 DOI: 10.3217/978-3-85125-663-5-39

GMM Interpolation for Blood Cell Cluster Alignment in Childhood Leukaemia

Roxane Licandro[1,2], Konstantin Miloserdov[2], Michael Reiter[2] and Martin Kampel[2]
licandro@cvl.tuwien.ac.at

***Abstract*— The accurate quantification of cancer (blast) and non-cancer cells in childhood leukaemia (blood cancer) is a key component in assessing the treatment response and to guide patient specific therapy. For this classification task, cell specific biomarker expression levels are estimated by using flowcytometry measurements of multiple features of single blood cells. For the automated distinction between blasts and non-blasts a main challenge are data shifts and variations in the high-dimensional dataspace caused by instrumental drifts, inter patient variability, treatment response and different machine characteristics. In this work we present a novel alignment scheme for stable (non-cancer) cell populations in flowcytometry using Gaussian Mixture Models (GMM) as data representation format for the cell clusters' probability density function and a Wasserstein interpolation scheme on the manifold of GMM. The evaluation is performed using a dataset of 116 patients with acute lymphoblastic leukaemia at treatment day 15. Classification results show an improved normalization performance using Wasserstein metric compared to two other metrics with a mean sensitivity of 0.97 and mean f-score of 0.95.**

I. INTRODUCTION

For the assessment of leukaemia (blood cancer) treatment response in the clinical routine, cell specific immunophenotypes of blood or bone marrow samples are measured using the FlowCytoMetry (FCM) technique [1], [4]. Therefor, in a staining step a combination of specific fluorescence-labeled antibodies is used to mark antigens on the surface of a cell. The expression of antigens on a surface varies among different blood cell types, which is used to identify blasts and non-blasts. In a subsequent step, every single cell is put in a fluid stream and FCM lasers of different wavelengths are used to measure physical properties (cell size and granularity) and the fluorescence pattern of the antibodies attached to the cell [12]. In Figure 1 a sample obtained by an FCM is illustrated. Every dot represents a cell (red cells refer to blast cells), where its position is determined by the features measured (in this case) for the expression of the antibodies CD 10 and CD 45. Up to 30 features can be measured and observed at once. Blasts are manually identified using a gating hierarchy, which consists of a defined procedure observing specific combinations of features in a two-dimensional representation and by drawing polygons (gates) to extract and mark cell clusters of interest in every step. The identified events of interest of a gating step serve as input of the subsequent gate in the hierarchy. This manual procedure strongly depends on the experience and skills of the operator, is highly time-consuming and subjective. [7]

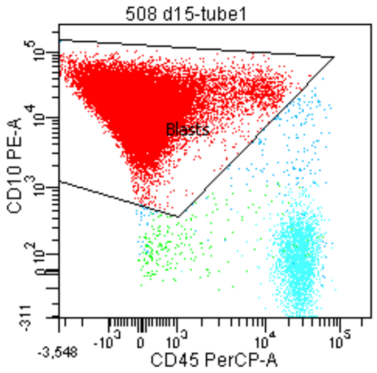

Fig. 1. Illustration of a flowcytometer sample. The labeling is obtained by manual gating (drawing polygons (gates) around populations of interest) in the two dimensional feature space (CD10, CD45). [Best viewed in color]

A. Challenges

The main challenge in gating FCM data is the detection of small cell populations, e.g. cancer cells in late therapy stages composing around 0.1% of all cells observed ($3*10^5$ - 10^6). This is further aggravated by the influence of treatment or age on the regeneration status of bone marrow precursors [5]. A further challenge forms shifts and deformations of the clusters observed in the high-dimensional feature space caused by different machine types, corresponding instrumental drifts and lab varying standard operating procedures [10]. Thus, it is particularly important to obtain a normalized representation of samples to be able to compare and identify clusters of interest and to perform patient or population based longitudinal analysis automatically over therapy time and multiple centers.

