensional phase space volume element ($d$, $p$, and $q$ are undefined in the original paper), $\kappa$ is the Boltzmann constant and $T$ is the temperature of the system, a control parameter which stems directly from thermodynamics. The phase space volume element is an interpretation from thermodynamics but for our purposes is not required. The proposed procedure to calculate this integral involves the movement of particles within a local region as shown in figure 4.1.

Let $\omega$ be the current configuration of the particle system. Let each particle’s location be represented by the standard coordinate system $(x, y)$. Choose a particle at random and shift its original position within a local $2\alpha \times 2\alpha$ square (denoted in blue in figure 4.1). The particle’s new position is $x' = x + \alpha u_x$.
and $y' = y + \alpha u_y$, where $u_x$ and $u_y$ are uniform random numbers on $[-1, 1]$. Let $\xi$ and $\xi'$ be the energy of the existing particle system and the proposed particle system respectively, with $\Delta \xi = \xi - \xi'$. Using the following criterion a decision is made if the existing particle arrangement should be retained or if the proposed arrangement, denoted $\omega'$, should be accepted,

$$
\omega = \begin{cases} 
\omega', & \text{if } \Delta \xi \geq 0, \\
\omega', & \text{if } \exp(\Delta \xi / \kappa T) > u_{\xi}, \\
\omega, & \text{otherwise,}
\end{cases}
$$

(4.1)

where $u_{\xi}$ is a uniform random variable on $[0, 1]$.

Figure 4.1: Proposed Metropolis algorithm movement steps. White circles indicate initial particle positions, while the grey circle is the initial location of the selected node. The blue square signifies the potential movement range of the grey node, with the black arrow and red circle showing a potential move.

If the proposed configuration decreases the system energy then the proposed configuration is accepted. If the proposed configuration results in an increase in energy then it is accepted with a probability based on the energy change $\Delta \xi$ and the temperature of the system, otherwise no change is made. The procedure is iterated until the particles are positioned in an equilibrium minimum energy state.

39
Intuitively, the original formulation of the Metropolis-Hasting algorithm is not fit for purpose when analysing lattice data since the movement of node positions is not possible. For binary lattices the value of the randomly chosen node is altered rather than being moved. The decision process for the retained lattice configuration follows as in the original formulation, subject to an appropriate energy function. The criterion in equation (4.1) is the Metropolis criterion and is the decision mechanism used in simulated annealing, the procedure detailed in the next section. While it is obvious that that minimal energy is achieved when all nodes are in agreement this is not the optimal solution for cluster identification. As such the energy is minimised subject to physical constraints imposed by the data. These constraints are not hard coded in but emerge naturally due to the data, the algorithmic setup and tuning parameters.

4.3 Simulated Annealing

One of the most commonly known combinatorial optimisation problems is that of the travelling salesman. The objective is to find the optimal travel route between each of \( N \) cities returning to the starting city, where the distance between each pair of cities, \( D_{ij} \), is known. The travelling salesman problem, along with many combinatorial optimisation problems, are among the \( NP \)-complete problems. An algorithm to solve such problems in computer time of a power of \( N \) is not possible, in fact the time increases superpolynomially with the number of cities. Simulated annealing has been proposed as one solution for combinatorial optimisation problems which can find close to optimal solutions in polynomial time.
Simulated annealing was proposed independently by Kirkpatrick et al. (1983) and Černý (1985). In the thermodynamics formulation of the problem simulated annealing can be seen as a method which heats up a solid such that the particles arrange themselves in a liquid phase then the solid is slowly cooled to find the equilibrium position of the particles. The outcome is that all particles are now arranged in a low energy state. For each temperature of the solid a thermal equilibrium state is reached, at which point the temperature is decreased. This thermal equilibrium is specified via the Boltzmann distribution

$$\Pr(\Omega = \omega) = \frac{1}{Z(T)} \exp \left( - \frac{\xi(\omega)}{\kappa T} \right),$$

where $Z$ is the normalising constant, $T$ the temperature of the system and $\kappa$ the Boltzmann constant. In fact the above equation collapses to the Gibbs distribution discussed in section 4.1 for a fixed value of the temperature.

From the Boltzmann distribution it is clear that for high temperatures most configurations will be accepted regardless of the change in energy. However, as the temperature decreases the focus is on states which reduce the energy within the system. As the temperature approaches zero only states with minimal energy will exhibit non-zero probabilities. The control of the temperature, or rate of annealing, is of particular importance to the implementation of the simulated annealing algorithm. The initial temperature $T_0$, the final temperature $T_k$ and the cooling schedule, all need to be specified appropriately. If the temperature is cooled too quickly then a global minimum may not be achieved, as thermal equilibrium would not have been reached at each temperature. Cool too slowly and the algorithm will take superpolynomial time to converge to the solution. Geman and Geman (1984) propose an initial temperature of 4. While the final temperature should be $T_k = 0$ such a
value of $T_k$ is unreachable in finite steps, and as such the final temperature is
decided based on the energy function not decreasing for a number of successive
temperature values. Kirkpatrick et al. (1983) proposed an exponential cooling
rate such that $T_{k+1} = cT_k$, with $c = 0.95$. However, $c$ is dependent on the
structure of the lattice energy and values in the range 0.98 to 0.999 have
been proposed for more complex situations (Kato and Zerubia, 2012). At
each temperature $T_k$ the algorithm computes $L_k$, fixed or variable, Metropolis
algorithms to reach a thermal equilibrium. In general, for higher temperature
values $L_k$ needs to be larger to allow for accepted proposals which increase the
energy. As the temperature decreases the number of steps required at each
$T_k$ is decreased. In practice, and in this thesis, $L_k \approx 0.6 \times N \times M$ (Kato
The identification of the optimal solution is recognised when the algorithm reaches a state whereby the change in energy over a number of temperature value is below a predefined threshold. In general the practice is to terminate the algorithm if the energy does not change over 5 consecutive temperature values (Kato and Zerubia, 2012). Algorithm 4.1 shows the simulated annealing algorithm.

### 4.4 Adapted MCMC

Since the simulated annealing approach is comprised of a sequence of Markov chains, it is governed by state transition probabilities. Mathematically these transition probabilities are stated as

$$P_{\omega\omega'} = \begin{cases} G_{\omega\omega'}A_{\omega\omega'}, & \text{if } \omega \neq \omega', \\ 1 - \sum_{\ell \notin \omega} G_{\omega\ell}A_{\omega\ell}, & \text{if } \omega = \omega', \end{cases} \quad (4.2)$$

where $G_{\omega\omega'}$ is the probability of generating configuration $\omega'$ from configuration $\omega$ and $A_{\omega\omega'}$ denotes the probability of accepting configuration $\omega'$ after it is generated from configuration $\omega$. In practice the generation probability which changes the initial lattice configuration to the proposed configuration is specified by a uniform, Gaussian or Gibbs distribution. Choosing the node to perturb in such an arbitrary manner can result in many proposal steps being rejected, thus extending completion times of the algorithm. An adaptation of the generation probability utilising the underlying distribution of the node energy function, shown in chapter 3, is the principal of the proposed adaptation.
Let $G_{ij}$ be the probability that node $(i, j)$ changes value between two successive lattice configurations generated such that high-energy nodes are identified. This is achieved by utilising the node energy probabilities defined in equation (3.6)-(3.8) and allowing $\theta$ to be node dependent.

Denote $\omega_{ij} = Y_{ij}$ as the node to be changed, $n^T_{ij}$ the total number of neighbours of $\omega_{ij}$ and $n^A_{ij}$ the number of neighbouring nodes of $\omega_{ij}$ which are active. Define a local region which surrounds the node $(i, j)$ such that $n^T_{S, ij}$ and $n^A_{S, ij}$ are the total number of nodes in the local region and the total number of active nodes in the region respectively. Under the constraint of a von Neumann and Burks (1996) first-order neighbourhood system then the local region is defined to be a first-order Moore (1962) neighbourhood. The probability that a node is active is then specified as

$$\hat{\theta}_{ij} = \frac{n^A_{S, ij} + \omega_{ij}}{n^T_{S, ij} + 1}. \quad (4.3)$$

With this information the estimation of the probability that a node takes a given energy value in a specific neighbourhood is possible, that is the value of $\Pr(\xi_{ij} = \epsilon_{ij}; \theta = \hat{\theta}_{ij}, n = n^T_{ij})$, which is derived in section 3.4.1.

Thus, $G_{ij}$, the probability of generating a configuration which differs from the initial configuration only at node $(i, j)$, is given by

$$G_{ij} = \frac{\Pr(\xi_{ij} = \epsilon_{ij}; \theta_{ij} = \hat{\theta}_{ij}, n = n^T_{ij})^{-1}}{\sum_{(i,j)} \Pr(\xi_{ij} = \epsilon_{ij}; \theta_{ij} = \hat{\theta}_{ij}, n = n^T_{ij})^{-1}}, \quad (4.4)$$

where the reciprocal prioritises those nodes for which their calculated energy is not in a low-energy state. The choice of which node’s value to change is now based on the probability of the observed energy occurring given the current system structure. This requires a calculation of the energy at each node across the lattice prior to initialising the simulated annealing algorithm, and at each
iteration which can be computationally expensive.

