
CYTOPT: OPTIMAL TRANSPORT WITH DOMAIN ADAPTATION FOR INTERPRETING FLOW CYTOMETRY DATA

Paul Freulon^{*,1,2}, Jérémie Bigot^{1,2}, and Boris P. Hejblum^{1,3,4}

¹*Université de Bordeaux, Bordeaux, 33000, France.*

²*Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*

³*Bordeaux Population Health Research Center Inserm U1219, Inria SISTM, 33000 Bordeaux, France.*

⁴*Vaccine Research Institute (VRI), 94010 Créteil, France.*

**corresponding author: paul.freulon@math.u-bordeaux.fr*

ABSTRACT

The automated analysis of flow cytometry measurements is an active research field. We introduce a new algorithm, referred to as CytOpT, using regularized optimal transport to directly estimate the different cell population proportions from a biological sample characterized with flow cytometry measurements. We rely on the regularized Wasserstein metric to compare cytometry measurements from different samples, thus accounting for possible mis-alignment of a given cell population across samples (due to technical variability from the technology of measurements). In this work, we rely on a supervised learning technique based on the Wasserstein metric that is used to estimate an optimal re-weighting of class proportions in a mixture model from a source distribution (with known segmentation into cell sub-populations) to fit a target distribution with unknown segmentation. Due to the high-dimensionality of flow cytometry data, we use stochastic algorithms to approximate the regularized Wasserstein metric to solve the optimization problem involved in the estimation of optimal weights representing the cell population proportions in the target distribution. Several flow cytometry data sets are used to illustrate the performances of CytOpT that are also compared to those of existing algorithms for automatic gating based on supervised learning.

Keywords Automatic gating · flow cytometry · Optimal Transport · Stochastic Optimization

1 Introduction

1.1 Flow Cytometry Data Analysis

Flow cytometry is a high-throughput biotechnology used to characterize a large amount of cells from a biological sample. Flow cytometry is paramount to many biological and immunological research with applications, for instance, in the monitoring of the immune system of HIV patients by counting the number of CD4 cells.

The first step to characterize cells from a biological sample with flow cytometry is to stain those cells. Specifically, cells are stained with multiple fluorescently-conjugated monoclonal antibodies directed to the cellular markers of interest. Then, the cells flow one by one through the cytometer laser beam. The scattered light is characteristic to the biological markers of the cells [N.Aghaeepour et al., 2013]. Thus, from a biological sample analyzed by a flow cytometer, we get a data set X_1, \dots, X_I where each observation X_i corresponds to a single cell crossing the laser beam. For an observation $X_i \in \mathbb{R}^d$, the coordinate $X_i^{(m)}$ corresponds to the light intensity emitted by the fluorescent antibody attached to the biological marker m . Interestingly, such a data set may also be considered as a discrete probability distribution $\frac{1}{I} \sum_{i=1}^I \delta_{X_i}$ with support in \mathbb{R}^d , this is the point of view taken in this paper.

With flow cytometry one can assess individual characteristics at a cellular level for samples composed of hundreds of thousands of cells. The development of this technology now leads to state-of-the-art cytometers that can measure up to 50 biological markers at once on a single cell [Y.Saeyns et al., 2016]. The constant increase of the number of measured markers allows to identify more precisely the different cell populations and subtypes present in the biological sample.

The analysis of cytometry data is generally done manually, by drawing geometric shapes (referred to as “gates”) around populations of interest in a sequence of two-dimensional data projections. Such manual analysis features several drawbacks: i) it is extremely time-consuming ; ii) manual gating lacks reproducibility across different operators [N.Aghaeepour et al., 2013]. To overcome these shortcomings, several automated methods have been proposed [N.Aghaeepour et al., 2013]. Those automated approaches aim at a clustering of the flow cytometry data to derive the proportions of the cell populations that are in the biological sample. Some methods follow an unsupervised approach. For instance, FLOWMEANS [Aghaeepour et al., 2011] is an automated method based on the K-means designed for flow cytometry data. We also mention Cytometree [D.Commenges et al., 2018], an algorithm based on the construction of a binary tree. Methods which perform model based clustering have also been proposed, see e.g. Y.Ge and S.C.Sealfon [2012], B.P.Hejblum et al. [2019]. To improve the accuracy of the classification, supervised machine learning techniques have been applied to flow cytometry data analysis. Among those techniques, one may cite DeepCyTOF introduced in H.Li et al. [2017] which is based on deep-learning algorithms to gate cytometry data. In M.Lux et al. [2018], the authors introduced flowlearn, a method that uses manually gated samples to predict gates on other samples. We also mention a new supervised approach named OptimalFlow developed by Barrio et al. [2019] that relies on the Wasserstein distance to quantify discrepancies between cytometry data sets.

In spite of numerous efforts to automate cytometry data analysis, manual gating remains the gold-standard for benchmarking. The fact that automated methods have not outdone manual gating can be explained, at least in part, by the significant variability of flow cytometry data. This variability is two fold; first, it is induced by biological heterogeneity across the samples analyzed [Hahne et al., 2010]. For instance, it is likely that subtypes proportions within the biological sample of a healthy patient will differ from the subtypes proportions within the biological sample of a sick patient. In addition to this variability of the biological phenomena of interest, technical variability appears during the process of flow cytometry analysis. For instance, differences in the staining procedure, in the data acquisition settings or cytometers performances are very likely to happen and to lead to undesirable variability between flow cytometry data [Maecker and McCoy, 2010].

1.2 Optimal transport in Statistics

To tackle the matter of Flow Cytometry data analysis, this work proposes to make use of tools that come from the theory of optimal transport. Optimal transport has recently gained interest in machine learning and statistics. Indeed, the introduction of approximate solvers that for large dimension problems allowed to move beyond the high computational cost of optimal transport. Thus, optimal transport has found various applications in machine learning for regression [H.Janati et al., 2018], classification [R.Flamaray et al., 2018] and generative modeling [M.Arjovsky et al., 2017]. Those computational progresses in optimal transport have also allowed its use in imaging sciences [J.Solomon et al., 2015]. Recent progresses in applied optimal transport has been fueled by the development of efficient, large-scale optimization algorithms for problems in this domain. In particular, our method rely on stochastic algorithms [A.Genevay et al., 2016, B.Bercu and J.Bigot, 2020] to alleviate the computational cost of optimal transport. In relationship with our field of research, and to make the point that optimal transport tools are now applied on real-life data, we mention two single cell algorithms based on optimal transport lately published. First Schiebinger et al. [2019] model the development of cells in order to understand their differentiation process. And in Li et al. [2019], the authors propose a method that deals with the issue of cluster alignment with the purpose to discriminate between intrinsic clusters and spurious clusters.

1.3 An illustrative data set

As an illustrative example, we shall analyze in this paper the flow cytometry data from the T-cell panel of the Human Immunology Project Consortium (HIPC) – publicly available on ImmuneSpace [R.Gottardo et al., 2014]. Seven laboratories stained three replicates (denoted A, B, and C) of three cryo-preserved biological samples denoted patient 1, 2, and 3 (e.g. cytometry measurements from the Stanford laboratory for replicate C from patient 1 will be denoted as the “Stanford1C” data set). After performing cytometry measurements in each center, the resulting FCS files were manually gated centrally for quantifying 10 cell populations: CD4 Effector (CD4 E), CD4 Naive (CD4 N), CD4 Central memory (CD4 CM), CD4 Effector memory (CD4 EM), CD4 Activated (CD4 A), CD8 Effector (CD8 E), CD8 Naive (CD8 N), CD8 Central memory (CD8 CM), CD8 Effector memory (CD8 EM) and CD8 Activated (CD8 A). Hence, for these data sets, a manual clustering is at our disposal to evaluate the performances of automatic gating methods. The flow cytometry data sets built from these 10 sub populations have a size that range from 15 554 observations for the smallest data set to 112 318 observations for the largest data set. For each cell, seven biological markers have been

measured; it leads to cytometry observations X_i that belong to \mathbb{R}^d with $d = 7$. A 3D projection using three markers is displayed in Figure 1 for the “Stanford1A” data set with the corresponding manual gating into 10 clusters. Additionally, we also benchmarked our method on the flow cytometry data sets used in Barrio et al. [2019].

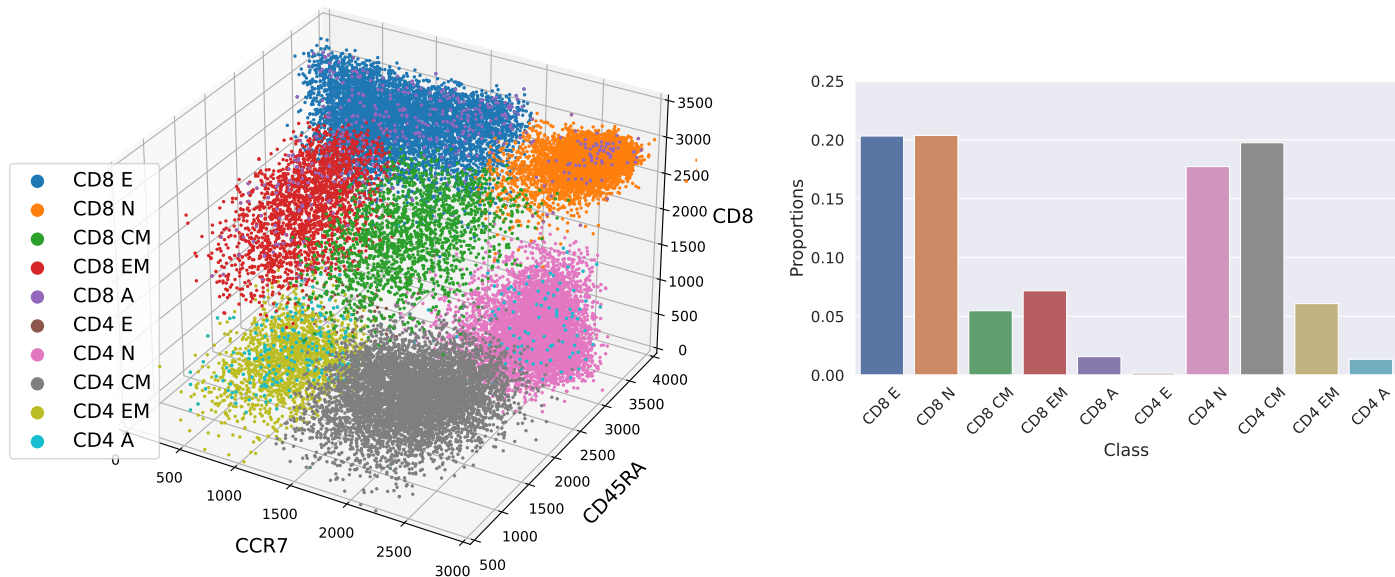


Figure 1: **Example flow cytometry data set of a biological sample from Stanford patient 1 replicate A.** Left: Manual clustering of the cytometry data. Existing automated methods target a clustering in order to derive the class proportions. Right: Class proportion derived from the manual gating. Our method CytOpT aims at going straight to the estimation of class proportions without clustering the cytometry data.

1.4 Main contributions

We propose a new supervised method to estimate the relative proportions of the different cell sub-types in a biological sample analyzed by a cytometer. Our approach aims at finding an optimal re-weighting of class proportions in a mixture model between a source data set (with known segmentation into cell sub-populations) to fit a target data set with unknown segmentation. This estimation of the class proportions is done without any preliminary clustering of the target cytometry data. To the best of our knowledge, all the previous automated methods are based on the classification of the cells of the biological sample to deduce a class proportions estimation. However, from a clinical perspective, the relevant information is the class proportions [Maecker and McCoy, 2010], while the clustering of individual cells is simply a means to an end to get there. Our method can nonetheless be extended in order to obtain a clustering of cytometry data.

We believe that going straight to the estimation of class proportions is an original approach to tackle the issue of flow cytometry data analysis. Hence, the result of our algorithm is not a clustering of the cells analyzed by the cytometer but a vector $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_K)$, where each coefficient $\hat{\pi}_k$ amounts to estimation of the percentage of the k^{th} population cell among all the cells analyzed in the target sample. To do so, we rely on tools derived from optimal transport and regularized Wasserstein distance between probability measures.

In this work, we propose to estimate the proportions $\pi = (\pi_1, \dots, \pi_K)$ of sub-populations cell in a target data set (whose gating is unknown) by minimizing the regularized Wasserstein distance between the target distribution, and a re-weighted source distribution (from a source data set whose gating into K sub-populations cell of interest is known). Since the Wasserstein distance is able to capture the underlying geometry of the measurements, it has the ability to make meaningful comparisons between distributions whose supports do not overlap and to quantify potential spatial shifts. Our work demonstrates the benefits of using Wasserstein distance for the analysis of the cytometry data as it allows to handle the technical variability induced by different settings of measurements.