B. Related Work

A key issue for multi-center studies as well as in studies where data is acquired with different flowcytometers, is the normalization and alignment of cell population clusters to improve the automated blast identification process. In recent approaches ([8], [11], [7]) Gaussian Mixture Models are

[1] Medical University of Vienna, Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research (CIR).
[2] TU Wien, Institute of Visual Computing and Human-Centered Technology, Computer Vision Lab (CVL).
*This work has been supported by the European Commission Project AutoFLOW (http://www.autoflow-project.eu/) and the Computer Vision Club. We thank Markus Diem and Florian Kleber at the CVL for their valuable input. We thank Michael Dworzak at St. Anna Childrens' Hospital Vienna and Angela Schumich and Susanne Suhendra at LabDIA GmbH, for their kind support and providing the flowcytometry data and annotations.

used to model the probability density function of blood cell clusters to assign a biologically meaningful label to every observed cell automatically. Various approaches for the registration and interpolation of GMM distributions have been proposed in the computer vision and pattern recognition community: Jian et al. [2] present a robust registration scheme of GMM for point sets using a L2 distance metric, since it provides a closed-form expression for GMM and needs no specification of tuning parameters. However, no intermediate evolution step guarantees that the interpolation between two GMMs lies in the manifold of K-component GMM as well as minimizing the L2 distance for registration is prone to instability due to many local optima. Kim et al. [6] address this problem by introducing cross-entropy instead of the L2 distance and provide a closed form solution with no numerical difficulties. In [9] Orlova et al. propose an alignment approach, based on quadratic formcluster matching *QF Match* for flow and mass cytometry data. They provide a multivariate extension of the quadratic form distance for cell cluster alignment. In [10] Orlova et al. used the Earth Mover's Distance (also known as Wasserstein Distance (WD)) as a metric for comparing the expression levels of biomarker in cell populations measured by FCM. Using WD as a metric allows to distinguish small shifts caused by e.g. instrument drifts from biologically significant differences. A key benefit of WD over other metrics (quadratic form, bipartite matching) is that the WD score between two samples involves the magnitude of change and the proportion of cells whose antigen expression has changed.

C. Contribution

The contribution of this work is three fold: Firstly, we provide an optimal transport technique for FCM data on the manifold of Gaussian Mixture Models using a Wasserstein metric. Kim et al. [6] propose an interpolation scheme on the manifold on k-component GMM using the Kullback Leibler Divergence (KLD) as a metric. We extended the GMM interpolation approach by introducing the Wasserstein metric as a distance measure between two GMMs. We were inspired in doing this by the approach of Chen et al. [3], who provided an interpolation scheme based on the Wasserstein metric for GMM on the space of Gaussian distributions, so it can be seen as a discrete measure and retains the Gaussian mixture structure. This formulation showed efficiency for high dimensional GMMs with a small number of components, but (to our knowledge) has not been used for the alignment of cell distributions acquired by FCM. Secondly, we provide an extensive evaluation scheme against two different metrics. Thirdly, we used the proposed alignment strategy on FCM data to improve the non-blast alignment in childhood Acute Lymphoblastic Leukaemia (ALL).

This paper gives an overview of the methodology proposed in Section II. The dataset used and evaluation results are presented in Section III and the conclusion of this work and possibilities for future work are summarized in Section IV.

II. METHODOLOGY

In this section the proposed normalization strategy for FCM data is presented. For the representation of the probability density function $p(x|.)$ of non blast cells x a N component Gaussian Mixture Model (GMM) $S(x)$ is used as a weighted sum of N Gaussian distributions $S_i(x) = \mathcal{N}(x|\mu_i, \sigma_i)$ as expressed in Equation 1. θ is the set of parameters required to parameterize a Gaussian component i in a GMM (cf. Equation 2), where w_i is the Gaussian weighting parameter, μ_i the mean (cf. Equation 3) and Σ_i the covariance (cf. Equation 4).

$$S(x) = p(x|\theta) = \sum_{i=1}^{N} w_i S_i(x) \tag{1}$$

$$\theta = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{N} \tag{2}$$

$$\mu_i = \mathbb{E}_{S_i(x)}[x] \tag{3}$$

$$\Sigma_i = \mathbb{E}_{S_{(x)}}[(x-\mu_i)-(x-\mu_i)^T] \tag{4}$$

A. GMM based Optimal Transport

The optimal transport problem is the transport of a source distribution S(x) of a mass x on a manifold $\mathcal{M}_{GMM}$ of N component GMMs, in a way that it is transformed into the target distribution $T(x) \in \mathcal{M}_{GMM}$ without loss of mass [3]. Therefore we can define an optimization function and transport plan γ. Optimizing γ refers to the finding of an optimal interpolation path on $\mathcal{M}_{GMM}$ by minimizing distances D between the Gaussian components $S_i(x), T_j(x)$ in $S(x)$ and $T(x)$ (cf. Equation 5). In Figure 2 a schematic illustration of the GMM based optimal transport problem for FCM data is visualised. Two physical features (SS INT and FS INT) are used as dimensions for visualisation. Cyan ellipses correspond to Gaussian components of the source GMM and black ellipses to the target GMM. In the optimal case, components of the source GMM overlay with the target GMM components after alignment (left image). Only non-cancer cells are visualised.

