**Automated Gating of Flow Cytometry Data via Markov Random Fields[[1]](#footnote-1)**

**Running Headline:** Markov Random Field Clustering for FCM Data

**Corresponding Author:**

Kevin Christopher Brosnan

Department of Mathematics and Statistics,

University of Limerick,

Limerick,

Ireland.

Email: [kevin.c.brosnan@ul.ie](mailto:kevin.c.brosnan@ul.ie)

Phone: +353 87 644 3584

**Co-Authors:**

Dr. Norma Bargary

Department of Mathematics and Statistics,

University of Limerick,

Limerick,

Ireland.

Email: [norma.bargary@ul.ie](mailto:norma.bargary@ul.ie)

Dr. Kevin Hayes

Department of Mathematics and Statistics,

University of Limerick,

Limerick,

Ireland.

Email: [kevin.hayes@ul.ie](mailto:kevin.hayes@ul.ie)

**Abstract**

Flow Cytometry 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. The methodology outlined in this paper utilises the Ising model as an image segmentation tool, allowing for a distribution free method for the identification of homogenous cell populations in cytometry data. The implementation of the methodology is discussed whereby simulated annealing and hierarchical approaches are applied to improve computational efficiency. Finally, the use of two publicly available datasets allows the methodology to be evaluated by experts and in comparison with other automated gating solutions in the literature.

**Introduction**

Although the instrumentation and underlying technologies supporting flow cytometry (FCM) have advanced significantly in recent years (1, 2), the development of automated data analysis tools remain elusive, the absence of which has perhaps restricted the deployment of this state-of-the-art sensor technology (3). The conventional method for analysing FCM data has relied heavily on manual expert-driven approaches rather than a unified automated statistical framework (4). The development of such a statistical framework is in strong demand (2, 4-9), and would, ideally, allow for reproducible and standardised analyses to be conducted while also reducing the considerable time-investment currently required for manual analysis in FCM.

FCM analysis involves two key stages: the gating stage, where sub-populations of interest are identified, and the tagging stage, where correlations between the characteristics of identified sub-populations are explored. The gating procedure currently involves the manual drawing of gates to specify regions of interest in a 2D graphical representation of a pair of FCM variables (7, 10-12). This manual expert-driven approach is highly subjective across laboratories and to a lesser degree across colleagues within individual laboratories. In addition, the projection of high-dimensional data to a 2D graphical representation can lead to substantial information loss (13, 14). This information loss can result in identified cell populations not being representative across all dimensions of the data. A reliable statistical methodology with appropriate software implementations would aid in reducing the subjectivity and hence variability associated with manual gating.

Gating FCM data is an application of statistical cluster analysis where the requirement is to identify cell populations that are homogenous in nature. A variety of statistical clustering techniques have been applied to gating FCM data (9, 15, 16). The k-means algorithm and its extensions have been utilised extensively (8, 17-19) but classical k-means methods only allow cells to belong to one single cluster. To overcome this restrictive feature, a fuzzy k-means clustering approach that allows cells to belong to multiple clusters with an associated probability of cluster membership, has been proposed (19). However, the criterion employed to select the ‘best’ clustering solution in these approaches can unduly restrict the shape, size and orientation of identified sub-populations, which can result in failure to identify “true” clusters. Some authors have attempted to use supervised learning algorithms such as neural networks (20, 21) and support vector machines (22, 23) to automate the gating procedure. These approaches require the availability of training data which often renders them unsuitable in a field where training data is not always available.

Model-based clustering has gained prominence as a statistical tool for the gating of FCM data (13,14). Model-based clustering methods assume that sample observations arise from a mixture of one or more probability densities where each probability density represents a unique sub-population or cluster (24-27). Typically it has been assumed that each mixture component follows a *p*-variate Gaussian distribution, with the number of clusters identified via a suitable information criterion (28). The implications of the Gaussian assumption are that the resulting sub-populations will be elliptical in shape which is not always true for FCM data. Lo et al. (14) considered model-based clustering with *p*-variate mixtures of *t*-distributions in FCM analysis. The larger tail of the *t*-distribution makes the approach more robust to outliers, a common feature of FCM data due to cell debris and doublets, but still retains the elliptical shape constraint of Gaussian mixtures. However in general, the identified sub-populations in FCM do not conform to elliptical clusters even after appropriate data transformation and/or when accommodation is made for outliers.

