IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 19, 2021, accepted April 5, 2021, date of publication April 12, 2021, date of current version April 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072613

Optimal Transport With Relaxed Marginal Constraints

JIA LI, (Fellow, IEEE), AND LIN LIN
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

Corresponding author: Jia Li (jiali@psu.edu)

This work was supported by the National Science Foundation under Grant DMS-2013905.

ABSTRACT Optimal transport (OT) is a principled approach for matching, having achieved success in diverse applications such as tracking and cluster alignment. It is also the core computation problem for solving the Wasserstein metric between probabilistic distributions, which has been increasingly used in machine learning. Despite its popularity, the marginal constraints of OT impose fundamental limitations. For some matching or pattern extraction problems, the framework of OT is not suitable, and post-processing of the OT solution is often unsatisfactory. In this paper, we extend OT by a new optimization formulation called *Optimal Transport with Relaxed Marginal Constraints* (OT-RMC). Specifically, we relax the marginal constraints by introducing a penalty on the deviation from the constraints. Connections with the standard OT are revealed both theoretically and experimentally. We demonstrate how OT-RMC can easily adapt to various tasks by three highly different applications in image analysis and single-cell data analysis. Quantitative comparisons have been made with OT and another commonly used matching scheme to show the remarkable advantages of OT-RMC.

INDEX TERMS Optimal transport, linear programming, pattern extraction, fixed target matching, single-cell data analysis, cluster alignment.

I. INTRODUCTION

Optimal transport (OT) has been successfully applied in diverse areas including machine learning, computer vision, and bioinformatics. A theoretical treatment of the topic is referred to the seminal book of [1]. From the perspective of probability measures, OT solves the Wasserstein metric between two probability distributions. A historical account on the Wasserstein metric is provided in [1], [2]. Consider two probability measures $\mathcal{P}_1, \mathcal{P}_2$ on the d-dimensional Euclidean space $\mathbb{R}^d$. For random vectors $X, Y \in \mathbb{R}^d$, suppose $X \sim \mathcal{P}_1$ and $Y \sim \mathcal{P}_2$. Let $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ be the collection of joint distributions for X and Y on $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal distribution of X is fixed at $\mathcal{P}_1$ and that of Y at $\mathcal{P}_2$. These requirements of $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ are called the *marginal constraints*. The p-Wasserstein distance $W_{\mathrm{p}}(\cdot, \cdot)$ between $\mathcal{P}_1$ and $\mathcal{P}_2$, $\mathrm{p} \geq 1$, is a true metric defined by

$$W_{\mathrm{p}}^{\mathrm{p}}(\mathcal{P}_1, \mathcal{P}_2) = \inf_{\gamma \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^{\mathrm{p}} d\gamma(\mathbf{x}, \mathbf{y}). \quad (1)$$

$\Pi(\cdot, \cdot)$ is often called the *coupling set*. The distribution $\gamma^* \in \Pi(\mathcal{P}_1, \mathcal{P}_2)$ achieving the infimum in the above equation is called *optimal coupling*. Here, $\| \cdot \|$ is a norm on $\mathbb{R}^d$. When the distributions are finite and discrete, Wasserstein distance is solved by linear programming. In this case, it will also become clear that OT fits well with our intuition of optimal matching. Detailed discussion is presented in Section II. Not surprisingly, in many applications, OT is used as an optimal matching scheme.

One enormously famous case of OT's usage is the Earth Mover's Distance in computer vision for comparing histograms or in general discrete distributions [3]. Recently, OT has been applied to track cells [4] or identify cell developmental trajectories [5]. Via the link with Wasserstein distance, OT has sparked much interest from the machine learning community in recent years. One active area focuses on solving the barycenter for a collection of distributions under the Wasserstein distance. Although the barycenter is a counterpart of average in the Euclidean space, the computational hurdle of solving the Wasserstein barycenter is high, attracting devoted work from applied mathematics [6]–[12], computer science, and signal processing [13]–[15].

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

It has been found that the Wasserstein barycenter is valuable for geometric interpolation in computer graphics [16], [17] and for synthesizing images in meteorology forecasting [18]. As a strong competitor of the Kullback-Leibler distance, OT has been used to define new distance between Gaussian mixture models or hidden Markov models [19]–[21]. Wasserstein distance has also been used for robust supervised learning [22]–[24]. In the case of unsupervised learning, OT readily applies to the issue of aligning clustering results (the consistent cluster labeling issue), which then forms the basis for ensemble clustering and uncertainty analysis for clustering [25], [26].

One profound limitation of OT in some applications is rooted in the marginal constraints in the optimization problem. For example, when OT is used to match (aka align) clusters in different clustering results and subsequently to consolidate clusters in those results, the marginal constraints imply that the proportions of the true clusters are fixed across the results. In another word, it is implicitly assumed that the variation observed in the results arises from the randomness in the data or other nuance factors. However, as is known in single-cell data analysis (details in Section IV), this assumption does not necessarily hold. The proportions of clusters can vary substantially across samples for biomedical reasons, and new clusters may exist in some samples but not others. We will also see from another two examples in image data analysis when OT is used to match points in two sets. If the sets are interfered with by noise, the marginal constraints render OT defenseless against such issues.

In this paper, we propose a novel principled approach to address the limitations posed by the marginal constraints. We extend the optimization problem of OT to the so-called *Optimal Transport with Relaxed Marginal Constraints* (*OT-RMC*). OT-RMC overcomes the aforementioned limitations and in the meanwhile ensures that the solution does not degenerate into a trivial scheme. In a nutshell, we introduce the gap variables such that deviation from the marginal constraints is allowed whereas the gap variables are subject to penalty and/or bounds. OT-RMC allows different schemes to control the gap variables, thus easily adaptable to various applications. We explore the application of OT-RMC to three quite different problems in Section IV.

The rest of the paper is organized as follows. We present notations and the basic OT problem in Section II. The new framework of OT-RMC and its theoretical connection to the standard OT are presented in Section III. In Section IV, we describe three example applications and provide experimental results. Finally, we conclude in Section V.

II. PRELIMINARIES

In this section, we introduce the notations and the formulation of OT between two finite discrete distributions. Suppose the lth distribution, denoted by $\mathcal{P}_l$, $l = 1, 2$, has support $\mathbf{X}_l = \{\mathbf{x}_i^{(l)}, i = 1, \ldots, n_l\}$, $\mathbf{x}_i^{(l)} \in \mathbb{R}^d$. The probability on $\mathbf{x}_i^{(l)}$ is $q_i^{(l)}$, $i = 1, \ldots, n_l$, $\sum_{i=1}^{n_l} q_i^{(l)} = 1$. Let $\mathbf{q}_l = (q_1^{(l)}, q_2^{(l)}, \ldots, q_{n_l}^{(l)})^t$, $l = 1, 2$. We sometimes write a distribution as a list of support points and their corresponding probabilities, $\mathcal{P}_l$: $\{(\mathbf{x}_1^{(l)}, q_1^{(l)}), (\mathbf{x}_2^{(l)}, q_2^{(l)}), \ldots, (\mathbf{x}_{n_l}^{(l)}, q_{n_l}^{(l)})\}$.

Define a cost function between two points $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(2)}$ by $c(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)})$, written in short as $c_{i,j}$. To solve the p-Wasserstein distance, $c(\cdot, \cdot)$ is the p-th power of a norm on $\mathbb{R}^d$. Common choices include the L_2 norm, its square, or the L_1 norm. Consider a joint distribution on the Cartesian product set $\mathbf{X}_1 \times \mathbf{X}_2$ with probability $w_{i,j}$ on $(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)})$. Let $\mathbf{c} = (c_{i,j})$ and $\mathbf{w} = (w_{i,j})$ be the matrices of size $n_1 \times n_2$ containing elements $c_{i,j}$ and $w_{i,j}$ respectively. We call $\mathbf{c}$ the cost matrix, and $\mathbf{w}$ the *matching weight matrix*. Denote the inner product by $\langle \cdot, \cdot \rangle$. For matrices, the inner product is applied to their vectorized versions. Hence $\langle \mathbf{c}, \mathbf{w} \rangle = \sum_{i,j} c_{i,j} w_{i,j}$. In the sequel, relational signs, e.g., $>$, $=$, between matrices (or vectors) or between a matrix and a constant, apply in an element-wise manner. For example $\mathbf{w} \geq 0$ means every element $w_{i,j} \geq 0$. Denote by $\mathbb{1}_n$ a vector of dimension n with every element being 1.

The OT problem is stated as follows.

$$
\begin{aligned}
R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w}}{min} & \langle \mathbf{c}, \mathbf{w} \rangle \\
s.t. \ & \mathbf{w} \geq 0 \\
& \mathbf{w} \cdot \mathbb{1}_{n_2} = \mathbf{q}_1 \\
& \mathbf{w}^t \cdot \mathbb{1}_{n_1} = \mathbf{q}_2 .
\end{aligned} \tag{2}
$$

The last two equality constraints are the marginal constraints. The first ensures that the row-wise sum of $\mathbf{w}$ is determined by the marginal distribution $\mathcal{P}_1$, while the second ensures that the column-wise sum is determined by $\mathcal{P}_2$. According to (2), we can view OT as an optimal matching scheme in which the matching weights are provided by $w_{i,j}$ and the goal is to minimize the weighted sum of the cost for matching any pair $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(2)}$. The total weight assigned to any $\mathbf{x}_i^{(1)}$ (or $\mathbf{x}_j^{(2)}$) reflects its overall influence on R and is fixed at its probability, as guaranteed by the marginal constraints. If the cost is defined by $c(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}) = \|\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)}\|^{\mathrm{p}}$, where $\| \cdot \|$ is a norm on $\mathbb{R}^d$, then the Wasserstein distance between $\mathcal{P}_1$ and $\mathcal{P}_2$, as defined by Eq. (1), is given by

$$
W_{\mathrm{p}}(\mathcal{P}_1, \mathcal{P}_2) = R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2)^{1/\mathrm{p}} . \tag{3}
$$

III. OPTIMAL TRANSPORT WITH RELAXED MARGINAL CONSTRAINTS

In some applications, the marginal constraints in problem (2) are too rigid or even conflict with the nature of the problem. For example, in unsupervised clustering, cluster labels are named arbitrarily only as symbols to distinguish groups. Since clusters generated in multiple results usually do not correspond to each other sharply, OT is used to match clusters in different results [25]. It is sometimes improper to assume that the clustering results are random realizations of one underlying "truth''. For instance, in single-cell data analysis, clusters can be formed from measurements of the same set of subjects at different time spots. The intrinsic groups of subjects are expected to evolve over time, and the proportion of each cluster can be an important source of change. It is thus unreasonable to enforce the marginal constraints. On the other hand, as the characteristics of clusters (e.g., mean feature vectors) are not fixed across datasets, it is difficult to discern to what extent the change in the overall clustering result comes from the cluster proportions and to what extent from the cluster characteristics. The need to overcome the restriction of the marginal constraints will become more vivid when we address applications in Section IV.