![Figure 4.2: Examples of initial configurations of a lattice: (a) shows a lattice with all nodes in agreement; (b) shows a lattice with one disagreeing node; (c) shows a lattice with alternating nodes.](image)

Consider the three initial configurations of the lattice shown in figure 4.2, and calculate each of their generation probability matrices. In a slight abuse of notation some scalar quantities defined earlier are utilised in this example section as a matrix, this is for demonstration purposes only. The configuration of figure 4.2(a) shows all nodes in the same state, thus $\omega$ is a $3 \times 3$ matrix comprised of entirely +1s. The total number of neighbours, $n_T$, and the total number of neighbours in the local region, $n_T^S$, are given by

$$n_T = \begin{pmatrix} 2 & 3 & 2 \\ 3 & 4 & 3 \\ 2 & 3 & 2 \end{pmatrix} \quad \text{and} \quad n_T^S = \begin{pmatrix} 3 & 5 & 3 \\ 5 & 8 & 5 \\ 3 & 5 & 3 \end{pmatrix}.$$  

For the system where all nodes are in the +1 state $n_A = n_T$ and $n_A^S = n_T^S$.

Using $\omega$, $n_T$, $n_A$, $n_T^S$, $n_A^S$ and equations (3.6)-(3.8), the generation probability for $\omega'$ differing from $\omega$ at node $(i, j)$ is

$$G_{\omega\omega'} = \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix}.$$  

Figure 4.2(b) shows a configuration where the centre node is different to all of its surrounding nodes, thus $\omega$ is all +1s with the exception of $\omega_{22} = 0$.  

45
Intuitively, the node to select for a proposed change of value is the node which disagrees with all of its neighbours. The total number of neighbours and the number of nodes in the local region is equivalent to the first example, while the number of active neighbours and active neighbours in then local region is

\[
\begin{align*}
n^A &= \begin{pmatrix} 2 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 2 \end{pmatrix} \quad \text{and} \quad n^A_S &= \begin{pmatrix} 2 & 4 & 2 \\ 4 & 8 & 4 \\ 2 & 4 & 2 \end{pmatrix}.
\end{align*}
\]

The resulting generation probabilities \(G\) give largest probability to the centre node as expected, the complete probabilities are

\[
G_{\omega\omega}' \approx \begin{pmatrix} 0.0621 & 0.0902 & 0.0621 \\ 0.0902 & 0.3907 & 0.0902 \\ 0.0621 & 0.0902 & 0.0621 \end{pmatrix}.
\]

The final configuration within the example set, figure 4.2(c), shows a lattice grid of alternating nodes. All nodes disagree with each of their neighbours and as such the number of active neighbours is

\[
\begin{align*}
n^A &= \begin{pmatrix} 0 & 3 & 0 \\ 3 & 0 & 3 \\ 0 & 3 & 0 \end{pmatrix} \quad \text{and} \quad n^A_S &= \begin{pmatrix} 1 & 3 & 1 \\ 3 & 4 & 3 \\ 1 & 3 & 1 \end{pmatrix}.
\end{align*}
\]

The resulting generation probabilities are

\[
G_{\omega\omega}' = \frac{1}{50039} \begin{pmatrix} 3510 & 6656 & 3510 \\ 6656 & 9375 & 6656 \\ 3510 & 6656 & 3510 \end{pmatrix} \approx \begin{pmatrix} 0.070 & 0.133 & 0.070 \\ 0.133 & 0.187 & 0.133 \\ 0.070 & 0.133 & 0.070 \end{pmatrix},
\]

with the node with the largest number of disagreeing neighbours taking the highest probability. From the three elementary examples it can be seen that the proposed change of the generation probabilities increases the chance of selecting configurations that substantially change the energy system.
4.5 Probability Map

For each configuration of the lattice a probability map $P_{\text{min}, ij}$ which quantifies the measure of local node agreement is specified. The probability map is given by

$$P_{\text{min}, ij} = \begin{cases} 
\theta^5 + (1 - \theta)^5, & \text{if } n_{ij} = 2 \text{ (corner node)}, \\
\theta^4 + (1 - \theta)^4, & \text{if } n_{ij} = 3 \text{ (edge node)}, \\
\theta^3 + (1 - \theta)^3, & \text{if } n_{ij} = 4 \text{ (inner node)}, 
\end{cases}$$

(4.5)

which is generated from section 3.4.1. The map is attained from the probability of each node being in a state of minimum energy. Defining $\theta$ as in equation (4.3) the probability map is restricted to a set of discrete probabilities for each type of node (i.e. corner, edge or inner). The probability map allows for a visual representation of the energy function across the full lattice.

4.6 Application to Binary Images

To test the accelerated simulated annealing approach two binary images, which are equivalent to binary observables on a lattice, are considered. The letter A and the letter G are shown in figure 4.3. Both images are $128 \times 128$ node lattices consisting of 0s and 1s. This section initially adds noise to the observed images of A and G. Then the simulated annealing algorithm and the proposed adaptation are used to reconstruct the original images. Finally, a comparison of solutions to the original images, and between the methods is presented.

Each of the images in figure 4.3 are distorted, where a node is retained with a probability of 0.8 or switched with probability 0.2. The resulting noisy images are shown in figure 4.4(a) and figure 4.5(a). The transformed image is
Figure 4.3: Sample binary images each comprised of $128 \times 128$ pixels.

denoted by $I'$, which is formulated as

$$I' = I - C(2I - 1),$$

where $C$ is a binary matrix with each element dependent on a uniform random variable $u_C \in (0, 1)$ such that

$$C_{ij} = \begin{cases} 1, & \text{if } u_C \geq 0.8, \\ 0, & \text{if } u_C < 0.8. \end{cases}$$

Figure 4.4: The noisy letter A with the solutions from a simulated annealing approach and the accelerated method proposed in this chapter.

Figure 4.4 shows the noisy image of the letter A, along with the final state from the simulated annealing approach and the approach with the adapted generation probabilities. The values of $T_k$ and $L_k$ utilised are those stated earlier in this chapter. The simulated annealing approach, using a uniform
generation distribution, and the accelerated method produce images which are similar to the original letter A, verified by the contingency matrix in table 4.1, with accuracy of 95.99% and 97.30% respectively. However, the number of runs required to attain the solution is 199560 for the standard simulating annealing approach, taking 297.48 seconds, and 85305 for the proposed methodology, taking 262.42 seconds.

<table>
<thead>
<tr>
<th></th>
<th>Simulated Annealing (297.48 seconds)</th>
<th>Proposed Adaptation (262.42 seconds)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>0</td>
<td>11536</td>
<td>396</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>261</td>
<td>4191</td>
</tr>
<tr>
<td>Total</td>
<td>11797</td>
<td>4587</td>
<td>11960</td>
</tr>
</tbody>
</table>

Table 4.1: Method comparison between the original image, the simulated annealing procedure and the proposed adaptation for the letter A.

Figure 4.5: The noisy letter G with the solutions from a simulated annealing approach and the accelerated method proposed in this chapter.

The methods were also applied to the distorted letter G shown in figure 4.5(a). The resulting images are shown in figure 4.5(b)-(c), with both methods returning images similar to the initial image shown in figure 4.3(b). The contingency matrix is given in table 4.2 from which accuracy for the simulated annealing and the accelerated approach is calculated as 94.89% and 96.64% respectively. While the difference in accuracy levels of the competing approaches is minimal the number of Metropolis-Hasting iterations computed
to reach such accuracy differs considerably with simulated annealing requiring 184096 iterations, taking 280.51 seconds of computing time, and the adapted generating approach requiring only 92391 iterations, requiring 289.23 seconds.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Simulated Annealing (280.51 seconds)</th>
<th>Proposed Adaptation (289.23 seconds)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10524 405</td>
<td>10718 211</td>
<td>10929</td>
</tr>
<tr>
<td>1</td>
<td>432 5023</td>
<td>339 5116</td>
<td>5455</td>
</tr>
<tr>
<td>Total</td>
<td>10956 5428</td>
<td>11057 5327</td>
<td>128^2</td>
</tr>
</tbody>
</table>

Table 4.2: Method comparison between the original image, the simulated annealing procedure and the proposed adaptation for the letter G.

The reconstruction of the distorted image of the letter A and the letter G by the standard simulated annealing approach and the accelerated method show little difference in accuracy levels. However, of particular importance is the decrease in the number of Metropolis-Hasting iterations required to reach an equilibrium state of the image. The letter images each comprise $2^{14}$ elements, in contrast to standard FCM images which contain $2^{20}$. The improved efficiency of the Metropolis-Hasting iterations will thus have an impact on the computation time required for FCM analysis.

4.7 Summary

The minimisation of the energy function across a lattice of nodes, which represents FCM data appropriately, is the focus of this chapter. Existing approaches are computationally expensive as many proposal steps are rejected due to the arbitrary generation of configurations. Utilising the novel probability densities of node energies outlined in chapter 3 an accelerated simulated annealing algorithm is developed. The improved scheme provides a computational advantage by allowing nodes not in a minimum energy state to be prioritised.
when generating proposed configurations. While the methodology improves
the sampling scheme when proposing an updated configuration, a hierarchical
approach would offer substantial advantages by identifying large regions of
agreement within the lattice, a typical feature of FCM data. Chapter 5
proposes a hierarchical gating strategy which relies on the statistical properties
derived in chapter 3 and incorporates the accelerated simulated annealing
approach.
5 Hierarchical Approach

The Bernoulli algebra derived in chapter 3 provided the basis for an accelerated MCMC procedure outlined in chapter 4. A hierarchical approach which incorporates the accelerated MCMC approach would allow large regions of node agreement to be examined at a high-resolution. Hierarchical analysis of lattice data has been used extensively to model both the local and global features (Bouman and Liu, 1991; Graffigne et al., 1995; Marques et al., 1992; Schneider et al., 2000). This chapter develops a pruned hierarchical approach which facilitates the detection of large regions in node agreement, while minimising the energy across the lattice.

