Chapter 6
Advances in Cytometry Gating Based on Statistical Distances and Dissimilarities

Hristo Inouzhe

Abstract In this chapter, we overview some recent and relevant applications of discrepancy measures (distances and dissimilarities) between statistical objects (such as random variables, probability distributions, samples) in the field of cytometry gating. Cytometry gating identifies cell populations in cytometry datasets, i.e., finds groups in multidimensional measurements of (hundreds of) thousands of single cells. From a statistical perspective, cytometry gating is a classification problem, and hence the applicable methods can be unsupervised, supervised, or semi-supervised. Since substantial amounts of variability are unavoidable in biological data, crucial tasks to help classification are establishing similarity between entire (or parts of) cytometry datasets and finding transformations between datasets that are optimal in some sense. A powerful approach to establish similarity between cytometry datasets is to model them as statistical objects and to use some distance or dissimilarity such as the Wasserstein distance, maximum mean discrepancy, and some f-divergence such as Kullback–Leibler or Hellinger or some statistic such as Friedman–Rafsky. We briefly overview the previous discrepancy measures and present how they are (or can be) used for grouping cytometry datasets, for producing templates from a group of datasets, or for interpolation between datasets. We provide instructive examples and further sources of information. The code for generating all figures is freely available at https://github.com/HristoInouzhe/Gating-with-Statistical-Distances.

Keywords Flow cytometry gating · Optimal transport · Statistical distances

6.1 Introduction

Cytometry is concerned with the measurement (“metry”) of physical, chemical, and other properties of a cell (“cyto”) and, therefore, offers a characterization

H. Inouzhe (✉)
Basque Center for Applied Mathematics, Bilbao, Spain
e-mail: hinouzhe@bcamath.org

Y. Larriba (ed.), *Statistical Methods at the Forefront of Biomedical Advances*,
https://doi.org/10.1007/978-3-031-32729-2_6

of biological samples at the single-cell level. Cytometry has diverse and relevant applications in clinical and research immunology and oncology, for example, for diagnosing a range of hematologic (blood) cancers and diseases such as AIDS. Two prevalent techniques for characterizing single cells are flow cytometry (FC) and time of flight (mass) cytometry (CyTOF). In FC, cells tagged with fluorescent antibodies are exposed to lasers with different wavelengths, and the resulting light spectrum is measured and used to characterize a cell. An extensive and comprehensible description of flow cytometry and its applications can be found in [1]. In CyTOF, cells tagged with heavy metal isotope-coupled antibodies go through a mass spectrometer, and the resulting mass spectrum is used for the characterization of a cell. An up-to-date review on CyTOF and its applications can be found in [2]. Currently, the number of different antibodies, also known as markers, used to characterize a cell with CyTOF can get as high as 100, while for FC it can be close to 50. The number of cells that can be characterized in a single sample goes from around hundred thousand to more than several millions. Hence, cytometry data, obtained after appropriate measurement and preprocessing, belong to high-dimensional spaces with big sample sizes. Throughout this chapter, we will use cytometry data, cytometric data, cytometric datasets, and cytometries interchangeably.

A crucial task in cytometry data analysis is to identify different cell populations, which amounts to discovering groups of cells that display some significant differences in one or a group of the measured markers. A reference for standardized cell types is [3]. This allows a variety of applications, for example, the cell types and their relative proportions, identified in each cytometry dataset corresponding to a different blood sample in a study, can be used to characterize an immune system reaction or an illness. The standard way of identifying cell populations is called manual gating. Examples of manual gating can be found in Figure 1 in [4] and Figure 169 in [1]. It consists of an expert selecting a pair of markers, then, in the corresponding bi-dimensional projection the expert selects a region where cells inside it are further inspected. That is, a new pair of markers is selected, the cells inside the previous region are represented in the corresponding bi-dimensional plot, and the expert selects a new region for further inspection. This continues until the cells inside the region of interest are considered to be of the same type. Hence, a single cell type is obtained by defining a sequence (hierarchy) of pairs of markers with corresponding regions of interest (also known as gates). Different cell types are defined by a different sequence of pairs of markers and the corresponding gates. This algorithmic procedure allows to identify the cell populations of interest or to discover new ones in a cytometry dataset.

Manual gating has been extremely successful, but it presents several drawbacks [1, 4, 5]. Firstly, it is subjective and time-consuming. The selected hierarchy of pairs of markers and the corresponding gates depend on the knowledge and dedication of the expert. Hence, reproducibility of results between different experts may be low. The bi-dimensional inspection can be very time-consuming when the number of markers (the dimension of the space) is high (for example, around 30). In consequence, the time required for an expert or a small group of experts to annotate (label, gate) hundreds of high-dimensional cytometry datasets is a major bottleneck for modern studies. Secondly, high-dimensional information is lost, due to the sequential exploration of bi-dimensional projections, which makes it impossible to find and use intricate correlations between multiple markers as a criterion for defining a cell type. To address some of these limitations, automated or semi-automated gating has been introduced using powerful tools based on the interplay of statistics and computer science commonly known as Machine Learning (ML) (see [5, 6]).

The main applications of ML to cytometry gating originate from the tools developed for classification tasks and can be divided into three broad categories: unsupervised, supervised, and semi-supervised. Unsupervised techniques try to extract structure from the raw data without requiring knowledge of any ground truth. The primary tool to consider is cluster analysis or clustering, where data are divided into groups (clusters) where elements in the same group are more similar to each other, in some predefined way, than members of different groups. Clustering algorithms can be split into partitioning and agglomerative ones. Partitioning algorithms try to divide the data into a number of clusters such that an optimality criterion is fulfilled, where k-means is the best known and most widely used. Agglomerative algorithms start with single observations and merge them into clusters according to some dissimilarity criteria; hierarchical clustering is probably the most popular example. For readers interested in the topic, a good source is [7]. Clustering is applied to cytometry data as a way of discovering cell populations in high-dimensional spaces in an automatic, time-efficient, unbiased, and reproducible manner. Typically, after clustering a cytometry dataset, clusters are assigned, by an expert or another algorithm, to previously known cell populations or are considered as a new cell population.

When using supervised learning, the approach to cytometry gating is fundamentally different to the unsupervised case. The task is to automatically learn a gating strategy from previously manually or otherwise gated cytometry datasets to gate a new ungated one. In this case, contrary to the unsupervised setting, the algorithm directly assigns each cell in an unlabeled cytometry dataset to a specific cell population. The available tools are many, and we highlight quadratic discriminant analysis, tree-based methods as random forest or any approach based on neural networks. It is out of the scope of this chapter to present in detail the many tools of supervised learning, so we refer the interested reader to [8, 9]. The main point of supervised learning applied to cytometry data is to use high-quality historical data, i.e., previously expertly manually gated cytometry datasets, to gate a new ungated cytometry dataset in a time-efficient way which uses intricate and high-dimensional correlations between markers. The clinical setting, where well-established protocols lead to good historical data, is especially well suited for supervised methods.

Semi-supervised learning can be considered a mixture between the previous settings where an unsupervised method requires some input from a human or a previous example for the gating task. This is a fairly common paradigm in cytometry gating since it offers a good trade-off between time efficiency and previously available or expert information. Examples of such applications can be found in [10, 11].

One of the main challenges with automatic or semi-automatic gating is the huge variability present in cytometric data, which ensues from a diversity of sources (see [12]). There is a natural biological variability, for example, the cytometry data from blood samples of the same individual in the same conditions measured on the same flow cytometer may present non-negligible differences. A technical source of variability, commonly referred to as batch effect, corresponds to measurements in different conditions (different days, locations, temperature, pressure, etc.), with different machines, with the same machine but different settings, with different staining antibodies, and so on. Another prominent source of variability comes from experts having different criteria for gating, i.e., different sequence of pairs of markers and gates, and the different level of completeness in a gated cytometry dataset, and it is common to gate only some cell populations of interest leaving the rest ungated. Therefore, any automatic gating method that deals with cytometry data from a variety of measurements must be robust to the previous types of variability while also correctly detecting meaningful variability coming from the natural response of the immune system, a cell population characteristic of a disease, a vaccine effect, and many others.

To address the previous difficulties and aid the automatic gating workflow, a successful strategy has been to quantify variability of cytometric data (see, for example, [11, 13, 14]). Such variability quantification has been based on mathematical tools that measure the difference between raw or gated cytometry datasets. In essence, when the signal of interest is stronger than natural biological variability or batch effects, one expects higher values for the measure of variability. Hence, with an appropriate measure of difference or similarity between cytometric data, one can detect meaningful and meaningless variability, and this can guide the learning strategy for automated gating. Additionally, the possibility of establishing similarity between cytometry datasets allows for matching and alignment which are of common use in gating workflows. Matching refers to the problem of how to optimally assign a group of cytometry datasets to another group of cytometry datasets, while alignment (interpolation) refers to the problem of transforming, in some predefined way, one cytometry dataset into another cytometry dataset.