1.5 Organization of the paper

In Section 2 we present the regularized Wasserstein distance and the stochastic procedure that we leverage to compute it. Section 3 details our estimation method named CytOpT that yields an estimate of the class proportions in an unsegmented data set. Section 4 is devoted to a thorough simulation study of CytOpT where it is compared to classic classification methods and its robustness is assessed. Finally, in Section 5 we demonstrate the performance of CytOpT on the T-cell panel proposed by the Human Immunology Project Consortium (HIPC) described previously, and we compare the performance of CytOpT, with existing methods.

2 Mathematical framework

Optimal transport allows the definition of a metric between two probability distributions α and β supported on \mathbb{R}^d . This metric is informally defined as the lowest cost to move the mass from one probability measure, the source measure α , onto the other, the target measure β . As optimal transport handles probability distributions, we must describe a cytometry data set X_1, \dots, X_I where each X_i belongs to \mathbb{R}^d as a probability distribution in \mathbb{R}^d . A natural choice to move from observations to probability measure is to use the empirical measure. Therefore, for a data set X_1, \dots, X_I , we consider its empirical measure defined as $\frac{1}{I} \sum_{i=1}^I \delta_{X_i}$.

2.1 Wasserstein distance and its regularized counterpart

In this work, we shall consider optimal transport between discrete measures on \mathbb{R}^d , and we denote by $\alpha = \sum_{i=1}^I a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^J b_j \delta_{y_j}$ two such measures. Note that a_i , the probability associated to the point x_i , is not necessarily equal to $1/I$ as a re-weighting of the source measure will be considered. Optimal transport then seeks a plan that moves the probability measure α to the probability measure β with a minimum cost. The transportation cost is encoded by a function $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ where $c(x, y)$ represents the cost to move one unit of mass from x to y . In the discrete setting the transportation cost is encoded by a matrix $C \in \mathbb{R}^{I \times J}$ where $C_{i,j} = c(x_i, y_j)$. Once set, the cost matrix C ensures that the transportation is carried out between a source distribution with support x_1, \dots, x_I and a target distribution with support y_1, \dots, y_J . Then, to ensure the conservation of mass, i.e. the transport of a distribution with weights a_1, \dots, a_I toward a distribution with weights b_1, \dots, b_J , we introduce the set $\Pi(a, b) = \{P \in \mathbb{R}^{I \times J} : P1_J = a \text{ and } P^T 1_I = b\}$ of all the coupling matrices between a and b . Thus, minimizing the transportation cost from α to β boils down to the following optimization problem:

$$W_c(\alpha, \beta) = \inf_{\pi \in \Pi(a,b)} \sum_{i=1}^I \sum_{j=1}^J C_{i,j} P_{i,j} = \inf_{\pi \in \Pi(a,b)} \langle C, P \rangle \quad (1)$$

All along this work the coefficient $C_{i,j}$ of the cost matrix will be defined as the squared Euclidean distance between x_i and y_j , i.e. $C_{i,j} = \|x_i - y_j\|_2^2$, and we use the notation $W(\alpha, \beta) = W_2^2(\alpha, \beta)$. In spite of appealing theoretical properties (see e.g. Santambrogio [2015]) Wasserstein metrics drag a computational burden. Indeed, the cost to evaluate the Wasserstein distance between two discrete probability distributions with support of equal size N is generally of order $O(N^3 \log(N))$ [M.Cuturi, 2013]. To allow the evaluation of the Wasserstein distance at a lower cost, M.Cuturi [2013] has proposed to add an entropic regularization term to the linear optimization problem (1) in order to reach an approximate solution with faster computations. This regularized optimal transport problem reads:

$$W^\varepsilon(\alpha, \beta) = \inf_{\pi \in \Pi(a,b)} \langle C, P \rangle + \varepsilon H(P), \quad (2)$$

where the entropy $H : \mathbb{R}_+^{I \times J} \rightarrow \mathbb{R}$ is defined for $P \in \mathbb{R}_+^{I \times J}$ by $H(P) = \sum_{i,j} (\log(P_{i,j}) - 1) P_{i,j}$, and $\varepsilon > 0$ is the regularization parameter. The entropic regularization of the Kantorovich problem (2) leads to an approximation of the Wasserstein distance which can be calculated in $O(N^2 \log(N))$ operations if the two distributions have a support of size N [G.Peyré et al., 2019]. In this work, the regularized Wasserstein distance is calculated with a statistical procedure based on the Robbins-Monro algorithm for stochastic optimization. This way to calculate the Wasserstein distance is investigated in A.Genevay et al. [2016] and B.Bercu and J.Bigot [2020]. The keystone of this approach is that the regularized Wasserstein problem can be written as the following stochastic optimization problem:

$$W^\varepsilon(\alpha, \beta) = \max_{u \in \mathbb{R}^I} \mathbb{E}_{Y \sim \beta} [g_{\varepsilon, \alpha}(Y, u)], \quad (3)$$

where Y is a random variable with distribution β , and $g_{\varepsilon, \alpha}(y, u)$ is easy to compute for any y in β 's support and any $u \in \mathbb{R}^I$. The expression of $g_{\varepsilon, \alpha}$ can be found in equation (B.10) of the Appendix. This formulation of the regularized Wasserstein distance as the maximum of an expectation allows the application of stochastic optimization methods.

We stress that the large number of observations in cytometry data sets makes those stochastic approaches particularly relevant. For a more detailed presentation of stochastic algorithms for optimal transport, we refer the reader to Section A of the Appendix.

2.2 Statistical model

Let us consider X_1^s, \dots, X_J^s the cytometry measurements from a first biological sample that is referred to as the source sample or source observations. The distribution of the source sample is modeled by a mixture of K distributions: $\alpha = \sum_{k=1}^K \rho_k \alpha_k$, where each term of the mixture corresponds to the cytometry measurements of one type of cell. For instance, α_k represents the underlying distribution behind the cytometry measurements of the CD4 effector T cells that are in the biological sample. The weights $\rho \in \Sigma_K$ are coefficients lying in the probability simplex $\Sigma_K = \{h \in \mathbb{R}_+^K : \sum_{k=1}^K h_k = 1\}$, and ρ_k represents the proportion of one cell sub-type among all the cells found in the sample.

We also consider a second set of cytometry measurements X_1^t, \dots, X_J^t that is referred to as the target sample or target observations, and that corresponds to a second biological sample that may come from another patient or a cytometry analysis performed in a different laboratory. We assume that the underlying distribution of the target observations is another mixture of K distributions: $\beta = \sum_{k=1}^K \pi_k \beta_k$, where $\pi \in \Sigma_K$ represents the unknown class proportions and β_k corresponds to the distribution of the k^{th} cell population. Note that, in the framework of flow cytometry, we cannot make the assumption that $\alpha_k = \beta_k$, neither that $\rho_k = \pi_k$. Indeed, ρ could differ from π due to biological differences. For instance if the source sample comes from a healthy patient and the target sample comes from a sick patient, π may feature significant differences from ρ . Moreover, technical variability of cytometry measurements could induce differences (e.g. location shift) between each component α_k and β_k whereas these two distributions represent the same biological phenomenon.

These potential differences between the distributions of two flow cytometry measurements make the development of supervised methods to infer the unknown proportions π in the target data from those of the source data a difficult task. In this work, we rely on the geometric properties of the Wasserstein distance to handle the differences between samples that are due to technical reasons or inter-variability between healthy and sick patients.

3 Class proportions estimation

We now detail a supervised algorithm that estimates the class proportions π in the target distribution. While state-of-the-art automated methods in cytometry data analysis attempt to classify the observations X_1^t, \dots, X_J^t , we go straight to the estimation of the class proportions in the unsegmented target data set. To do so, we borrow ideas from the domain adaptation technique proposed in I.Redko et al. [2018] where it is proposed to re-weight the source observations by searching for weights minimizing the regularized Wasserstein distance between the re-weighted source measure and the target measure. Contrary to I.Redko et al. [2018], we handle the regularized Wasserstein distance and the minimization problem by a stochastic approach. Indeed the large number of observations produced by flow cytometry makes stochastic techniques competitive to handle the high-dimensionality of such data sets.

3.1 A new estimator of the class proportions

We now specify the definition of the estimator of the class proportions in the target data set. For the target sample X_1^t, \dots, X_J^t , the segmentation into various cell sub-types is not available, hence we define the empirical target measure as $\hat{\beta} = \frac{1}{J} \sum_{j=1}^J \delta_{X_j^t}$. From the source observations X_1^s, \dots, X_I^s , we define the empirical source measure as $\hat{\alpha} = \frac{1}{I} \sum_{i=1}^I \delta_{X_i^s}$. Then, the knowledge of the gating of the source data allows to re-write the measure $\hat{\alpha}$ as a mixture of probability measures where each component corresponds to a known sub-population of cells in the data

$$\hat{\alpha} = \sum_{k=1}^K \frac{n_k}{I} \left(\sum_{i: X_i^s \in C_k} \frac{1}{n_k} \delta_{X_i^s} \right) = \sum_{k=1}^K \frac{n_k}{I} \hat{\alpha}_k, \quad (4)$$

where $n_k = \#C_k$ and $\hat{\alpha}_k = \sum_{i: X_i^s \in C_k} \frac{1}{n_k} \delta_{X_i^s}$. Namely, the component $\hat{\alpha}_k$ is the empirical measure of the observations that belong to the sub-population C_k (known class). Then, instead of only considering the true class proportions $(n_1/I, \dots, n_K/I)$ in the source data set, we can re-weight the clusters C_k in the empirical distribution as desired. Indeed, for a probability vector $h = (h_1, \dots, h_K) \in \Sigma_K$ we can define the measure $\alpha(\hat{h})$ that corresponds to the re-weighted

measure $\hat{\alpha}$ such that for all $k \in \{1, \dots, K\}$ the component $\hat{\alpha}_k$ amounts for h_k in the measure $\hat{\alpha}(h)$. In mathematical terms, the measure $\hat{\alpha}(h)$ is thus defined by:

$$\hat{\alpha}(h) = \sum_{k=1}^K h_k \hat{\alpha}_k. \quad (5)$$

Then, to derive the class proportions in the target data, we minimize the regularized Wasserstein distance (2) between the re-weighted source empirical distribution (5) and the target empirical distribution. The main idea is that the source distribution will get closer to the target distribution as the class proportions in its re-weighted version get closer to the class proportions of the target distribution. Thus, we propose to estimate the weights $\pi = (\pi_1, \dots, \pi_K) \in \Sigma_K$ of the underlying distribution β behind the observations X_1^t, \dots, X_J^t of the unlabelled target data set by

$$\hat{\pi} \in \arg \min_{h \in \Sigma_K} W^\varepsilon(\hat{\alpha}(h), \hat{\beta}). \quad (6)$$

The combination of the estimator $\hat{\pi}$ and the algorithms described in Subsection 3.2 to solve the associated minimization problem (6) will be referred to as CytOpT.

3.2 A brief overview of the minimization procedure

The optimization problem (6) leading to the estimator $\hat{\pi}$ of the class proportions does not have a solution in a closed form expression, and a numerical approximation is needed. We propose two methods to solve this optimization problem. In this section, we only give some insights of these two methods without focusing on the technical details. For a more thorough presentation, we refer the reader to the appendix. Our first strategy to tackle Problem (6) consists in using a descent-ascent algorithm where the inner loop yields a stochastic approximate of $\nabla_h W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$. For the second procedure, we borrowed ideas from M.Ballu et al. [2020]. In this second strategy, we add a regularizing term to Problem (6) that permits to swap the min and the max in order to rewrite this problem as a simple expectation maximization problem. Therefore, this regularized version of Problem (6) could be solved with a straightforward stochastic gradient ascent. From our numerical experiments we have found that the second strategy is ten times faster than the descent-ascent procedure. As in practice both procedures seem to provide close estimates, all the results reported were produced with the second strategy, referred to as the "min-max swapping" strategy.

3.3 Some computational considerations

The constant increase of the number of cellular markers used in cytometry analysis raises the question of the impact of such higher-dimensions, both for manual gating (dramatically increasing the resources necessary) and automated approaches (impacting the computational cost and efficiency). As mentioned earlier, Wasserstein metrics drag a computational burden. But regarding CytOpT, the increase in the number of markers does not extend the computational cost by much. Indeed, the dimension only impacts the initial computation of the distance matrix between the observations. Moreover, the random procedures allow a computational cost independent of the number of observations in the target data set. To be more specific, we study the evolution of the computational cost depending on the variables of interest that are; the number of markers measured d , the number of clusters K , and the number of observations I in the source and in the target J . For the descent ascent optimization strategy, which is a double loop algorithm, the computational cost is of order $O(n_{out} n_{in} I d K^2)$ where n_{out} denotes the number of iterations of the outer loop and n_{in} the number of iterations of the inner loop. In the case where the second strategy is preferred, a single loop algorithm can be applied at a computational cost of order $O(n_{iter} I d K)$. Here, n_{iter} denotes the number of iterations of the stochastic gradient algorithm. Notice that in both cases, the random aspect of the optimization procedures provides a computational cost that is independent of the number of observations J in the target data set.