5.1 Hierarchical Methods

Hierarchical approaches have been used in numerical analysis for solving systems of differential equations. The principle of these approaches is to use both coarse problem solutions, achieved by solving a simplified system, and fine-resolution solutions, attained by solving the complete problem. The coarse problem solution, which is computationally easier to solve, is used to correct the fine-resolution solution. Correcting the fine-resolution solution allows a computationally efficient approach, while achieving global solutions. The human vision system acquires images at multiple resolutions simultaneously to produce the images we see. The image processing community adopted hierarchical approaches in the mid-70's (Kato and Zerubia, 2012).
Multi-grid and hierarchical methods are often described via the principle of a pyramid scheme depicted in figure 5.1. When each of the layers in the pyramid are unconnected the methods are referred to as a multi-grid approaches, where the underlying model is represented by a bundle of smaller and smaller lattices. Those methods which rely on communication between layers are referred to as hierarchical approaches. This communication property of hierarchical models is of particular importance for this research, however a detailed overview of the multi-grid approaches can be found in Jolion and Rosenfeld (2012).

Hierarchical MRF models (Kato et al., 1993, 1996) rely on the connection between each pyramid layer, whereby each node in a coarse grid represents a $2 \times 2$ sub-lattice at the next resolution lattice. The model assumes that the stack of random fields, from the coarsest grid scale (one node) to the finest resolution (complete lattice), forms a Markov chain. The coarse solutions in the hierarchical lattice model, generated using a standard simulated annealing algorithm or otherwise, represent a simplified field of vision. The coarse solution is then projected onto the lattice at the next resolution level, whereby the process is iterated (see figure 5.2) until the complete lattice is estimated.
5.2 Pruning

Statistics involves the exploration of data with an aim of ultimately identifying a model which captures the underlying features in the data. Specifying a model that is too large or contains many variables risks overfitting to the observed data, often resulting in high error rates for the prediction of future observations. A simplistic model which does not capture the underlying features, a phenomenon known as underfitting, also exhibits low predictive power for future observations. The identification of the optimal statistical model is a well-studied area of statistics known as model selection, used across all aspects of statistics.

![Decision Trees](image)

Figure 5.3: A decision tree model for identifying survivors on board the Titanic using the individual’s sex, class, and age. (a) shows the complete decision tree model, which is overfitting the training data. (b) shows the pruned decision tree model, which produces less decision nodes.

Pruning is an approach to model selection used in the context of decision trees or more generally tree-based classification procedures (Breiman et al., 1984). In the classification tree setting the idea is to grow the tree minimising the misclassification error at each step (see figure 5.3(a)). Fitting a decision
tree in such a fashion leads to overfitting. To correct for this the tree is pruned using a cost complexity pruning criterion (Friedman et al., 2001). The pruning criterion works in a similar manner to an information criterion whereby the accuracy of the tree model is balanced with the number of decision nodes (coefficients) in the tree. This pruning procedure results in the classification tree being less complex and, more importantly, robust to variations in future observations (see figure 5.3(b)).

The use of pruning as a model selection criterion has not been restricted to tree-based classification methods, nor has it been restricted to minimising the effects of over-fitting. Killick et al. (2012) present a pruning procedure for the detection of changepoints. They show that their Pruned Exact Linear Time method under certain conditions can locate the global minimum of the associated cost function in a computational time linear in the number of observations. Gowaikar and Hassibi (2003) proposed a pruning based method for maximum likelihood decoding. The method achieved performance arbitrarily close to maximum likelihood, but which is computationally more efficient than a sphere decoder, a particular algorithm which can solve the maximum likelihood problem exactly. Pruning is used in both contexts to reduce the computational burden of the associated procedures.

A novel hierarchical pruning approach is presented to identify clusters in lattice data. The pruning will allow regions of organised nodes to be identified at a relatively large sub-lattice dimension, thus allowing the algorithm to focus only on regions where disorganisation of adjoining nodes is evident. The clustering algorithm relies on the properties of the lattice energy derived in chapter 3.
5.3 Pruned Hierarchical Method

Chapter 4 proposes an adapted simulated annealing approach for lattice data, which relies on the underlying properties of the node energies detailed in chapter 3. The proposed adaptation showed a reduction in the number of iterations over classical annealing methods, while achieving solutions of comparable accuracy. FCM data requires the exploration of lattices of dimensions in the order $2^{10} \times 2^{10}$, which even with an adapted simulated annealing scheme is computationally challenging. Hierarchical approaches can be used to restrict the dimension of the lattices that require exploration. Wavelet theory (Bruce and Gao, 1996) uses a hierarchical methodology to compute the wavelet transforms allowing global coarse features and local fine resolution components of an image to be captured. A hierarchical procedure for lattice data would improve the accuracy of identified regions while removing the computational complexity by reducing the number of required iterations.

![Hierarchical Approach Diagram](image)

Figure 5.4: Hierarchical approach to identifying clusters within an image.

The hierarchical structure utilised for the analysis of FCM data in this research is shown in figure 5.4, using the example of a simplistic $4 \times 4$ lattice. Let the complete lattice of dimension $2^n \times 2^n$, a dyadic structure, be denoted as $R_0$ the lattice at resolution zero. While the approach is not restricted
to dyadic square lattices, non-dyadic non-square lattices pose a considerable computational challenge. At resolution level $R_r$ there exists a maximum of $4^r$ sub-lattices each of dimension $2^{n-r} \times 2^{n-r}$, for $r \in \{0, 1, \ldots, n\}$. Each sub-lattice at resolution level $R_r$ can be combined to reconstruct each sub-lattice at resolution level $R_{r-1}$.

A pruning procedure will be used to reduce the number of sub-lattices which need to be explored. Many lattices will contain large regions of active or inactive nodes. Exploring regions of this nature is time consuming without providing an improvement on the accuracy of the identified regions. If a region is disorganised then the region warrants further investigation through the exploration of sub-lattices at additional resolutions. If a branch of the hierarchical tree is pruned at resolution $R_k$ then this represents the relevant section of the lattice. On completion of the pruning procedure the sub-lattices at the base of each branch are collated up the tree to reconstruct the complete lattice. The remainder of this section presents a number of decisions which govern the pruning procedure.

Given the proportion of active nodes $\theta$ and the dimension of the lattice, the expected value of the energy and its variance can be calculated using the results derived in chapter 3. Additionally, chapter 3 provides the probability of observing a minimum and maximum energy state of a lattice for a given resolution level $R_r$. These quantities provide the basis for the decision theory of the hierarchical pruning methodology. The pruning methods depend on four decisions. The two initial decisions focus on states where the lattice is in a completely organised or disorganised energy state. Both decisions remove the need for exploration of large regions of the lattice. The third decision
utilises the expected value and variance of the energy function. The final decision implements the accelerated simulated annealing approach of chapter 4 on smaller sub-lattices.

**Decision 1: Minimum Energy**

If the probability that the lattice is in a state of minimum energy, where \( \theta \) is the proportion of active nodes in the given lattice, exceeds a defined threshold \( \tau_{\text{min}} \), then the lattice is stopped at the current resolution. If \( \theta \geq 0.5 \), then all nodes in the lattice are set to 1, otherwise all nodes are set to 0. The threshold parameter \( \tau_{\text{min}} \) must lie between \([0,1]\), where a value of zero means every sub-lattice is deemed to be in minimum energy while \( \tau_{\text{min}} = 1 \) means only a completely organised lattice is selected. A threshold value of \( \tau_{\text{min}} = 0.1 \) is considered in this research, as this equates to approximately 2 to 3 nodes being out of sync per lattice.

**Decision 2: Maximum Energy**

A state of maximum energy suggests that the lattice is in a complete state of disorganisation. Generally, moving to the next resolution level will retain the same problem. The proposal is to skip a resolution level when it is known that the lattice is in a state of maximum energy. Let \( \tau_{\text{max}} \) be a threshold parameter which, if exceeded, results in a resolution level being skipped. A rescaled probability of observing maximum energy under the assumption of square lattices of dimension \( 2^r \times 2^r \), a simplification of equation (3.11), is given by

\[
\frac{2 \theta^{\frac{r}{2}} (1 - \theta)^{\frac{r}{2}}}{2 \left(\frac{1}{2}\right)^\alpha},
\]

where \( \theta \) is the proportion of active nodes in the sub-lattice under exploration and \( \alpha \) is the number of nodes on the lattice. The scaling factor \( 2\left(\frac{1}{2}\right)^\alpha \) is chosen
as it maximises the probability of observing minimum energy, this occurs at $\theta = 0.5$. Thus, $\tau_{\text{max}}$ close to one suggests complete disorganisation while $\tau_{\text{max}}$ close to zero is evident of an organised system. The rescaled probability measure is shown in figure 5.5. The threshold parameter is set as $\tau_{\text{max}} = 0.8$ for the analysis in this thesis, as this allows only those lattice that are highly disorganised to be captured.