The aim of this chapter is to introduce the reader to some of the main aspects of the cytometry gating workflow where statistical dissimilarities and distances, that is, measures of discrepancy (difference) between statistical objects, are used. Our goal is to provide some basic notions, while the interested reader can find much more details in the references. We presuppose that the reader has basic notions of probability, if it is not the case we refer to the chapters on random variables, integration (expectation), and joint distributions in the introductory books [15, 16]. The first section of this chapter is dedicated to introducing the mathematical modelling of cytometric data and to presenting some popular statistical discrepancy measures used in automated flow cytometry gating. In the second section, we present how statistical measures of dissimilarity can be used in the gating workflow, particularly: for grouping cytometric datasets, for producing template cytometry data, and for interpolation between cytometries. We finish this chapter with some brief concluding remarks.

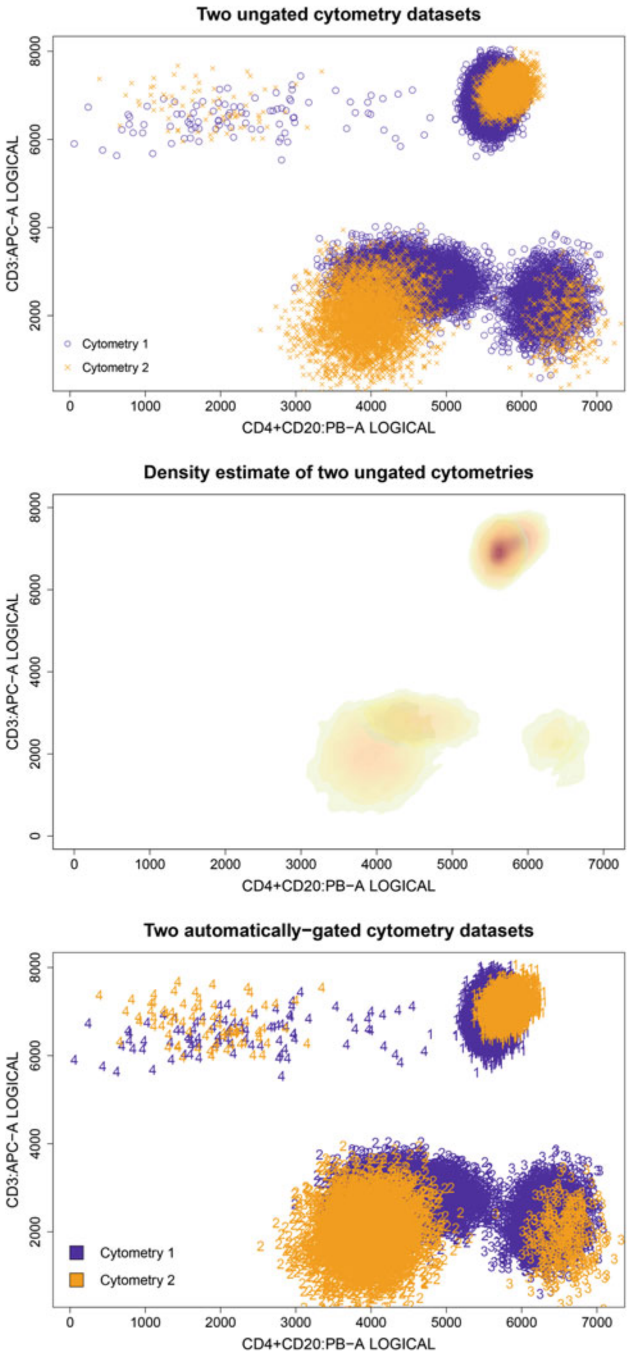
6.2 Dissimilarities and Distances

In this section, the mathematical formalism for dealing with cytometric data is provided. Firstly, several useful ways of describing a cytometry dataset are presented. Secondly, we provide definitions for the notions of dissimilarity and distance between cytometric data. Finally, some of the most popular dissimilarities and distances used in gating are overviewed.

A raw cytometry dataset X can be viewed as a collection of single-cell measurements $X = \{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^m$ or equivalently as a matrix $X \in \mathbb{R}^{N\times m}$, with N the number of cells in the measured sample and m the used markers. An example of two raw cytometries for two markers can be seen in the top of Fig. 6.1. A cytometric dataset can be viewed as an empirical probability distribution

$$\eta = \sum_{i=1}^{N} \frac{1}{N}\delta_{x_i}, \tag{6.1}$$

i.e., a probability distribution giving weights $1/N$ to each $x_i \in X$, or alternatively, as some probability distribution η_X estimated from the raw sample X. As was noted in Sect. 6.1, N can be in the millions and m close to hundred, and therefore cytometry data can be considered high-dimensional and sample sizes are not particularly small. A gated cytometry is a cytometry dataset with labels for each cell, i.e., $\tilde{X} = \{(x_i, l_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^m$ and $l_i \in \mathcal{L}$, where $\mathcal{L}$ is some finite set of labels, usually the names of the cell populations from a manual or supervised gating or the label of the cluster from an unsupervised gating. An example of two gated cytometries can be seen in the bottom of Fig. 6.1, where the set of labels is $\mathcal{L} = \{1, 2, 3, 4\}$. A gated cytometry can also be viewed as a collection of probability distributions and some associated weights. Let us say that the number of cell populations or the number of clusters in the gated cytometry $\tilde{X}$ is K; equivalently, we can write $|\mathcal{L}| = K$. Then, one can take $\tilde{X} = \{(\mu_k, w_k)\}_{k=1}^K$ with weights $w_k > 0$ such that $\sum_{k=1}^K w_k = 1$ and with μ_k representing some probability distribution. Usually, μ_k is obtained from a model-based or a non-parametric fit to the cells belonging to the k-th cell population or cluster, while w_k is the relative frequency of the cells in the k-th group with respect to the total amount of cells. For instance, one can fit a multivariate normal distribution to each group in the bottom of Fig. 6.1, and the collection of normal distributions and the relative frequencies of the points in the clusters are a


Two ungated cytometry datasets

Density estimate of two ungated cytometries

Two automatically−gated cytometry datasets

Fig. 6.1 *Top:* Two ungated cytometries $X_1 = \{x_{1,i}\}_{i=1}^{43156}$ and $X_2 = \{x_{2,i}\}_{i=1}^{10660}$ with $X_1, X_2 \subset \mathbb{R}^2$. *Middle:* Non-parametric density estimation with kernel smoothing of X_1 and X_2. *Bottom:* Unsupervised gating of X_1 and X_2, which we denote $\tilde{X}_1$ and $\tilde{X}_2$, with a clustering method called tclust (see [17]) looking for four different cell types

good representation of the cytometries at hand. Notice that the different ways of modelling cytometry data can be suitable for different purposes. Equipped with a formal definition of cytometry data, we can provide tools for comparing different datasets.

A dissimilarity (discrepancy) between two cytometry datasets X and Y is a measure of how different the two objects are, with values close to zero if the two datasets are similar and high values if they are very different. Exactly the same concept can be applied to the gated versions $\tilde{X}, \tilde{Y}$ which we omit for simplicity of exposition. Formally, a dissimilarity d is a symmetric divergence, and hence it fulfils:

1. $d(X, Y) \geq 0$ for any two cytometry datasets X and Y and $d(X, Y) = 0$ if and only if the datasets are the same, $X = Y$ (definition of divergence).
2. $d(X, Y) = d(Y, X)$ for any X, Y (symmetry).

A distance (or metric) between cytometry datasets, d, is a dissimilarity with an additional property:

3. $d(X, Z) \leq d(X, Y) + d(Y, Z)$ for any cytometry datasets X, Y, Z (triangle inequality).

A question that arises naturally is what formalism to use for cytometry data and what is an appropriate dissimilarity or distance for different contexts. These questions are not independent, and in the next sections, we briefly present some of the most useful dissimilarities and their applications. In Table 6.1, one can find a summary with some general information about the discrepancy measures introduced in the next sections.

6.2.1 *Wasserstein Distance*

Wasserstein distance, sometimes referred to as the Earth mover's distance, is a distance between probability distributions which is well suited for cytometric data since it is robust against small translations and small changes in probability [18]. Even more, it can handle distributions with non-overlapping supports, which is a typical situation in cytometry datasets. Broadly speaking, the Wasserstein distance measures the cost of optimally transporting one distribution into the other. The following are good references for the theory of optimal transport [19] and for the theory and computation [20].