3.4 Extension of CytOpT to a soft assignment method

While our method aims at estimating the class proportions and not to classify the target observations, regularized optimal transport offers a natural soft assignment method which can be used to derive a soft classification of the target data set, as illustrated in Figure 7. Assuming that we have access to the optimal transport plan P_ε with respect to the regularized problem (2), the coefficient $(P_\varepsilon)_{i,j}/b_j$ can be interpreted as the probability that X_i^s is assigned to X_j^t . Here, b_j is the weight associated to the observation X_j^t . Thus, the probability $\gamma_j^{(k)}$ that X_j^t belongs to the class C_k is $\gamma_j^{(k)} = \frac{1}{b_j} \sum_{i=1}^I 1_{X_i^s \in C_k} (P_\varepsilon)_{i,j}$. By choosing the class with highest probability, we can derive a classification for the observation X_j^t . To compute an approximate \hat{P} of P_ε one can plug \hat{U} , the Robbins-Monro approximation of the optimal dual vector u^* , in formula (A.3) of the supplementary. Thus, it is possible to derive an automatic gating of the target data

thanks to optimal transport. However, the transfer of the classification from the source data set toward the target data set requires to compute all the columns of the optimal transport plan. Therefore, obtaining a classification in the target data calls for additional calculations, while in clinical applications, the useful information is very often the relative proportions of the different cell sub-types. That is why this work focuses on the estimation of class proportions.

3.5 Measure of performance

To evaluate our approach on real flow cytometry data, the benchmark class proportions $\pi = (\pi_1, \dots, \pi_K) \in \Sigma_K$, will be defined thanks to the manual segmentation of the target observations. We recall that our algorithm does not make use of the segmentation of the target data set, and that it is only used to evaluate our method. In flow cytometry data analysis the F -measure is a popular tool to assess the performances of the clustering methods. As our method does not yield a clustering but an estimate $\hat{\pi}$ of the class proportions π , we cannot rely on this measure. Indeed, CytOpT yields a probability vector $\hat{\pi} \in \Sigma_K$ that we wish to be the closest to the class proportions $\pi \in \Sigma_K$ of the target data set. Hence, a natural way to measure the discrepancy between $\hat{\pi}$ and π , is the Kullback-Leibler divergence. Thus, with an estimation $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_K)$ and a benchmark $\pi = (\pi_1, \dots, \pi_K)$, to assess the quality of the estimator $\hat{\pi}$, we compute the Kullback-Leibler divergence KL defined as $\text{KL}(\hat{\pi}|\pi) = \sum_{k=1}^K \hat{\pi}_k \log\left(\frac{\hat{\pi}_k}{\pi_k}\right)$.

We also relied on graphical diagnoses from Bland-Altman plots (JM.Bland and DG.Altman [1986]) to visually assess and compare the performance of CytOpT for class proportions estimation. Those diagnoses provide an overview of CytOpT behavior and performance on a collection of several cytometry data sets at once. For one target data set, a Bland-Altman plot compares the estimation $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_K)$ with the benchmark $\pi = (\pi_1, \dots, \pi_K)$ by plotting the difference $\hat{\pi} - \pi$ against the mean $(\hat{\pi} + \pi)/2$. To simultaneously visualize the results from the analysis of two data sets, one just adds on to the graph the points defined by $\hat{\pi}^{(2)} - \pi^{(2)}$ for the y-axis and $(\hat{\pi}^{(2)} + \pi^{(2)})/2$ for the x-axis (with $\hat{\pi}^{(2)}$ denoting the estimation on the second data set and $\pi^{(2)}$ the benchmark on that data set). This way, one can actually represent the results from the analysis of several data sets in a single plot. We also display on the Bland-Altman plots the mean of the difference with a solid horizontal line, and with dashed horizontal lines, ± 1.96 times the standard deviation of the difference. As we aim for $\hat{\pi}$ to be as close as possible to π , the closer to the x -axis the points are, the better the performance.

4 Simulation Study

4.1 Comparison with classification methods when targeting the weights of a mixture

To evaluate the performance of our method we generate two data sets from two different mixtures of $K = 10$ components each. The first data set is drawn from a Gaussian mixture distribution in dimension $d = 10$ with a vector of proportions $\rho \in \Sigma_K$. Thus, $\alpha = \sum_{k=1}^{10} \rho_k \alpha_k$ where the α_k are Gaussian distributions. The second data set is drawn from the same Gaussian components α_k but with different mixture proportions $\pi \neq \rho$. Then, we add a non-linear mapping $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ mimicking domain-shift. Thus, $\beta = \sum_{k=1}^{10} \pi_k \beta_k$ where $\beta_k = T\#\alpha_k$ (that is the push-forward of the measure α_k by the mapping T). In this simulation study, we apply the same pre-processing of the data as the one applied to the real flow cytometry data. That is, we threshold the negative values at zero and re-scale the observations such that every observation X belongs to $[0, 1]^{10}$. See Figure 2 for an illustration of a generated data-set. We compare our method with three state-of-the-art classification methods: one unsupervised method – the k -means ; and two supervised methods – quadratic discriminant analysis (QDA), and random forest (RF). After a training step on the source data set, each supervised method yields a classification of the target data set. Hence, after this classification step by either QDA or RF, we can derive an estimation of the class proportions in the target data set. We indicate, that in this framework (namely with a single learning data set), the QDA algorithm corresponds to the `OptimalFlow` method [Barrio et al., 2019], that is a state-of-the-art method designed to cluster flow cytometry data. As the k -means is an unsupervised method, we apply it straight to the target data set. Since the k -means does not yield a labelled classification, it is necessary to add an annotation step for each cluster C_k returned by the k -means algorithm. To annotate a cluster C_k , we retrieve the majority population according to the true labels and then we label all the observations that belong to C_k with this label. Therefore, we provide a significant upper hand to this method with this additional information. To study the performance and the stability of these various methods, we sampled 100 source data sets with respect to the source model (with a fixed number $I = 115,783$ of observations), and sampled 100 target data sets with respect to the target model (with a fixed number $J = 68,981$ of observations). Then, for each pair of source and target data sets we apply the supervised methods, that is CytOpT, QDA and RF, and on each target data set, we apply the k -means algorithm followed by the annotation step described above. Hence, for each method we obtain 100 estimates of the class proportions. Figure 2 displays an example of a pair of source and target data sets with their corresponding class

proportions. Figure 3 displays the median class proportions retrieved with these 100 estimates and the variability of these estimates. In this scenario, CytOpT is the method which offers the best trade-off between accuracy and stability.



Figure 2: **Simulated source data set and Simulated target data set.** Using the segmentation of the source data set (Left) CytOpT estimates the class proportions in a target data set (Right) without making use of the classification of the target data set.

4.2 Robustness evaluation

This section is devoted to the robustness analysis of CytOpT. To do so, we disrupted the ideal situation where the source mixture $\alpha = \sum_{k=1}^{10} \rho_k \mathcal{N}(\mu_k, \sigma I)$ and the target mixture $\beta = \sum_{k=1}^{10} \pi_k \mathcal{N}(\mu_k, \sigma I)$ share the same normal components but have different weights in three different ways. First, we added a uniform noise in $[0, 1]^d$ to the target data set, to mimic outliers observations that can often be observed in practice. In our experiments, the noise amounted from 10% of the target data to 70% of the target data. This first scenario is displayed in the upper panels of Figure 4. Second, we added a $K + 1^{th}$ class in the target data set that could not be found in the source data set. In this more involved situation, the additional class was also sampled from a Gaussian distribution with its center randomly chosen in $[0, 1]^d$. This additional class represented up to 20% of the target data set in some envisioned scenarios. This second scenario is displayed in the lower panels of Figure 4. Third, we remove from 1 to 5 components in the target mixture. Hence, in this last scenario more classes were presented in the source data set than in the target data set. For each of these three cases, we sampled 10 couples (source data set, target data set) to get 10 estimates of the class proportions for a given framework. Figure 5 represent the performance of CytOpT in terms of Kullback-Leibler divergence in these three robustness trials. Unsurprisingly, the more we disrupt the initial setting, the less accurate the estimation. However, one can observe in Figure 5 that the increase of the Kullback-Leibler divergence is proportional to the number of classes removed or to the proportions of the noisy observations in the target. Those results are reassuring for practical applications as it shows that some reasonably mild disturbances in the target will not lead to meaningless results.

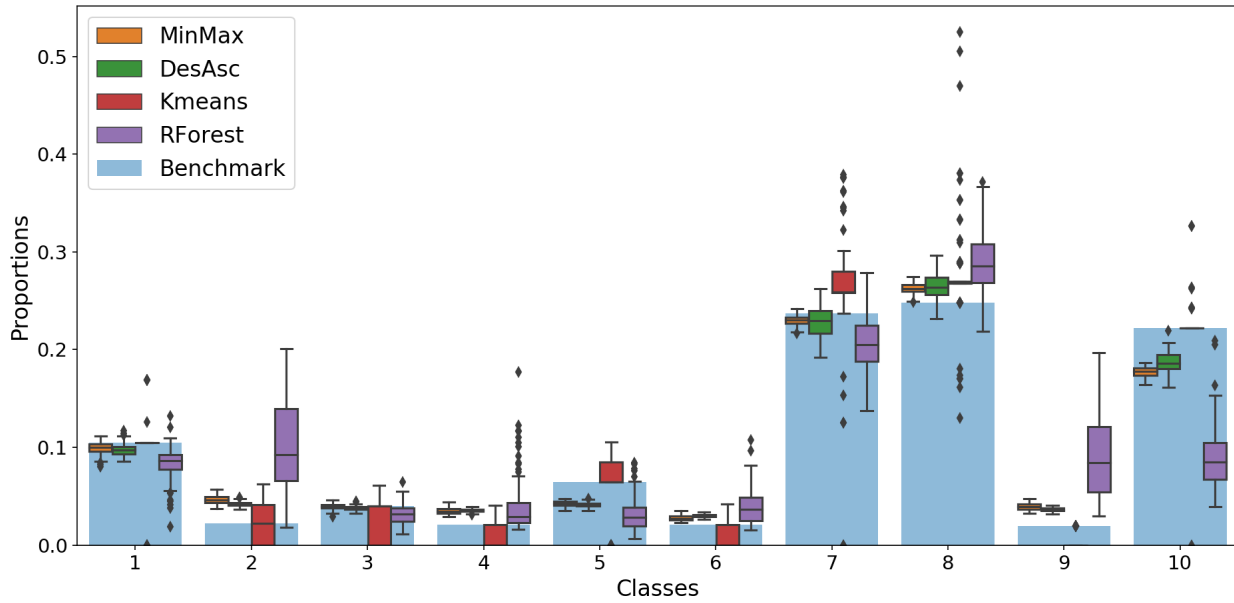


Figure 3: **Performances of the automated methods tested to estimate the class proportions.** For each method, one boxplot displays the median and the variation of the estimation of the weight of one component.

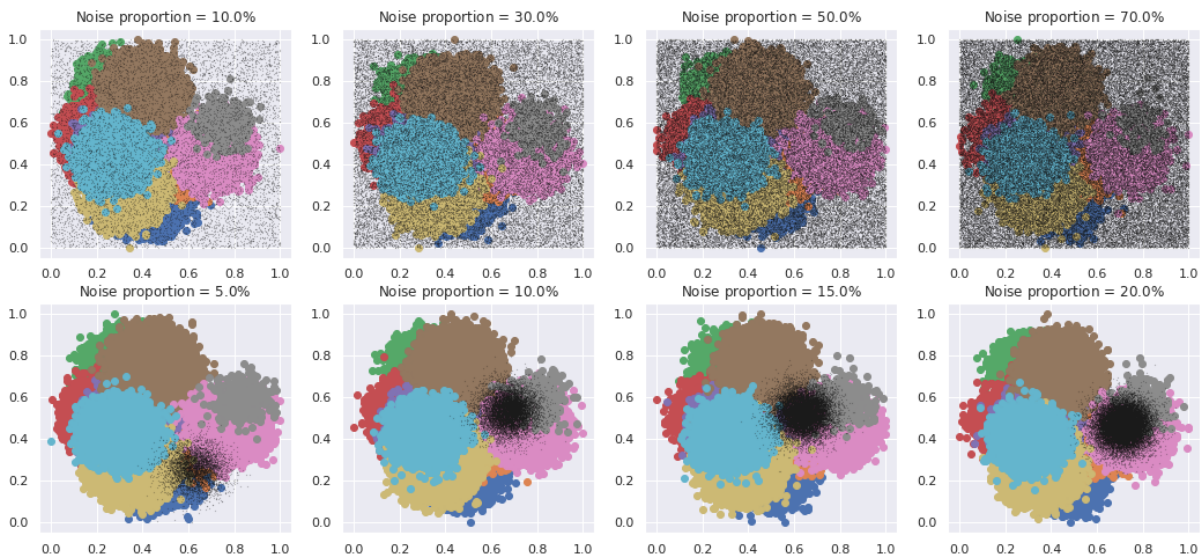


Figure 4: **Scenarios with additional observations in the target.** Top: the additional class (in black) is uniformly distributed in $[0, 1]^d$. Bottom: the additional class (in black) is distributed with respect to a Normal distribution.