A. OT-RMC FORMULATION

We extend the OT framework in Eq. (2) by introducing two *marginal gap* vectors $\mathbf{g}_l = (g_1^{(l)}, g_2^{(l)} .., g_{n_l}^{(l)})^t$, $l = 1, 2$. Let $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$ be the column-wise concatenated vector of $\mathbf{g}_1$ and $\mathbf{g}_2$. Let $L(\mathbf{g})$ be a loss function that penalizes non-zero gap vectors, for example, the p-th power of the L_p norm, $L(\mathbf{g}) = \|\mathbf{g}\|_p^p$. Let λ be a hyperparameter. Let $\mathbf{g}_{up} \in \mathbb{R}^{n_1+n_2}$, $\mathbf{g}_{up} \geq 0$, be an upper bound vector. For brevity of notation, denote the first n_1 dimensions of $\mathbf{g}_{up}$ by $\mathbf{g}_{up,1}$, and the rest n_2 dimensions by $\mathbf{g}_{up,2}$. That is, $\mathbf{g}_{up} = (\mathbf{g}_{up,1}, \mathbf{g}_{up,2})$ (column-wise concatenation). The bounds in $\mathbf{g}_{up}$ are pre-given for the optimization problem below.

The problem of OT-RMC is stated as follows.

$$\begin{aligned} R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w},\mathbf{g}}{min} \langle \mathbf{c}, \mathbf{w} \rangle + \lambda L(\mathbf{g}) & \\ s.t.\ \mathbf{w} \geq 0 & \\ \mathbb{1}_{n_1}^t \cdot \mathbf{w} \cdot \mathbb{1}_{n_2} = 1 & \\ \mathbf{g} \leq \mathbf{g}_{up} & \\ \mathbf{q}_1 - \mathbf{g}_1 \leq \mathbf{w} \cdot \mathbb{1}_{n_2} \leq \mathbf{q}_1 + \mathbf{g}_1 & \\ \mathbf{q}_2 - \mathbf{g}_2 \leq \mathbf{w}^t \cdot \mathbb{1}_{n_1} \leq \mathbf{q}_2 + \mathbf{g}_2 . & \quad (4) \end{aligned}$$

The first two constraints ensure that $\mathbf{w}$ specifies a valid joint distribution on $\mathbf{X}_1 \times \mathbf{X}_2$. In OT, the marginal constraints on $\mathbf{w}$ imply the unit sum requirement (thus omitted). We call the last two inequality constraints the *relaxed marginal constraints*.

Problem (4) is called the *canonical* form of OT-RMC. It is interesting to point out three variations from the canonical form. The usage of a particular formulation depends strongly on the application in consideration. In this paper, we focus on the usage of the canonical form.

1) *Regularized OT-RMC*: We can extend the objective function in Eq. (4) to include a regularization function G on $\mathbf{w}$:

$$R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w},\mathbf{g}}{min} \langle \mathbf{c}, \mathbf{w} \rangle + \lambda L(\mathbf{g}) + \eta G(\mathbf{w}) .$$

The design of $G(\mathbf{w})$ depends on the application. For example, if the support points correspond to pixels in an image, it might be desirable to enforce smoothness in the induced distributions through $G(\mathbf{w})$. Specifically, we may favor similar values for the induced probabilities on neighboring pixels in the image plane. The increase in the complexity or memory load of the regularized OT-RMC depends on $G(\mathbf{w})$.

2) *Non-homogeneously penalized OT-RMC*: Suppose the loss $L(\mathbf{g})$ is additive in the elements of $\mathbf{g}$ (true for the p-th power of the L_p norm). Let $\tilde{L}$ be a loss function defined on one variable. An additive $L(\mathbf{g})$ can be written as $L(\mathbf{g}) = \sum_{i=1}^{n_1} \tilde{L}(g_i^{(1)}) + \sum_{j=1}^{n_2} \tilde{L}(g_j^{(2)})$. We can then penalize the gap variables by different λ's. Consider hyperparameter vector $\vec{\lambda} = (\lambda_1, \ldots, \lambda_{n_1+n_2})^t$. Let $\vec{L}(\mathbf{g}) = (\tilde{L}(g_1^{(1)}), \ldots, \tilde{L}(g_{n_1}^{(1)}), \tilde{L}(g_1^{(2)}), \ldots, \tilde{L}(g_{n_2}^{(2)}))^t$. The objective function of (4) becomes

$$R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w},\mathbf{g}}{min} \langle \mathbf{c}, \mathbf{w} \rangle + \langle \vec{\lambda}, \vec{L}(\mathbf{g}) \rangle .$$

Non-homogeneously penalized OT-RMC does not cause particular computational difficulty. If $\tilde{L}$ is the L_1 norm (or square of L_2 norm), the optimization is linear programming (or convex quadratic programming).

3) *Asymmetrically bounded OT-RMC*: The upper bound on the gap vector $\mathbf{g}$ in (4) ensures that the induced distributions $\tilde{\mathbf{q}}_1$ and $\tilde{\mathbf{q}}_2$ (defined below) are bounded symmetrically around $\mathbf{q}_1$ and $\mathbf{q}_2$. If the desired bounds are not symmetric, we can directly impose constraints on $\tilde{\mathbf{q}}_1$ and $\tilde{\mathbf{q}}_2$. Note that $\tilde{\mathbf{q}}_1 = \mathbf{w} \cdot \mathbb{1}_{n_2}$, $\tilde{\mathbf{q}}_2 = \mathbf{w}^t \cdot \mathbb{1}_{n_1}$, $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$, where $\mathbf{g}_1 = \left|\tilde{\mathbf{q}}_1 - \mathbf{q}_1\right|$ and $\mathbf{g}_2 = \left|\tilde{\mathbf{q}}_2 - \mathbf{q}_2\right|$. Then the problem is stated as

$$\begin{aligned} R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w},\mathbf{g}}{min} \langle \mathbf{c}, \mathbf{w} \rangle + \lambda L(\mathbf{g}) & \\ s.t.\ \mathbf{w} \geq 0\,, \text{ and } \mathbb{1}_{n_1}^t \cdot \mathbf{w} \cdot \mathbb{1}_{n_2} = 1 & \\ \mathbf{g}_{low,1} \leq \tilde{\mathbf{q}}_1 \leq \mathbf{g}_{up,1} & \\ \mathbf{g}_{low,2} \leq \tilde{\mathbf{q}}_2 \leq \mathbf{g}_{up,2} . & \quad (5) \end{aligned}$$

B. CONNECTIONS WITH OT

Suppose $\mathbf{w}^*$, $\mathbf{g}^*$ achieve the minimum in (4). Let the two new marginal probability vectors induced by $\mathbf{w}^*$ be $\tilde{\mathbf{q}}_1 = \mathbf{w}^* \cdot \mathbb{1}_{n_2}$ and $\tilde{\mathbf{q}}_2 = \mathbf{w}^{*t} \cdot \mathbb{1}_{n_1}$. Denote the corresponding marginal distributions by $\tilde{\mathcal{P}}_l$, $l = 1, 2$. We can then define a total cost of matching or a distance between the two distributions by

$$D(\mathcal{P}_1, \mathcal{P}_2) = W_{\mathrm{p}}(\tilde{\mathcal{P}}_1, \tilde{\mathcal{P}}_2)\,, \quad (6)$$

where W_{p} is defined by (3) for discrete distributions. The above definition is equivalent to

$$D(\mathcal{P}_1, \mathcal{P}_2)^{\mathrm{p}} = \langle \mathbf{c}, \mathbf{w}^* \rangle\,. \quad (7)$$

We formally state the equivalence of Eq. (6) and Eq. (7) by Theorem 1 below. We first prove the following lemma.

Lemma 1: Assume $L(\mathbf{g})$ is an increasing function in every variable in $\mathbf{g}$. Suppose the optimal solution of (4) is $(\mathbf{w}^*, \mathbf{g}^*)$. Let $\tilde{\mathbf{g}}^* = (\tilde{\mathbf{g}}_1^*, \tilde{\mathbf{g}}_2^*)$, where $\tilde{\mathbf{g}}_1^* = \left|\mathbf{w}^* \cdot \mathbb{1}_{n_2} - \mathbf{q}_1\right|$ and $\tilde{\mathbf{g}}_2^* = \left|\mathbf{w}^{*t} \cdot \mathbb{1}_{n_1} - \mathbf{q}_2\right|$. Then $(\mathbf{w}^*, \tilde{\mathbf{g}}^*)$ is also an optimal solution of (4). If $L(\mathbf{g})$ is strictly increasing in every variable in $\mathbf{g}$, then $\mathbf{g}^* = \tilde{\mathbf{g}}^*$.

Proof: By the last two inequality constraints in (4),

$$\begin{aligned} \mathbf{g}_1^* &\geq \left|\mathbf{w}^* \cdot \mathbb{1}_{n_2} - \mathbf{q}_1\right| = \tilde{\mathbf{g}}_1^* \\ \mathbf{g}_2^* &\geq \left|\mathbf{w}^{*t} \cdot \mathbb{1}_{n_1} - \mathbf{q}_2\right| = \tilde{\mathbf{g}}_2^* \end{aligned}$$

Since $L(\cdot)$ is increasing in every variable, $L(\mathbf{g}^*) \geq L(\tilde{\mathbf{g}}^*)$. Thus

$$\langle \mathbf{c}, \mathbf{w}^* \rangle + \lambda L(\mathbf{g}^*) \geq \langle \mathbf{c}, \mathbf{w}^* \rangle + \lambda L(\tilde{\mathbf{g}}^*) .$$

Since $\tilde{\mathbf{g}}^* \leq \mathbf{g}^* \leq \mathbf{g}_{up}$, and by construction, $(\mathbf{w}^*, \tilde{\mathbf{g}}^*)$ satisfies the last two inequalities in (4), we conclude that $(\mathbf{w}^*, \tilde{\mathbf{g}}^*)$ is an optimal solution of (4) and $L(\mathbf{g}^*) = L(\tilde{\mathbf{g}}^*)$ must hold. If $L(\cdot)$ is strictly increasing in every variable, $L(\mathbf{g}^*) = L(\tilde{\mathbf{g}}^*)$ and $\tilde{\mathbf{g}}^* \leq \mathbf{g}^*$ imply $\mathbf{g}^* = \tilde{\mathbf{g}}^*$. □

Theorem 1: Assume $L(\mathbf{g})$ *is an increasing function in every variable in* $\mathbf{g}$*. Let the optimal solution of problem (4) be* $(\mathbf{w}^*, \mathbf{g}^*)$*. Then* $\mathbf{w}^*$ *is an optimal solution of problem (2) when the two distributions in (2) are the induced distribution* $\tilde{\mathbf{q}}_1$ *and* $\tilde{\mathbf{q}}_2$ *by* $\mathbf{w}^*$.