While several approaches for the automation of FCM gating have been proposed, none to date have exploited the inherent structure of the underlying data. The use of an analogue-to-digital converter (ADC) to process the individual wavelength intensities in FCM scanners means that the output data are in reality discretised values. The range of discrete integer values observed in the resulting 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 (29). This underlying structure provides a lattice grid of unique pairs of discretised wavelength intensities in a 2D analysis, thus the 2D graphical representation can be expressed as an image, where each pixel has a Bernoulli outcome. This paper develops a methodology based on the Ising model, which has been used extensively in the field of image processing (30-34) for removal of noise/outliers and the identification of sub-images, that provides a solution for automatic gating by exploiting this unexplored structural layer embedded in FCM data.

The proposed methodology will follow closely the FCM data analysis framework proposed by Bashashati and Brinkman (9). It will be assumed that the quality assessment and normalisation components of the framework will be completed by the experimentalist, and hence the focus will be on the succeeding three components in the pipeline; outlier removal, automated gating and cluster labelling. This paper considers the three aforementioned components as a single overarching component rather than three individual elements of the analysis. As such the methodology outlined in this paper will address all three components by providing a single algorithmic solution which combines methodology from statistical literature and the field of image processing and segmentation.

**Materials and Methods**

***Flow Cytometry Data***

For expository purposes, the proposed automated gating solution will be exhibited by using two publicly available FCM datasets: the rituximab data from (35); and the Graft-versus-Host-Disease (GvHD) data from (36). The use of these two FCM experiments will demonstrate the utility of the proposed methods and additionally permit comparison to previous published automated gating solutions (14, 37).

The Rituximab data (35) were collected in an experiment to identify the enhancement of antilymphoma activity of rituximab by different agents. Sixteen hundred varying compounds were distributed to 96-well plates and underwent an initial incubation period. Following this, Rituximab was added to one of the duplicate plates, and both plates were incubated for a further period. The experiment thus contained cells treated with the compound alone, cells treated with the compound and Rituximab, and as controls, untreated cells and cells treated with Rituximab alone. Throughout the process cells were stained with BrdU which allowed new synthesised DNA to be labelled. In addition, following the completion of culture the samples were dyed with anti-BrdU and 7-AAD, a DNA binding dye. The cells were analysed using flow cytometric high-content screening recording forward-light scatter (FSC), sideward-light scatter (SSC) and fluorescent markers of anti-BrdU and 7-AAD, resulting in 1545 cells across the four variables.

The GvHD data (36) were collected from stem cell transplant recipients. GvHD is one of the most common complications observed in clinical transplantation of bone marrow, whereby the tissues of the recipient, often the liver and gut, are attacked by donor-immune cells in the graft. The GvHD experiment was initiated to identify the key biomarkers which lead to the development of GvHD in patients. Blood samples from 31 patients were taken and assigned to 96-well plates with 10,000-100,000 cells per well. The well plates were then dyed with 10 different four-colour antibody combinations. Samples from two patients both containing physical property variables FSC and SSC, and four additional fluorescent markers, anti-CD4, anti-CD8β, anti-CD3 and anti-CD8, are considered in this paper. One patient later developed acute GvHD while the second was taken from a control group. Both samples contain the six aforementioned variables measured for greater than 12,000 cells.

***Cytometry in Practice***

In practice the analysis of FCM data is rarely carried out on the entire cell population recorded. Instead an initial partition of the data is produced by gating on FSC and SSC. This initial partition provides an appropriate way of removing outliers and doublets from the analysis as FSC and SSC correspond to physical properties of the cells, namely shape and granularity. Similar to the work of Lo et al. (14) and Hahne et al. (15), we focus on this structured approach to gating throughout this paper. The initial step subsets the recorded observations into two groups, a group which will be used for further analysis and a group which will be considered to be cell debris and not relevant to the analysis. However, the proposed gating strategy can be applied to any pair of recorded FCM variables.