Figure 5: **Robustness assessment of CytOpT performance according to Kullback-Leibler divergence of the estimated proportion.**

5 Application to real flow cytometry data analysis

5.1 Illustration of the methodology with a two classes example

In this section, we illustrate our proposed method in the setting where the cytometry data from HIPC described in Section 1.3 are divided only among two broad classes: the CD4 cells and the CD8 cells. For ease of visualization, we use only two markers: the CD4 marker and the CD8 marker – that is $d = 2$. This basic case, where two-dimensional data are divided into two classes, is a first illustration of our method in a favorable situation. We aim to demonstrate that our approach may reach a good approximation of the CD4 proportion and the CD8 proportion in a target cytometry data set with unknown gating.

We consider the two data sets from the HIPC T-cell panel that are displayed in Figure 6. The first data set is a series of cytometry measurements performed in a Stanford laboratory on a biological sample that comes from a patient identified as patient 1. This first data set is chosen as the segmented source data. The second data set is a series of cytometry measurements done in the same laboratory on a biological sample that comes from an other patient identified as patient 3. This second data set will be the target data set. The manual gating classification, and thus the class proportions are available for both samples. Nevertheless, we only use the classification available for the source (i.e. Stanford1A) data set in order to estimate the class proportions in the target (i.e. Stanford3A) data set. The manual gating classification for the target data set is only used as a benchmark to evaluate our method.

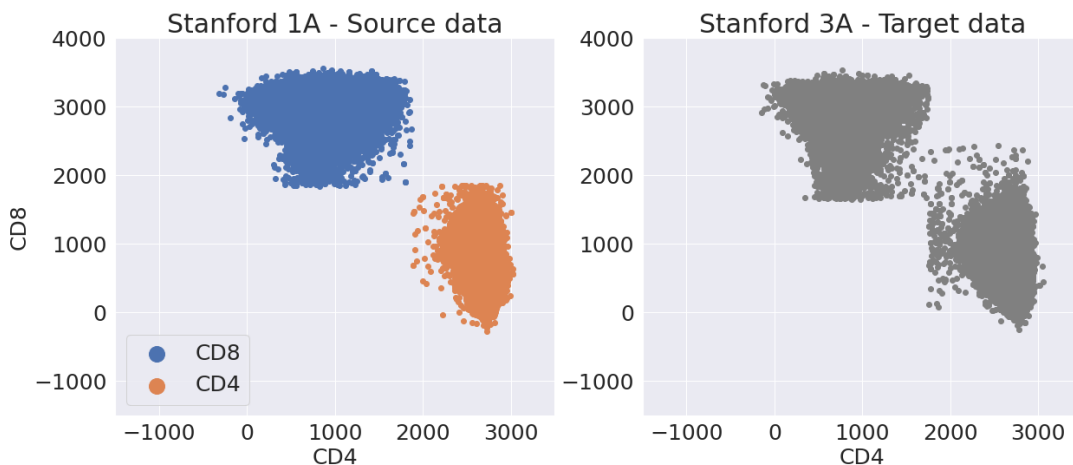


Figure 6: **Illustration of the CytOpT framework.** CytOpT estimates the class proportions in an unclassified data set (Stanford3A, right) from one classified data set (Stanford1A, left) without clustering the observations of Stanford3A.

First, to assess the relevance of our method, we present in Figure C.2 the evolution of the regularized Wasserstein distance as a function of the weights associated to each class in the source distribution. To this end, we evaluate the function $F : h_1 \mapsto W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$, where $h = (h_1, 1 - h_1)$, on a finite grid $\mathcal{H} = \{h^{(1)}, \dots, h^{(m)}\}$. For $h_1 \in \mathcal{H}$ we

approximate $W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$ by the estimator \widehat{W}_n defined in equation (B.7). It can be observed that the regularized Wasserstein distance decreases as the class proportions of the source data set get closer to the class proportions of the target data set. Using the segmentation of a cytometry data set, CytOpT adequately retrieves the true class proportions of an unlabelled cytometry data set. Even if the two classes situation is somewhat an easy scenario, one can observe the significant gap in the class proportions between the source data set and the target data set. In the Stanford1A data set the CD4 cells constitute 45.1 % of the cells and the CD8 cells 54.9 % of the cells, whereas in the Stanford3A data set the CD4 cells constitute 73.9 % (estimated at 73.3 % by CytOpT) of the cells and the CD8 cells 26.1 % (estimated at 26.7 % by CytOpT).

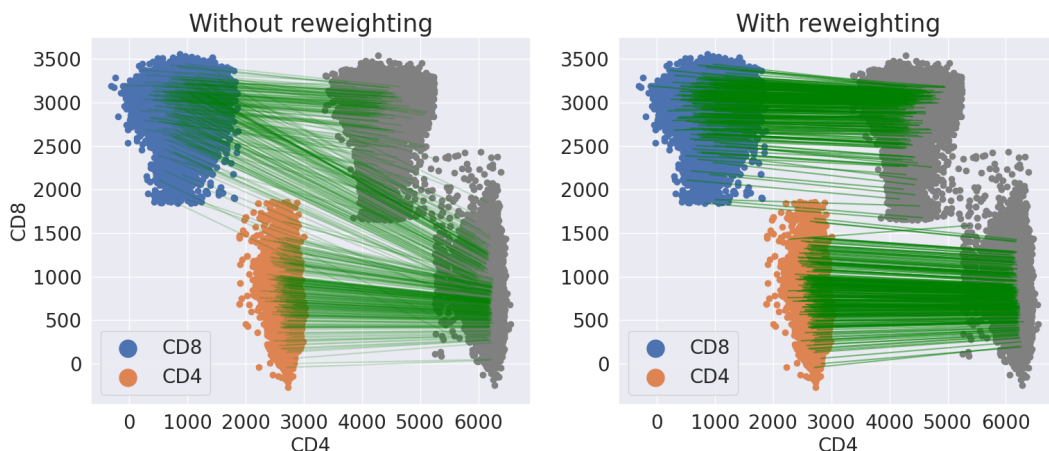


Figure 7: **Optimal transport plan P_ε between the source and target distribution.** A green line between X_i^s and X_j^t indicates that the optimal transport plan P_ε moves some mass from X_i^s to X_j^t . To facilitate readability, the target data set has been shifted, and only 500 coefficients of P_ε have been represented. Without re-weighting, mass from the CD8 class in the source data set is sent toward the CD4 class in the target data set. With re-weighting, the mass from one class in the source data set is sent to the corresponding class in the target data set.

The results presented in Figure 7 and Figure 8 represent the use of the soft-clustering and classification methods described in Section 3.4, and they outline the importance of re-weighting the source data. Without re-weighting, the classification obtained by optimal transport is really unlike the manual gating classification. However, one can re-weight the observations with the transformation $a(\pi) = \Gamma\pi$ in order to match the class proportions π in the target data set. Here, Γ is the linear operator defined in the equation (B.8) of the appendix. And once the source data are re-weighted with the estimated class proportions $\hat{\pi}$, the classification by regularized transport is very similar to the manual gating classification.

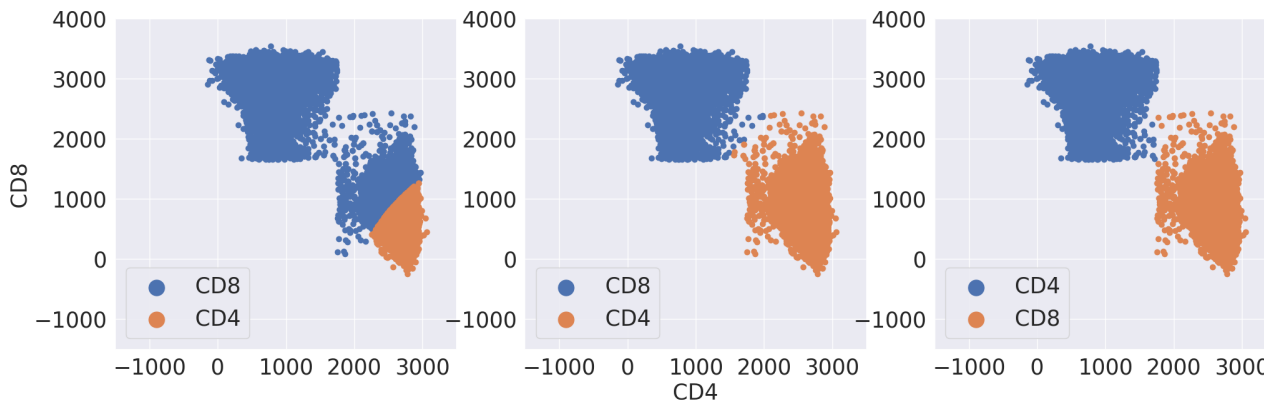


Figure 8: **Results of the soft assignment method and comparison with the manual clustering on the data set Stanford3A.** From left to right: Label transfer without reweighting - Label transfer with reweighting - Manual gating benchmark

5.2 Estimating 10 cell sub-type proportions from 7 cellular markers

We apply our method to the full T-cell panel of the HIPC flow cytometry data described in Section 1.3. Contrary to the illustrative results presented in Section 5.1, we now use all of the $d = 7$ markers at hand to estimate the proportions of $K = 10$ cell populations. This represents a more involved classification task than the one considered in Section 5.1 with $K = 2$. We arbitrarily chose the data set "Stanford1A" as the source measure to derive the class proportions in all the other data sets as targets. A comprehensive evaluation of CytOpT performance is provided in Figure 11, which features a Bland-Altman plot [JM.Bland and DG.Altman, 1986] displaying all the cell sub-types proportions estimations by CytOpT when targeting all available data sets across all the seven different centers available. In this example, we solely used the information from one reference classification from manual gating of one data set (namely Stanford1A) to estimate the class proportions in each of the 61 unsegmented data sets targeted. In more than 90% of the cases the absolute difference between the estimated proportion and the manual gating gold-standard proportion is below 5%. And in more than 99% of the cases the absolute error is no more than 10%. Due to the stochastic nature of our algorithm, a new call to CytOpT would lead to a slightly different estimate $\hat{\pi}$ compare to a former estimation. However, the results displayed in Figure 11 and Figure 13 are representative of the general quality of the estimation produced by CytOpT in such settings.

Note that when applying our method to estimate class proportions on real flow cytometry data, a simple pre-processing is required. First, the signal processing of the cytometer can induce some contrived negative values of light intensity [Tung et al., 2007]. To undo this effect, we merely threshold those few negative values at zero. Second, to set the parameters of our algorithms, in particular ε , we need to bound the displacement cost. To do so, we scale the data such that: $\forall i \in \{1, \dots, I\}, X_i^s \in [0, 1]^d$ and $\forall j \in \{1, \dots, J\}, X_j^t \in [0, 1]^d$.

5.2.1 Comparison with a straight application of one Manual Gating

As previously stated, technical variability across samples can lead to misalignment of flow cytometry data. For instance, spatial shift is displayed on Figure 9 with only CD4 cells and the corresponding manual gating established on the basis of the Stanford1A measurements alone. While the four classes displayed on Figure 9 represent the same biological phenomenon, it appears that the boundaries differ among the different centers. To illustrate the interest of CytOpT when technical variability induces a shift between the source and the target data set we compare CytOpT with the mere application of the Stanford 1A manual gates to the ungated data set. This comparison is performed on the 61 remaining ungated data sets. The bland-Alman displayed on Figure 10 shows that CytOpT provides a significant improvement over the direct application of Stanford1A's gating. We will now compare CytOpT with state-of-the-art automated approaches designed to analyze flow cytometry data.