Proof: Consider the standard OT problem (2) with marginal constraints: $\mathbf{w} \cdot \mathbb{1}_{n_1} = \tilde{\mathbf{q}}_1$, $\mathbf{w}^t \cdot \mathbb{1}_{n_2} = \tilde{\mathbf{q}}_2$. By the definitions of $\tilde{\mathbf{q}}_1$ and $\tilde{\mathbf{q}}_2$, $\mathbf{w}^*$ is a feasible solution of (2). Suppose $\mathbf{w}^*$ is not an optimal solution of (2) and $\mathbf{w}^\dagger$ is an optimal solution of (2). Then

$$\langle \mathbf{c}, \mathbf{w}^* \rangle > \langle \mathbf{c}, \mathbf{w}^\dagger \rangle \tag{8}$$

Because $\mathbf{w}^\dagger$ and $\mathbf{w}^*$ are feasible solutions of (2),

$$\mathbf{w}^* \cdot \mathbb{1}_{n_2} = \mathbf{w}^\dagger \cdot \mathbb{1}_{n_2} = \tilde{\mathbf{q}}_1$$
$$\mathbf{w}^{*t} \cdot \mathbb{1}_{n_1} = \mathbf{w}^{\dagger^t} \cdot \mathbb{1}_{n_1} = \tilde{\mathbf{q}}_2$$

Hence, $(\mathbf{w}^\dagger, \mathbf{g}^*)$ satisfies the constraints of (4), and thus is a feasible solution of (4). By (8), $(\mathbf{w}^\dagger, \mathbf{g}^*)$ achieves a smaller value of the objective function of (4) than $(\mathbf{w}^*, \mathbf{g}^*)$. This conflicts the assumption that $(\mathbf{w}^*, \mathbf{g}^*)$ is an optimal solution of (4). Hence we conclude that $\mathbf{w}^*$ is an optimal solution of problem (2) with marginal distributions $\tilde{\mathbf{q}}_1$ and $\tilde{\mathbf{q}}_2$. □

Based on Theorem 1, we can interpret OT-RMC as an approach to solve the two induced distributions $\tilde{\mathcal{P}}_1$ and $\tilde{\mathcal{P}}_2$ followed by solving a standard OT using the induced distributions (as replacement for the original $\mathcal{P}_1$ and $\mathcal{P}_2$). This viewpoint helps us decide whether an induced distribution is practically acceptable, which in turn informs us how to set $\mathbf{g}_{up}$ in (4).

C. PRACTICAL ISSUES

The purpose of the loss $L(\mathbf{g})$ is to penalize large deviation from the marginal constraints. Common choices for $L(\mathbf{g})$ can be the p-th power of the L_p norm: $\|\mathbf{g}\|_p^p$, $p \geq 1$. When $p = 1$, OT-RMC is a linear programming. At $p = 2$, OT-RMC is convex quadratic programming. We experimented with both L_1 and L_2 norm, and found that the difference in results is insignificant. With L_1 norm, the solution tends to be more sparse, which we adopt in the paper.

The relaxed marginal constraints in (4) imply that $\mathbf{g} \geq 0$. The upper bound $\mathbf{g}_{up}$ can be set to $+\infty$ if we do not want to bound $\mathbf{g}$ from above. A trivial upper bound naturally satisfied by $\mathbf{g}$ is 1 since $\mathbf{w}$ is a joint probability mass function. The purpose of setting a non-trivial upper bound on $\mathbf{g}$ is to ensure that no point in $\mathbf{X}_1$ or $\mathbf{X}_2$ is assigned with too big or too small a probability according to the induced distributions. For instance, we may require that the marginal probability assigned to any point according to $\mathbf{w}$ cannot exceed twice the original probability on the point given by $\mathbf{q}_1$ or $\mathbf{q}_2$. Then we can set $\mathbf{g}_{up,1} = \mathbf{q}_1, \mathbf{g}_{up,2} = \mathbf{q}_2$. The upper bound $\mathbf{g}_{up}$ also provides the convenience to impose asymmetric restrictions on the gap vectors $\mathbf{g}_1$ and $\mathbf{g}_2$. For instance, if deviation from the marginal constraint is allowed only for the first distribution, but not the second, we can simply set $\mathbf{g}_{up,2} = 0$.

The hyperparameter λ controls the tolerance allowed to deviate from the marginal distributions $\mathbf{q}_1$ and $\mathbf{q}_2$. In the extreme case of $\lambda \to +\infty$, $\mathbf{g} \to 0$, and OT-RMC is reduced to OT. In the other extreme, when $\lambda = 0$, if trivial bounds $\mathbf{g}_{up} = +\infty$ are used, the solution is degenerated to the following. Suppose $c_{i^*,j^*} = \min_{i,j} c_{i,j}$, then $w_{i,j} = 1$ if $i = i^*$ and $j = j^*$, 0 otherwise. We simply have $R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = c_{i^*,j^*}$.

As will be discussed in Section IV-A, OT-RMC can be used to select relevant points from the support sets to address the issue that $\mathbf{X}_l$ (the support of $\mathcal{P}_l$), $l = 1, 2$, are corrupted by noise. Depending on the task at hand, instead of using the induced distributions by $\mathbf{w}^*$ to compute the distance in Eq. (6), it can be preferable to use the truncated conditional distributions of $\mathcal{P}_1$ and $\mathcal{P}_2$. Denote the truncated conditional distributions by $\breve{\mathcal{P}}_1$ and $\breve{\mathcal{P}}_2$. For instance, if the first $n_1' < n_1$ points are selected from $\mathbf{X}_1$, then $\breve{q}_i^{(1)} = q_i^{(1)} / \sum_{i'=1}^{n_1'} q_{i'}^{(1)}$, $i = 1, \ldots, n_1'$. We then use the following distance:

$$D(\mathcal{P}_1, \mathcal{P}_2) = W_{\mathrm{p}}(\breve{\mathcal{P}}_1, \breve{\mathcal{P}}_2) . \tag{9}$$

When OT-RMC is used to select points, based on which distance between two sets is computed, it is feasible to examine a range of values for λ and select the best result based on the given distance. This is intrinsically different from choosing hyperparameters for a machine learning algorithm where the tuning requires an extra validation set, and consequently, the hyperparameters cannot be adjusted for every sample during usage. In the case here, we can select different values of λ for every pair of distributions since the distance between them needs no validation from other data. The only cost of this practice is the increase in computation. It is worthy to point out that the extent of penalty $L(\mathbf{g})$ on the deviation from the marginal constraints is not controlled by λ in an absolute sense. The severity of the penalty depends on the relative value of the transport cost. If the transport cost is scaled, to achieve the same amount of penalty, λ should be scaled accordingly. As a result, we cannot recommend a range of λ values independently from the particular problem. Roughly speaking, smaller variation in transport costs would require smaller λ. On the other hand, we have observed that for the same problem, the range of λ needed to consider for different distributions is stable and the results are often similar for a wide range of λ.

IV. EXPERIMENTS

We illustrate the usage of OT-RMC with three example applications. Although they all follow the general optimization problem (4), the framework is flexible to address the specific

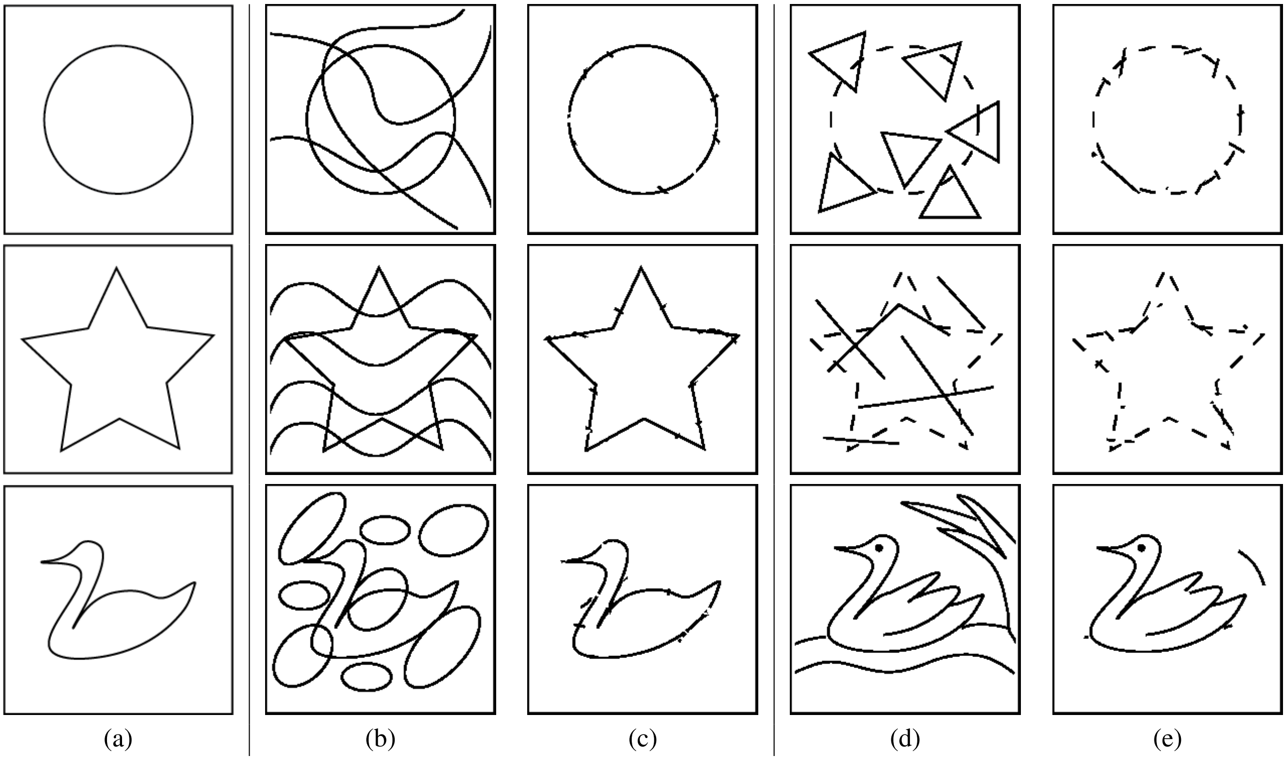


FIGURE 1. Extract patterns from corrupted images. For each target pattern image, two source images and the corresponding extracted patterns by OT-RMC are shown. (a) Target pattern; (b), (c): the first pair of source image and the extracted pattern; (d), (e): the second pair of source image and the extracted pattern.

needs. We will see that for different applications, the amount of relaxation on the marginal constraints of the two distributions can be different, which is realized by setting different bounds on the gap variables.