Definition 1 (Optimal Transport Cost) Let X and Y be random variables on the spaces $\mathcal{X}$ and $\mathcal{Y}$ where $\text{law}(X) = \mu$ $(X \sim \mu)$ and $\text{law}(Y) = \nu$ $(Y \sim \nu)$. Let $\pi(\mathcal{X} \times \mathcal{Y})$ be the space of all joint probability measures on the product space $\mathcal{X} \times \mathcal{Y}$ with first marginal μ and second marginal ν, i.e., such that for any joint probability $\pi \in \pi(\mathcal{X} \times \mathcal{Y})$, $\int_{\mathcal{Y}} d\pi(x, y) = \mu$ and $\int_{\mathcal{X}} d\pi(x, y) = \nu$. Finally, let $c(x, y)$ be a cost function representing the cost of transporting a unit of probability mass from

Table 6.1 Summary of the statistical measures of discrepancy between probability distributions presented in this chapter. *Type* indicates if one is dealing with a distance or a dissimilarity as defined in the beginning of Sect. 6.2. *Distributions* indicates if both discrete and continuous probability distributions can be handled. *Density estimation* refers to the need of density estimation to compute the respective measure of discrepancy in the cytometry gating setting. The fourth column presents if barycenters (also known as Frechet means) can be computed in an efficient way. The last column presents some freely available software

	Type	Distributions	Density estimation	Efficient barycenter computation	Software
Wasserstein	Distance	Discrete, continuous	No	Yes	R: transport
					Python: POT
Maximum mean discrepancy	Distance	Discrete, continuous	No	Yes	R: kernlab
					Python: Easy to implement
Symmetric Kullback-Leibler	Dissimilarity	Discrete, continuous	Yes	No	Easy to implement
Hellinger	Distance	Discrete, continuous	Yes	No	Easy to implement
Friedman-Rafsky	Dissimilarity	Discrete	No	No	Python: PyTorch
					R (Bioconductor): flowMap

x to y. The optimal transport (OT) cost is defined as the solution of the following optimal transport problem:

$$\begin{aligned}\mathrm{OT}_c(\mu, \nu) &= \min_{\pi \in \pi(\mathcal{X} \times \mathcal{Y})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \\ &= \min_{(X,Y)} \{E_{(X,Y)} c(X, Y) : \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\}. \end{aligned} \tag{6.2}$$

Notice that for discrete measures, that is, when μ and ν give masses to finite collections of points $x_i, \dots, x_n$ and $y_1, \dots, y_{n'}$, respectively, the OT cost (6.2) becomes

$$\mathrm{OT}_c(\mu, \nu) = \min_{\pi \in \pi(\mathcal{X} \times \mathcal{Y})} \sum_{i=1}^{n} \sum_{j=1}^{n'} \pi_{i,j} c(x_i, y_j), \tag{6.3}$$

where $\pi(\mathcal{X} \times \mathcal{Y})$ is the set of $n \times n'$-matrices with row sums equal to $(\mu(x_1), \dots, \mu(x_n))$ and column sums equal to $(\nu(y_1), \dots, \nu(y_{n'}))$. Let us stress that the discrete OT problem can be viewed as a soft assignment problem, where the mass of an origin point x_i is assigned in different proportions to the corresponding y_j points. This is in contrast with a hard assignment where origin points are assigned exclusively to one destination point. When $n = n'$, the OT problem is equivalent to a hard assignment problem, i.e., a bijection between the sets $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$.

The p-Wasserstein distance is an Optimal Transport Cost where the cost function is a p-power of the usual Euclidean distance.

Definition 2 (p-Wasserstein Distance) Let $\mathcal{X} = \mathcal{Y} \subseteq \mathbb{R}^m$, EX^p, $EY^p < 0$ (finite p-moments), and $c(x, y) = \|x - y\|^p$, with $\| \cdot \|$ representing the usual Euclidean distance. Then, the p-Wasserstein distance is defined as

$$\begin{aligned} d_{W_p}(\mu, \nu) &= \left(\min_{\pi \in \pi(\mathcal{X} \times \mathcal{X})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p d\pi(x, y) \right)^{1/p} \\ &= \left(\min_{(X,Y)} \{E_{(X,Y)} \|X - Y\|^p : \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\} \right)^{1/p}. \end{aligned} \tag{6.4}$$

Associated with $d_W(\mu, \nu)$, under some mild conditions, there is a (not-necessarily unique) optimal coupling T, which encodes an optimal way of transforming μ into ν. For example, when both probabilities are discrete, T is a matching indicating how much mass to send from each origin point to its corresponding matched points in the destination. A particularly important case is the 2-Wasserstein distance, where, when at least one of the probability distributions has a density, there is a unique optimal map T such that

$$d_{W_2}^2(\mu, \nu) = \int_{\mathcal{X}} \|x - T(x)\|^2 \, d\mu(x).$$

Intuitively, a coupling T provides a mean to do interpolation (alignment) between probability distributions, and this interpolation is unique when at least one of the probabilities has a density and T is obtained through the 2-Wasserstein distance.

A very attractive property of the Wasserstein distance is the fact that it allows to produce a "sort of average," which respects geometrical properties of the underlying data. In applications to gating, this means that it is possible to obtain a template from a group of cytometry datasets. The tool we refer to is the Wasserstein barycenter.

Definition 3 (Wasserstein Barycenter) Let $\mu_1, \ldots, \mu_n$ be a set of probability distributions belonging to $\mathcal{P}_p(\mathbb{R}^m)$, the set of probability distributions with finite p-moment. Let $\{w_i\}_{i=1}^n$ be weights such that $\sum_{i=1}^n w_i = 1$. Then, the p-Wasserstein barycenter is the measure μ^* that solves the following optimization problem:

$$\mu^*_{W_p} = \text{argmin}_{\mu \in \mathcal{P}_p(\mathbb{R}^m)} \sum_{i=1}^n w_i d^p_{W_p}(\mu, \mu_i). \tag{6.5}$$

In the case of the 2-Wasserstein barycenter, if one of the probability distributions μ_i has a density, then the barycenter is unique. Furthermore, if all μ_i have densities, there are unique optimal interpolations between each of them and the barycenter.

For practical purposes, it is essential to be able to compute Wasserstein distances and barycenters efficiently. The task is relatively simple when dealing with location-scatter families such as the (multivariate) normal distribution, but for more general distributions there are only approximate algorithms. Furthermore, to improve efficiency, it is common to use some (entropically) regularized versions of the Wasserstein distance which are only approximations of it. For details on computation, we refer to [20]. However, this allows many real-world applications and in particular its application to different stages of cytometry gating workflows.

6.2.2 *Maximum Mean Discrepancy*

Maximum mean discrepancy (MMD) is a popular measure of difference between probability distributions. It is of great interest since it allows the use of many standard ML algorithms such as Support Vector Machines and Gaussian Process Regression with probability distributions. For the interested reader, extensive details can be found in [21].

Definition 4 (Maximum Mean Discrepancy) Let μ and ν be probability distributions on $\mathcal{X}$, and let $\mathcal{F}$ be a class of real-valued functions defined on $\mathcal{X}$. The MMD with respect to $\mathcal{F}$ is defined as

$$\begin{aligned} \text{MMD}(\mu, \nu, \mathcal{F}) &= \sup_{f \in \mathcal{F}} \left(\int_{\mathcal{X}} f(x) d\mu(x) - \int_{\mathcal{X}} f(x) d\nu(x) \right) \\ &= \sup_{f \in \mathcal{F}} \left(E_{X\sim\mu} f(X) - E_{X\sim\nu} f(X) \right). \end{aligned} \tag{6.6}$$

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with kernel κ, i.e., $\mathcal{H} = \text{closure}(\text{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\})$ and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive definite function. Then (6.6) becomes

$$\begin{aligned}\text{MMD}^2(\mu, \nu, \mathcal{H}) &= \int_{\mathcal{X}\times\mathcal{X}} \kappa(x, x')d\mu(x)d\mu(x') - 2\int_{\mathcal{X}\times\mathcal{X}} \kappa(x, y)d\mu(x)d\nu(y)\\ &\quad + \int_{\mathcal{X}\times\mathcal{X}} \kappa(y, y')d\nu(y)d\nu(y')\\ &= E_{X,X'\sim\mu}\kappa(X, X') - 2E_{X\sim\mu,Y\sim\nu}\kappa(X, Y) + E_{Y,Y'\sim\nu}\kappa(Y, Y').\end{aligned} \tag{6.7}$$

If κ is a universal kernel, that is, the corresponding RKHS $\mathcal{H}$ is dense in the space of continuous and bounded functions in $\mathcal{X}$, then

$$\text{MMD}(\mu, \nu, \mathcal{H}) = 0 \quad \text{if and only if} \quad \mu = \nu. \tag{6.8}$$

Therefore, in that setting, the MMD is a distance, and we will denote it

$$d_{\text{MMD},\mathcal{H}}(\mu, \nu) = \text{MMD}(\mu, \nu, \mathcal{H}). \tag{6.9}$$

A well-known universal kernel in $\mathbb{R}^m$ is the Gaussian kernel with parameter λ, given by

$$\kappa(x, y) = \exp\left(\frac{-\|\mathrm{x} - \mathrm{y}\|^2}{2\lambda^2}\right). \tag{6.10}$$

We stress that there are other universal kernels in $\mathbb{R}^m$ and that the choice of kernel is non-trivial and it is problem specific.

There is an equivalent barycenter problem for MMD for a set of measures $\{\mu_i\}_{i=1}^n$, which can be written as

$$\mu^*_{\text{MMD},\mathcal{H}} = \text{argmin}_{\mu\in\mathcal{P}(\mathbb{R}^m)} \sum_{i=1}^n w_i d^2_{\text{MMD},\mathcal{H}}(\mu, \mu_i), \tag{6.11}$$

where the barycenter has the following expression (see [22]):

$$\mu^*_{\text{MMD},\mathcal{H}} = \sum_{i=1}^n w_i \mu_i. \tag{6.12}$$

In the setting of MMDs, optimal interpolation between measures, i.e., following geodesics, is straightforward since a geodesic between μ and ν is given by $\mu_t = (1-t)\mu+t\nu$ for $0 \leq t \leq 1$. It is worth mentioning that barycenters coming from Eq. (6.5) with $p > 1$, particularly the 2-Wasserstein distance, have significantly different geometrical properties than barycenters coming from (6.11). A clear example can be seen in the top of Fig. 6.3.