![Figure 5.5: Probabilities of maximum energy state for $\theta$ across varying lattice sizes. The red curve is for a $2 \times 2$ lattice, while the purple curve is $2^6 \times 2^6$, the remaining lines represent the dyadic structures between these two dimensions.](image)

**Decision 3: Expected Energy**

While decision 1 and 2 provide simple metrics for identifying lattices which are completely organised or disorganised, in practice the lattice structure will lie somewhere between the two edge states. Let $\xi^{(i)}_r$ be the observed energy of the $i^{\text{th}}$ sub-lattice at resolution $r$ calculated by summing the energy of all nodes (equation (3.3)). The associated expected energy calculated from equation (3.4) is given by

$$E[\xi^{(i)}_r] = -4(\alpha - \sqrt{\alpha})(2\theta_{r-1} - 1)^2$$
for square dyadic lattices, where $\alpha = 2^{2(n-r)}$, the dimension of the lattice at resolution $r$ and $\theta_{r-1}$ represents the proportion of active nodes at resolution $r-1$. Let $\Delta \xi^{(i)}_r = \xi^{(i)}_r - \mathbb{E}[\xi^{(i)}_r]$ be the difference between the observed and expected energy. Define the scaled difference as

$$\tau_{var} = \frac{\Delta \xi^{(i)}_r}{\sqrt{\text{Var}[\xi^{(i)}_r]}},$$

(5.2)

where $\tau_{var} < 0$ implies that the lattice is achieving a lower energy than expected, while $\tau_{var} > 0$ implies the opposite. The variance, $\text{Var}[\xi^{(i)}_r]$, is derived in equation (3.5) and expressed for a square dyadic lattice as

$$\text{Var}[\xi^{(i)}_r] = 16\theta(1-\theta)[(14\alpha - 26\sqrt{\alpha} + 8)(2\theta - 1)^2 + (2\alpha - 2\sqrt{\alpha})].$$

Decisions are made based on the number of standard deviations by which the observed and expected energy differ. The following rules govern the process:

- if $\tau_{var} \leq -2$ then the accelerated simulated annealing approach of chapter 4 is employed as the lattice is already approaching minimum energy, thus the number of annealing iterations should be small;
- if $-2 < \tau_{var} \leq 0$ then the resolution level is changed from the current level $R_k$ to $R_{k+1}$ since the lattice is approaching an organised state;
- while if $\tau_{var} > 0$ then the resolution level is changed from $R_k$ to $R_{k+2}$ as the lattice is in a state of disorganisation.

To examine the proposed decision, four example $4 \times 4$ lattices are explored. The four example lattices are shown in figure 5.6. The proportion of active nodes across each lattice is set at $\theta = 0.5$, as such the expected value and variance of the energy on each lattice are $\mathbb{E}[\xi^{(i)}_r] = 0$ and $\text{Var}[\xi^{(i)}_r] = 96$. The first example (figure 5.6(a)) shows a system which is well organised, its scaled difference is $-3.27$. Thus, the decision is to apply the accelerated simulated
Figure 5.6: A set of example 4 × 4 lattices: (a) and (b) two lattices which are organised; (c) a lattice with one outlying node; (d) is a disorganised lattice.

annealing approach to the sub-lattice which would target isolated disagreeing nodes and then terminate the procedure. The lattice shown in figure 5.6(b) and 5.6(c) return a value of −1.63. Both structures are thus explored at the next resolution level, where figure 5.6(b) will achieve minimum energy in all quadrants and figure 5.6(c) sub-lattices A and C will require the accelerated simulated annealing approach which will flip the blue and red in sub-lattices A and C respectively. The procedure will then terminate as the constrained minimum state has been achieved. Finally, the disorganised lattice of figure 5.6(d) has a scaled difference in energy of 3.27. The lattice is thus explored at resolution level $R_{k+2}$. In this case where the resolution has reached individual nodes the simulated annealing approach will automatically suggest it is in a minimum energy state and terminate, decision 4 removes this possibility.

**Decision 4: Simulated Annealing**

Reducing the size of the lattice to a single node is impractical from a computational and statistical viewpoint as single nodes will not be changed by the simulated annealing procedure. Thus, if a sub-lattice reaches a dimension of $2^3 \times 2^3$, that is, it is at resolution level $R_n$, the accelerated simulated annealing approach discussed in chapter 4 is automatically employed. The
requirement is to select a lattice dimension whereby the number of simulated annealing iterations required is small while not sacrificing the outcome of the algorithm. The exact choice of lattice dimension to utilise needs further exploration.

5.4 Application to Binary Images

The binary images utilised in chapter 4 are used to explore the advantages of the hierarchical approach presented in this chapter. Figure 5.7 shows the two letter binary images, $A$ and $G$, along with the disturbed images. The disturbance to the images is formulated as

$$I' = I - C(2I - 1),$$

where $I$ and $I'$ are the original and disturbed images respectively. The matrix $C$ controls the level of disturbance using an uniform random variable $u_c \in (0, 1)$ such that

$$C_{ij} = \begin{cases} 
1, & \text{if } u_c \geq 0.8 \\
0, & \text{if } u_c < 0.8.
\end{cases}$$

![Figure 5.7: The letter A and letter G images, along with the noisy images.](image)

The resulting images after applying the hierarchical pruning method proposed in this chapter is shown in figure 5.8 for the letter $A$ and the letter $G$. For the letter $A$ the accelerated simulated annealing and hierarchical...
approach achieve similar results. However on closer inspection it can be seen that the hierarchical method provides a set of straight line edges, due to the deconstruction and reconstruction of the lattice, which is an unattractive feature imposed by the algorithm. Similar results can be seen for the letter $G$.

![A][A] ![G][G]

(a) ASA  (b) HPA  (c) ASA  (d) HPA

Figure 5.8: The letter $A$ and letter $G$ binary images solutions for the accelerated simulated annealing (ASA) and the hierarchical pruning algorithm (HPA) approaches.

Table 5.1 and table 5.2 compares the accelerated simulated annealing, and the hierarchical procedures to the original images. Marginal improvements are observed in the denoising of the letter $G$ when using the hierarchical approach, with an increase of accuracy from 95.24% to 95.30% compared with the accelerated simulated annealing method. The accuracy levels for the letter $A$ are 97.46% and 96.77% for the simulated annealing and the hierarchical model respectively.

<table>
<thead>
<tr>
<th></th>
<th>Chapter 4 Method (256.76 seconds)</th>
<th>Chapter 5 Method (184.41 seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1</td>
<td>0 1</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>11762 170</td>
<td>11668 264</td>
</tr>
<tr>
<td></td>
<td>246 4206</td>
<td>265 4187</td>
</tr>
<tr>
<td>Total</td>
<td>12008 4376</td>
<td>11933 4451</td>
</tr>
</tbody>
</table>

Table 5.1: The letter A: chapter 4 approach versus chapter 5 approach.

While the resulting reconstruction of the binary images included in this chapter do not improve on those from the accelerated simulated annealing approach of chapter 4, the method does provide a computational advantage.
The hierarchical approach provides a reduction in the time taken on both images, saving in excess of 1 minute of computational time on each image. No large sparse regions, a feature of FCM data, exist in the binary images due to the randomness of the added noise. In such circumstances the hierarchical pruning approach will provide a significant computational over the accelerated simulated annealing algorithm.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Chapter 4 Method (299.03 seconds)</th>
<th>Chapter 5 Method (238.59 seconds)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10596 333</td>
<td>10613 316</td>
<td>10929</td>
</tr>
<tr>
<td>1</td>
<td>447 5008</td>
<td>454 5001</td>
<td>5455</td>
</tr>
<tr>
<td>Total</td>
<td>11043 5341</td>
<td>11067 5317</td>
<td>128^2</td>
</tr>
</tbody>
</table>

Table 5.2: The letter G: chapter 4 approach versus chapter 5 approach.

5.5 Summary

The hierarchical procedure outlined in this chapter relies on the Bernoulli algebra derived in chapter 3 and the accelerated simulated annealing procedure provided in chapter 4. The pruned hierarchical procedure allows large regions exhibiting a minimum energy state to be identified early in the procedure reducing the dimension of the lattice requiring exploration. The hierarchical procedure provides improvements in computational efficiency versus the adapted simulated annealing approach, with little change in accuracy for the examples provided. However, since FCM data exhibits regions of node agreement more frequently than the noisy binary images presented, it is anticipated that the hierarchical approach will provide a computational advantage over the accelerated simulated annealing procedure. One limiting characteristic of the hierarchical approach is the straight line boundaries at predefined locations on the lattice. This algorithmic artifact could be overcome by
exploring overlapping boundaries within the hierarchical procedure. However, the methodology provides a novel statistical framework for the gating of FCM data.
6 Application to FCM

This chapter explores the application of the proposed method of chapter 5 to FCM data. Initially, an additional step required to derive the final gates is discussed. The proposed method will then be compared to flowClust, an existing technique within the FCM literature on two data sets. The rituximab data (Gasparetto et al., 2004), and a Graft-versus-host disease data set from Brinkman et al. (2007), both often used when proposing new methods within the field of cytometry, are analysed.