A. PIXEL PATTERN EXTRACTION

We demonstrate in this experiment that OT-RMC can be used to extract pixel patterns from images. Here, a pixel pattern, referred to simply as pattern when the context is clear, means a collection of pixels in an image plane. As shown in Figure 1, pixel patterns are viewed as black-and-white images, where a black pixel indicates inclusion in the pattern. Suppose the black pixels in an image have coordinates $\mathbf{x}_i = (x_{i,1}, x_{i,2})$, $i = 1, \ldots, n$, where $x_{i,1}$ is the vertical position and $x_{i,2}$ is the horizontal position of the ith pixel. The top left corner of an image plane has coordinate (0, 0). We represent the pattern by the distribution $\mathcal{P} = \{(\mathbf{x}_1, q_1), (\mathbf{x}_2, q_2), \ldots, (\mathbf{x}_n, q_n)\}$, where the probabilities are uniform $q_i = 1/n$, $i = 1, \ldots, n$.

We consider the problem of finding a given pattern from an image. The given patterns, e.g., those in the left column of Figure 1, are called the *target*, whereas, the images, e.g., those in the second and forth columns of the figure, are called the *source*. The source images are also black-and-white and can be represented by distributions over the pixel coordinates. A source image is generally "noisy" in the sense that it contains black pixels other than the pattern. Moreover, the pattern itself may have been altered, e.g., partially missing or with a somewhat different appearance. For example, in column (d) of Figure 1, the embedded circle and star are in dash line, and the drawing of the swan is more detailed.

We first discuss a toy example with results in Figure 1. Then we present a more challenging experiment, in which the target is the hand drawn swan image (the bottom left image of Figure 1) but the source images are edge maps of real photos. The edge maps of photos are harder to handle because there are many background noises and the edges of swans in a picture, if successfully captured by edge detection, can differ a lot from the target pattern.

We apply OT-RMC to extract pixels in the source to best match with the target. The cost function between two support points is the square of the Euclidean distance. In this case, the marginal constraints on the target distribution and the source distribution are treated fully asymmetrically. Without loss of generality, to discuss in the context of Eq. (4), let the source be the first distribution and the target the second. The induced distribution by the matching weight matrix $\mathbf{w}^*$ for the target is fixed at the given distribution, that is, the gap variables for the marginal constraints of the target distribution are set to zero. In problem (4), this is realized by setting $\mathbf{g}_{up,2} = 0$. On the other hand, no upper bound is set on the gap variables corresponding to the source distribution, that is, $\mathbf{g}_{up,1} = +\infty$. We call this problem the *fixed target matching problem*, the properties of which are discussed in Appendix A.

Following the notations in Eq. (4), suppose the source contains n_1 black pixels and the target contains n_2 pixels. After solving OT-RMC, the induced distribution over the source pixels is $\tilde{\mathbf{q}}_1 = \mathbf{w} \cdot \mathbb{1}_{n_2}$. Due to the limited numerical precision, although some elements of $\tilde{\mathbf{q}}_1$ can be nearly zero, we usually do not obtain precise zero. We thus use a non-zero threshold ϵ to select pixels. The threshold ϵ is chosen such that the total probability over the chosen pixels according to $\tilde{\mathbf{q}}_1$ reaches a set level, e.g., 95%, and the probability of each chosen pixel is at least 20% of the probability assigned to a pixel in the target. Let us call the image containing the selected pixels the *extracted pattern*. In Figure 1, the extracted patterns for the source images in the second and forth columns are shown to the right of the source. In each row, the target image is shown in the first column. As demonstrated by the figure, the extracted patterns are crisp even when the embedded patterns vary from the target considerably. The extracted pattern is particularly interesting for the picture of a swan with waves and grass (third row and column (d)). Details are retained, e.g., the eye and the wing, although the swan in the target pattern is simpler and also slightly rotated. The reason is that OT-RMC penalizes deviation from the original distribution of the source and thus tends to keep fidelity to the source.

Although the patterns in the target images seem to be at the same positions as the embedded patterns in the source images, this information is not used in the experiment. Neither OT nor OT-RMC is translation invariant. In our experiments, a target pattern is shifted at a given step size to different locations to match with the source image. At any location, a patch (a rectangle area) from the source image is used to match with the target. The patch is larger than the bounding box of the pattern but smaller than the whole image. We found that it is unnecessary to use too large a patch because OT-RMC will not select pixels that are sufficiently far from those in the target pattern. Reducing the number of pixels in matching can speed up computation substantially. Another approach to reduce computation is to subsample the pixels in the target and the source. At each location, we compute the distance between the target and the patch based on the pixels selected by OT-RMC using Eq (9), and we denote this distance by D_W when discussing the experiments. The extracted pattern from the patch with the minimum D_W is taken as the best match identified in that image. The influence of translation on the extraction of patterns is illustrated by Figure 11 in Appendix B, where more detailed discussion is provided.

To examine the effect of pattern extraction in a more real situation, we experimented with extracting the swan pattern from 254 swan photos downloaded from Google's image site. For these real photos, the swans in the pictures can be surrounded by a highly noisy background, be rotated, or have very different sizes. The target swan pattern is still the image in the third row and first column of Figure 1. As the target swan image is of size 230 × 230, the real photo images were scaled such that the shorter of the two sides of an image is 230. The color images were then converted to grayscale, and Canny edge detection was applied to obtain edge maps. Pixels identified as edge are the black pixels in our setting, and the rest are the white background. Example images and their edge maps are shown in Figure 2. We generated 3 swan target patterns by scaling the support points (i.e., the coordinates of black pixels) in the target distribution at three levels: 1.0, 0.85, 0.65. As the majority of the patches do not match well with any of the target, for clarity of the study, we only select at most one patch from each image across all the three scale target patterns to conduct extensive experiments and make comparisons with other approaches. Details about how the patches are selected are presented in Appendix C. Among the 254 images, 137 images have a patch selected. Our results below are based on these 137 patches.

We compare OT-RMC with two other approaches: the regular OT (with strict marginal constraints) and the so-called nearest neighbor (NN) match. For NN match, each pixel in the target is matched with the closest pixel in the source (according to the Euclidean distance of their coordinates). In our setup, as the marginal constraints on the target are strict and there are no bound constraints on the source distribution, NN match is equivalent to OT-RMC at $\lambda = 0.0$. We prove this fact in Appendix A. In the discussion below, OT-RMC means OT-RMC with $0 < \lambda < +\infty$. If $\lambda = +\infty$, OT-RMC becomes the regular OT. We compare OT-RMC with NN and OT to demonstrate its effectiveness at selecting pixels to match with the target. Note that for OT, no selection of pixels is performed. To compare on a common ground, we use OT-RMC and NN only to select pixels. Once the pixels are selected, uniform probabilities are assigned to them to compute various distances. To remove the effect of translation or rotation on the distance, we compute the *Translation-Rotation Adjusted Wasserstein distance* (TRA-Wasserstein), denoted by D_{TRA}, between the selected subset of pixels and the target pattern. This is motivated by the observation that translation and rotation (at least to a moderate extent) do not affect our perception of a pixel pattern. We also compute a so-called *angle distance*, denoted by D_{ANG}. Definitions of these two distances are described in details in Appendix C.

For OT-RMC, we experimented with $\lambda = 0.05$, 0.1, 1.0, 5.0, 10.0, 20.0. For each patch, we report the minimum D_{TRA} or D_{ANG} obtained by OT-RMC across the λ's. In Figure 2(a)-(b), we compare OT-RMC, NN, OT in terms of D_{TRA} and D_{ANG}. For clarity of visualization, we sorted the images according to D_{TRA} (or D_{ANG}). As shown in the plots, OT yields higher D_{TRA} and D_{ANG} than OT-RMC across all the images. In fact, the difference is remarkable. For the vast majority of images, NN performs between OT and OT-RMC. The histograms for the best λ values for either D_{TRA} and D_{ANG} are shown in Figure 2(c). When NN outperforms OT-RMC, $\lambda = 0$ is the best. In Figure 2(d), we show the running time to finish OT-RMC with respect to the support sizes of the source distributions. The target distribution has a fixed support size of 223. A quadratic regression curve is fitted to the running time in seconds. We can see that the variation in running time (e.g., due to different λ's) at any fixed support size of the source distribution is small. In Figure 3, we show example

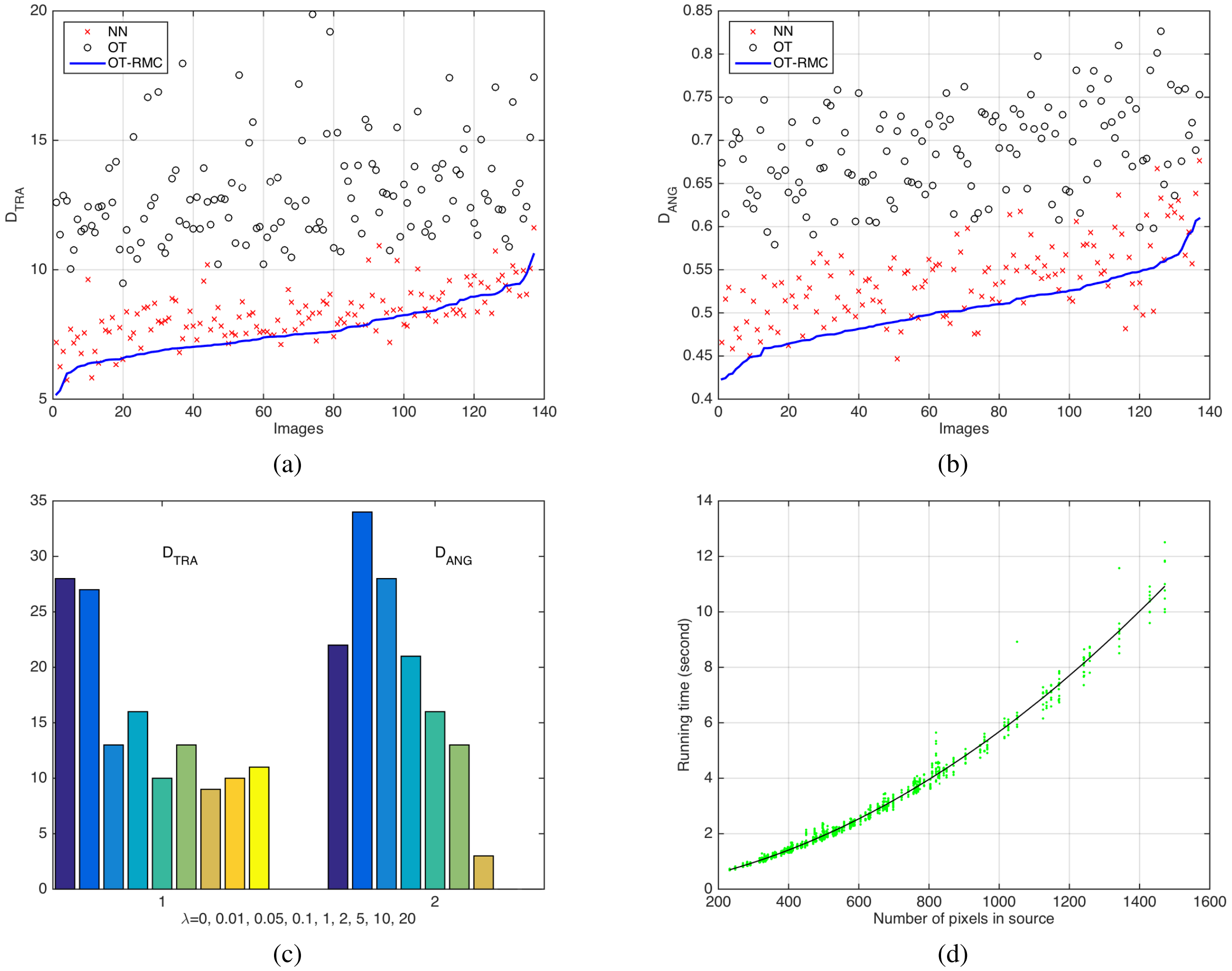


FIGURE 2. Compare the performance of OT-RMC, OT, and NN in the pixel pattern extraction experiment. (a) D_{TRA} for 137 source images. (b) D_{ANG} for the source images. (c) Histograms for the best chosen λ to minimize D_{TRA} and D_{ANG} respectively. (d) Running time in seconds. The running time increases approximately in quadratic order with respect to the number of pixels in the source image. The black line is a fitted second order polynomial.