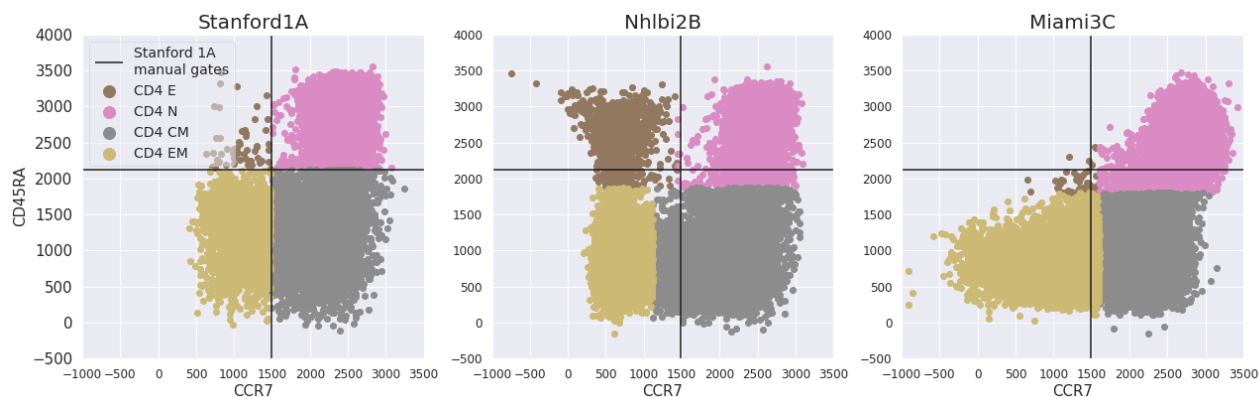


Figure 9: **2D projection of three cytometry data sets.** Left: CD4 cell measurements performed in Stanford. Middle: CD4 cell measurements performed in Nhlbi. Right: CD4 cell measurements performed in Miami.

5.2.2 Comparison with other methods

On the HIPC data, we compare CytOpT with 3 state-of-the-art automated-gating approaches specifically designed to analyze flow cytometry data. First, `flowMeans` [Aghaeepour et al., 2011] which is a variation of the k -Means algorithm designed to cluster flow cytometry data. As `flowMeans` is an unsupervised method, we will label every observation of a cluster retrieved by `flowMeans` with the major population according to the reference manual clustering (Note that

this in theory should provide a substantial advantage over the other methods). We also apply `Cytometree` to cluster the HIPC data. This method offers an automated annotation of its estimated clustering. Therefore, with `Cytometree`, we can directly retrieve an estimation of the class proportions. Finally, we compare the performances of our method with `OptimalFlow`, recently developed by Barrio et al. [2019] and which performs supervised automatic gating. We mention that it is the first method to make use of the Wasserstein distance in the context of automatic-gating of flow cytometry data analysis. We stress that, even though both our method and `OptimalFlow` are supervised approaches relying on the Wasserstein distance, they greatly differ. First, they have different objectives: `OptimalFlow` aims at classifying the cells, while `CytOpT` directly aims at estimating the cellular sub-types proportions. Second, `OptimalFlow` uses the Wasserstein distance to cluster a compendium of pre-gated cytometry data sets and uses the notion of Wasserstein barycenter to produce a prototype data set for each cluster. Then, using the Wasserstein distance, `OptimalFlow` browses among the prototypes to extract the most relevant prototype to classify a test data set. But ultimately, the classification is performed with tools such as `tclust` [B.Dost et al., 2010], or Quadratic Discriminant Analysis [T.Hastie et al., 2009] that do not belong to the field of optimal transport. On the contrary, `CytOpT` only requires one pre-gated data set and the estimator of the class proportions is fully based on regularized optimal transport. In this comparison, `OptimalFlow` had a learning database composed of the nine Stanford cytometry data sets, and it was set to produce one prototype from the database (Note that `OptimalFlow` thus leverages the information from multiple source data sets, not just one). For the rest of the method, we applied `OptimalFlow` with the default settings proposed in the vignette¹.

Figure 11 displays a Bland-Altman plot for each of the compared method. `flowMeans` and `Cytometree` clearly stand-out with less accurate class proportions estimation that cannot compete with the good performance of `CytOpT` and `OptimalFlow`. Although visually `CytOpT` and `OptimalFlow` seem to yield close results, one can notice that `OptimalFlow` estimations are systematically biased as it underestimates the rarest cell populations proportions while overestimating the most represented cell populations frequencies.

To get a more quantitative criterion to compare `CytOpT` and `OptimalFlow`, we used the Kullback-Leibler divergence to assess the discrepancy between the gold-standard class proportions derived manually and the estimations of the two methods. Figure 12 shows that on almost every data set the quality of estimation is rather close. However when the target data set comes from Miami, the Kullback-Leibler divergence between the benchmark and `CytOpT` estimations is significantly lower than the Kullback-Leibler divergence between the benchmark and `OptimalFlow` estimations. This means that in these cases `CytOpT` performs significantly better than `OptimalFlow`. We account for these results by the spatial shift between the source and the target. To end this comparison, we present in Figure 13 a comparison of the class proportions estimation between `CytOpT` and `OptimalFlow` in two cases. The first case is the worst estimation for `OptimalFlow`, namely, when the target data set is `Miami3C`. In this case, `OptimalFlow` overestimates very much the proportion of CD4 effector and the proportion of CD4 activated. The second case corresponds to the worst estimation for `CytOpT`, namely when the target data set is `Nhlbi3B`. If `OptimalFlow` performs slightly better in this case, `CytOpT` still offers an acceptable and competitive estimation.

5.3 Application to the `OptimalFlow` data

A second evaluation of our method was performed on the cytometry data used in Barrio et al. [2019]. This second cytometry panel is available on GitHub², and we refer to this panel as the `OptimalFlow` data sets. First, in order to run a sensible comparison between `CytOpT`, `OptimalFlow` and `FlowMeans` we selected nine cell populations that were present in the first 21 data sets. As `FlowMeans` is an unsupervised method, it was directly applied to the 21 data sets. For `OptimalFlow`, it was set to use the 3 first data sets of the database, to build one template, and then to apply the classification step. Hence `OptimalFlow` yielded a class proportions estimation for the 18 remaining data sets. For `CytOpT`, we used the first data set as a source data set and we estimated the class proportions in the 20 remaining data sets. Figure 14 shows the Bland-Altman plots of the results for those three methods. One can notice that `CytOpT` is the method where the estimation of the class proportions $\hat{\pi}_k$ is overall the closest to the manual gold-standard π_k .

6 Discussion

We have introduced `CytOpT`, a supervised method that differs from existing method as it directly estimates the cell population proportions, without a preliminary clustering or classification step. While obtaining a classification of the observation is often a mathematical tool for achieving automatic gating, we emphasize that from a clinical perspective, the proportions of the different cell populations is the quantity of interest [Henel and Schmitz, 2007]. Thus, we have

¹at <https://github.com/HristoInouzhe/optimalFlow>

²at <https://github.com/HristoInouzhe/optimalFlowData>

proposed efficient numerical schemes that rely on stochastic algorithms to address the delicate issue of computing and minimizing the regularized Wasserstein distance. We have demonstrated empirically, both on simulated data and on real flow cytometry data, the outstanding performance of CytOpT to tackle this question of cell population proportion estimation, especially in difficult situations where the technical variability of flow cytometry induces spatial shifts between samples. We also think that with higher-dimensional data, our method could perform even better as the Wasserstein distance would take advantage of all the dimensions available at once.

We present a few perspectives to further improve the performance of CytOpT. First, an additional pre-processing step to handle outliers could make CytOpT more robust to extreme cellular observations that have little biological meaning. Indeed, due to the needed scaling standardization of the data, one outlier observation can change the layout of the source data in comparison with the layout of the target data. A data driven strategy to choose the regularization parameter ε could also help to tackle this outlier issue. Following the supervised approach of Barrio et al. [2019], one could also propose an extension of CytOpT that would leverage the information from several labeled source data sets. A potential approach would be to select the most relevant labeled data set to estimate the class proportions in an unclassified target data set. To be more specific, let us assume that we have M classified data sets X^1, \dots, X^M . For each data set X^m , we can define the re-weighted empirical measure $\alpha^m(h)$ such that the weight of the k^{th} class equals h_k for all $k \in \{1, \dots, K\}$, which would yield an optimal re-weighting $\hat{\pi}^m$ that corresponds to an estimation of the class proportions π in the target data set. Finally, to select the most relevant estimator $\hat{\pi}^m$, we choose the one leading to the smallest Wasserstein cost between the re-weighted empirical measure $\hat{\alpha}^m(\hat{\pi}^m)$ and the empirical target distribution $\hat{\beta}$ that is :

$$\hat{\pi} = \arg \min_{\hat{\pi}^1, \dots, \hat{\pi}^M} \{W^\varepsilon(\hat{\alpha}^1(\hat{\pi}^1), \hat{\beta}), \dots, W^\varepsilon(\hat{\alpha}^M(\hat{\pi}^M), \hat{\beta})\} \quad (7)$$

Another extension of this work could also consider the Sinkhorn divergence $S^\varepsilon(\alpha, \beta) = W^\varepsilon(\alpha, \beta) - \frac{1}{2}W^\varepsilon(\alpha, \alpha) - \frac{1}{2}W^\varepsilon(\beta, \beta)$, instead of $W^\varepsilon(\alpha, \beta)$. Indeed, the authors of J.Feydy et al. [2018] showed that S^ε has theoretical properties that make it a suitable loss for applications in machine learning.

Codes and data availability

The illustrative HIPC data sets described in Subsection 1.3, as well as Python Notebooks to reproduce the Figures presented in this paper are available at <https://github.com/Paul-Freulon/CytOpT>. A user-friendly Python package is available from pypi with source code at <https://pypi.org/project/CytOpT/>. A user-friendly R package, relying on the reticulate framework [Ushey et al., 2022] to encapsulate the above Python code, will be available from CRAN with source code at <https://github.com/sistm/CytOpT-R>.

Acknowledgments

The authors would like to thank the associate editor and the three referees for their suggestions and constructive comments which helped to improve the paper substantially. The author would also like to thank Kalidou Ba for developing the CytOpT packages. Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine. Jérémie Bigot is a member of Institut Universitaire de France (IUF), and this work has been carried out with financial support from the IUF.

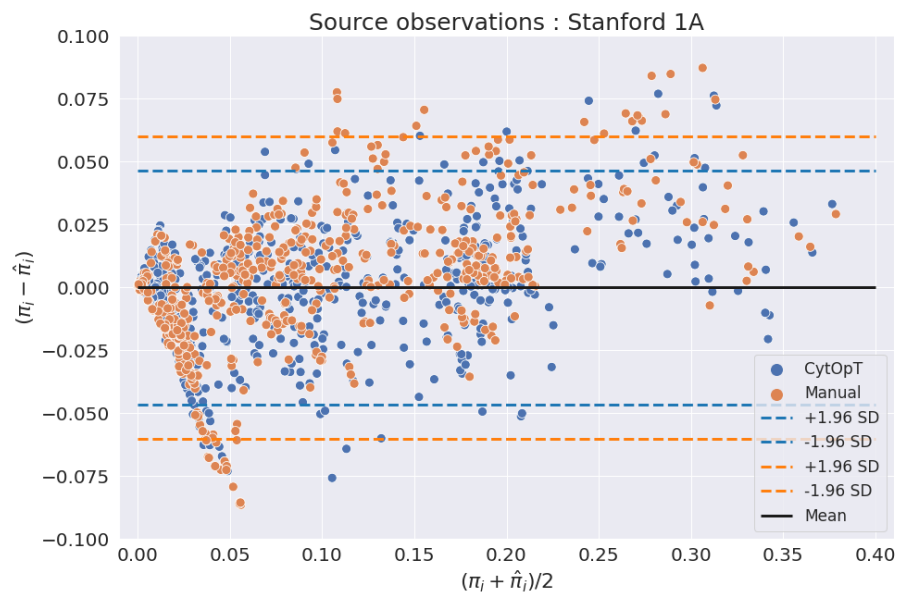


Figure 10: Bland-Altman of CytOpT and Stanford1A's gating compared to specific manual gating benchmark.