Computations of $d_{\text{MMD},H}$ based on (6.7) are fairly straightforward since for samples $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_i\}_{i=1}^{n'}$ the following is an unbiased estimator:

$$d^2_{\text{MMD},\mathcal{H},n} = \frac{1}{n(n-1)} \sum_{x \neq x' \in X} k(x, x') - \frac{2}{nn'} \sum_{x \in X, y \in Y} k(x, y) + \frac{1}{n'(n'-1)} \sum_{y \neq y' \in Y} k(y, y').$$

The computation of the barycenter in the setting of MMD can be found in [22].

6.2.3 Kullback–Leibler Divergence

The Kullback–Leibler (KL) divergence, which fulfils only the first criteria of a dissimilarity and is also called relative entropy, is a very popular way of measuring the difference between two probability distributions. The KL divergence is probably the most notorious representative of a family of difference measures between probability distributions known as f-divergences. More information about the topic can be found in Section 8.1 in [20].

Definition 5 (Kullback–Leibler Divergence) For discrete distributions μ and ν defined on the same space $\mathcal{X}$, the Kullback–Leibler divergence is defined as

$$\text{KL}(\mu, \nu) = \sum_{x \in \mathcal{X}} \mu(x) \log \frac{\mu(x)}{\nu(x)}. \tag{6.13}$$

When μ and ν have densities f_μ and f_ν and are defined on $\mathcal{X} \subseteq \mathbb{R}^m$, the Kullback–Leibler divergence takes the form

$$\text{KL}(\mu, \nu) = \int_{\mathcal{X}} f_\mu(x) \log \frac{f_\mu(x)}{f_\nu(x)} dx. \tag{6.14}$$

It is important to notice that if μ assigns probability to a set (region, points) where ν does not assign any probability, $\text{KL}(\mu, \nu) = \infty$, and if it is the other way around, then $\text{KL}(\mu, \nu) = -\infty$. Therefore, the KL divergence is better suited for absolutely continuous measures, that is, for measures that assign zero probability to the same sets. One possible way to obtain a dissimilarity from the KL divergence is the Symmetric KL divergence defined as

$$d_{KL}(\mu, \nu) = \text{KL}(\mu, \nu) + \text{KL}(\nu, \mu). \tag{6.15}$$

The computation of the KL divergence in the continuous case requires an estimation of the densities and then a numerical computation of the integrals which makes it very sensible to the curse of dimensionality. Additionally, when absolute continuity is not strictly fulfilled, one may need to allow some tolerance for the KL divergence to return meaningful results.

6.2.4 Hellinger Distance

The Hellinger distance is another popular measure of the difference between probability distributions. As the KL divergence it also belongs to the family of f-divergences. It presents some desirable properties, for example, it can be easily used to define kernel functions (see [23]), something that is not the case with the Wasserstein distance or the KL divergence. It is also more computationally amenable than the KL divergence.

Definition 6 (Hellinger Distance) For discrete distributions μ and ν defined on the same space $\mathcal{X}$, the Hellinger distance is defined as

$$d_H^2(\mu, \nu) = \frac{1}{2} \sum_{x \in \mathcal{X}} \left(\sqrt{\mu(x)} - \sqrt{\nu(x)} \right)^2 . \tag{6.16}$$

For μ and ν with densities f_μ, f_ν defined on $\mathcal{X} \subseteq \mathbb{R}^m$ the Hellinger distance takes the form

$$d_H^2(\mu, \nu) = \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{f_\mu(x)} - \sqrt{f_\nu(x)} \right)^2 dx. \tag{6.17}$$

Notice that in the continuous case (6.17), the Hellinger distance is the L^2 distance between the square roots of the density functions.

As for the KL divergence, when applied to the cytometry gating setting, the computation of the Hellinger distance requires estimating densities and then numerically computing an integral. Hence, it presents similar difficulties as the KL divergence case, although it handles better probabilities that assign zero measure to different sets.

6.2.5 Friedman–Rafsky Statistic

The Friedman–Rafsky (FR) statistic [24] was conceived as a statistic for testing if two multivariate samples came from the same distribution. The basic concept is the following, if the two samples come from the same distribution, they should be well mixed in the space of markers $\mathbb{R}^m$. The technically difficult aspect is how to measure the “mixedness” in space.

Broadly speaking, to measure how well mixed are two samples $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_{n'}\}$ with $x_i, y_j \in \mathbb{R}^m$, one creates a complete graph considering $\{z_k\}_{k=1}^{n+n'} = \{x_1, \dots, x_n, y_1, \dots, y_n\}$ as the vertices and the respective Euclidean distance $\|z_k - z_{k'}\|$ as the weight for the edge connecting $z_k, z_{k'}$. From the complete graph, one extracts the minimum spanning tree (MST), which is a subgraph of the complete graph that connects all vertices, without cycles and with the minimum total edge weight. Once the MST is obtained, all its edges connecting points from the two different samples are removed. The number of remaining subgraphs, r, is an indication of how well mixed the data are. An insightful example of this procedure can be found in [25]. For example, $r = 2$ means that there was only one edge in the MST connecting the two samples, and this can be interpreted as the samples being not well mixed. If the value of r is high, many edges in the MST were connecting points from the different samples and this can be interpreted as well mixedness. The FR statistic compares r to its expected value and normalizes with the standard variance. The formal definition is the following.

Definition 7 (Friedman–Rafsky Statistic) In the setting and notation of the previous paragraph, let us define $N = n + n'$, and

$$\mathfrak{m} = \frac{2nn'}{N} + 1,$$

$$\sigma^2 = \frac{2nn'}{N-1}\left(\frac{2nn' - N}{N} + \frac{(\mathfrak{c} - N + 2)(n' + N(N-1) - 4nn' + 2)}{(N-2)(N-3)}\right),$$

where $\mathfrak{c}$ is the total number of edge pairs sharing common nodes in the MST. Then, the Friedman–Rafsky statistic is defined as

$$\mathrm{FR}(X, Y) = \frac{r - \mathfrak{m}}{\sigma}. \tag{6.18}$$

From the FR statistic (6.18), one can define a dissimilarity between two cytometries X and Y, or the associated empirical distributions, as

$$d_{FR}(X, Y) = |\mathrm{FR}(X, Y)|\,. \tag{6.19}$$

The main computational challenge for obtaining the FR statistic is the computation of the MST. This can be done with standard tools in popular libraries such as *igraph* in R and *scipy* in Python.

6.3 Applications to the Gating Workflow

In this section, we present some fundamental applications of statistical distances and dissimilarities to the different gating workflows. In such workflows, the objective is to gate big amounts of cytometric data with no or minimum amount of expert intervention. As is expected, this is where automatic gating is the most useful. Since distances and dissimilarities allow to compare cytometric data, their main applications are the following:

- Group cytometry datasets into homogeneous groups, which reduces variability [11, 14, 25]
- Produce templates, through barycenters and other techniques, that can summarize the information in a group of cytometric datasets [13, 14, 26]
- Interpolate between cytometric datasets, allowing for gates in one cytometry dataset to be transferred to another or allowing for mitigation of batch effects [11, 26–29]

6.3.1 Grouping Cytometric Datasets

The idea behind grouping cytometry data is straightforward, since variability is so high it is useful to form groups of cytometric data where variability is lower and then work on these more homogeneous groups of datasets. Therefore, the objective is to do clustering on cytometric datasets. A simple procedure is the following, for a set of cytometry datasets $\{X_1, \ldots, X_n\}$, gated or ungated, choose a distance or dissimilarity d and produce a distances (dissimilarities) matrix D_d such that $[D_d]_{ij} = d(X_i, X_j)$. Then, one can use D_d for hierarchical clustering, although other clustering options are also possible, and obtain a partition of the cytometry datasets. The main difficulty here is to produce a distance matrix D_d according to how cytometry data are modelled.

6.3.1.1 Ungated Cytometry Datasets

The first case to consider is when data are samples, i.e., $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\} \subset \mathbb{R}^m$ for $1 \leq i \leq n$, with an associated empirical probability distribution to each X_i. This means that suitable candidates for a measure of similarity are the Wasserstein distance (6.4), the maximum mean discrepancy distance (6.7), and the dissimilarity based on the Friedman–Rafsky statistic (6.19). For example, one can consider the cytometry datasets in the top of Fig. 6.1 and compute the respective distance. Depending on the sample size of the cytometries involved, to lower the computational and memory cost of the distance calculation for d_{W_p}, (6.4), and $d_{\mathrm{MMD},\mathcal{H}}$, (6.9), multivariate adaptative extensions of histograms can be used (see, for example, [30, 31] and the references therein). Hence, a dataset X_i is approximated as $\hat{X}_i = \{(c_j, p_j)\}_{j=1}^{n'_i}$, where $c_j \in \mathbb{R}^m$ is a centroid corresponding to the points in the hyperbox j, p_j is the relative weight of the points in the same box, and $n'_i << n_i$ is the number of hyperboxes. Therefore, the original D_d is approximated by $[\hat{D}_d]_{ij} = d(\hat{X}_i, \hat{X}_j)$. Notice that non-parametric multivariate density estimation when the number of markers m is high is quite hard, and therefore using directly the symmetric KL divergence (6.15) or the Hellinger distance (6.17) is not a good idea in this situation. However, for some small number of markers or for some projections into low-dimensional subspaces, one can use those distances. With the density estimations of the middle of Fig. 6.1, one can compute the Hellinger distance or the symmetric KL divergence between the two cytometries X_1 and X_2.