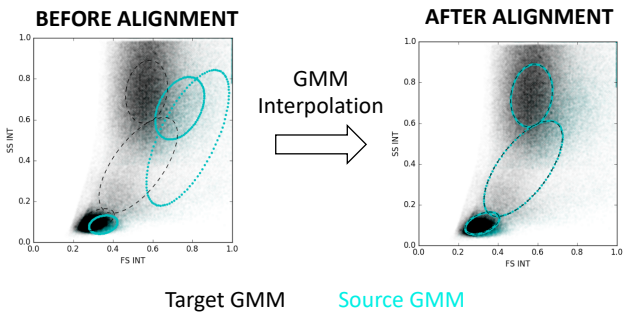

Fig. 2. Schematic illustration of the GMM based Optimal Transport problem with flowcytometry data. Only 2 measured features (SS INT and FS INT) are visualised. Black ellipses correspond to Gaussian components of the target GMM and cyan to the source GMM. Only non-cancer cells are visualised. [Best viewed in color]

$$S^* = argmin_{S(x)\in\mathscr{M}_{GMM}} \sum_{j=1}^{M} D(S(x)||T_j(x)) \tag{5}$$

In this work the Wasserstein metric (cf. Equation 6 [3]) is minimized between the Gaussian components of the source and target GMMs. In the following, i = 1,...N and j = 1,...,M denote the numbers of Gaussian components in the source and in the target GMM respectively. We evaluate the performance of the WD against the Kullback Leibler Divergence (KLD) (cf. Equation 7) and Bhattacharyya Bound (BB) (cf. Equation 8).

$$W^2 = \|\mu_i - \mu_j\|^2 + tr[\Sigma_i + \Sigma_j - 2\sqrt{(\sqrt{\Sigma_i})\Sigma_j(\sqrt{\Sigma_i})}] \tag{6}$$

$$KL = \frac{1}{2}k\log 2\pi + \log|\Sigma_j| + tr[\Sigma_j^{-1}\Sigma_i] + (\mu_i - \mu_j)^T\Sigma_j^{-1}(\mu_i - \mu_j) \tag{7}$$

$$BB = \frac{1}{8}(\mu_j - \mu_i)^T(\frac{\Sigma_i + \Sigma_j}{2})^{-1}(\mu_j - \mu_i) + \frac{1}{2}\ln\frac{|\frac{\Sigma_i+\Sigma_j}{2}|}{|\sqrt{|\Sigma_i||\Sigma_j|}} \tag{8}$$

Kim et al. [6] propose Expectation Maximization (EM) to estimate the parametrization θ of the transported GMM S^* by minimizing the KL-divergence. We adapted this approach for FCM data by estimating the likelihood of the WD (*responsibility*) between every Gaussian component $S_i(x)$ and $T_j(x)$ (cf. Equation 9) in the E-step to obtain the transportation matrix $\gamma_{ij} \in \mathbb{R}^{N\times M\times d}$, where d refers to the number of features observed per cell.

$$\gamma_{ij} = \frac{\delta_{i,j}}{\tau_i} \text{ where } \delta_{ij} = w_j \exp -W^2\left(S_i(x), T_j(x)\right) \text{ and } \tau_i = \sum_{j=1}^{M}\delta_{ij} \tag{9}$$

In the M-step the parameter set θ is updated using Equations 10, 11 and 12.

$$w_j = \frac{w_j}{\sum_{j=1}^{M}\sum_{i=1}^{N} w_j} \text{where } w_j = \sum_{i=1}^{N} w_i\gamma_{ij} \tag{10}$$

$$\mu_j = \sum_{i=1}^{N}\pi_i\mu_i \text{ where } \pi_i = \frac{w_i\gamma_{ij}}{\sum_i w_i\gamma_{ij}} \text{for fixed j} \tag{11}$$

$$\Sigma_j = \sum_{i=1}^{N}\pi_i\Sigma_i + \sum_{i=1}^{N}\pi_i(\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{12}$$

III. EVALUATION RESULTS

In this section first the dataset used is introduced, as well as the parametrisation of the algorithm and non-cancer cell classification procedure is described.