6.1 Gating

Due to sparsity evident in FCM data, a multi-resolution approach to the analysis is carried out. The data is aggregated to a $128 \times 128$ lattice where the hierarchical pruning procedure is applied. The resulting gating solution is then spatially smoothed which in turn informs the next level lattice ($256 \times 256$) continuing until the gates are identified on the complete lattice. The spatial smoothing of the data accounts for the boundaries between sub-lattices imposed by the hierarchical algorithm. The function used in this work is given by

$$x_{ij} = \frac{x_{ij} + \frac{1}{2}n^A_{ij} + \frac{1}{2\sqrt{2}}(n^A_{S,ij} - n^A_{ij})}{3 + \sqrt{2}}$$

where $n^A_{ij}$ is the number of active nodes a distance of 1 unit away and $(n^A_{S,ij} - n^A_{ij})$ is the number of active nodes which are exactly a distance of $\sqrt{2}$ from node $(i, j)$, thus weighting the nodes according to their distance.
from the node of interest. The resulting value is rounded to a zero or one outcome. After analysing data with the proposed method, the identification of boundary regions of clusters in the resulting probability map is addressed. Connected component labelling (Dillencourt et al., 1992) is an established pattern recognition tool used extensively in the identification of disjointed regions in binary, and with adaptations non-binary, lattices (Brummer et al., 1993; Moftah et al., 2010). Connected component labelling works by traversing the binary lattice resulting from analysing FCM data, node-by-node, to assign unique labels to each disjoint region within the lattice. Following a complete traversal of the lattice, the resultant lattice contains a set of disjoint regions, each corresponding to a unique cluster or gate.

6.2 Application to Rituximab Data

The initial gating step was replicated using the proposed methodology of chapter 5, with the resulting probability map shown in figure 6.1(b). The initial clustering on the FSC and SSC variables is shown in 6.1(c) which identifies the group of cells of interest for the second gating stage of the rituximab data. The initial gating solutions provided by flowClust (see figure 6.1(a)) and the proposed methodology exhibit both similarities and differences. The identification of the initial gates using flowClust produces a region whereby the density varies significantly, largely due to the restriction of elliptical cluster shapes imposed under the model-based clustering paradigm. In contrast, the proposed method identifies the region of high density and provides a probabilistic map where the colour spectrum of red, orange, yellow indicates varying probability levels of cell activity.
CHAPTER 6. APPLICATION TO FCM

Figure 6.1: (a)-(c) The initial gating stage based on FSC and SSC variables. (a) The t-mixture approach to gating. (b) The probability map generated from the proposed methodology. (c) The gates identified from the probability map, blue and grey points denote gates and unclassified points respectively. (d)-(f) The second gating stage based on 7 AAD v Anti-BrdU FITC. (d) The t-mixture approach to gating. (e) The probability map whereby red, orange, yellow and white regions denote high to low active probabilities. (f) The gates identified from the probability map denoted by blue points, while grey points denote unclassified points.
The restriction of size, shape and orientation of identified gates by *flowClust* is further exhibited in the second stage of the gating procedure (see figure 6.1(d)). Three clusters are identified using *flowClust* while the proposed method identifies a two cluster solution (see figure 6.1(f)). The lower cluster identified by the proposed methodology is largely equivalent to a combination of the two adjoining clusters in the competing approach. The upper gate in the *flowClust* solution captures a significant quantity of empty space, with the included points sparsely scattered in the region. In comparison the proposed methodology only captures a dense group of points, which exhibits more homogeneity within the gate.

### 6.3 Application to GvHD Data

The GvHD data is analysed using *flowClust* and the methodology proposed in this thesis. The GvHD data available has already been gated using the FSC and SSC variables and, as such, only stage two of the gating procedure is possible. The GvHD data contains 6809 cells in the control sample and 9083 cells in the positive sample. As the initial gating has already been completed (with doublets already removed) all cells are included in the second gating stage.

Figure 6.2 and figure 6.3 present the gating solutions on the control sample and positive sample of the GvHD data respectively. The gating of CD4 against CD8β by *flowClust* and the corresponding approach proposed are shown in figure 6.2(a)-(c) and figure 6.3(a)-(c) for the respective GvHD samples. Similarly, the gating of CD4 and CD3 by the aforementioned methods for the control and positive samples are shown in figure 6.2(d)-(f) and figure 6.3(d)-(f)
CHAPTER 6. APPLICATION TO FCM

Figure 6.2: GvHD Control: (a)-(c) Gating based on CD4 FITC and CD8β PE variables. (a) The $t$-mixture approach to gating. (b) The hierarchical model probability map showing high, medium and low active probabilities in red, orange and yellow respectively. (c) The clustering solution from the proposed methodology. (d)-(f) Gating solutions for CD4 FITC against CD3 PerCP. (d) The $t$-mixture approach to gating. (e) The probability map showing regions of probabilities where red, orange and yellow denote high, medium and low probabilities. (f) The sub-populations identified from the probability map.
Figure 6.3: GvHD Positive: (a)-(c) Gating based on CD4 FITC and CD8β PE variables. (a) The $t$-mixture approach to gating. (b) The probability map of the proposed methodology. (c) The clustering solution from the proposed methodology. (d)-(f) Gating solutions for CD4 FITC against CD3 PerCP. (d) The $t$-mixture approach to gating. (e) The probability map showing regions of probabilities where red, orange and yellow denote high, medium and low probabilities. (f) The sub-populations identified by the proposed methodology.
respectively. The solutions provided \textit{flowClust}, similar to the solution in the earlier rituximab gating solution, produces multiple overlapping clusters across both samples and both pairs of cytometry variables. However the proposed method is adaptive to unusual shapes of cell populations and thus produces solutions which define regions with irregularly shaped boundaries. The GvHD control sample produces two unique sub-populations for CD4 FITC with CD3 PerCP and CD4 FITC with CD8\(\beta\) PE under the proposed gating methodology. In contrast, \textit{flowClust} generates five and four cluster solutions to produce similar results.

The GvHD positive sample does not produce definitive clusters in either pairing of cytometry variables. However, the hierarchical model generates a probabilistic map around both regions of high density, producing a single cluster for for CD4 FITC with CD3 PerCP and a two cluster solution for CD4 FITC with CD8\(\beta\) PE. In contrast, \textit{flowClust} requires more than five components to produce a mirage of overlapping clusters.

6.4 Summary

The gating solution presented moves beyond the current standard of using manually drawn gates on 2D graphical representations. The proposed methodology is time efficient, reducing the considerable time-investment of experts required for manual gating. In addition, the reproducibility of gating solutions across laboratory experts moves FCM analysis in line with the reproducible science paradigm which is now paramount in applied research. The methodology automates the statistical analysis, while establishing a grounding in statistical inference rather than expert intuition.
While the work of Lo et al. (2008) moved beyond existing automated gating solutions resolving many of the issues with manual gating strategies, the methodology is restrictive in the gating solutions produced. The restriction on the size, shape and orientation of the resulting clusters produced under the \( t \)-mixture model-based clustering approach is overcome using the proposed methodology, allowing the data to inform the shape of the identified regions. In addition, the processing of the data by Lo et al. (2008) involving outlier removal and Box-Cox transformation is removed under the proposed methodology allowing the data to be used in its raw form.

To date the underlying structure of cytometry data, recorded in machines using analogue-to-digital converters, has been ignored in the modelling techniques. The modelling approach proposed in this thesis hinges on this underlying structure by exploiting the distance between the discrete values recorded. Using the data in its discrete untransformed structure is not only appropriate for the modelling of FCM data in this setting, but also allows a simplistic model which provides a probability map across the entire space on which the data lie. A naive approach within the proposed methodology is the active/inactive state of aggregated grid data, however this limitation can be overcome by extending the mathematical formulation of the underlying model. While the overall objective is to provide a completely automated solution to the gating of FCM data, the methodology proposed here provides a probabilistic grid which can be used to better inform manual gating strategies.

While the proposed methodology overcomes many of the issues that exist when utilising manual gating strategies, it cannot make decisions based on biological expertise. As such, it is advisable that the solutions produced by the
methodology should be interpreted by a FCM expert to ensure the underlying biological knowledge is adhered to.
7 Summary & Future Work

7.1 Summary

Chapter 1 and chapter 2 explored the scientific aspects of the data collection process in FCM. The mistreatment of the data in existing methodology, and the biased nature of manual gating motivated the requirement for a statistically appropriate method for gating. Chapter 3 specified the nodes of a lattice as Bernoulli random variables. Combined with the energy function of Ising (1925) a novel probabilistic quantification of the lattice energy was derived. The quantities of chapter 3 were utilised in chapter 4 to adapt simulated annealing so that nodes that when changed provide a significant decrease in energy are selected with a higher probability. Chapter 5 incorporated the approach of chapter 4 into a novel hierarchical pruning procedure in an effort to account for large regions of inactive nodes while retaining information on localised regions of active nodes. The thesis concluded by gating FCM data utilising the proposed methodology.

In summary, the novel methodology provided in this chapter moves beyond the current standard for gating FCM data. It respects the structure in the observed data, while providing reproducible gates across FCM laboratories and colleagues. Most importantly the approaches detailed are grounded in statistics, thus removing the requirement for biased manual gating to be used in practice. The mathematical details provided within this thesis can be extended
to clustering in multiple dimensions, however care must taken when defining
the neighbourhood energy function.

7.2 Further Work

7.2.1 Neighbourhood Systems

Figure 7.1: Lattice grid showing the first (red), second (red & blue) and
third (red, blue & green) order neighbourhoods of a node (black) for the von
Neumann and Moore systems.