6.3.1.2 Gated Cytometry Datasets

A second important case is when (manually, with supervised or unsupervised methods) gated cytometry data are available, i.e., one has cell measurements and their labels for different samples. Hence, the setup is $\{\tilde{X}_1, \dots, \tilde{X}_n\}$ where $\tilde{X}_i = \{(x_{i,1}, l_{i,1}), \dots, (x_{i,n_i}, l_{i,n_i})\}$ with $x_{i,j} \in \mathbb{R}^m$ (the measurements) and $l_{i,j} \in \mathcal{L}_i = \{\ell_{i,1}, \dots, \ell_{i,k_i}\}$ (the labels). Here, we are allowing different cytometries to have different cell types, different names for the same cell type, or different numbers of clusters. In the bottom of Fig. 6.1, we have two gated cytometries $\tilde{X}_1$ and $\tilde{X}_2$, each with four different clusters with the same space of labels, $\mathcal{L} = \{1, 2, 3, 4\}$. An alternative description of a gated cytometric dataset is given by $\tilde{C}_i = \{\{C_{i,l}\}_{l \in \mathcal{L}_i}, \mu_i\}$, where $C_{i,l} = \{x : (x, l) \in \tilde{X}_i\}$ is a grouping of all cells with label l and μ_i is a discrete probability distribution on the clusters, i.e., $\sum_{j=1}^{k_i} \mu_i(C_{i,\ell_j}) = 1$ and $\mu_i(C_{i,\ell_j}) \geq 0$. Therefore, $C_{i,l}$ is the collection of all points in the dataset $\tilde{X}_i$ that have label l, i.e., the cluster corresponding to label l in $\tilde{X}_i$. A gated cytometry is a collection of the clusters corresponding to each label and a measure that associates weights to each cluster. In this setting, one can compute the discrete optimal transport cost (6.3) between two cytometries $\tilde{C}_i, \tilde{C}_j$ as

$$\mathrm{OT}_c(\mu_i, \mu_j) = \min_{\pi \in \pi\left(\{C_{i,l}\}_{l \in \mathcal{L}_i} \times \{C_{j,l}\}_{l \in \mathcal{L}_j}\right)} \sum_{i'=1}^{k_i} \sum_{j'=1}^{k_j} \pi_{i,j} c(C_{i,\ell_{i'}}, C_{j,\ell_{j'}}). \tag{6.20}$$

There is a naive transport cost given by

$$\mathrm{NT}_c(\mu_i, \mu_j) = \sum_{i'=1}^{k_i} \sum_{j'=1}^{k_j} \mu_i(C_{i,\ell_{i'}}) \mu_j(C_{j,\ell_{j'}}) c(C_{i,\ell_{i'}}, C_{j,\ell_{j'}}).$$

This allows to introduce the following distances matrix between gated cytometries:

$$[D_d]_{i,j} = d_{sim,c}\left(\tilde{C}_i, \tilde{C}_j\right) = \frac{OT_c(\mu_i, \mu_j)}{NT_c(\mu_i, \mu_j)}. \tag{6.21}$$

To fully define the similarity distance, $d_{sim,c}$, which was introduced in [32], one has to specify c, the cost function between clusters. Equivalently, one can provide a cost matrix $[c_{ij}]_{i',j'} = c(C_{i,\ell_{i'}}, C_{j,\ell_{j'}})$. This is fairly straightforward, good candidates for the cost function are the distances that we have already discussed in Sect. 6.2. If clusters are modelled as discrete samples, one can use the Wasserstein or MMD distances or the Friedman–Rafsky dissimilarity. If one uses some density estimation for each cluster, typically one assumes that clusters have a multivariate normal shape (or are members of some other location-scale family), and the KL divergence and the Hellinger distance are also available. For more details, we refer to [14]. An example of different cost matrices c_{12} between two gated cytometries $\tilde{X}_1$ and $\tilde{X}_2$ is given in Fig. 6.2.

The solution of the OT problem (6.20) provides a soft assignment of the clusters of one cytometry to the clusters of the other. This is just one of the possible assignment strategies. Another one is to use the solution of a generalized edge covering (GEC) problem, where it is allowed for some origin and end points not to be matched. For more details, we refer to [13]. Hence, a dissimilarity measure can be the cost of the GEC problem, which yields

$$\begin{aligned}[D_d]_{i,j} &= d_{\mathrm{GEC},c,\lambda}(\tilde{C}_i, \tilde{C}_j) \\ &= \min_{BG \in \text{ bigraphs between} \{C_{i,l}\}_{l \in \mathcal{L}_i} \text{ and } \{C_{j,l}\}_{l \in \mathcal{L}_j}} \sum_{\{C,C'\} \in BG} c(C, C') + \lambda |V_{uc}|,\end{aligned} \tag{6.22}$$

where $|V_{uc}|$ is the number of clusters (vertices of the bigraph) that are unassigned, λ is a free parameter that penalizes unassigned clusters (vertices), and $c(C, C')$ is a cost between clusters of the two cytometries. Notice that the cost c can be chosen in the same fashion as for the similarity distance (6.21).

6.3.2 Template Production

A prominent feature of using statistical distances in cytometry gating workflows is that they allow the production of a synthetic cytometry dataset, which we will also call a template, that encapsulates the information of a set of cytometry datasets. This is useful since a template of a group of cytometry datasets can be manually or otherwise gated and then used to gate the cytometry datasets from which it was obtained. This can result in a significant reduction in the amount of expert input in the workflow. Notice that having a template that resembles the group of datasets that it represents facilitates comparison between a new cytometry dataset and the group itself, and it boils down to a comparison between the new cytometric data and the

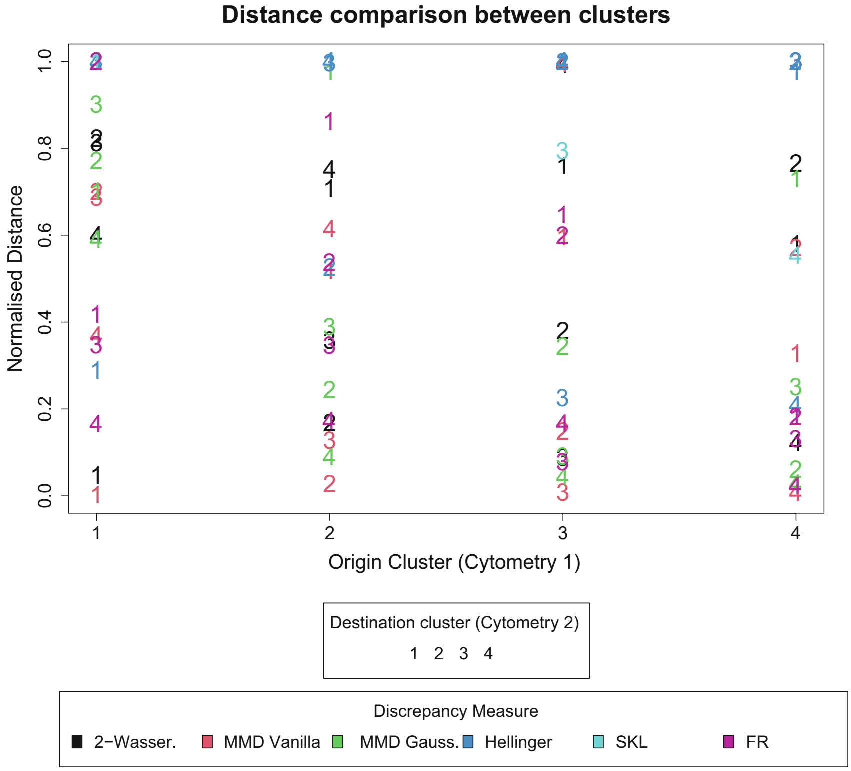


Fig. 6.2 Representation of the normalized cost matrix, $c_{12}/\max c_{12}$, between the cell types (clusters) found in the two gated cytometries $\tilde{X}_1$ and $\tilde{X}_2$ in the bottom of Fig. 6.1, for different distances between cell types. Notice that $[c_{12}]_{ij} = d(C_{1,\ell_i}, C_{2,\ell_j}) = d(C_{1,i}, C_{2,j})$, with $\mathcal{L}_1 = \mathcal{L}_2 = \{1, 2, 3, 4\}$. The distance d can be the 2-Wasserstein distance (6.4), the squared MMD distance (6.7) with the vanilla kernel $\kappa(x, y) =< x, y >$ (the usual scalar product), the squared MMD distance (6.7) with the Gaussian kernel (6.10) with $\lambda = 10/\sqrt{2}$, the Hellinger distance (6.17), and the symmetric KL divergence (6.15) and the FR dissimilarity (6.19). The plot's interpretation is the following: the black 1 (Cluster 1 in cytometry 2) over the x-label value 1 (cluster 1 in cytometry 1) is the lowest of all the other black numbers (rest of clusters in cytometry 2) at the same x-label value. Hence, in 2-Wasserstein distance, cluster 1 in Cytometry 1 is the closest to cluster 1 in Cytometry 2. This yields that only the 2-Wasserstein distance, the MMD with vanilla kernel, and the Hellinger distance have that the closest to clusters 1, 2, 3, and 4 of Cytometry 1 are clusters 1, 2, 3, and 4 of Cytometry 2, respectively, which correctly captures which clusters represent the same cell types

template. This reduces the amount of comparisons required and facilitates assigning a new cytometry to a group of datasets that is most similar to it. From the previous sections, it is clear that a good candidate for a template, but not the only possible one, is the barycenter of a group of cytometric datasets. Below we present some strategies on how to produce template cytometries.