A. Dataset

In this work we use a FCM dataset of 116 patients with B-ALL acquired at treatment day 15. The patient's therapy was guided by the AIEOP-BFM 2009 protocol[1]. At the national diagnostic reference center all samples have been prepared and annotated according to the international standard operating procedure for 6 color FCM. For every cell ten FCM based features are measured (3 optical features (FSC-A, FSC-W, SSC-A), 7 fluorescence based features (CD20, CD10, CD45, CD34, SYTO41, CD19, CD38). For every patient in average 300 000 cells are measured. In FCM partial overlapping of fluorescence spectra of different fluorochromes is removed by spillover compensation. The preprocessing concludes with a normalization of the measured parameter values between 0 and 1.

B. Experimental Setup

The methodology proposed is evaluated in the following way: As comparable performance measures for non-cancer cell identification after data normalization, the mean sensitivity, accuracy, precision, accuracy and f-score over all patients in the test fold are computed. We used a 4 fold cross validation setup, where the test fold consisted of 29 patients and the training fold of 87. The measured cells in the training data were subsampled in a random way, by extracting 100 000 non blast cells per patient, resulting in a total of $8.7*10^6$ non blast cells. Subsequently we learned a GMM representation of the non-blast cells of the training set (Source) and test set (Target) separately, using the toolbox *mixture* integrated in the python framework *sklearn*[2]. The parameter set of the transformed Source GMM is estimated by using 1000 iterations of Expectation Maximization (EM) and the methodology introduced in Section II. Subsequently, the transformed model is used to predict a score for every cell of a test patient's sample to identify non-cancer cells.

C. Discussion

In Table I the non-blast classification performance is summarized using three different alignment metrics and 3 different setups of component numbers to represent the source GMM and target GMM. Non-blasts are classified based on WD best with a precision of 0.9534, accuracy of 0.9402 and f-score of 0.9585 in the experiment with a number of components (3,4) and (4,6) compared to the KLD and BB metric. This confirms the observations of Chen et al. [3], that an GMM interpolation scheme based on WD is efficient for high-dimensional GMM with a small number of components. For (4,5) number of components the KL divergence performs better compared to WD and BB with an accuracy of 0.9366 and f-score of 0.9582. The worst results are achieved by the BB metric with an average difference of -0.2 of the f-score and -0.3 difference in the sensitivity for N_{NB} = 3 and M = 4 Gaussian components.

IV. CONCLUSION

We propose an optimal transport scheme for GMM representation of FCM data by minimizing the Wasserstein

[1]AIEOP-BFM 2009 is a conducted randomized clinical trial for ALL between age 1-18 years in 10 countries in- and outside Europe, with approximately 1000 patients observed per year (Dworzak, 2013)) https://bfminternational.wordpress.com/ [accessed 2019-03-15]

[2]PythonToolbox:https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture [accessed 2019-03-15]

TABLE I

ILLUSTRATION OF THE NON-BLAST CLASSIFICATION PERFORMANCE USING GMM INTERPOLATION WITH WASSERSTEIN DISTANCE (WD), KULLBACK LEIBLER DIVERGENCE (KLD) AND BHATTACHARYYA BOUND (BB) FOR DIFFERENT THERAPY TIME POINTS, AND WITH DIFFERENT NUMBER OF COMPONENTS N_{NB} FOR THE SOURCE NON BLAST GMM, AND M FOR THE TARGET SAMPLE.