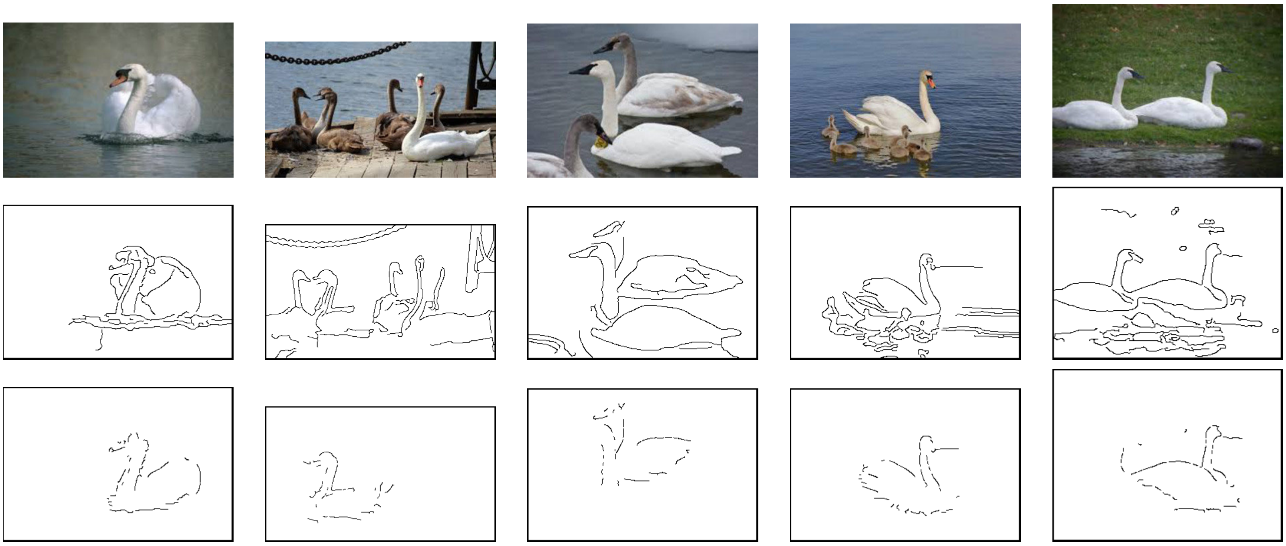


FIGURE 3. Example images, their edge maps, and the extracted patterns in the pixel pattern extraction experiment.

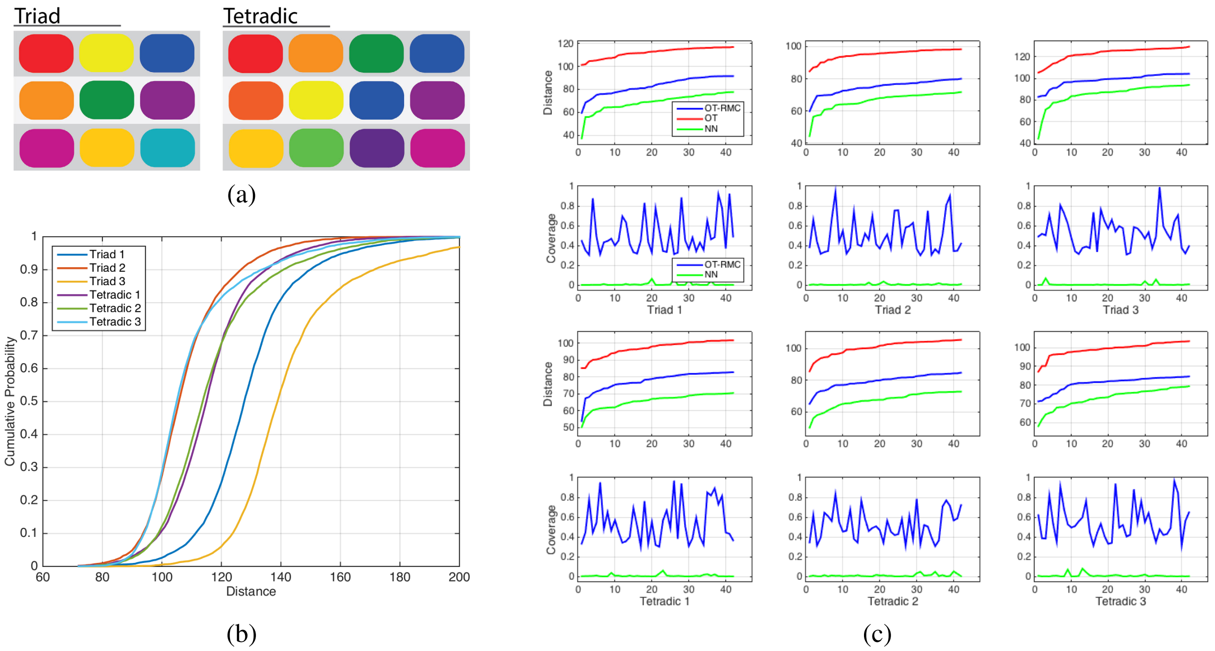


FIGURE 4. Results for the color scheme matching experiment. (a) The six color schemes in the study. (b) The cumulative distribution of D^*_W obtained by OT-RMC for each color scheme. (c) Comparison of the distances and coverage ratios obtained by OT-RMC, NN, and OT. For each color scheme, the top plot shows the 1% lowest distances obtained by each method over the 4, 238 paintings; the bottom plot shows the coverage ratios obtained by OT-RMC and NN for the corresponding images in their respective 1% set.

source images, their edge maps, and the extracted patterns by OT-RMC. For each image, only the best matched extracted pattern is shown.

B. COLOR SCHEME MATCHING

We now demonstrate an application of OT-RMC for identifying color schemes used in paintings. A color scheme is the co-existence of multiple colors that are positioned on the color wheel in a particular way. A color wheel represents the hue of a color by the angular position ranging from 0° to 360°. Examples include triad color schemes (three colors equally apart in angles on the color wheel), analogous color schemes, and complementary color schemes [27]. In this experiment, we examine six color schemes shown in Figure 4(a), three of which are triad schemes and the other three tetradic. For a triad scheme, the three hues are 120° apart, e.g., red, blue, and yellow. A tetradic scheme contains four color hues that form a rectangle on the color wheel. Each hue is 60°, 120°, and 180° apart from the other three hues.

Our dataset contains 4, 238 fine art paintings of over 30 artists, e.g., Michelangelo, Caravaggio, Vermeer, Van Gogh, Gauguin, Cezanne, Konchalovsky, Matisse. To summarize the colors used in an image, the 3D color vectors at each pixel containing red, green, and blue color components, are clustered by the modal clustering algorithm [28]. Then the representative colors and the proportions of pixels in each cluster form a color distribution for the image. On average, about 77 distinct colors are extracted for every painting. In this study, the color distribution of a painting is treated as the source, while the color distribution derived from every color scheme is the target. The target distribution for a color scheme is generated by assigning uniform probabilities to the colors contained in the scheme. As mentioned previously, we experimented with six color schemes shown in Figure 4(a). The cost function between two support points is the Euclidean distance. We did not use the square of the Euclidean distance because it tends to exaggerate the difference when two color vectors are not close.

Different from the pixel pattern extraction problem studied in the previous subsection, we do not impose strict marginal constraints on the target distribution in this problem. The induced distribution of the target is allowed to differ from the pre-set uniform distribution. However, the deviation is bounded. Specifically, we allow the gap variables for the marginal constraints of the target distribution to be non-zero, but upper bound them to ensure that each color in a scheme is guaranteed with a minimum value of probability. This setup is motivated by the fact that when a set of colors are used in a painting to achieve the visual effect of a color scheme, it is unnecessary to cover the same amount of area by each color. On the other hand, every color in the scheme needs to account for an adequate amount of proportion in order not to be ignored. In the extreme case, if a color has zero proportion, the color scheme essentially no longer exists. In particular, we set $\mathbf{g}_{up,2} = \delta$. For the triad schemes, we set $\delta = \frac{2}{3} \cdot \frac{1}{n_2}$. This implies that in the given color scheme, each color is assigned with a probability of at least 11%. For the tetradic schemes, we set $\delta = \frac{2}{5} \cdot \frac{1}{n_2}$, ensuring that each color has probability no less than 15%. The minimum probability per color is set higher for the tetradic schemes than triad schemes because a