Throughout this thesis the focus has been on utilising the first-order von
Neumann neighbourhood system (see equation (3.1)). One area of further work
is to expand this methodology to different neighbourhood systems. As before,
at each node \((i, j)\) of a regular lattice \(L\) let there exist a neighbourhood system
of range \(h\) (see figure 7.1), which connects \((i, j)\) to nodes within its locality.
For integer values of the range \(h\) the von Neumann neighbourhood system (von
Neumann and Burks, 1996) is observed, with the local neighbourhood system
at node \((i, j)\) given by

\[
\eta_{ij}(h) := \{(i', j') : 0 < \sqrt{(i - i')^2 + (j - j')^2} \leq h\}, \text{ if } h \in \mathbb{N}^+,
\]
where $i' \in \{1, \ldots, N\}$ and $j' \in \{1, \ldots, M\}$. For $h \in \mathbb{N}^+$ the square Moore neighbourhood (Moore, 1962) is constructed with the local neighbourhood system at node $(i, j)$ given as

$$\eta_{ij}(h) := \{(i', j') : 0 < \max(|i - i'|, |j - j'|) \leq h\}, \text{ if } h \in \mathbb{N}^+,$$

where $i' \in \{1, \ldots, N\}$ and $j' \in \{1, \ldots, M\}$. The calculation of the expected energy remains trivial due to the monotone convergence theorem allowing the summation over all nodes to be taken outside the expectation. The variance, under higher-order neighbourhood systems, requires additional covariance calculations than when using the first-order von Neumann.

### 7.2.2 Non-Binary Observables

While binary observables are of particular interest in FCM, an extension of this work to non-binary observables would be of considerable interest for the evolving FCM technology. A lattice of non-binary observables is one where the value at each node is a count of the data points at that location. It further extends beyond the binary observables (active and inactive) by allowing values outside 0 and 1 to exist. The methodology proposed would require substantial changes to accommodate such an extension as the underlying assumption of Bernoulli random variables at each node would need to be replaced with a more general distribution (i.e. binomial, Poisson or beta), which in turn would require a change in the distribution of each node’s neighbourhood. In the application to FCM this would accommodate nodes where more than one recording has been observed. While these counts are minimal currently, the further development of FCM technology, allowing the analysis of more cells, would benefit from a methodology which exploits multiple counts at an individual node.
7.2.3 Implementation

The clock speeds discussed in chapter 4 and chapter 5 do not appear to offer a significant advantage over existing methods. However, the programming of the proposed routines is naive and with further consideration could be greatly improved in terms of computational overhead and memory management. One of the main improvements would be to implement the proposed methods in a low-level language like C++, rather than in R. The computational engine of the competing algorithms are programmed in C++ or Fortran.
A Bernoulli Algebra

A.1 The Generalised Binomial Distribution

Let $Y$ and $X$ be Bernoulli and binomial random variables respectively with the following properties:

$$Y \sim \text{Bernoulli}(\pi)$$

$$\Pr(Y = y) = \pi^y(1 - \pi)^{1-y}$$

$$\mathbb{E}[Y] = \pi$$

$$\mathbb{E}[Y^2] = \pi$$

$$\text{Var}[Y] = \pi(1 - \pi)$$

$$X \sim \text{Binomial}(\theta, n)$$

$$\Pr(X = x) = \left(\begin{array}{c} n \\ x \end{array}\right) \theta^x (1 - \theta)^{n-x}$$

$$\mathbb{E}[X] = n\theta$$

$$\mathbb{E}[X^2] = n\theta(n\theta + 1 - \theta)$$

$$\text{Var}[X] = n\theta(1 - \theta)$$

Let $G = Y X \sim \text{GenBinom}(\pi, \theta, n)$ represent a generalised binomial where $Y$ and $X$ are independent random variables. The sample space of $G$ is \{0, 1, 2, \ldots, n\}. Consider first the case where $G = 0$:

$$\begin{cases}
(1 - \pi)(1 - \theta)^n, & \text{if } X = 0 \& Y = 0, \\
\pi(1 - \theta)^n, & \text{if } X = 0 \& Y = 1, \\
(1 - \pi) \sum_{i=1}^{n} \left(\begin{array}{c} n \\ i \end{array}\right) \theta^i (1 - \theta)^{n-i}, & \text{if } X \neq 0 \& Y = 0,
\end{cases}$$

resulting in $\Pr(G = 0) = (1 - \theta)^n + (1 - \pi) \sum_{i=1}^{n} \left(\begin{array}{c} n \\ i \end{array}\right) \theta^i (1 - \theta)^{n-i}$. For $G$ to be non-zero it requires $Y = 1$ and $X \neq 0$. Thus for $G = \{1, 2, \ldots, n\}$ the probability is given by

$$\Pr(G = g; g \neq 0) = \pi^{n-g} \left(\begin{array}{c} n \\ g \end{array}\right) \theta^g (1 - \theta)^{n-g}.$$
Finally, the complete probability density function of $G = YX$ is given by

$$
\Pr(G = g) = \begin{cases} 
(1 - \theta)^n + (1 - \pi) \sum_{i=1}^{n} \binom{n}{i} \theta^i (1 - \theta)^{n-i}, & g = 0 \\
\pi \binom{n}{g} \theta^g (1 - \theta)^{n-g}, & g > 0.
\end{cases}
$$

### A.2 Expected Value and Variance

From the $k^{th}$ moments of the generalised binomial, the expected value and variance can be calculated. The solution for the symmetric generalised binomial requires $\theta = \pi$ in the following.

$$
\begin{align*}
\mathbb{E}[G^k] &= \sum_{g=0}^{n} g^k \Pr(G = g) \\
\mathbb{E}[G^k] &= 0 \cdot \Pr(G = 0) + \sum_{g=1}^{n} g^k \left[ \pi \binom{n}{g} \theta^g (1 - \theta)^{n-g} \right] \\
\mathbb{E}[G^k] &= \pi \left[ \sum_{g=0}^{n} g^k \binom{n}{g} \theta^g (1 - \theta)^{n-g} \right] \\
\mathbb{E}[G^k] &= \pi \mathbb{E}[X^k]
\end{align*}
$$

Whence

$$
\mathbb{E}[G] = n\pi\theta,
$$

and

$$
\text{Var}[G] = n\pi\theta(1 - \theta + n\theta(1 - \pi)).
$$

### A.3 Covariance

Let $R = YX Y Z$ where $Y \sim \text{Bern}(\pi_Y)$, $X \sim \text{Binom}(\theta, n_X)$ and $Z \sim \text{Bern}(\pi_Z)$. The calculation of $\mathbb{E}[R^k]$ will result in the expected value and variance of $R$. Define $S = Y^2 \sim \text{GenBinom}(\pi_Y, 1, 1)$ and $Q = XZ \sim \text{GenBinom}(\pi_Z, \theta, n_X)$. 

80
Since $S$ and $Q$ are independent it can be stated that

\[
\Pr(R = r) = \begin{cases} 
(1 - \theta)^n + (1 - \pi_Y \pi_Z) \sum_{i=1}^{n} \binom{n}{i} \theta^i (1 - \theta)^{n-i}, & r = 0 \\
\pi_Y \pi_Z \binom{n}{r} \theta^r (1 - \theta)^{n-r}, & r > 0.
\end{cases}
\]

The calculation of the $k^{th}$ moment is given by $E[R^k] = \pi_Y \pi_Z E[X^k]$. Thus the expected value is stated as $E[R] = n_X \pi_Y \pi_Z \theta$ and the variance is given by $\text{Var}[R] = n_X \pi_Y \pi_Z \theta (1 - \theta + n_X \theta (1 - \pi_Y \pi_Z))$.

Now, let $U = YX \sim \text{GenBinom}(\pi_Y, \theta, n_X)$ and $V = YZ \sim \text{GenBinom}(\pi_Y, \pi_Z, 1)$, where $U$ and $V$ are not independent as they share the common variable $Y$. The covariance between $U$ and $V$ is given by $\text{Cov}[U, V] = E[UV] - E[U]E[V]$. Using the expected value of $R = UV$ and the properties of the generalised binomial the result is

\[
\text{Cov}[U, V] = n \pi_Y \pi_Z \theta (1 - \pi_Y).
\]

### A.4 Node Energy Function

The energy function of Ising (Ising, 1925) is specified in the Bernoulli setting of $\{-1, +1\}$. The energy function is defined as

\[
\xi_{ij} = -\ell_{ij} \sum_{n_{ij}} \ell_{n_{ij}}.
\]

Specifying the energy function on the standard Bernoulli support of $\{0, 1\}$ results in the energy function of the form

\[
\xi_{ij} = -(2\ell_{ij} - 1) \sum_{n_{ij}} (2\ell_{n_{ij}} - 1).
\]

Now, let $X \sim \text{Binom}(\theta, n_X)$ and $Y \sim \text{Bern}(\theta)$ re-express the energy function in terms of the random variables as

\[
\xi_{ij} = -(2Y - 1)(2X - n_X) = -4XY + 2n_X Y + 2X - n_X.
\]
A.5 Overlapping Neighbourhood Covariances

This section outlines the calculations required for the covariance between node energies on a discrete lattice. In particular the focus will be on the calculation of the covariance of the structures shown in figure 3.3 and the results of lemmas (3.2) – (3.4).