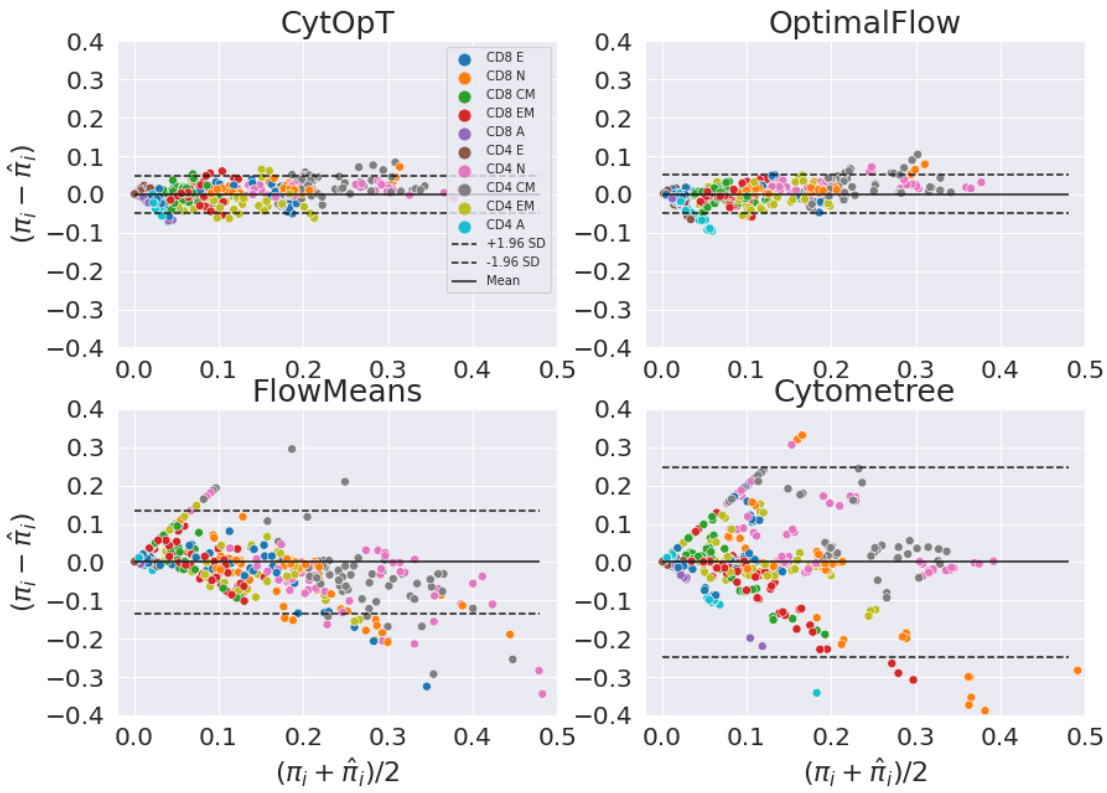


Figure 11: Comparison of the proportions $\hat{\pi}$ estimated by four automated methods and the manual benchmark π on the HIPC database. By plotting the difference against the mean of the results of two different methods, the Bland-Altman plot allows to assess the agreement between an automated method and the manual benchmark.

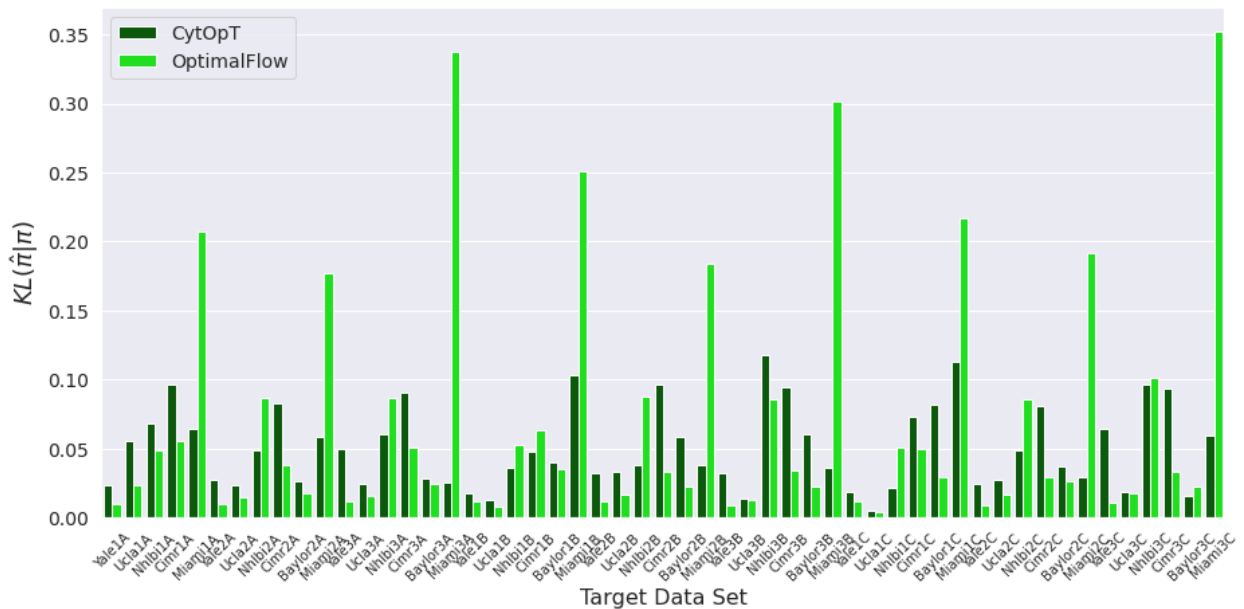


Figure 12: Comparison between our algorithm CytOpT and one of the state of the art automated method for cytometry data analysis: OptimalFlow. The comparison was performed on the data sets of the HIPC panel.

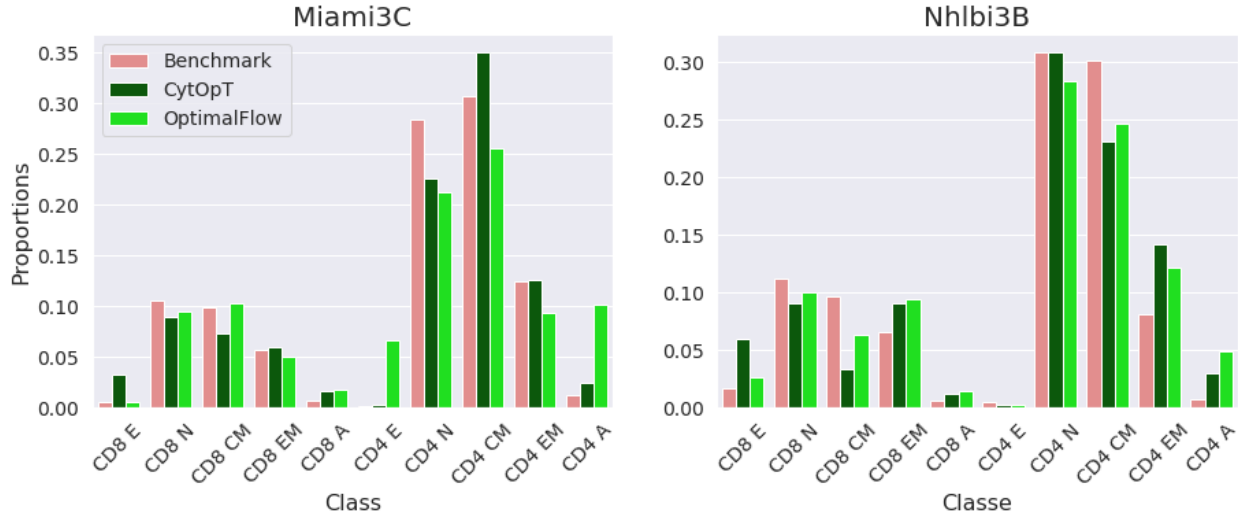


Figure 13: Comparison of the estimated proportions $\hat{\pi}$ by CytOpT and OptimalFlow with the manual gating benchmark π . Left: The target data set is Miami3C. Right: The target data set is Nhlbi3B. In both cases, the data set Stanford1A has been used as a source data set for CytOpT.

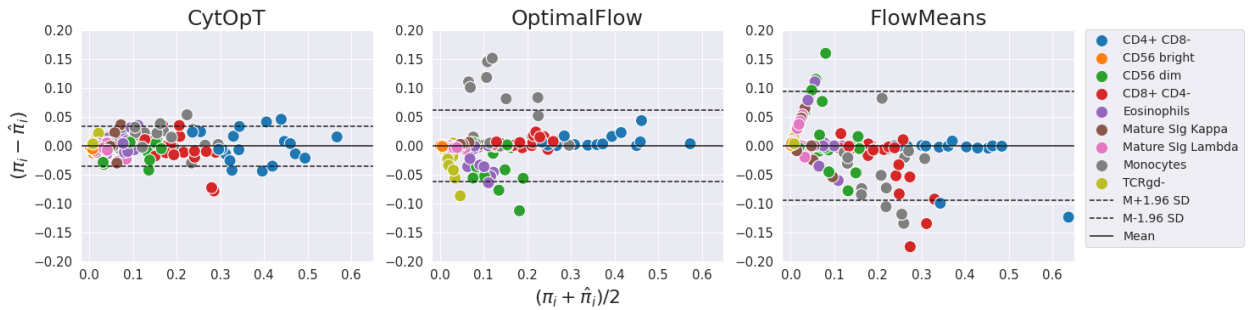


Figure 14: Comparison of the proportions $\hat{\pi}$ estimated with CytOpT and the manual benchmark π on the OptimalFlow database.

A Stochastic algorithm for the regularized Wasserstein distance

In this work, the regularized Wasserstein distance is calculated with a statistical procedure based on the Robbins-Monro algorithm for stochastic optimization. This way to calculate the Wasserstein distance is investigated in A.Genevay et al. [2016] and B.Bercu and J.Bigot [2020]. The keystone of this approach is that the regularized Wasserstein problem can be written as the following stochastic optimization problem:

$$W^\varepsilon(\alpha, \beta) = \max_{u \in \mathbb{R}^I} \mathbb{E}[g_{\varepsilon, \alpha}(Y, u)], \quad (\text{A.1})$$

where Y is a random variable with distribution β , and $g_{\varepsilon, \alpha}$ is defined as:

$$g_{\varepsilon, \alpha}(y, u) = \sum_{i=1}^I u_i a_i + u_{c, \varepsilon}(y) - \varepsilon, \quad (\text{A.2})$$

where $u_{c, \varepsilon}(y_j) = \varepsilon \left(\log(b_j) - \log \left(\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right) \right) \right)$. We point out that the solution P_ε of the regularized primal problem (2) is recovered from any $u^* \in \mathbb{R}^I$ solution of the semi-dual problem (3) as:

$$(P_\varepsilon)_{i,j} = \exp \left(\frac{u_i^* + u_{c, \varepsilon}^*(y_j) - c(x_i, y_j)}{\varepsilon} \right). \quad (\text{A.3})$$

Formulation (A.1) and the fact that for all $y \in \mathbb{R}^d$, the function $g(y, \cdot)$ is concave lead us to estimate the vector u^* by the Robbins-Monro algorithm H.Robbins and S.Monro [1951] given, for all $n \geq 0$, by:

$$\widehat{U}_{n+1} = \widehat{U}_n + \gamma_{n+1} \nabla_u g_\varepsilon(Y_{n+1}, \widehat{U}_n), \quad (\text{A.4})$$

where the initial value \widehat{U}_0 is a random vector which can be arbitrarily chosen, Y_1, \dots, Y_{n+1} is an i.i.d. sequence of random variables sampled from the distribution β , and $(\gamma_n)_{n \geq 0}$ is a positive sequence of real numbers decreasing toward zero satisfying

$$\sum_{n=1}^{\infty} \gamma_n = +\infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^2 < +\infty. \quad (\text{A.5})$$

It follows from M.Cuturi [2013] that the maximizer u^* of (A.1) is unique up to a scalar translation of the form $u^* - t1_I$ for any $t \in \mathbb{R}$. Throughout this paper, we shall denote by u^* the maximizer of (A.1) satisfying $\langle u^*, 1_I \rangle = 0$ which means that u^* belongs to $\langle 1_I \rangle^\perp$ where $\langle 1_I \rangle$ is the one-dimensional subspace of \mathbb{R}^I spanned by 1_I . As shown in B.Bercu and J.Bigot [2020], to obtain a consistent estimator of u^* , one must slightly modify the Robbins-Monro algorithm by requiring that \widehat{U}_0 belongs to $\langle 1_I \rangle^\perp$. To approximate the regularized Wasserstein distance $W^\varepsilon(\alpha, \beta)$ B.Bercu and J.Bigot [2020] proposed the recursive estimator:

$$\widehat{W}_n = \frac{1}{n} \sum_{k=1}^n g_\varepsilon(Y_k, \widehat{U}_{k-1}). \quad (\text{A.6})$$

Throughout this work, we will use this estimator.

B Two different numerical optimization schemes

In this supplementary to the article, we present the technical details of the optimization strategies to compute the estimator of the class proportions which is given by:

$$\hat{\pi} \in \arg \min_{h \in \Sigma_K} W^\varepsilon(\hat{\alpha}(h), \hat{\beta}). \quad (\text{B.7})$$

In this optimization problem (B.7) that yields an estimator $\hat{\pi}$ of the class proportions, the weights of the source distribution are modified, but the support is fixed. Thus, the parameter h only impacts the reweighted measure $\hat{\alpha}(h)$ through its weights $a(h) = (a_1(h), \dots, a_I(h))$. We therefore introduce the linear operator $\Gamma \in \mathbb{R}^{I \times K}$ that maps the class proportions h to the weight vector $a(h)$ with the constraint that among a given class C_k every observation have the same weight $\frac{h_k}{n_k}$, where $n_k = \#C_k$. This linear operator Γ is defined by:

$$\forall (i, k) \in \{1, \dots, I\} \times \{1, \dots, K\}, \Gamma_{i,k} = \begin{cases} \frac{1}{n_k} & \text{if } X_i^s \in C_k, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{B.8})$$

Thanks to this transformation, the vector $a(h) = \Gamma h$ is such that the class proportions in the discrete measure $\sum_{i=1}^I a_i(h) \delta_{X_i^s}$ equals the vector h .