6.3.2.1 Ungated Cytometry Datasets

To obtain a template sample $T = \{t_1, \dots, t_{n_T}\} \subset \mathbb{R}^m$ from a group of raw cytometry samples $\{X_1, \dots, X_n\}$ (as in Sect. 6.3.1.1), one can solve a barycenter problem as the one introduced in Eqs. (6.5) and (6.11). One gets T as a sample from the distribution $\mu^*_{W_p}$ or $\mu^*_{MMD,\mathcal{H}}$, respectively. The main difficulty with this approach is computation. For low-dimensional subspaces of markers, $m \leq 3$, and for relatively small cell counts (sample sizes), it can be done in reasonable computation time. However, to tackle more realistic situations further work in the field of barycenter computation is required. A toy example is given in the top of Fig. 6.3. The synthetic cytometries, i.e., the plotted samples of size 1500, encapsulate the common information present in X_1 and X_2 which are plotted in the top of Fig. 6.1. These templates, or barycenters, can be used as a representation of the set of cytometric datasets $\{X_1, X_2\}$.

6.3.2.2 Gated Cytometry Datasets

A workaround to the problem faced in the ungated setting is to try to work with gated datasets. Since there are many efficient unsupervised gating procedures, this is a viable option. For extensive reviews on such methods, we refer to [4, 6, 33–35]. Hence, we are in the setting of Sect. 6.3.1.2 and there is a collection of cytometry datasets where each individual dataset is modelled as a collection of clusters. Therefore, one has $\{C_i = \{C_{i,l}\}_{l \in \mathcal{L}_i}\}_{i=1}^n$. This is the setting represented in the bottom of Fig. 6.1, where we have two cytometries each formed by four clusters labelled from one to four. There are different approaches on how to obtain a template in this setting.

One way, which is used in [14], is to pool all clusters together, hence obtaining the set $\{C_{1,\ell_{1,1}}, \dots, C_{1,\ell_{1,k_1}}, \dots, C_{n,\ell_{n,1}}, \dots, C_{n,\ell_{n,k_n}}\} = \{\mathcal{C}_i\}_{i=1}^{k_1+\dots+k_n}$, and try to group elements in this set. The rationale behind this is that similar clusters, with respect to some dissimilarity (distance) measure, will represent the same, or at least similar, cell types, and hence grouping them together will allow to separate different cell types. Once different cell types are separated, one can obtain a template for each cell type. The collection of cell type templates is the template for the group of cytometry datasets. Once more, a viable strategy is to obtain a distance matrix $[D_d]_{ij} = d(\mathcal{C}_i, \mathcal{C}_j)$ and to use hierarchical clustering. As previously, the distance between clusters d can be chosen as in Sect. 6.3.1.2. In the particular case when each cluster is modelled as a member of a location-scale family, there is an efficient extension of the k-means algorithm known as k-barycenter (for details see [36]), which produces a template for $\{\mathcal{C}_i\}_{i=1}^{k_1+\dots+k_n}$ with k different cell types.

Another possible strategy, the one followed in [13], is to start with individual cytometric datasets and produce a template of the two closest ones. Then, since templates can be handled exactly the same as the other cytometric datasets, one can continue merging the two closest cytometric datasets until there is only one

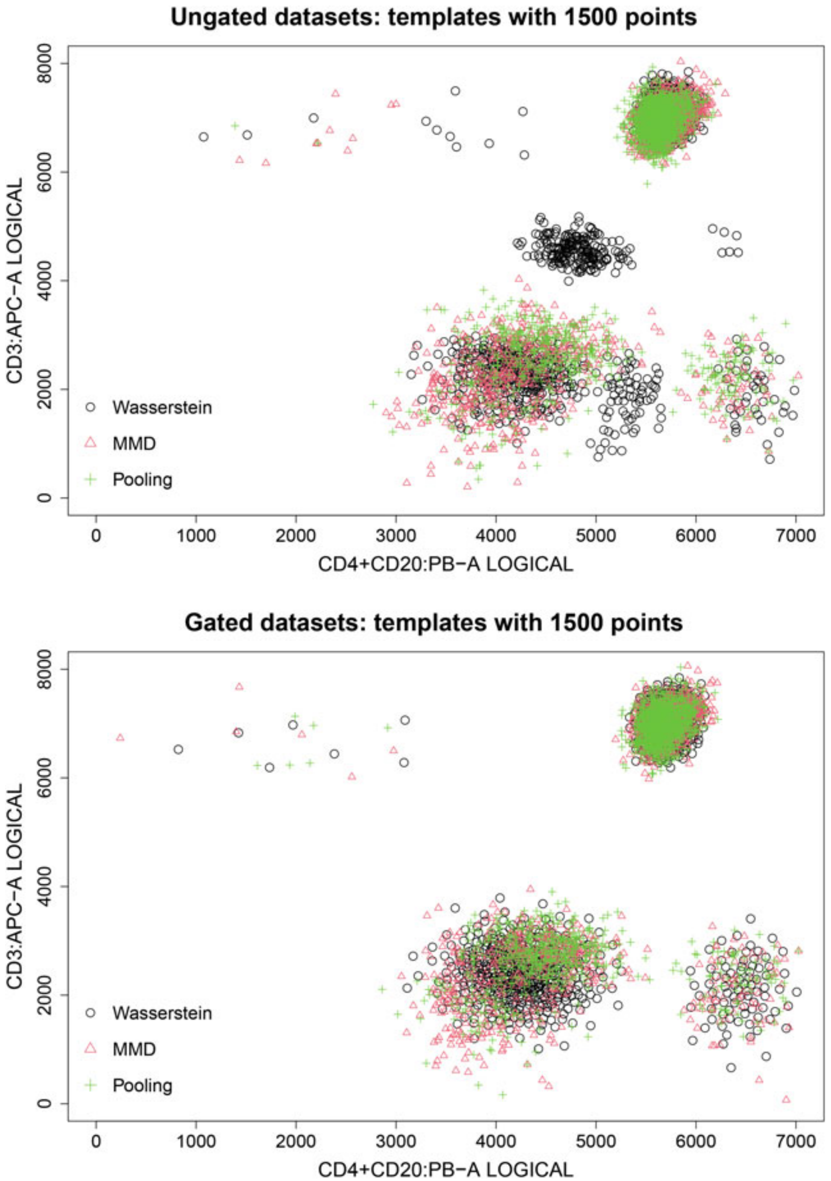


Fig. 6.3 Examples of different templates for a group of cytometric datasets. The goal is to obtain a synthetic cytometry that captures the information of both cytometries in Fig. 6.1. For ease of computation, we select the template to have 1500 cells. *Top:* Templates obtained from the ungated cytometry datasets X_1 and X_2 in the top of Fig. 6.1. We see that the most relevant information of both cytometries is well represented, with the 2-Wasserstein barycenter producing some spurious clusters. *Bottom:* Templates obtained from the gated versions of X_1 and X_2, $\tilde{X}_1$ and $\tilde{X}_2$, plotted in the bottom of Fig. 6.1. We considered clusters that are closest in 2-Wasserstein distance to correspond to the same cell type (see Fig. 6.2) and obtained a barycenter for each cell type. Again, templates seem to represent the original information well, with the 2-Wasserstein template producing the more homogeneous cell types

last template of the whole group of cytometric datasets. Hence, in this approach it is important to have a distance between gated cytometries such as $d_{sim,c}$ (6.21) or $d_{\mathrm{GEC},c,\lambda}$ (6.22). Notice that the last one comes with a hard assignment which can be used as a recipe for which clusters to merge together and which ones to leave unmerged.

To complete the template production, one needs to describe a method for obtaining a template from the clusters that have been grouped together as representing the same or similar cell types. A straightforward cell template can be achieved by pooling together all the points of the clusters grouped together. A more sophisticated approach is to solve a barycenter problem. Here, solving the barycenter problem (6.11) to obtain $\mu^*_{\mathrm{MMD},\mathcal{H}}$ will give a template which will be fairly similar to the one obtained by pooling. By solving a Wasserstein barycenter problem (6.5), one can obtain a different result. As mentioned in the previous section, when the space is high dimensional and the involved clusters have hundreds of thousands of points, one may need to fit location-scale models with densities to the clusters and then solve a 1-barycenter problem. In the bottom of Fig. 6.3, we have the templates obtained, from grouping together each cluster from $\tilde{X}_1$ to its closest counterpart in $\tilde{X}_2$, and, then, the barycenters for each cell type are obtained by a 2-Wasserstein barycenter, an MMD barycenter or by pooling. We see that the resulting templates are a sensible representation of the information stored in the two original cytometries.