Lemma 3.2. The energy at node $Y$ and node $Z$ are given by

$$\xi_Y = -4Y(X + Z) + 2Y(n_X + 1) + 2(X + Z) - (n_X + 1)$$

and

$$\xi_Z = -4Z(W + Y) + 2Z(n_W + 1) + 2(W + Y) - (n_W + 1).$$

On expansion the covariance becomes


Using the fact that $\text{Var}[Y] = \text{Var}[Z]$ and $\text{Cov}[Y, YZ] = \text{Cov}[Z, YZ]$ the covariance is reduced to

$$\text{Cov}[\xi_Y, \xi_Z] = 16\text{Var}[YZ] + 16\text{Cov}[YX, ZY] + 16\text{Cov}[YZ, ZW] - 8\text{Cov}[YX, Y] - 8\text{Cov}[ZW, Z] - 8(n_X + n_W + 4)\text{Cov}[Y, Z] + 4(n_X + n_W + 2)\text{Var}[Y];$$

for which each covariance and variance term is known from the generalised binomial distribution and lemmas (3.2)–(3.4). As such the covariance between $\xi_Z$ and $\xi_Y$ is

$$\text{Cov}[\xi_Y, \xi_Z] = 4\theta(1 - \theta)((1 + n_X + n_W)(2\theta - 1)^2 + 1).$$
Lemma 3.3. The energy at node $Y$ and node $Z$ are given by

$$\xi_Y = -4Y(X + P) + 2Y(n_X + 1) + 2(X + P) - (n_X + 1)$$

and

$$\xi_Z = -4Z(W + P) + 2Z(n_W + 1) + 2(W + P) - (n_W + 1).$$

On expansion the covariance becomes

$$\text{Cov}[\xi_Y, \xi_Z] = 16\text{Cov}[YP, ZP] - 8\text{Cov}[Y, YP] - 8\text{Cov}[Z, ZP] + 4\text{Var}[P].$$

Using the fact that $\text{Cov}[Y, YP] = \text{Cov}[Z, ZP]$ the covariance is reduced to

$$\text{Cov}[\xi_Y, \xi_Z] = 16\text{Cov}[YP, ZP] - 16\text{Cov}[Y, YP] + 4\text{Var}[P];$$

for which each covariance and variance term is known from the generalised binomial distribution and lemmas (3.2)–(3.4). As such the covariance between $\xi_Z$ and $\xi_Y$ is

$$\text{Cov}[\xi_Y, \xi_Z] = 4\theta(1 - \theta)(2\theta - 1)^2.$$

\hfill \Box

Lemma 3.4. The energy at node $Y$ and node $Z$ are given by

$$\xi_Y = -4Y(X + Q + P) + 2Y(n_X + 2) + 2(X + Q + P) - (n_X + 2)$$

and

$$\xi_Z = -4Z(W + Q + P) + 2Z(n_W + 2) + 2(W + Q + P) - (n_W + 2).$$

On expansion the covariance becomes

$$\text{Cov}[\xi_Y, \xi_Z] = 16\text{Cov}[YP, ZP] + 16\text{Cov}[YQ, ZQ] - 8\text{Cov}[YQ, ZQ] - 8\text{Cov}[YP, ZP] + 4\text{Var}[P] + 4\text{Var}[Q].$$
Using the fact that $\text{Cov}[P,YP] = \text{Cov}[P,ZP] = \text{Cov}[Q,ZQ] = \text{Cov}[Q,YQ]$ and $\text{Cov}[YP,ZP] = \text{Cov}[YQ,ZQ]$ the covariance is reduced to

$$\text{Cov}[\xi_Y,\xi_Z] = 32\text{Cov}[YP,ZP] - 32\text{Cov}[Q,ZQ] + 8\text{Var}[P];$$

for which each covariance and variance term is known from the generalised binomial distribution and lemmas (3.2)–(3.4). Then the covariance between $\xi_Z$ and $\xi_Y$ is

$$\text{Cov}[\xi_Y,\xi_Z] = 8\theta(1-\theta)(2\theta - 1)^2.$$

\[ \square \]

### A.6 Properties of Lattice Energy

The expected value of the energy across the complete lattice can be calculated as the sum of each of the node expectations. The expected value is

$$E\left[\sum \xi_{ij}\right] = E\left[\sum -4XY\right] + E\left[\sum 2nXY\right] + E\left[\sum 2X\right] + E\left[\sum -nX\right],$$

then since $X$ and $Y$ are both non-negative random variables the monotone convergence theorem allows an exchange of the summation and expectation operators resulting in

$$E\left[\sum \xi_{ij}\right] = \sum (-4E[XY] + 2nXE[Y] + 2E[X] - nX).$$

The expected value of $X$, $Y$ and $XY$ are known through the expectation of the binomial, Bernoulli and generalised binomial distributions respectively. The expected energy across the lattice is thus

$$E\left[\sum \xi_{ij}\right] = \sum (-4nX\theta^2 + 2nX\theta + 2nX\theta - nX) = \sum -nX(2\theta - 1)^2.$$

Under the first-order von Neumann neighbourhood system corner, edge and inner nodes each have 2, 3 and 4 neighbours each. The number of corner, edge
and inner nodes is 4, 2N + 2M − 8 and NM − 2N − 2M + 4 respectively. Therefore, the expectation of the complete lattice energy is

\[ \mathbb{E}[\xi_L] = -(4NM - 2N - 2M)(2\theta - 1)^2. \]

The variance of the energy across the lattice is specified as

\[ \text{Var}\left[ \sum \xi_{ij} \right] = \sum \left\{ \text{Var}[\xi_{ij}] + \sum_{(\ell,m) \neq (i,j)} \text{Cov}[\xi_{ij}, \xi_{\ell,m}] \right\}. \]

The node level variance terms can be calculated similar to the expected energy by exchanging the position of the variance and summation operators and accounting for the required covariance terms, such that

\[ \text{Var}[\xi_{ij}] = 16\text{Var}[XY] + 4n_X^2\text{Var}[Y] + 4\text{Var}[X] - 16n_X\text{Cov}[XY,Y] - 16\text{Cov}[XY,X]. \]

Using the results defined earlier the node variance is

\[ \text{Var}[\xi_{ij}] = 4n_X\theta(1 - \theta)(n_X(2\theta - 1)^2 + 1). \]

The corner, edge and inner nodes each have varying number of neighbours as before and thus the summed variance at the individual nodes is expressed as

\[ \sum \text{Var}[\xi_{ij}] = 8\theta(1 - \theta)((8NM - 7N - 7M + 4)(2\theta - 1)^2 + (2NM - N - M)]. \]

The covariance terms are non-zero for nodes that overlap with a nodes neighbour, via the Markov property. The summed covariance term is constructed by a combination of the covariances shown in lemmas (3.2)–(3.4). Table A.1 gives the quantity of each of the structures shown in figure 3.3 required for a particular dimension of lattice grids. Of particular note is that the general \( N \times M \) solution for the covariance terms holds for all \( N, M \geq 3 \), however this is not true for any \( 2 \times M \) lattice. As alluded to in the main body of this thesis, lattice structures of dimension \( 1 \times M \) are not explored, however
APPENDIX A. BERNOULLI ALGEBRA

Table A.1: The number of each of the lemma structures from section 3.1 required for the variance of given sized lattices.

<table>
<thead>
<tr>
<th>Size</th>
<th>Lemma 3.2</th>
<th>Lemma 3.3</th>
<th>Lemma 3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>4</td>
<td>6M - 16</td>
</tr>
<tr>
<td>N</td>
<td>M</td>
<td>16</td>
<td>14M - 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n_x</th>
<th>n_y</th>
<th>n_W</th>
<th>d_1</th>
<th>d_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>16</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td>6M - 16</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>6M - 16</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

a similar approach of exploring possible structures can be used to quantify the energy associated with them.

While the covariance structures do not follow a consistent structure for all lattices of interest the complete measure of variance does. As such the variance of the energy for the complete lattice of dimension $N \times M$, $N, M \geq 2$ is

$$\text{Var}[\xi_L] = 16\theta(1 - \theta)[(14NM - 13N - 13M + 8)(2\theta - 1)^2 + (2NM - N - M)].$$

A.7 Probability of Node Energy

Corner Nodes

For each corner node the number of neighbouring nodes is two, and thus the energy function is a discrete outcome from the set $\{-2, 0, +2\}$. The possible configurations for a corner node are given in figure A.1, along with the associated probability of observing the configuration and its contribution to the energy. Summing over the probability of observing configurations for each of the discrete energy values gives the probability density function as

$$\text{Pr}(\xi = \epsilon) = \begin{cases} 
\theta^3 + (1 - \theta)^3, & \epsilon = -2 \\
2\theta(1 - \theta)^2 + 2\theta^2(1 - \theta), & \epsilon = 0 \\
\theta^2(1 - \theta) + \theta(1 - \theta)^2, & \epsilon = +2.
\end{cases}$$
APPENDIX A. BERNOULLI ALGEBRA

\[ \xi = -2 \quad \xi = -2 \quad \xi = 0 \quad \xi = 0 \quad \xi = +2 \quad \xi = +2 \]

\[ (1 - \theta)^3 \quad \theta^3 \quad \theta(1 - \theta)^2 \theta^2 (1 - \theta) \theta^2 (1 - \theta) \theta (1 - \theta)^2 \]

Figure A.1: Configurations of corner nodes with associated energy and the observation probability. Red is an active node observed with probability \( \theta \), while blue is an inactive node which occurs with probability \( 1 - \theta \).