Experiment	N_{NB}	M	Sensitivity	Precision	Accuracy	f-score
KLD	3	4	**0.9780**	0.9488	0.9339	**0.9564**
WD	3	4	0.9732	**0.9505**	**0.9347**	0.9558
BB	3	4	0.6904	0.9441	0.6798	0.7188
KLD	4	5	**0.9821**	0.9489	**0.9366**	**0.9582**
WD	4	5	0.9702	**0.9503**	0.9316	0.9540
BB	4	5	0.8587	**0.9506**	0.8274	0.8703
KLD	4	6	**0.9814**	0.9495	0.9376	**0.9585**
WD	4	6	0.9755	**0.9534**	**0.9402**	**0.9585**
BB	4	6	0.8711	0.9499	0.8419	0.8835

distance between GMMs using Expectation Maximization. Results suggest that the WD performs best for non-blast identification and normalization compared to KLD and BB, for a low number of Gaussian mixture components. The focus of this work is to align stable populations (non cancer cells), which is a limit of the approach. For future work we plan to extend this approach for normalized non-cancer cell populations to perform anomaly detection for blast identification of multi-center data and of data from different treatment time points.

REFERENCES

[1] G. Basso, M. Veltroni, M. Valsecchi, M. Dworzak, R. Ratei, D. Silvestri, A. Benetello, B. Buldini, O. Maglia, G. Masera, *et al.*, "Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow," *Journal of Clinical Oncology*, vol. 27, no. 31, pp. 5168–5174, 2009.

[2] Bing J. and B. Vemuri, "Robust Point Set Registration Using Gaussian Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1633–1645, aug 2011.

[3] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Optimal transport for Gaussian mixture models," *arXiv preprint arXiv:1710.07876*, oct 2017.

[4] M. Dworzak, G. Fröschl, D. Printz, G. Mann, U. Pötschger, N. Mühlegger, G. Fritsch, and H. Gadner, "Prognostic significance and modalities of flow cytometric minimal residual disease detection in childhood acute lymphoblastic leukemia," *Blood*, vol. 99, no. 6, pp. 1952–1958, 2002.

[5] G. Gaipa, G. Cazzaniga, M. Valsecchi, R. Panzer-Grümayer, B. Buldini, D. Silvestri, L. Karawajew, O. Maglia, R. Ratei, A. Benetello, S. Sala, A. Schumich, A. Schrauder, T. Villa, M. Veltroni, W.-D. Ludwig, V. Conter, M. Schrappe, A. Biondi, M. Dworzak, and G. Basso, "Time point-dependent concordance of flow cytometry and real-time quantitative polymerase chain reaction for minimal residual disease detection in childhood acute lymphoblastic leukemia." *Haematologica*, vol. 97, no. 10, pp. 1582–93, October 2012.

[6] H. J. Kim, N. Adluru, M. Banerjee, B. C. Vemuri, and V. Singh, "Interpolation on the Manifold of K Component GMMs," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015, pp. 2884–2892.

[7] R. Licandro, M. Reiter, M. Diem, M. Dworzak, A. Schumich, and M. Kampel, "Application of machine learning for automatic mrd assessment in paediatric acute myeloid leukaemia," in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, INSTICC. SciTePress, 2018, pp. 401–408.

[8] I. Naim, S. Datta, J. Rebhahn, J. Cavenaugh, T. Mosmann, and G. Sharma, "Swift - scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design," *Cytometry Part A*, vol. 85, no. 5, pp. 408–421, 2014.

[9] D. Y. Orlova, S. Meehan, D. Parks, W. A. Moore, C. Meehan, Q. Zhao, E. E. B. Ghosn, L. A. Herzenberg, and G. Walther, "QFMatch: multidimensional flow and mass cytometry samples alignment," *Scientific Reports*, vol. 8, no. 1, p. 3291, dec 2018.

[10] D. Y. Orlova, N. Zimmerman, S. Meehan, C. Meehan, J. Waters, E. E. B. Ghosn, A. Filatenkov, G. A. Kolyagin, Y. Gernez, S. Tsuda, W. Moore, R. B. Moss, L. A. Herzenberg, and G. Walther, "Earth movers distance (emd): A true metric for comparing biomarker expression levels in cell populations," *PLOS ONE*, vol. 11, no. 3, pp. 1–14, 03 2016.

[11] M. Reiter, P. Rota, F. Kleber, M. Diem, S. Groeneveld-Krentz, and M. Dworzak, "Clustering of cell populations in flow cytometry data using a combination of gaussian mixtures," *Pattern Recogn.*, vol. 60, no. C, pp. 1029–1040, Dec. 2016.

[12] P. Rota, M. Reiter, S. Groeneveld-Krentz, and M. Kampel, "The role of machine learning in medical data analysis. a case study: Flow cytometry," in *Proceedings of the Internaction Conference on Computer Vision Theory and Applications*, 2016.