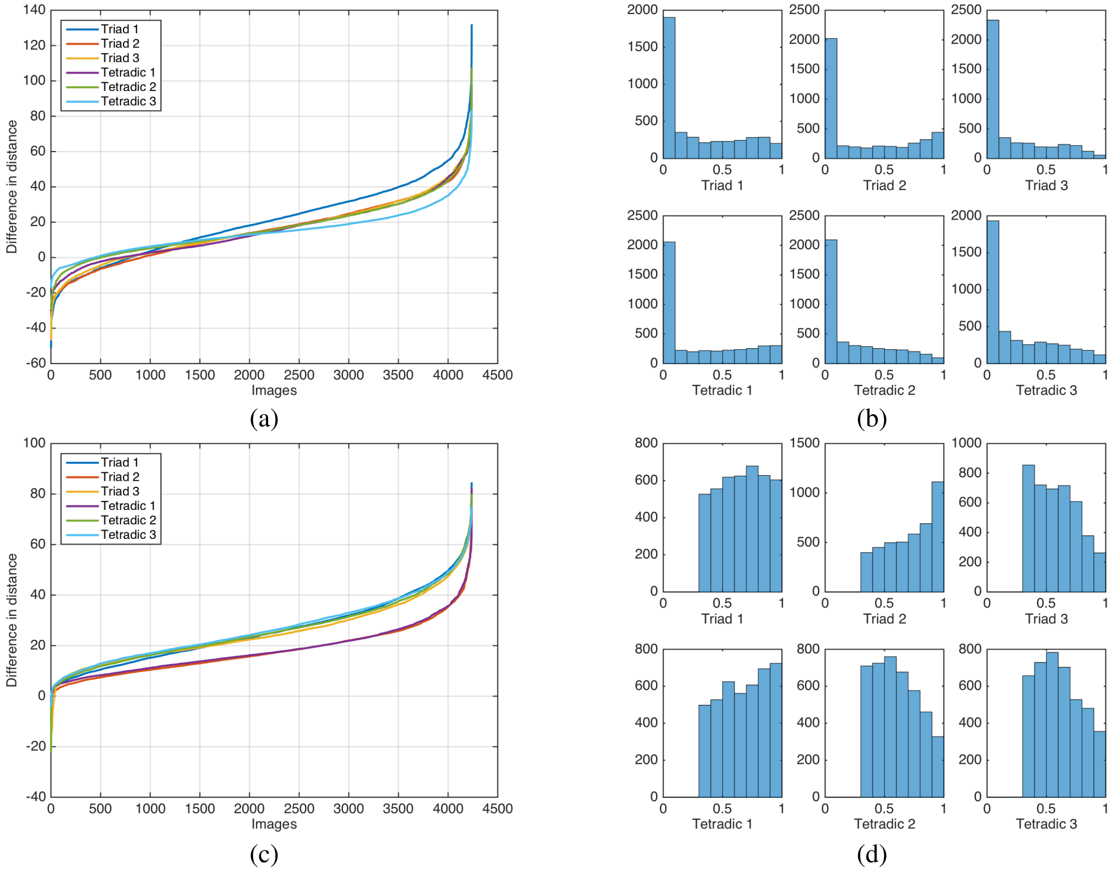


FIGURE 5. Compare the performance of OT-RMC, OT, and NN for the color scheme matching experiment. (a) Difference in distance between NN and OT-RMC, that is, $D_W^{NN} - D_W^*$. (b) Histograms for the coverage ratios obtained by OT-RMC at the λ that yields D_W^*. Every panel corresponds to one color scheme. (c) Difference in distance between OT and OT-RMC under the condition that the coverage ratio obtained by OT-RMC at the chosen λ is at least 30%, that is, $D_W^{OT} - \tilde{D}_W^*$. (d) Histograms for the coverage ratios obtained by OT-RMC at the λ that yields $\tilde{D}_W^*$.

tetradic scheme may appear like a triad scheme unless all the four colors are significant in proportion.

Again, we use OT-RMC and NN to choose colors from a source image. After the colors have been selected from a source, the source distribution is changed to the truncated conditional distribution on the selected colors. We then compute the Wasserstein distance, denoted by D_W, between the source distribution and the induced target distribution. We also compare the methods based on *coverage ratio*, which is defined as the total proportion of the selected colors. If the coverage ratio is very low, e.g., 5%, even if the match with a color scheme is nearly perfect, the painting cannot effectively convey the impression of the color scheme. In another word, for viewers to perceive a color scheme in a painting, the area occupied by the color scheme must be large enough.

One by-product of OT-RMC is the induced target distribution, which is optimized to best capture the proportions of the target colors that actually appear in the source image. In contrast, NN or OT cannot dynamically determine the target distribution. Hence the target distribution used for OT or NN will remain as the original uniform distribution. For OT-RMC we tested $\lambda = 0, 5, 20, 35, 50, 52, 53, 55, 65$. We would like to remark that as the marginal constraints on the target distribution are not strict in this problem, OT-RMC at $\lambda = 0$ is not equivalent to NN because the induced target distribution can be different from the original distribution.

To compare with NN, we compute the minimum D_W distance obtained by OT-RMC across the λ's, denoted by D_W^*. We also record the coverage ratio obtained at the optimal λ chosen for each image. We then compute the difference between D_W^{NN}, the D_W distance obtained by NN, and D_W^* by OT-RMC: $D_W^{NN} - D_W^*$. Figure 5(a) shows the sorted difference in distance. A negative value indicates that NN achieves smaller distance. The figure shows that for the vast majority of the images, for every color scheme, OT-RMC achieves lower distance. Moreover, the coverage ratio by NN is too low to be practically useful. On average, colors selected by NN only cover 2% to 5% of an image for any of the six color schemes. As mentioned previously, when the coverage ratio is too small, a good match with a color scheme extracted from a source image is not really meaningful. In light of this, NN has failed to yield useful results for a very high percentage of

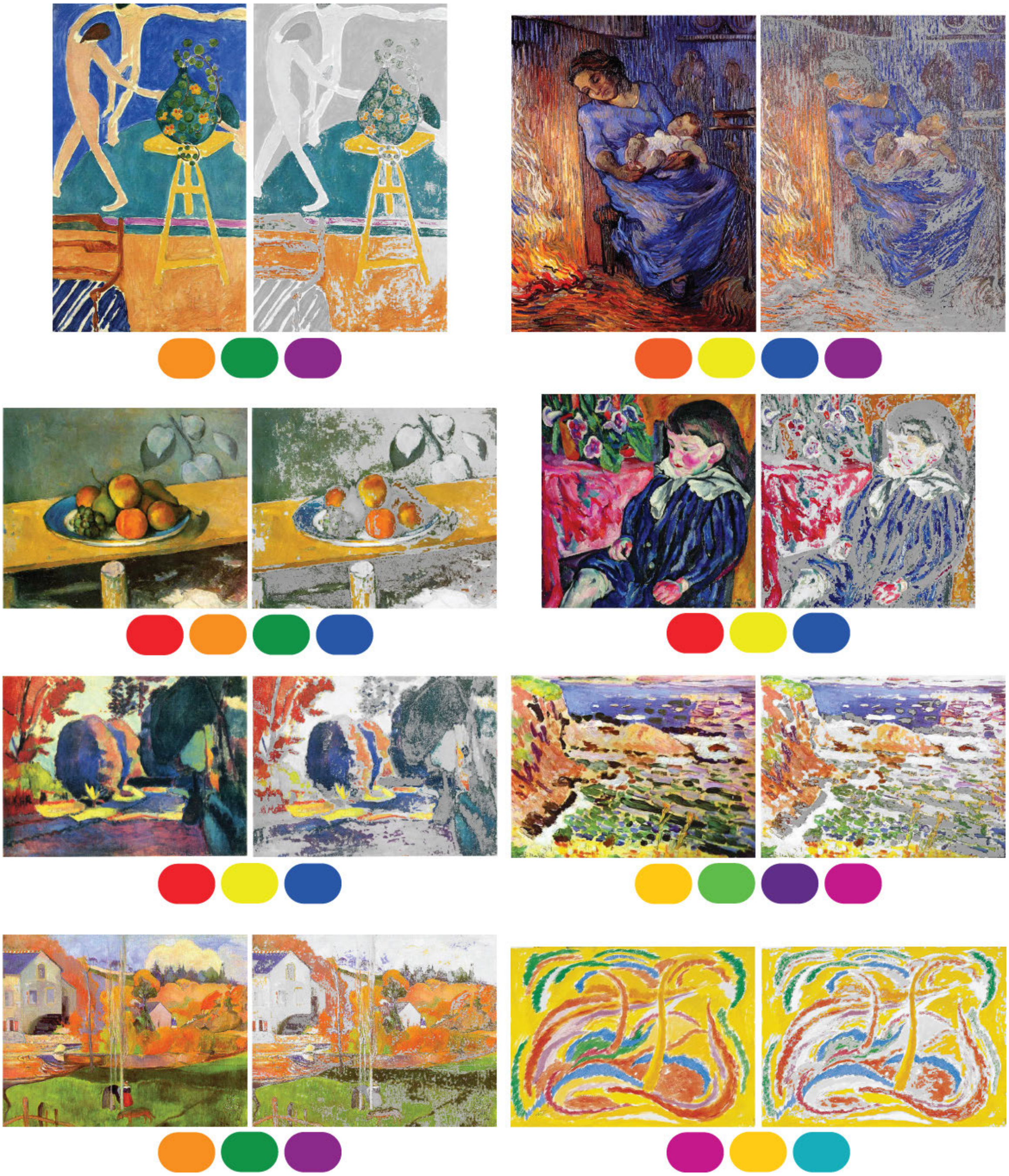


FIGURE 6. Example paintings and their well-matched color schemes. The original image is shown on the left. The image on the right shows the pixels selected by OT-RMC for a particular color scheme (shown below the image) in color while those not selected in gray scale.

paintings. We show the histogram of the coverage ratio by OT-RMC in Figure 5(b). As we can see, the percentage of images with very low coverage ratios is not negligible if we select λ based only on D_W. Next, we select the best λ by minimizing D_W under the condition that the coverage ratio is at least 30%. The corresponding distance at the chosen λ is denoted by $\tilde{D}_W^*$. We then compare $\tilde{D}_W^*$ with D_W^{OT}, the D_W distance obtained by OT. We find that, for any of the six color schemes, the best λ values are among the three values $[20, 35, 50]$ for 75% to 90% of the images. Figure 5(c) shows the sorted difference in D_W by OT and $\tilde{D}_W^*$. Again, regardless of the color scheme, for the vast majority of the images, OT-RMC achieves lower distance. The histograms of the coverage ratios obtained at the best λ that yields $\tilde{D}_W^*$ are shown in Figure 5(d). Every plot corresponds to one color scheme. By setup, the coverage ratio is at least 30%.

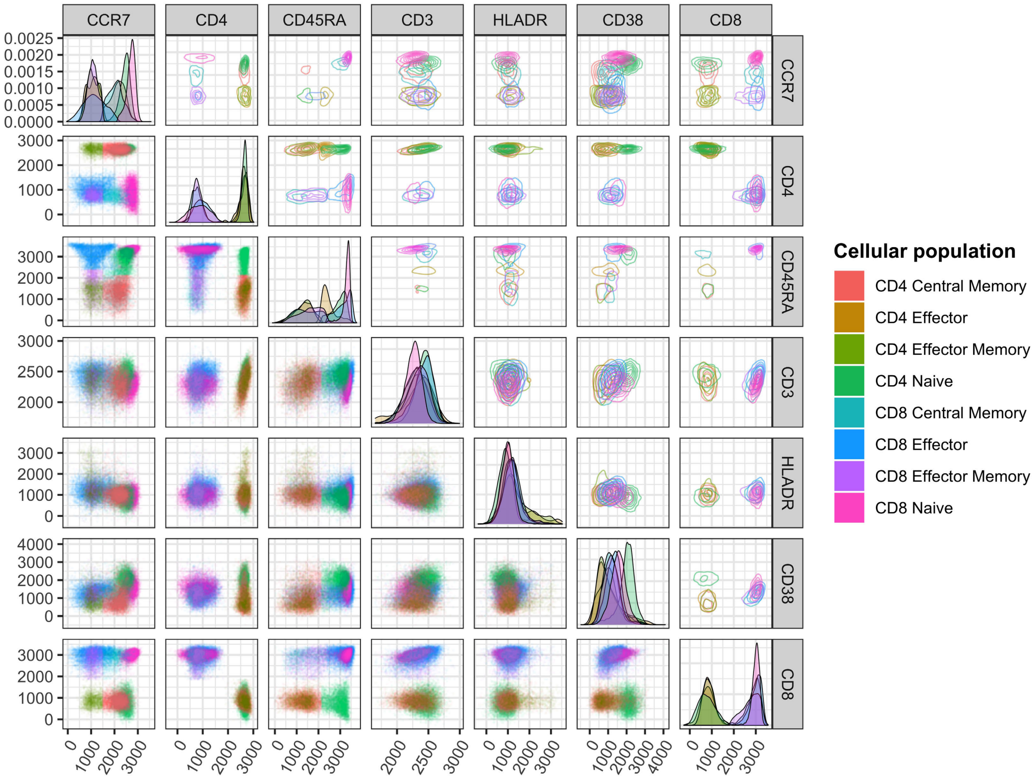


FIGURE 7. Visualization of 1228R1 FCM data which contain 30, 427 cells. Eight manually gated cell populations are color coded. The diagonal plots show the estimated marginal density functions of each cellular marker (i.e., feature). The lower triangle plots show 2D scatter plots of the cells, and the upper triangle plots show the estimated bi-variate densities. In every plot, visualization is provided for the eight clusters simultaneously which are distinguished by different colors.