6.3.3 Interpolation Between Cytometry Datasets

The ability to transform one cytometry dataset into another in some controlled fashion is very desirable. Two major consequences are the following: firstly, one can translate gates used to gate one of the cytometric datasets to gate the other one, and secondly, one can transform several cytometry datasets to try to reduce batch effects in a procedure known as normalization. In this section, we describe several methods based on statistical distances.

6.3.3.1 Gate Transportation

Typically, manual gating is a one- or a two-dimensional hierarchical procedure, and therefore to use gates from a gated cytometry, one needs to be able to interpolate in one or two dimensions and not in the full space of m markers. This is a considerable reduction in dimension and makes computation far easier. A good tool for interpolation in this setting is the transport map T, whenever it exists, associated with the solution of the optimal transport problem (6.4). Hence, in order for the transport maps to exist and be unique, we will assume that we are working with the 2-Wasserstein distance and that the cytometric datasets are samples from probability distributions with densities. Although transport maps in two dimensions can be used to transport two-dimensional gates from one cytometry dataset to another one, it is far more technical and it is out of the scope of this chapter. On the other hand, the one-dimensional counterpart provides good intuition and can be broadly applied. The optimal map between two measures μ and ν defined in $\mathbb{R}$ is given by

$$T(x) = F_\nu^{-1} \circ F_\mu(x), \tag{6.23}$$

where F_μ is the cumulative distribution function (CDF) of μ and F_ν^{-1} is the quantile function (QF), also known as the generalized inverse, of ν. For a sample $X = \{x_1, \dots, x_{n_X}\}$ from μ and $Y = \{y_1, \dots, y_{n_Y}\}$ from ν, a plug-in estimator is obtained by

$$T_n(x) = F_{n,\nu}^{-1} \circ F_{n,\mu}(x), \tag{6.24}$$

where $F_{n,\mu}$ is the empirical CDF associated with the sample X and $F_{n,\nu}^{-1}$ is the empirical QF associated with the sample Y. Notice that when dealing with one-dimensional projections, a gate associated with marker m_i is just a value $\theta_i \in \mathbb{R}$. Therefore, the transported version of the gate is $T_n(\theta_i)$, and it is the one to be used for gating Y with respect to marker m_i. Examples of this gate transportation can be seen in Fig. 6.4. Let us stress that here the samples are just the one-dimensional projections into marker m_i. We want to point out that this alignment method can replace or be used alongside the alternatives in the workflows presented in [11] and [26].

A different approach, also based on the OT problem and introduced in [27], tries to reweight the learning sample, the one that is gated, in order to minimize the OT cost to the ungated cytometry dataset. The optimal weights can be understood as the relative frequencies of the original gated cell types in the new cytometry dataset. Since, usually, the relative frequencies are relevant for diagnosis, the previous procedure can be good enough in many practical situations. Being that the origin cytometry is gated, one can write its empirical distribution function as

$$\eta = \sum_{k=1}^{K} \frac{|C_{1,\ell_k}|}{\sum_{j=1}^{K} |C_{1,\ell_j}|} \left(\sum_{x \in C_{1,\ell_k}} \frac{1}{|C_{1,\ell_k}|} \delta_x \right) = \sum_{k=1}^{K} \frac{n_{\ell_k}}{n_1} \eta_{\ell_k}, \tag{6.25}$$

where $|C_{1,\ell_k}| = n_{\ell_k}$ is the number of points that have labels ℓ_k, $\sum_{j=1}^{K} |C_{1,\ell_j}| = n_1$ is the total number of cells in C_1, and η_{ℓ_k} is the empirical distribution of the points with label ℓ_k. A reweighting of η with weights $w = \{w_k\}_{k=1}^{K}$, with $w_k > 0$ and $\sum_{k=1}^{K} w_k = 1$, is given by

$$\eta(w) = \sum_{k=1}^{K} w_k \eta_{\ell_k}. \tag{6.26}$$

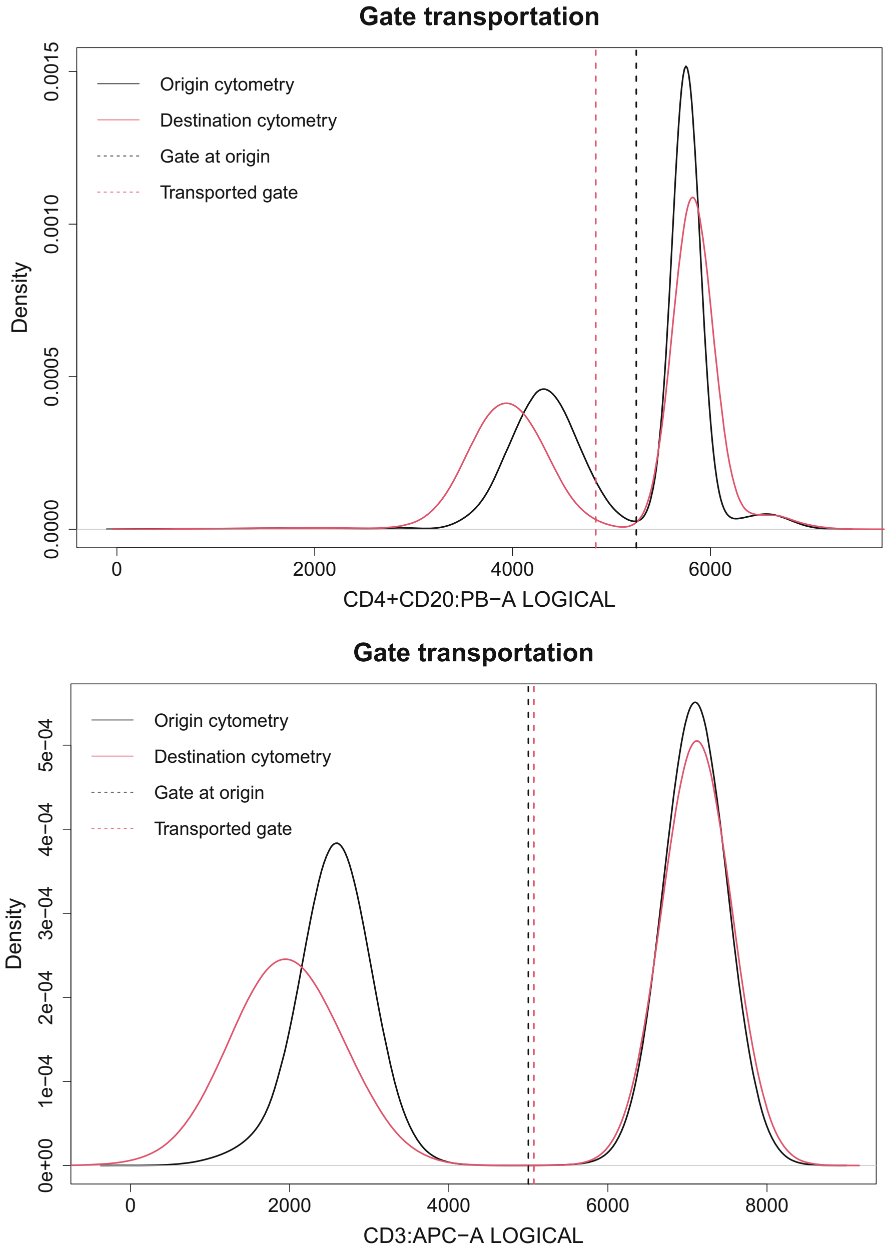


Fig. 6.4 Examples of gate transportation using the OT map (6.24). In solid black, we have the density estimation of projections into two markers of an origin cytometry X. Some one-dimensional gates are given in dashed black, which separate low and high values in the respective marker. Transported gates, which do not require any human input, are presented in dashed red.

Let us call the empirical distribution associated with the ungated cytometry $X_2 = \{x_{2,1}, \ldots, x_{2,n_2}\}$, $\eta' = \sum_{x \in X_2} \frac{1}{n_2} \delta_x$. Then, one has the following minimization problem:

$$w^* = \arg\min_{w} d^2_{W_2}(\eta(w), \eta'). \tag{6.27}$$

An approximate solution of the regularized version of this problem, i.e., with the entropically regularized Wasserstein distance as approximator of the Wasserstein distance, can be computed efficiently (see details in [27]) and its solution w^* represents the relative weights in the new cytometry X_2 of the cell types present in C_1. It is also possible to obtain a full gating of X_2 if one has access to the optimal coupling.

6.3.3.2 Reduction of Batch Effects

When the problem of interest is the reduction of batch effects, there are two distinct strategies. One is to look for cell type-dependent normalizations, and therefore one assumes that batch effects are not independent from the cell types. Another is to try to treat batch effects as a common perturbation to the whole dataset and therefore as cell type independent. The main setting is the following: there are batches $\{B_j\}_{j=1}^{N_B}$, where each batch is a collection of cytometric datasets $B_j = \{X_{j,1}, \ldots, X_{j,N_j}\}$ such that $X_{j,i} = \{x_{j,i,1}, \ldots, x_{j,i,n_{j,i}}\} \subset \mathbb{R}^m$.