***Outlier Removal***

The removal of outliers is carried out by removing all cells which record FSC or SSC values of maximum intensity. These measurements are often doublets, exhibiting cell area and cell density much larger than evident throughout the sample of cells.

***Automated Gating***

Markov Random Fields (MRF). The Ising Model (38) is a special case of MRF where the pixels of the underlying latticetakes Bernoulli outcomes, The Markov property infers that for a given pixel it’s probability of taking either Bernoulli outcome is dependent only on its four nearest neighbours, as shown in Figure 1. Letbe the neighbours ofand define as the number of neighbours ofwith colour The Ising model is then defined mathematically via full probability conditionals as

whereis an interaction strength such thatencourages neighbours of the same pixel value.

Model Fitting. The Metropolis algorithm (39), introduced in statistical physics, is a Monte-Carlo simulation procedure with convergence quicker than the standard Gibbs sampling (40). The Metropolis criterion relies on the calculation of the change in energy from changing a single element to the current configuration. If the change in energy is negative the new configuration is accepted, however if the energy change is positive the new configuration can be accepted with a probability governed by the current temperature of the system (41).

Simulated Annealing (SA) allows for a computationally efficient approach to solving the problem of identifying a global minimum in the global energy function defined on the lattice (42, 43). The approach is to begin with a relatively high initial temperature whereby almost all proposed changes are accepted. Upon achieving an equilibrium state at the temperature is decreased and the model is iterated until a new equilibrium state is achieved at the new temperature level. Following the work of Geman and Geman (42), however this initialisation is subject to the underlying energy function (43-45). In addition, the decrease in temperature will be treated as proposed by Van Laarhoven and Aarts (45).

A hierarchical MRF approach (46-49) is used to extract features of the data at varying resolutions. In this methodology the hierarchical approach to MRF will be utilised to infer the initial configuration of when beginning the SA approach for a defined resolution. This overcomes the issue of sparsity in the lattice grid whereby the number of cells recorded is small.

***Cluster Labelling***

Connected component labelling (50) is an established pattern recognition tool used extensively in the identification of disjointed regions in binary, and with adaptations non-binary, images (51-53). Connected component labelling works by traversing an image, pixel-by-pixel, to identify regions of connected pixels based on their intensity values. In the binary setting, utilised in this methodology, the intensity values can be viewed as an active (intensity of 1) or inactive pixel (intensity of 0) in the image. The algorithm traverses the pixels row-by-row until an active pixel is found, the algorithm then proceeds as follows:

1. If all four neighbours of the pixel are inactive, assign a new label to the pixel;
2. If only one neighbour is active, assign the label of the neighbouring pixel;
3. If more than one neighbours are active, assign one of the labels and record label equivalences.

Following the complete traversal of the image, assign a single label to all members of equivalence classes. The resultant image will contain G disjoint regions within the image.

**Results**

The methodology introduced in the Material and Methods section was applied to the two publicly available FCM datasets discussed earlier. The gating solutions produced from the proposed methodology are compared to the model based clustering with t-distributions methodology of Lo et al. (14). The approach by Lo et al. (14) has previously been compared to manual gating by FCM experts, performing in line with manual gating solutions and has also been shown to be less restrictive than k-mean clustering solutions.

***Application to Rituximab data***

The Rituximab data has been analysed following the standard FCM data analysis practice described earlier. As mentioned earlier the data consisted of 1545 observations across four FCM variables, however 36 observations which recorded maximum intensities in either FSC or SSC were removed prior to analysis, similar to the approach by Lo et al. (14). Figure 1 (a) displays the initial clustering solution generated from a t mixture model with Box-Cox transformation restricted to selecting only one cluster. This initial gating was replicated using the MRF approach outlined in the Materials and Methods section of this paper, the resulting solution is shown in Figure 1 (b).