The histograms show that the average coverage ratio is always above 50%.

In Figure 4(b), the cumulative probability distribution is shown for D_W^* for every color scheme. In Figure 4(c), we compare OT-RMC, OT, and NN in the following way. For each of the three methods, we select the lowest 1% of the distances (a set of 42) over all the images. As these 42 images are generally different for different methods, we cannot interpret the difference in any measure as performance difference on any particular image. For OT-RMC, $\tilde{D}_W^*$ is used for selection (thus guarantee coverage ratio of at least 30%). For each color scheme, a pair of plots (arranged vertically) are shown. The top plot shows the sorted distances separately for each method. The bottom plot shows the coverage ratios obtained by OT-RMC and NN for their respective set of 42 images. As OT does not select colors, the coverage ratio is always 1, thus not shown in the plots. We see that for each color scheme, the lowest distances by OT are considerably higher than that of OT-RMC and NN. The difference between the latter two is relatively small comparing with OT. However, there is significant difference between the coverage ratios of OT-RMC and NN. For OT-RMC, the coverage ratio is ensured to be at least 30%, with the average in the range of [0.50, 0.55]. For NN, the coverage ratio is nearly zero for a very high percentage of images. Specifically, for the six color schemes, the percentage of images with coverage ratio below 2% by NN ranges from 88% to 95%. Hence, although NN yields lower distances on its lower end images, the coverage ratios are too small to be useful.

In Figure 6, we show some example paintings and the color schemes that matched well with the paintings. Here, a good match means the distance yielded by OT-RMC is small. The color scheme is shown below the images. To show the pixels that are selected by OT-RMC, meaning that the representative colors of the clusters which the pixels belong to are selected, we show these pixels in the original color, while the pixels not selected are shown in brightened gray scale. It is interesting that the selected pixels help us better see a certain color scheme by separating the scheme from the rest of the painting. This separation helps viewers focus their attention on the usage of certain colors, which might be overlooked in a painting rich in color, and perhaps better understand and appreciate the artistic choices of the painter.

C. CLUSTER ALIGNMENT FOR SINGLE-CELL DATA

Single-cell technologies, including flow cytometry (FCM), Cytometry by Time-Of-Flight mass spectrometry (CyTOF), single-cell RNA-seq, have revolutionized biology through transcriptomic profiling at the single-cell level [29]. Single-cell analysis is key to a better understanding of cellular

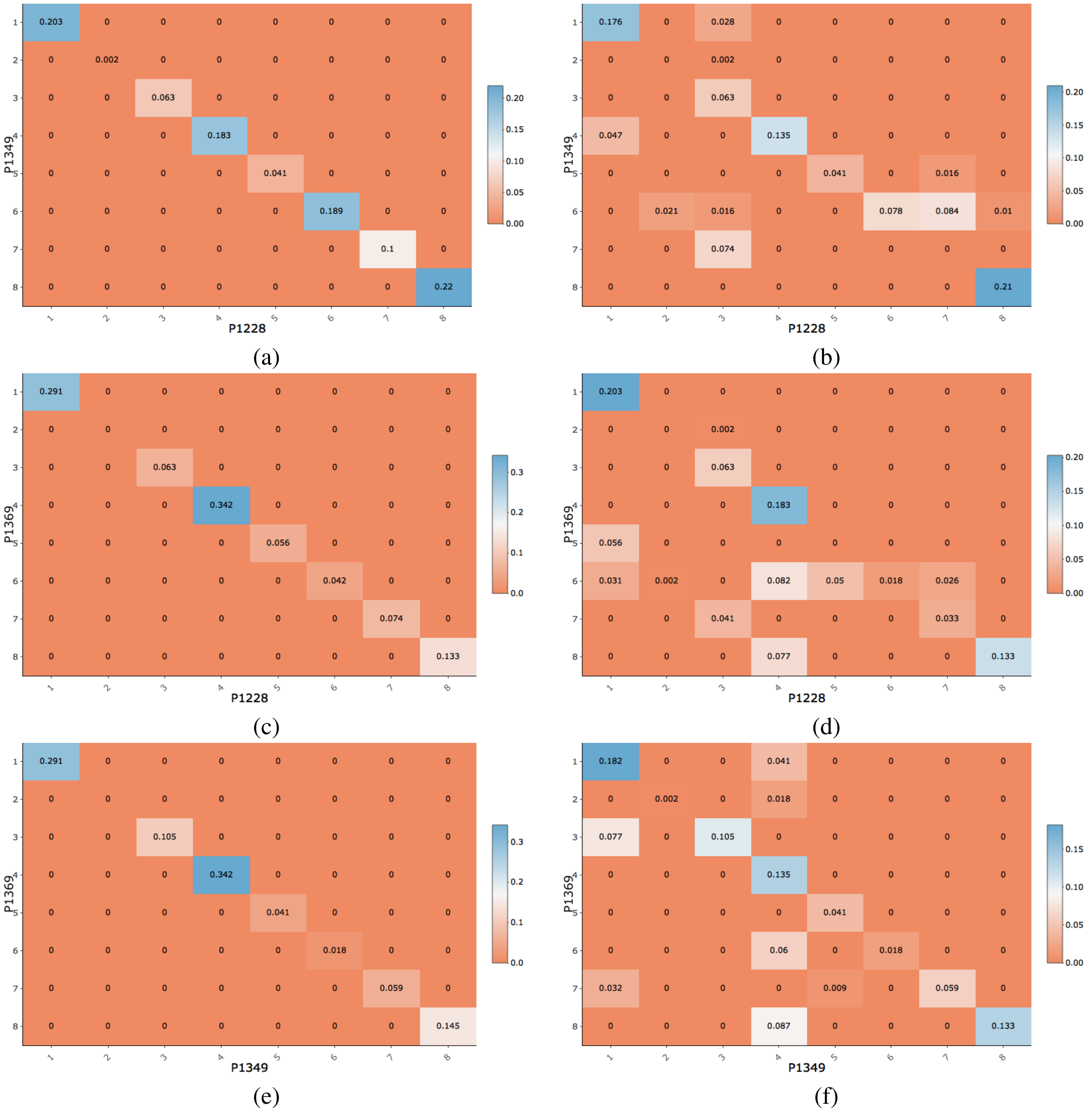


FIGURE 8. Compare OT-RMC with OT for cluster alignment of the three FCM datasets. Results for OT-RMC are shown in the plots (a), (c), and (e) on the left, while those for OT are (b), (d), and (f) on the right. Each plot is a heatmap with every box corresponding to one entry in the matching weight matrix. The value of each entry is indicated in its box. The two datasets aligned in each plot are indicated along the horizontal and vertical axis.

plasticity, stem cell biology, immunology, and cancer heterogeneity. Clustering is one of the most commonly applied analyses of single-cell RNA-seq data. In various disease studies and vaccine development, cell clusters identified computationally help reveal different patterns across conditions (e.g., healthy vs. diseased; placebo vs. vaccinated; different tumor samples), based on which existing biomedical conjectures may be substantiated, or new hypotheses/experiments on potentially important biomarkers, disease monitoring and treatments, and vaccine development may be inspired [30]–[37].

One key challenge encountered by single-cell clustering analysis is how to relate clustering results for datasets acquired from multiple sources. For reasons such as privacy of data, limitation of communication systems, and storage restrictions, it can be impractical or impossible to access all the datasets at a central site for simultaneous analysis. It is thus necessary to consolidate clustering results obtained from different datasets. We call this the cluster alignment problem. We apply OT-RMC with no upper bounds on the gap variables. Different from the first two applications, the two distributions in problem (4) are treated symmetrically here.

To demonstrate the application of OT-RMC to cluster alignment across different datasets, we first analyze a set of three FCM samples. These three samples correspond to the first replicates of Stanford center in the T-cell Lyoplate