A cell type-dependent normalization may proceed by previously gating in an unsupervised fashion the available datasets, and then producing a normalization dependent on the groups (as in [26]), or, alternatively, it can affect only some gates in a gating hierarchy (as in [37]). The main tools required in those settings are the production of an "average" element for a group of 1D samples and the interpolation between two 1D samples. When possible, reducing batch effects is helped if one can have a control sample at each batch (see [26]). Therefore, one can find the empirical quantile function of the barycenter of the 1D projections onto the marker m_q of the control sample in the different batches, which we denote as $F^{-1}_{n,*,q}$. For details on the computation, see Remark 9.6 in [20]. Hence, for each batch j, there is a transport map

$$T_{n,j,q}(x) = F^{-1}_{n,*,q} \circ F_{n,j,q}(x), \tag{6.28}$$

where $F_{n,j,q}$ is the empirical CDF associated with the projection onto the marker m_q of the control sample in batch j. A normalization of marker m_q for batch j corresponds to applying $T_{n,j,q}$ to the m_q projection of the other cytometric datasets of batch j. The full normalization is the resulting data for the correction of all (or some) markers. When normalization is cell type specific, this is done for each cell type and the corresponding markers. We stress that in [26, 37] methods not directly related to statistical measures of discrepancy are used. However, adopting the techniques presented in this section is fairly simple.

A different approach consists of finding an approximation of a function $g : \mathbb{R}^m \to \mathbb{R}^m$ such that cytometry datasets in different batches are closer after transforming them by g. This problem can be situated in the fields of domain adaptation of transfer learning in ML (for some comprehensive reviews on the topics see [38, 39]). The idea is to choose a reference cytometry dataset X^*, usually the one that will be gated, and try to find a function g that brings all other cytometries closer to X^*. This can be done using generative adversarial networks (GANs) where a loss function is based on a distance d between cytometric datasets. Examples of this procedure can be found in [28, 29]. The distance d can be any of the ones we have discussed so far, but usually the most efficient are the Wasserstein and the one based on MMD since they can be computed from samples efficiently.

6.4 Conclusions

The main point that the reader should take from this chapter is that working with cytometry datasets as statistical objects, particularly through the lenses of statistical distances, is very helpful in the whole gating workflow. Some popular gating methods can be readily adapted to incorporate (or already do) a measure of discrepancy or an alignment method between cytometry datasets based on the tools discussed in this chapter. Furthermore, in supervised settings, preprocessing steps based on reducing variability have proven effective in improving performance as shown in [14]. Therefore, any practitioner or interested researcher should have at least a basic knowledge of the topics presented in this chapter. This knowledge can be very helpful since many of the discussed topics are very active fields of research and innovation which can have further positive impact in the cytometry gating workflow.

Acknowledgments This research was partially supported by project MTM2017-86061-C2-1-P (AEI/FEDER, UE), by grant PID2021-128314NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE), by the Basque Government through the BERC 2018-2021 and the Basque Modelling Task Force (BMTF) program, by the Spanish Ministry of Science and Innovation (MCIN) and AEI (BCAM Severo Ochoa accreditation CEX2021-001142-S), and by grants VA005P17 and VA002G18 (Junta de Castilla y León).

References

1. Cossarizza, A., Chang, H.D., Radbruch, A., Acs, A., Adam, D., Adam-Klages, S., Agace, W.W., Aghaeepour, N., Akdis, M., Allez, M., et al.: Guidelines for the use of flow cytometry and cell sorting in immunological studies. European journal of immunology **49**(10), 1457–1973 (2019)

2. Iyer, A., Hamers, A.A., Pillai, A.B.: Cytof® for the masses. Frontiers in Immunology **13**, 815828 (2022)
3. Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., Raddassi, K., Devine, L., Obermoser, G., Pekalski, M.L., et al.: Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. Scientific reports **6**(1), 1–11 (2016)
4. Saeys, Y., Van Gassen, S., Lambrecht, B.N.: Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature Reviews Immunology **16**(7), 449–462 (2016)
5. Hu, Z., Bhattacharya, S., Butte, A.J.: Application of machine learning for cytometry data. Frontiers in Immunology p. 5703 (2021)
6. Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H.: Critical assessment of automated flow cytometry data analysis techniques. Nature methods **10**(3), 228–238 (2013)
7. Hennig, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of cluster analysis. CRC Press (2015)
8. Alpaydin, E.: Introduction to machine learning. MIT press (2020)
9. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)
10. Hu, Z., Jujjavarapu, C., Hughey, J.J., Andorf, S., Lee, H.C., Gherardini, P.F., Spitzer, M.H., Thomas, C.G., Campbell, J., Dunn, P., et al.: Metacyto: a tool for automated meta-analysis of mass and flow cytometry data. Cell reports **24**(5), 1377–1388 (2018)
11. Lux, M., Brinkman, R.R., Chauve, C., Laing, A., Lorenc, A., Abeler-Dörner, L., Hammer, B.: flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. Bioinformatics **34**(13), 2245–2253 (2018)
12. Maecker, H.T., McCoy, J.P., Nussenblatt, R.: Standardizing immunophenotyping for the human immunology project. Nature Reviews Immunology **12**(3), 191–200 (2012)
13. Azad, A., Pyne, S., Pothen, A.: Matching phosphorylation response patterns of antigen-receptor-stimulated t cells via flow cytometry. In: BMC Bioinformatics, vol. 13, pp. 1–8. Springer (2012)
14. Del Barrio, E., Inouzhe, H., Loubes, J.M., Matrán, C., Mayo-Íscar, A.: optimalFlow: optimal transport approach to flow cytometry gating and population matching. BMC bioinformatics **21**(1), 1–25 (2020)
15. Klenke, A.: Probability theory: a comprehensive course. Springer Science & Business Media (2013)
16. Ross, S.M.: A first course in probability. Pearson (2014)
17. García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. The Annals of Statistics **36**(3), 1324–1345 (2008)
18. Orlova, D.Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E.E., Filatenkov, A., Kolyagin, G.A., Gernez, Y., Tsuda, S., et al.: Earth mover's distance (EMD): a true metric for comparing biomarker expression levels in cell populations. PLoS one **11**(3), e0151859 (2016)
19. Villani, C.: Optimal transport: old and new, vol. 338. Springer (2009)
20. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
21. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al.: Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning **10**(1-2), 1–141 (2017)
22. Cohen, S., Arbel, M., Deisenroth, M.P.: Estimating barycenters of measures in high dimensions. arXiv preprint arXiv:2007.07105 (2020)
23. Haasdonk, B., Bahlmann, C.: Learning with distance substitution kernels. In: Joint pattern recognition symposium, pp. 220–227. Springer (2004)
24. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics pp. 697–717 (1979)
25. Hsiao, C., Liu, M., Stanton, R., McGee, M., Qian, Y., Scheuermann, R.H.: Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman–Rafsky test statistic as a distance measure. Cytometry Part A **89**(1), 71–88 (2016)

26. Van Gassen, S., Gaudilliere, B., Angst, M.S., Saeys, Y., Aghaeepour, N.: Cytonorm: a normalization algorithm for cytometry data. Cytometry Part A **97**(3), 268–278 (2020)
27. Freulon, P., Bigot, J., Hejblum, B.P.: CytOpT: Optimal transport with domain adaptation for interpreting flow cytometry data. arXiv preprint arXiv:2006.09003 (2020)
28. Li, H., Shaham, U., Stanton, K.P., Yao, Y., Montgomery, R.R., Kluger, Y.: Gating mass cytometry data by deep learning. Bioinformatics **33**(21), 3423–3430 (2017)
29. Shaham, U., Stanton, K.P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., Kluger, Y.: Removal of batch effects using distribution-matching residual networks. Bioinformatics **33**(16), 2539–2546 (2017)
30. Ram, P., Gray, A.G.: Density estimation trees. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 627–635 (2011)
31. Roederer, M., Moore, W., Treister, A., Hardy, R.R., Herzenberg, L.A.: Probability binning comparison: a metric for quantitating multivariate distribution differences. Cytometry: The Journal of the International Society for Analytical Cytology **45**(1), 47–55 (2001)
32. Coen, M.H., Ansari, M.H., Fillmore, N.: Comparing clusterings in space. In: ICML (2010)
33. Cheung, M., Campbell, J.J., Whitby, L., Thomas, R.J., Braybrook, J., Petzing, J.: Current trends in flow cytometry automated data analysis software. Cytometry Part A **99**(10), 1007–1021 (2021)
34. Liu, P., Liu, S., Fang, Y., Xue, X., Zou, J., Tseng, G., Konnikova, L.: Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. Frontiers in cell and developmental biology **8**, 234 (2020)
35. Montante, S., Brinkman, R.R.: Flow cytometry data analysis: Recent tools and algorithms. International Journal of Laboratory Hematology **41**, 56–62 (2019)
36. Alvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matran, C.: Wide consensus aggregation in the Wasserstein space. application to location-scatter families. Bernoulli **24**(4A), 3147–3179 (2018)
37. Finak, G., Jiang, W., Krouse, K., Wei, C., Sanz, I., Phippard, D., Asare, A., De Rosa, S.C., Self, S., Gottardo, R.: High-throughput flow cytometry data normalization for clinical trials **85**(3), 277–286 (2014)
38. Kouw, W.M., Loog, M.: A review of domain adaptation without target labels. IEEE transactions on pattern analysis and machine intelligence **43**(3), 766–785 (2019)
39. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)