B.1 Descent-Ascent procedure

Formulation of the Problem To present the ideas of our first way to numerically solve (6), we re-write this optimization problem as:

$$\min_{h \in \Sigma_K} W^\varepsilon(\hat{\alpha}(h), \hat{\beta}) = \min_{h \in \Sigma_K} \max_{u \in \mathbb{R}^I} \mathbb{E}[g_{\varepsilon, \alpha}(Y, u, h)], \quad (\text{B.9})$$

where Y is a random variable with distribution $\hat{\beta}$, and for $y_j \in \mathbb{R}^d$ an observation of the random variable Y ,

$$g_{\varepsilon, \alpha}(y_j, u, h) = \sum_{i=1}^I u_i a_i(h) + \varepsilon \left(\log(b_j) - \log \left(\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right) \right) \right) - \varepsilon. \quad (\text{B.10})$$

We mention that contrary to expression A.2, the function $g_{\varepsilon, \alpha}$ that appears in expression B.10 only depends of α through its support $X^s = (X_1^s, \dots, X_I^s)$. For ease of notation we stick to the notation $g_{\varepsilon, \alpha}$.

Due to the min-max formulation of Problem (B.9), a descent-ascent strategy is a natural way to solve this optimization problem. At each iteration, our method runs multiple stochastic gradient ascent steps to estimate the solution of the inner maximization problem. Thanks to this estimation, we can approximate the gradient of the function $h \mapsto W(\hat{\alpha}(h), \hat{\beta})$ that we wish to minimize. We now go into the detail of our stochastic descent-ascent optimization procedure.

A new parameterization of the optimization of our problem First, in order to avoid projecting h_n on the simplex Σ_K at each step of the procedure, we re-parameterized our problem with a soft-max function. Our new optimization problem is now:

$$\min_{z \in \mathbb{R}^K} W^\varepsilon(\Gamma(\sigma(z)), b) = \min_{z \in \mathbb{R}^K} F(z), \quad (\text{B.11})$$

where $F(z) = W^\varepsilon(\Gamma(\sigma(z)), b)$, and $\sigma : \mathbb{R}^K \rightarrow \Sigma_K$ is the soft-max function defined as

$$\sigma(z)_l = \frac{\exp(z_l)}{\sum_{k=1}^K \exp(z_k)}. \quad (\text{B.12})$$

Then, if we denote \hat{z} a minimizer of (B.11), we will derive an estimator of π as $\hat{\pi} = \sigma(\hat{z})$.

An approximation of the gradient of the objective function F The gradient of the function $a \mapsto W^\varepsilon(a, b)$ is given by:

$$\frac{\partial}{\partial a} W^\varepsilon(a, b) = u^* \quad (\text{B.13})$$

where u^* is the unique solution to (A.1) centered such that $\sum_{i=1}^I u_i^* = 0$. This result is established in Proposition 4.6 of [G.Peyré et al., 2019]. Applying the chain rule of differentiation to the objective function F of problem (B.11), we obtain that:

$$\nabla_z W^\varepsilon(\Gamma(\sigma(z)), b) = (\Gamma J_\sigma(z))^T u_z^*, \quad (\text{B.14})$$

where Γ is the linear operator defined in (B.8), u_z^* denotes a maximizer of (A.1) when the weights of the distribution $\hat{\alpha}$ equal $a = \Gamma\sigma(z)$, and $J_\sigma(z)$ is the Jacobian matrix of σ . For $z \in \mathbb{R}^K$, $J_\sigma(z)$ is given by:

$$\forall (i, j) \in \{1, \dots, K\}^2, J_\sigma(z)_{i,j} = \begin{cases} \frac{\exp(x_j) (\sum_{k=1}^K \exp(x_k))^{-\exp(2x_j)}}{(\sum_{k=1}^K \exp(x_k))^2} & \text{if } i = j \\ -\frac{\exp(x_i + x_j)}{(\sum_{k=1}^K \exp(x_k))} & \text{otherwise.} \end{cases} \quad (\text{B.15})$$

Using formula (B.14), the gradient of the objective function F at the point \hat{z}_n reads $\nabla_z F(\hat{z}_n) = (\Gamma J_\sigma(z))^T u_{\hat{z}_n}^*$. We propose to plug an estimate of $u_{\hat{z}_n}^*$ in this last formula where the estimate of $u_{\hat{z}_n}^*$ is computed with the Robbins-Monro [H.Robbins and S.Monro, 1951] algorithm described in Section A. Therefore, the estimate $\hat{U}_{m+1}^{(n+1)}$ of $u_{\hat{z}_n}^*$ is calculated with the recursive algorithm:

$$\hat{U}_{k+1}^{(n+1)} = \hat{U}_k^{(n+1)} + \gamma_{k+1} \nabla_u g_\varepsilon(Y_{k+1}^{(n+1)}, \hat{U}_k^{(n+1)}, \hat{z}_n), \quad (\text{B.16})$$

where $\hat{U}_0^{(n+1)}$ is an arbitrary vector in \mathbb{R}^I such that $\langle \hat{U}_0^{(n+1)}, 1_I \rangle = 0$, $Y_1^{(n+1)}, \dots, Y_{m+1}^{(n+1)}$ are i.i.d random variables sampled from $\hat{\beta}$, and $(\gamma_n)_{n \geq 0}$ is a positive sequence of real numbers decreasing toward zero satisfying condition (A.5). Finally, our estimate of $\nabla_z W^\varepsilon(\Gamma(\sigma(\hat{z}_n)), b)$ is given by:

$$\hat{\omega}(\hat{z}_n) = (\Gamma J_\sigma(\hat{z}_n))^T \hat{U}_{m+1}^{(n+1)}. \quad (\text{B.17})$$

An approximated gradient step With this stochastic approximation $\hat{\omega}(\hat{z}_n)$ of $\nabla_z F(\hat{z}_n)$, we propose an approximated gradient descent to minimize $F : z \mapsto W^\varepsilon(\Gamma\sigma(z), b)$. Thus, the algorithm considered is given by the recursive procedure

$$\hat{z}_{n+1} = \hat{z}_n - \eta \hat{\omega}(\hat{z}_n). \quad (\text{B.18})$$

Once our minimization procedure is over and we have computed \hat{z} a minimizer of (B.11), we compute an estimate of the class proportions in the target observations. To do so, we set

$$\hat{\pi} = \sigma(\hat{z}). \quad (\text{B.19})$$

Algorithm 1: Solving $\min_{z \in \mathbb{R}^K} \max_{u \in \mathbb{R}^I} \mathbb{E}_{Y \sim \hat{\beta}}[g_{\varepsilon, \alpha}(h)(Y, u, \sigma(z))]$

```

 $z \leftarrow \mathbf{1}_K$ 
for  $l \leftarrow 1$  to  $n_{out}$  do
   $U \leftarrow$  an arbitrary vector in  $\mathbb{R}^I$ 
  for  $k \leftarrow 1$  to  $n_{in}$  do
     $Y \sim \hat{\beta}$ 
     $U \leftarrow U + \gamma_k \nabla_u g_\varepsilon(Y, U, \sigma(z))$ 
  end
  /* Approximation of the gradient of  $z \mapsto W^\varepsilon(\hat{\alpha}(\sigma(z)), \hat{\beta})$  */
   $\hat{\omega}(z) \leftarrow (\Gamma J_\sigma(z))^T U$ 
   $z \leftarrow z - \eta \hat{\omega}(z)$ 
end
/* Computation of the estimator of the class proportion  $\hat{\pi}$  */
return  $\hat{\pi} = \sigma(z)$ 

```

Choice of the parameters for the descent-ascent procedure For this Descent-Ascent procedure, we need to set several parameters. For the step size policy $(\gamma_k)_{k>0}$ of the inner loop, we rely on the recommendation proposed in [B.Bercu and J.Bigot, 2020]. Therefore, we chose $\gamma_k = \gamma/n^c$ where $c = 0.51$ and $\gamma = J\varepsilon/1.9$ with J the cardinal of $\hat{\beta}$. We set the other parameters experimentally. Thus, we chose $n_{out} = 10000$, $n_{in} = 10$, $\eta = 10$ and $\varepsilon = 0.0001$.

B.2 Minmax swapping procedure

An additional regularization term In this section we use the ideas of [M.Ballu et al., 2020] to propose an alternative scheme for solving minimization problem (6). To this end, we slightly modify problem (6) by adding the entropic term:

$$\varphi(h) = \sum_{k=1}^K h_k \log(h_k). \quad (\text{B.20})$$

Thus, our new problem is:

$$\min_{h \in \Sigma_k} W^\varepsilon(\hat{\alpha}(h), \hat{\beta}) + \lambda \varphi(h) = \min_{h \in \Sigma_k} \max_{u \in \mathbb{R}^I} \mathbb{E}[g_{\varepsilon, \alpha}(Y, u, h)] + \lambda \varphi(h), \quad (\text{B.21})$$

where Y is a random variable with distribution $\hat{\beta}$, $g_{\varepsilon, \alpha}$ is defined in (B.10) and λ is a regularization parameter such that $\lambda \geq \varepsilon > 0$.

Swapping the minimum and the maximum By swapping the minimum and the maximum according to Fan's minimax theorem [J.M.Borwein and D.Zhuang, 1986] we get:

$$\begin{aligned}
\min_{h \in \Sigma_k} \max_{u \in \mathbb{R}^I} \mathbb{E}[g_{\varepsilon, \alpha}(Y, u, h)] + \lambda \varphi(h) &= \max_{u \in \mathbb{R}^I} \min_{h \in \Sigma_k} \mathbb{E}[g_\varepsilon(Y, u, h)] + \lambda \varphi(h) \\
&= \max_{u \in \mathbb{R}^I} \min_{h \in \Sigma_k} \sum_{i=1}^I u_i (\Gamma h)_i + A(u) - \varepsilon + \lambda \varphi(h) \\
&= \max_{u \in \mathbb{R}^I} A(u) + \min_{h \in \Sigma_k} \sum_{i=1}^I u_i (\Gamma h)_i + \lambda \varphi(h) - \varepsilon \\
&= \max_{u \in \mathbb{R}^I} A(u) + J_\lambda(u) - \varepsilon,
\end{aligned} \quad (\text{B.22})$$

where $A(u) = \varepsilon \sum_{j=1}^J \left(\log(b_j) - \log \left(\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right) \right) \right) b_j$ and

$$J_\lambda(u) = \min_{h \in \Sigma_k} u_i(\Gamma h)_i + \lambda \varphi(h). \quad (\text{B.23})$$

Arguing e.g. as in the proof of [C.Boyer et al., 2014, Proposition 4.1], the solution $h(u)$ of (B.23) is defined by:

$$\forall k \in \{1, \dots, K\}, (h(u))_k = \frac{\exp \left(-\frac{(\Gamma^T u)_k}{\lambda} \right)}{\sum_{l=1}^K \exp \left(-\frac{(\Gamma^T u)_l}{\lambda} \right)}. \quad (\text{B.24})$$

By plugging (B.24) in (B.22), problem (B.21) boils down to the following maximization problem with respect to $u \in \mathbb{R}^d$:

$$\max_{u \in \mathbb{R}^d} \varepsilon \sum_{j=1}^J \left(\log(b_j) - \log \left(\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right) \right) \right) b_j - \lambda \log \left(\sum_{l=1}^K \exp \left(-\frac{(\Gamma^T u)_l}{\lambda} \right) \right) - \varepsilon. \quad (\text{B.25})$$

An expectation formulation This problem can be rewritten as the maximization of an expectation:

$$\max_{u \in \mathbb{R}^d} \mathbb{E}[f_{\varepsilon, \lambda, \alpha}(Y, u)], \quad (\text{B.26})$$

where Y is a random variable with distribution $\hat{\beta}$, and for $y_j \in \mathbb{R}^d$ an observation of Y , and $u \in \mathbb{R}^d$,

$$f_{\varepsilon, \lambda, \alpha}(y_j, u) = \varepsilon \left(\log(b_j) - \log \left(\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right) \right) \right) - \lambda \log \left(\sum_{l=1}^K \exp \left(-\frac{(\Gamma^T u)_l}{\lambda} \right) \right) - \varepsilon. \quad (\text{B.27})$$

The fact that for all $y \in \mathbb{R}^d$ the function $f_{\varepsilon, \lambda}(y, \cdot)$ is concave and the expectation form (B.26) of problem (B.25) lead us to estimate the optimal vector u^* by the Robbins-Monro algorithm given, for all $n \geq 0$ by

$$\hat{U}_{n+1} = \hat{U}_n + \gamma_{n+1} \nabla_u f_{\varepsilon, \lambda, \alpha}(Y_{n+1}, \hat{U}_n), \quad (\text{B.28})$$

where the initial value \hat{U}_0 is a random vector which can be arbitrarily chosen, Y_1, \dots, Y_{n+1} are i.i.d random variables sampled from $\hat{\beta}$, and $(\gamma_n)_{n \geq 0}$ is a positive sequence of real numbers decreasing toward zero satisfying condition (A.5). Moreover, the gradient of $f_{\varepsilon, \lambda}$ can be easily calculated. Indeed, for $y_j \in \mathbb{R}^d$ an observation of Y , and $u \in \mathbb{R}^d$, $\nabla_u f_{\varepsilon, \lambda}(y_j, u)$ is defined by $\forall i_0 \in \{1, \dots, I\}$,

$$(\nabla_u f_{\varepsilon, \lambda}(y_j, u))_{i_0} = \frac{\sum_{l=1}^K \Gamma_{i_0, l} \exp \left(-\frac{(\Gamma^T u)_l}{\lambda} \right)}{\sum_{k=1}^K \exp \left(-\frac{(\Gamma^T u)_k}{\lambda} \right)} - \frac{\exp \left(\frac{u_{i_0} - c(x_{i_0}, y_j)}{\varepsilon} \right)}{\sum_{i=1}^I \exp \left(\frac{u_i - c(x_i, y_j)}{\varepsilon} \right)}. \quad (\text{B.29})$$

Once the algorithm has converged, and we get \hat{U} , a satisfactory approximation of a maximizer of problem B.26, one can compute an estimate of the class proportions by setting:

$$\hat{\pi} = h(\hat{U}), \quad (\text{B.30})$$

where $h(u)$ is defined in B.24.