***Application to Graft-versus-host-disease data***

**Discussion**

**Literature Cited**

1. ­Perfetto SP, Chattopadhyay PK, Roederer M. Seventeen-colour flow cytometry: unravelling the immune system. Nature Reviews Immunology. 2004 Aug 1;4(8):648-55.
2. De Rosa SC, Brenchley JM, Roederer M. Beyond six colors: a new era in flow cytometry. Nature medicine. 2003 Jan 1;9(1):112-7.
3. Lizard G. Flow cytometry analyses and bioinformatics: interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research. Cytometry Part A. 2007 Sep 1;71(9):646-7.
4. Eudey TL. Statistical considerations in DNA flow cytometry. Statistical Science. 1996 Nov 1:320-34.
5. Braylan RC. Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias. Cytometry Part A. 2004 Mar 1;58(1):57-61.
6. Herzenberg LA, Parks D, Sahaf B, Perez O, Roederer M, Herzenberg LA. The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clinical chemistry. 2002 Oct 1;48(10):1819-27.
7. Bagwell CB. DNA histogram analysis for node‐negative breast cancer. Cytometry Part A. 2004 Mar 1;58(1):76-8.
8. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: the future just started. Cytometry Part A. 2010 Jul 1;77(7):705-13.
9. Bashashati A, Brinkman RR. A survey of flow cytometry data analysis methods. Advances in bioinformatics. 2009 Dec 6;2009.
10. Suni MA, Dunn HS, Orr PL, De Laat R, Sinclair E, Ghanekar SA, Bredt BM, Dunne JF, Maino VC, Maecker HT. Performance of plate-based cytokine flow cytometry with automated data analysis. BMC immunology. 2003 Sep 2;4(1):1.
11. Finak G, Langweiler M, Jaimes M, Malek M, Taghiyar J, Korin Y, Raddassi K, Devine L, Obermoser G, Pekalski ML, Pontikos N. Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. Scientific reports. 2016;6.
12. Parks DR. Data processing and analysis: Data management. Current Protocols in Cytometry. 1997;10:1-0.
13. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL. Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences. 2009 May 26;106(21):8519-24.
14. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model‐based clustering. Cytometry Part A. 2008 Apr 1;73(4):321-32.
15. Hahne F, Arlt D, Sauermann M, Majety M, Poustka A, Wiemann S, Huber W. Statistical methods and software for the analysis of highthroughput reverse genetic assays using flow cytometry readouts. Genome Biology. 2006 Aug 17;7(8):1.
16. Chester C, Maecker HT. Algorithmic tools for mining high-dimensional cytometry data. The Journal of Immunology. 2015 Aug 1;195(3):773-9.
17. Schut TC, De Grooth BG, Greve J. Cluster analysis of flow cytometric list mode data on a personal computer. Cytometry. 1993 Aug 1;14(6):649-59.
18. Wilkins MF, Hardy SA, Boddy L, Morris CW. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry. 2001 Jul 1;44(3):210-7.
19. Rousseeuw PJ, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices. Computational Statistics & Data Analysis. 1996 Nov 15;23(1):135-51.
20. Kothari R, Cualing H, Balachander T. Neural network analysis of flow cytometry immunophenotype data. IEEE Transactions on Biomedical Engineering. 1996 Aug;43(8):803-10.
21. Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Marine Ecology Progress Series. 2000 Mar 31;195:47-59.
22. Rajwa B, Venkatapathi M, Ragheb K, Banada PP, Hirleman ED, Lary T, Robinson JP. Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier. Cytometry Part A. 2008 Apr 1;73(4):369-79.
23. Morris CW, Autret A, Boddy L. Support vector machines for identifying organisms—a comparison with strongly partitioned radial basis function networks. Ecological Modelling. 2001 Dec 1;146(1):57-67.
24. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association. 2002 Jun 1;97(458):611-31.
25. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics. 1993 Sep 1:803-21.
26. McLachlan GJ, Basford KE. Mixture models. Inference and applications to clustering. Statistics: Textbooks and Monographs, New York: Dekker, 1988.
27. Titterington DM, Smith AF, Makov UE. Statistical analysis of finite mixture distributions. Wiley 1985.
28. Schwarz G. Estimating the dimension of a model. The annals of statistics. 1978;6(2):461-4.