**Edge Nodes**

In a similar manner to the corner nodes, the probability density function for the energy at edge nodes can be calculated given that each node has three neighbours. The energy function is restricted to \( \xi \in \{-3, -1, +1, +3\} \), with the configurations shown in figure A.2. The probability density function, formulated by summing the probability of observing each configuration with respect to the energy function values is

\[
\Pr(\xi = \epsilon) = \begin{cases} 
\theta^4 + (1 - \theta)^4, & \epsilon = -3 \\
3\theta(1 - \theta)(1 - 2\theta + 2\theta^2), & \epsilon = -1 \\
6\theta^2(1 - \theta)^2, & \epsilon = +1 \\
\theta(1 - \theta)(1 - 2\theta + 2\theta^2), & \epsilon = +3.
\end{cases}
\]

Figure A.2: Configurations of edge nodes with the associated energy and the observation probability. Red is an active node, while blue is an inactive node.
Interior Nodes

The probability density of the energy for each interior node is given by:

\[
Pr(\xi = \epsilon) = \begin{cases} 
\theta^5 + (1 - \theta)^5, & \epsilon = -4 \\
4\theta(1 - \theta)(\theta^3 + (1 - \theta)^3), & \epsilon = -2 \\
6\theta^2(1 - \theta)^2, & \epsilon = 0 \\
4\theta^2(1 - \theta)^2, & \epsilon = +2 \\
\theta(1 - \theta)(\theta^3 + (1 - \theta)^3), & \epsilon = +4.
\end{cases}
\]

Each interior node has four neighbouring nodes, and thus the energy function is a discrete outcome from the set \{-4, -2, 0, +2, +4\}. The possible configurations for an interior node are given in figure A.3, along with the associated probability of observing the configuration and its contribution to the energy.

Figure A.3: Configurations of inner nodes with the associated energy and the probability of observing such configurations. Red indicates an active node, while blue indicates an inactive node.
A.8 Probability of Lattice Energy

The 2 × 2 Lattice

The 2 × 2 lattice comprises of four corner nodes with the energy restricted to the discrete set \{-8, 0, +8\}. The possible configurations are shown in figure A.4 summing the probabilities over the unique energy values gives the 2 × 2 lattice energy density as

\[
\Pr(\xi = \epsilon) = \begin{cases} 
\theta^4 + (1 - \theta)^4, & \epsilon = -8 \\
4\theta(1 - \theta)(\theta^2 - \theta + 1), & \epsilon = 0 \\
2\theta^2(1 - \theta)^2, & \epsilon = +8.
\end{cases}
\]

The expected energy on the 2 × 2 lattice is given by

\[
\mathbb{E}[\xi] = -8(2\theta - 1)^2,
\]

which is as stated in the general solution in equation (3.4). The variance stated in equation (3.5) agrees with the variance calculated directly from the density function as

\[
\text{Var}[\xi] = \theta(1 - \theta)(192(2\theta - 1)^2 + 64).
\]

Figure A.4: Configurations of 2 × 2 lattice with the associated energy and the probability of observing such configurations. Red indicates an active node, while blue indicates an inactive node. The circled numbers denotes the number of each type of configuration.
The 3 × 3 Lattice

The 3 × 3 lattice consists of four corner nodes, four edge nodes and one interior node, with 512 possible configurations of the lattice. The set of configurations which can represent all 512 possible configurations are shown in figure A.5. The probability density function for the energy defined on the set \{-24, -16, -12, -8, -4, 0, +4, +8, +12, +16, +24\}, calculated by summing the probabilities of observing the configurations for each energy level, is given by

\[
\Pr(\xi = \epsilon) = \begin{cases} 
\theta^9 + (1 - \theta)^9, & \epsilon = -24 \\
4\theta(1 - \theta)(\theta^7 + (1 - \theta)^7), & \epsilon = -16 \\
4\theta(1 - \theta)(\theta^5 + (1 - \theta)^5), & \epsilon = -12 \\
\theta(1 - \theta)(-23\theta^6 + 69\theta^5 - 81\theta^4 + 47\theta^3 - 11\theta^2 - \theta + 1), & \epsilon = -8 \\
12\theta^2(1 - \theta)^2(3\theta^2 - 3\theta + 1), & \epsilon = -4 \\
20\theta(1 - \theta)^3(4\theta^4 - 8\theta^3 + 16\theta^2 - 12\theta + 5), & \epsilon = 0 \\
12\theta^3(1 - \theta)^3, & \epsilon = +4 \\
\theta^3(1 - \theta)^3(17\theta^2 - 17\theta + 10), & \epsilon = +8 \\
4\theta^3(1 - \theta)^3, & \epsilon = +12 \\
4\theta^4(1 - \theta)^4, & \epsilon = +16 \\
\theta^4(1 - \theta)^4, & \epsilon = +24.
\end{cases}
\]

The expected value and variance follow from equation (3.4) and (3.5) such that

\[
\mathbb{E}[\xi] = -24(2\theta - 1)^2
\]

and the variance is equal to

\[
\text{Var}[\xi] = \theta(1 - \theta)(896(2\theta - 1)^2 + 192).
\]
Figure A.5: Configurations of $3 \times 3$ lattice with the energy and the probability of observing such configurations. Red indicates an active node, while blue is an inactive node. The circled numbers denotes the number of each configuration.
Maximum Lattice Energy

The maximum energy is achieved when all nodes are in the opposite state to all nodes within their neighbourhood. The number of nodes on the lattice governs the probability of achieving maximum energy which is given as

\[
\Pr(\text{max} \, \xi) = \theta \left\lfloor \frac{NM}{2} \right\rfloor (1 - \theta) \left\lceil \frac{NM}{2} \right\rceil + \theta \left\lceil \frac{NM}{2} \right\rceil (1 - \theta) \left\lfloor \frac{NM}{2} \right\rfloor,
\]

where the floor and ceiling operators account for grids with an odd number of nodes.

Minimum Lattice Energy

For a general \( N \times M \) lattice it is known that the minimum energy is achieved when all of the nodes in the lattice are in agreement. The probability of achieving minimum energy is given by

\[
\Pr(\text{min} \, \xi) = \theta^{NM} + (1 - \theta)^{NM}.
\]

Decision 1 of the proposed hierarchical pruning approach given in section 5.3 states that if

\[
\tau_{\text{min}} = \theta^\alpha + (1 - \theta)^\alpha
\]

is less than a threshold all nodes in the lattice are set to 0 or 1. The value of \( \theta \) which satisfies this condition can be explored via asymptotics. The two values of \( \theta \) which satisfy the criterion are \( \theta \approx 0 \) and \( \theta \approx 1 \). Consider the case of \( \theta \approx 0 \) then \( (1 - \theta)^\alpha \) is significantly larger than \( \theta^\alpha \), thus

\[
(1 - \theta)^\alpha \left( \frac{\theta^\alpha}{(1 - \theta)^\alpha} + 1 \right) = \tau_{\text{min}},
\]

where the term in red is transcendentally small and can be ignored. Performing a change of variable such that \( \theta = \frac{1}{\alpha} x \), where \( x \sim \mathcal{O}(1) \), the criterion can be expressed as

\[
\alpha \log \left( 1 - \frac{x}{\alpha} \right) = \log(\tau_{\text{min}}),
\]
after taking the natural logarithm of the expression. Taking the Taylor expansion of the left hand side of the equation gives a second order solution of

\[ \theta \approx \frac{\log\left(\frac{1}{\tau_{\text{min}}}\right)}{\alpha} - \frac{\log\left(\frac{1}{\tau_{\text{min}}}\right)^2}{2\alpha^2}. \]

By the symmetry of the underlying problem $1 - \theta$ is also a solution.
B FCM Data File Standard

The Data File Standard for Flow Cytometry, is at the time of publication in Version 3.1 (Spidlen et al., 2010). The data standard allows for collected data (FCS files) to follow principled guidelines thus allowing for the development of software to analyse these high-dimensional, highly complex data sets. It also allows data collected on multiple cytometers using different data collection processes to be compared and analysed using common analysis software without the need for pre-processing the data files.

While the data standard for flow cytometry is extensive, there are general conventions to which the data standard does not require compliance. One such convention is that all numerical values in the data are base 10 unless otherwise stated. While these conventions are adhered to by the vast majority of the cytometry community, their exclusion from the data standard policy excludes their inclusion when developing analysis software for flow cytometry data.

The data files for flow cytometry stored in binary, under the data standard policy, conform to having 4 required sections and any number of additional sections required to fully explain the data. A full analysis of the entire data should be possible based on the 4 required sections and each additional section should only be used for details of the collection of the data that would not affect the analysis.
B.1 Header Section

This section provides references to the start and end of each section in the FCS file and provides a reference to the version of the data standard which the collected data adheres to. A general example of a header section within an FCS file is as follows

\[ FCS3.1\cdots256\cdots1545\cdots1792\cdots202455\cdots0\cdots0 \]

The example above indicates that the data adheres to data standard version 3.1, the text section starts at byte 256 and ends at byte 1545, the data section starts at byte 1792 and continues until byte 202455 and the remaining zeros indicate that no analysis section is provided in this data file (the * above are replaced by the spaces in a real header section).

B.2 Text Section (Metadata)

The text section of the data file provides the metadata produced during the collection of data from a flow cytometer. This metadata is stored as ASCII encoded keyword-value pairs, an example of which is “$TOT/5000/”. The example states that the total number of observations in the data collected is 5000. Under the data standard policy there are several keyword-value pairs that are required in the text section and several more exist that can be supplied if deemed necessary by the data collector.

B.3 Data Section

This segment contains the raw data which will be used in the analysis. The data is stored in binary form and first has to be converted before it can be
analysed. In general this data is the output of the flow cytometer, as required by the data standard policy, however it has been seen that some collectors pre-process their data and then include it in this section. Throughout this research, the assumption will be made that the data in this section has not been processed in any way.

B.4 Analysis Section

While this section is a required section under the data standard policy it can be left empty if no analysis has yet been carried out on the data. The analysis included here can vary from the results of a complete analysis to the initial pre-processing of the data. Due to the variety of information that could be included here it will be ignored even if supplied throughout this research.
Bibliography