Algorithm 2: Solving $\max_{u \in \mathbb{R}^d} \mathbb{E}_{Y \sim \hat{\beta}} [f_{\varepsilon, \lambda}(Y, u)]$

$U \leftarrow 0_I$

for $l \leftarrow 1$ **to** n **do**

$Y \sim \hat{\beta}$
 $U \leftarrow U + \gamma_l \nabla_u f_{\varepsilon, \lambda}(Y, U)$

end

*/** Computation of the estimator of the class proportion \hat{p} **/*

$\hat{p} \leftarrow h(\hat{U})$

return \hat{p}

Choice of the parameters for the min-max swapping procedure For this second procedure we have to set several parameters. To satisfy condition (A.5) for the step size policy $(\gamma_n)_{n>0}$ of the Robbins-Monro algorithm, we chose $\gamma_n = \gamma/n^c$ where $\gamma = 5$ and $c = 0.99$. The other parameters have been set empirically. We chose $\lambda = 0.0001$ and $\varepsilon = 0.0001$. The number of iterations for this stochastic ascent algorithm has been set to $n = 10000$.

B.3 Preprocessing of the cytometry data

Finally, note that when applying our method to estimate the class proportions on real flow cytometry data, a simple pre-processing is required. First, the signal processing of the cytometer can induce some contrived negative values of light intensity. To undo this effect, we merely threshold those few negative values at zero. Second, to settle the parameters of our algorithms, in particular ε , we need to bound the displacement cost. To do so, we scale the data such that: $\forall i \in \{1, \dots, I\}, X_i^s \in [0, 1]^d$ and $\forall j \in \{1, \dots, J\}, X_j^t \in [0, 1]^d$.

B.4 Comparison of the minimization procedures

As said in Subsection 3.5, we assess the performance of our estimator $\hat{\pi}$ by computing the Kullback-Leibler divergence between $\hat{\pi}$ and π the benchmark class proportions. Figure B.1 displays the evolution of the Kullback-Leibler divergence along the iterations of the two minimization procedures. For the descent-ascent procedure, the iterations displayed correspond to the iterations of the outer loop. This comparison has been realized when the source data set is Stanford1A segmented into 10 classes. The target data set is Stanford3A where we try to estimate the proportions of the 10 cell sub-populations.

Both algorithms have been implemented in the Python package available at <https://github.com/sistm/CytOpT-python>. From our numerical experiments, we have found that the sequence of estimates $\hat{\pi}$ produced by the second algorithm described in Section B.2 levels off approximately ten times faster than the sequence of estimates produced with the descent-ascent procedure described in Section B.1. This computational time gain can be accounted by the simple loop complexity of this second algorithm. As in practice both procedures seem to provide close estimates, all the results reported were produced with the second strategy, referred to as the "min-max swapping" strategy.

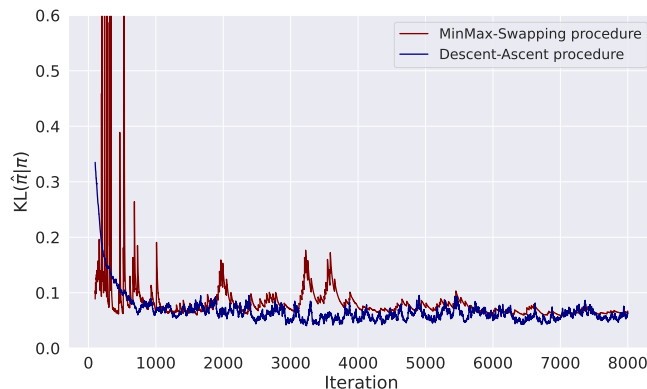


Figure B.1: Evolution of the Kullback-Leibler divergence between the estimated proportions $\hat{\pi}$ and the manual gating benchmark π .

C Additional figures

To assess the relevance of our method, we present in Figure C.2 the evolution of the regularized Wasserstein distance as a function of the weights associated to each class in the source distribution. To this end, we evaluate the function $F : h_1 \mapsto W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$, where $h = (h_1, 1 - h_1)$, on a finite grid $\mathcal{H} = \{h^{(1)}, \dots, h^{(m)}\}$. For $h_1 \in \mathcal{H}$ we approximate $W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$ by the estimator \widehat{W}_n defined in equation (B.25). It can be observed that the regularized Wasserstein distance decreases as the class proportions of the source data set get closer to the class proportions of the target data set.

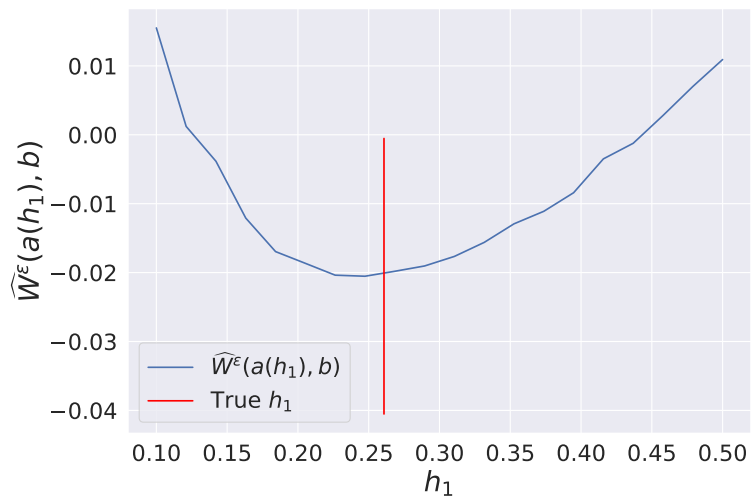


Figure C.2: **Approximation of the function $h_1 \mapsto W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$, for $h = (h_1, 1 - h_1)$, h_1 represents the weight associated to the CD8 cells.** The approximation of $W^\varepsilon(\hat{\alpha}(h), \hat{\beta})$ is produced using the estimator (A.6).

References

- N.Aghaeepour, G.Finak, H.Hoos, T.R.Mosmann, R.Brinkman, R.Gottardo, R.H.Scheuermann, FlowCAP Consortium, Dream Consortium, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228, 2013.
- Y.Saeyns, S.Van Gassen, and B.N.Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449, 2016.
- Nima Aghaeepour, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1):6–13, 2011.
- D.Commenges, C.Alkhassim, R.Gottardo, B.P.Hejblum, and R.Thiébaud. cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry Part A*, 93(11):1132–1140, 2018.
- Y.Ge and S.C.Sealfon. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics*, 28(15):2052–2058, 05 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts300. URL <https://doi.org/10.1093/bioinformatics/bts300>.
- B.P.Hejblum, C.Alkhassim, R.Gottardo, F.Caron, R.Thiébaud, et al. Sequential dirichlet process mixtures of multivariate skew t -distributions for model-based clustering of flow cytometry data. *The Annals of Applied Statistics*, 13(1): 638–660, 2019.
- H.Li, U.Shaham, K.P.Stanton, Y.Yao, R.R.Montgomery, and Y.Kluger. Gating mass cytometry data by deep learning. *Bioinformatics*, 33(21):3423–3430, 07 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx448. URL <https://doi.org/10.1093/bioinformatics/btx448>.
- M.Lux, R.R.Brinkman, C.Chauve, A.Laing, A.Lorenc, L.Abeler-Dörner, and B.Hammer. flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics*, 34(13):2245–2253, 02 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty082. URL <https://doi.org/10.1093/bioinformatics/bty082>.
- E.del Barrio, H.Inouzhe, JM.Loubes, C.Matrán, and A.Mayo-Íscar. optimalflow: Optimal-transport approach to flow cytometry gating and population matching. *arXiv preprint arXiv:1907.08006*, 2019.
- Florian Hahne, Alireza Hadj Khodabakhshi, Ali Bashashati, Chao-Jen Wong, Randy D Gascoyne, Andrew P Weng, Vicky Seyfert-Margolis, Katarzyna Bourcier, Adam Asare, Thomas Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 77(2):121–131, 2010.
- Holden T Maecker and J Philip McCoy. A model for harmonizing flow cytometry in clinical trials. *Nature immunology*, 11(11):975–978, 2010.
- H.Janati, M.Cuturi, and A.Gramfort. Wasserstein regularization for sparse multi-task regression. *arXiv preprint arXiv:1805.07833*, 2018.
- R.Flamary, M.Cuturi, N.Courty, and A.Rakotomamonjy. Wasserstein Discriminant Analysis. *Machine Learning*, 107(12):1923–1945, December 2018. doi: 10.1007/s10994-018-5717-1. URL <https://hal.archives-ouvertes.fr/hal-01377528>.
- M.Arjovsky, S.Chintala, and L.Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- J.Solomon, F.de Goes, G.Peyre, M.Cuturi, A.Butscher, A.Nguyen, T.Du, and L.Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4): 1–11, 2015.
- A.Genevay, M.Cuturi, G.Peyré, and F.Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.
- B.Bercu and J.Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *Annals of Statistics*, To be published, 2020.
- Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.
- Jia Li, Beomseok Seo, and Lin Lin. Optimal transport, mean partition, and uncertainty assessment in cluster analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(5):359–377, 2019.
- V.Brusic R.Gottardo, S.H.Kleinstein, M.M.Davis, D.A.Hafler, H.Quill, A.K.Palucka, G.A.Poland, B.Pulendran, E.L.Reinherz, et al. Computational resources for high-dimensional immune analysis from the human immunology project consortium. *Nature biotechnology*, 32(2):146, 2014.

- Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.
- M.Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.
- G.Peyré, M.Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.
- I.Redko, N.Courty, R.Flamary, and D.Tuia. Optimal transport for multi-source domain adaptation under target shift. *arXiv preprint arXiv:1803.04899*, 2018.
- M.Ballu, Q.Berthet, and F.Bach. Stochastic optimization for regularized wasserstein estimators. *arXiv preprint arXiv:2002.08695*, 2020.
- JM.Bland and DG.Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.
- James W Tung, Kartoosh Heydari, Rabin Tirouvanziam, Bitu Sahaf, David R Parks, Leonard A Herzenberg, and Leonore A Herzenberg. Modern flow cytometry: a practical approach. *Clinics in laboratory medicine*, 27(3): 453–468, 2007.
- B.Dost, C.Wu, A.Su, and V.Bafna. Tclust: A fast method for clustering genome-scale expression data. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(3):808–818, 2010.
- T.Hastie, R.Tibshirani, and J.Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- Gabriella Henel and John L Schmitz. Basic theory and clinical applications of flow cytometry. *Laboratory Medicine*, 38(7):428–436, 2007.
- J.Feydy, T.Séjourné, FX.Vialard, SI.Amari, A.Trouvé, and G.Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.
- Kevin Ushey, JJ Allaire, and Yuan Tang. *reticulate: Interface to 'Python'*, 2022. URL <https://CRAN.R-project.org/package=reticulate>. R package version 1.23.
- H.Robbins and S.Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- J.M.Borwein and D.Zhuang. On fan’s minimax theorem. *Mathematical programming*, 34(2):232–234, 1986.
- C.Boyer, P.Weiss, and J.Bigot. An algorithm for variable density sampling with block-constrained acquisition. *SIAM Journal on Imaging Sciences*, 7(2):1080–1107, 2014.