29. Shapiro HM. The evolution of cytometers. Cytometry Part A. 2004 Mar 1;58(1):13-20.
30. Parisi G. Statistical Field Theory. Am J Phys. 1989;57(3):286.
31. Li SZ. Markov Random Field Modeling in Image Analysis - Third Edition. Advances in Pattern Recognition. 2009. 372 p.
32. Derin H, Elliott H. Modeling and segmentation of noisy and textured images using Gibbs random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1987 Jan(1):39-55.
33. Elliott H, Derin H, Cristi R, Geman D. Application of the Gibbs distribution to image segmentation. ICASSP ’84 IEEE Int Conf Acoust Speech, Signal Process. 1984;9:1–4.
34. Besag J. On the statistical analysis of dirty pictures. J R Stat Soc Ser B. 1986;48(3):259–302.
35. Gasparetto M, Gentry T, Sebti S, O'Bryan E, Nimmanapalli R, Blaskovich MA, Bhalla K, Rizzieri D, Haaland P, Dunne J, Smith C. Identification of compounds that enhance the anti-lymphoma activity of rituximab using flow cytometric high-content screening. Journal of immunological methods. 2004 Sep 30;292(1):59-71.
36. Brinkman RR, Gasparetto M, Lee SJ, Ribickas AJ, Perkins J, Janssen W, Smiley R, Smith C. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. Biology of Blood and Marrow Transplantation. 2007 Jun 30;13(6):691-700.
37. Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. Combining mixture components for clustering. Journal of Computational and Graphical Statistics. 2012.
38. Chandler D. Introduction to modern statistical mechanics. Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771. 1987 Sep.
39. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6):1087–92.
40. Liu JS. Peskun's theorem and a modified discrete-state Gibbs sampler. Biometrika. 1996 Sep 1;83(3):681-2.
41. Kato Z, Zerubia J. Markov random fields in image segmentation. Now Publishers Incorporated; 2012 Oct 5.
42. Černý V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. J Optim Theory Appl. 1985;45(1):41–51.
43. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Pattern Anal Mach Intell. 1984;PAMI-6(6):721–41.
44. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–80.
45. Van Laarhoven PJ, Aarts EH. Simulated annealing. In Simulated Annealing: Theory and Applications 1987 (pp. 7-15). Springer Netherlands.
46. Graffigne C, Heitz F, Perez P, Preteux FJ, Sigelle M, Zerubia JB. Hierarchical Markov random field models applied to image analysis: a review. InSPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation 1995 Aug 11 (pp. 2-17). International Society for Optics and Photonics.
47. Schneider MK, Fieguth PW, Karl WC, Willsky AS. Multiscale methods for the segmentation and reconstruction of signals and images. IEEE Trans Image Process. 2000;9(3):456–68.
48. Marques F, Cunillera J, Gasull A. Hierarchical segmentation using compound Gauss-Markov random fields. InAcoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on 1992 Mar 23 (Vol. 3, pp. 53-56). IEEE.
49. Bouman C, Liu B. Multiple Resolution Segmentation of Textured Images. IEEE Trans Pattern Anal Mach Intell. 1991;13(2):99–113.
50. Dillencourt MB, Samet H, Tamminen M. A general approach to connected-component labeling for arbitrary image representations. Journal of the ACM (JACM). 1992 Apr 1;39(2):253-80.
51. Harauz G, Fong-Lochovsky A. Automatic selection of macromolecules from electron micrographs by component labelling and symbolic processing. Ultramicroscopy. 1989 Dec 1;31(4):333-44.
52. Hossam MM, Hassanien AE, Shoman M. 3D brain tumor segmentation scheme using K-mean clustering and connected component labeling algorithms. In2010 10th International Conference on Intelligent Systems Design and Applications 2010 Nov 29 (pp. 320-324). IEEE.
53. Brummer ME, Mersereau RM, Eisner RL, Lewine RR. Automatic detection of brain contours in MRI data sets. IEEE transactions on Medical Imaging. 1993 Jun;12(2):153-66.

1. Kevin C. Brosnan was supported by an Irish Research Council Government of Ireland Scholarship (GOIPG/2014/19). Kevin Hayes was supported by the Science Foundation Ireland Research Investigator's Award (SFI-12/IA/1683). [↑](#footnote-ref-1)