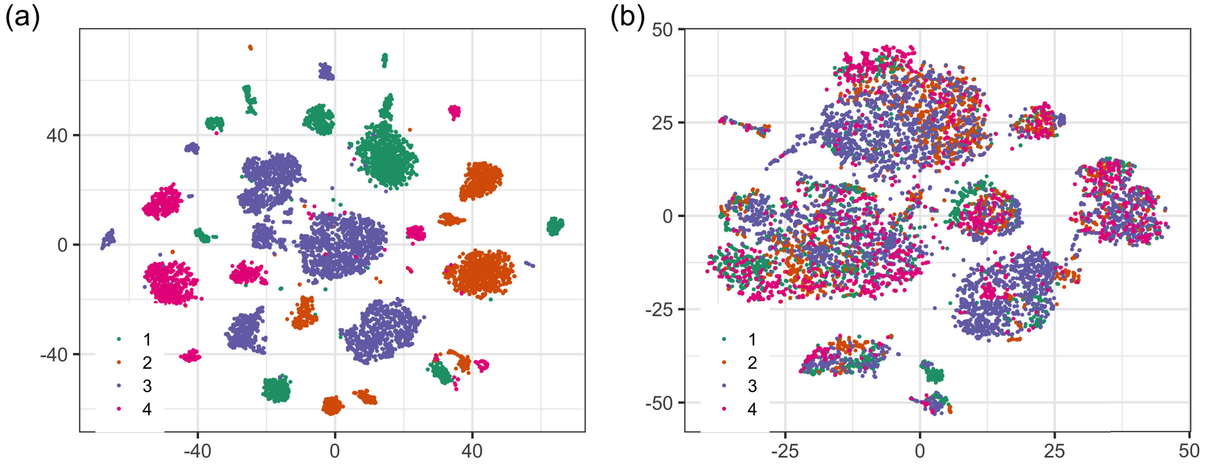


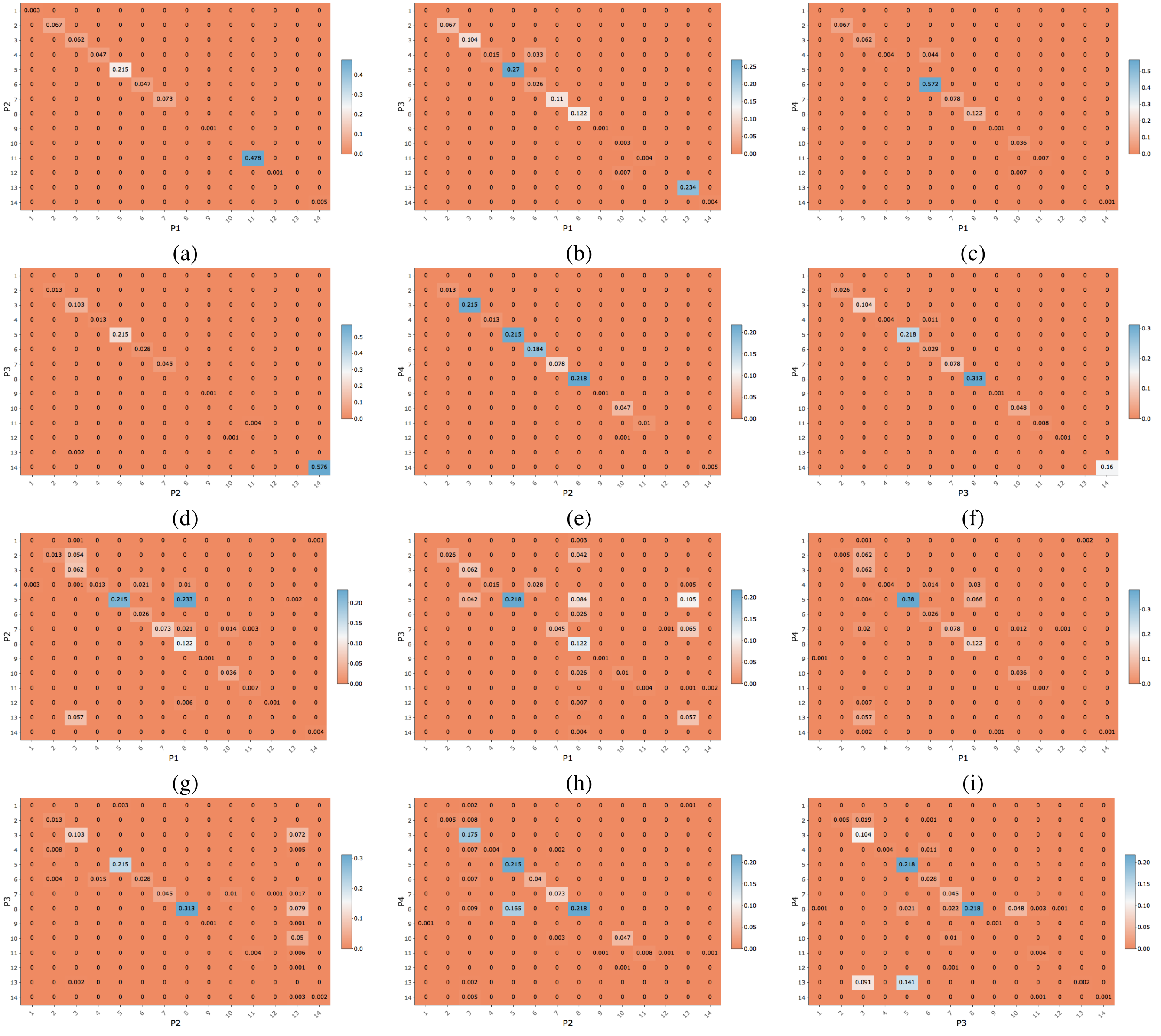
FIGURE 9. t-SNE plot of individual cells color coded by four human donors. (a): t-SNE is performed on the raw data; (b): t-SNE is performed on the 13-dimensional canonical vectors output from Seurat CCA.

panel of the SeraCare cell HIPC study for each of the three available patients denoted "P1228R1", "P1349R1" and "P1369R1" [38]–[40]. The original raw data files are available on the immunspace platform. In addition, we have a reference manual gating (i.e., manually identified) of those cells for each patient into 8 mutually exclusive cell populations. We first remove dead cells and doublets to focus on manually gated cells. The resulting datasets P1228R1, P1349R1 and P1369R1 contain 30, 427, 31, 228 and 32, 948 cells, respectively. In addition, each cell was characterized by 7 cellular markers (i.e. features), namely CCR7, CD4, CD45RA, CD3, HLADR, CD38, and CD8. Figure 7 shows a descriptive representation of the 1228R1 FCM data. The proportion of the largest cell population (aka, cell cluster) is about 30% and the smallest being 0.2%. The proportion of each cell cluster can vary dramatically across the three patients. For instance, for any cluster across two patients, suppose the proportions of this cluster are a_1 and a_2 respectively. We compute the ratio $|a_1-a_2|/\max(a_1, a_2)$, which is referred to as the *disparity ratio of proportions*. This ratio ranges from 4.5% to 91.6% with median 38.2%, the first quartile 24.6%, and the third quartile 61.6%. The large variation in the proportions of the cell clusters poses difficulty in alignment.

In this example, the 8 cell populations are one-to-one matched across the three datasets according to the manually gated results (ground truth). We will examine the alignment results based on OT-RMC and OT, specifically, the matching weight matrices obtained by each, and see whether the matrices indicate one-to-one match between the clusters. To perform OT-RMC, we first represent the cell populations for each dataset by the distribution $\mathcal{P}_k = \{(\mathbf{x}_1^{(k)}, q_1^{(k)}), (\mathbf{x}_2^{(k)}, q_2^{(k)}), \ldots, (\mathbf{x}_8^{(k)}, q_8^{(k)})\}$, $k = 1, 2, 3$, where the probabilities are estimated by the empirical proportions of the cell populations. The support point for the ith cluster, $\mathbf{x}_i^{(k)} \in \mathbb{R}^7$, is the mean vector for cell population i in the kth dataset. Since there is no target distribution (dataset) in this example, we perform pairwise alignment among the three datasets. We use the square of Euclidean distance as cost between two support points and set $\lambda = 0.5$ for OT-RMC. Figure 8 displays the matching weight matrices for both OT-RMC and OT as heatmaps. Figure 8(a), (c) and (e) show that the resulting matching weight matrices by OT-RMC are all diagonal, indicating that the 8 cell populations are one-to-one matched across any pair of datasets. In contrast, OT tends to split or merge several cell populations from one dataset and align them with a cluster from another dataset. In fact, even if we apply post-processing to the matching weights to enforce one-to-one matching, OT would still result in some wrong matches. For example, cluster 3 in P1228 would match with cluster 7 in P1349 since either would be the best-matched cluster of the other. Similarly, cluster 7 in P1228 would match with cluster 6 in P1349.

Next, we demonstrate OT-RMC with a more challenging example. The datasets are single-cell RNA-seq of pancreatic islets and they contain gene expression data from the same tissue of four human donors with 20215 genes (features) detected in total [41]. All four datasets contain the same 14 cell populations which are identified independently based on a droplet-based single-cell RNA-seq method called inDrop. The proportion of the largest cell population is about 45% and the smallest being 0.07%, a very wide range. In addition, comparing with the FCM datasets, the same cell population can differ more dramatically in proportion across the four datasets. The disparity ratio of proportions ranges from 1.4% to 99.3% with median 50.4%, the first quartile 28.8%, and third quartile 72.8%.

Figure 9(a) shows the two-dimensional t-SNE plot [42] of single-cell RNA-seq data with cells color coded by the four human donors. Further, it shows that cells tend to cluster not only by cell type but also by donor, which suggests the existence of batch effects among the four datasets. Thus, we first apply Seurat CCA (Canonical Correlation Analysis) [43] to remove batch effects before we align clusters across these four datasets. Seurat CCA performs dimension reduction and



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

FIGURE 10. Compare OT-RMC with OT for cluster alignment of the four single-cell RNA-seq datasets. Results for OT-RMC are shown by (a)-(f) in the top two rows, while those for OT are (g)-(l) in the bottom two rows. Each plot is a heatmap with every box corresponding to one entry in the matching weight matrix. The value of each entry is indicated in its box. The two datasets aligned in each plot are indicated along the horizontal and vertical axis.

aligns the subspaces of different datasets using dynamic time warping. Figure 9(b) shows the t-SNE plot obtained from the 13-dimensional canonical vectors which are output by Seurat CCA. It verifies that cells are no longer grouped by donor. Next, as with the FCM datasets, we represent the cell populations for each donor by a discrete distribution over the derived 13-dimensional canonical vectors. Every support point corresponds to one cluster and is given by the average of the 13-dimensional vectors in that cluster. Comparing with the FCM datasets, the cost given by the square of Euclidean distance is much smaller between any of the clusters. Thus we use L_1 norm as cost instead, and set $\lambda = 0.01$. In Figure 10, the top two rows show the pairwise matching weight matrices obtained by OT-RMC, and the bottom two rows show the results of OT. The matching weight matrices by OT-RMC are all nearly diagonal, indicating one-to-one correspondence of cell populations across the four donors. However, OT again fails to detect the one-to-one correspondence between many clusters.

V. CONCLUSION

In this paper, we propose the framework of OT-RMC, which is motivated by intrinsic limitations of OT in applications. OT-RMC can be solved efficiently by linear programming.

The framework is flexible for a range of application scenarios. OT-RMC enables us to extract patterns that are embedded with noise and to better handle the case of varying object proportions when aligning two sets. We illustrate the usage of OT-RMC with three example applications, each with a different setup of the optimization problem. These examples show the diverse usages of OT-RMC and its advantages over OT and the nearest neighbor matching scheme. Given that it is an extension of OT, we expect OT-RMC to have wide potential applications.

APPENDIX A PROPERTIES OF THE FIXED TARGET MATCHING PROBLEM

Let the source distribution be $\mathcal{P}_1$: $\{(\mathbf{x}_i^{(1)}, q_i^{(1)}), i = 1, \ldots, n_1\}$ and the target distribution be $\mathcal{P}_2$: $\{(\mathbf{x}_i^{(2)}, q_i^{(2)}), i = 1, \ldots, n_2\}$. For the pixel pattern extraction application in Subsection IV-A, we use OT-RMC formulated by problem (4) with $\mathbf{g}_{up,1} = +\infty$ and $\mathbf{g}_{up,2} = 0$. Let the support sets of the two distributions be $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. This case of problem (4) can be stated equivalently as the following problem:

$$\begin{aligned} R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \underset{\mathbf{w},\mathbf{g}}{min} &\langle \mathbf{c}, \mathbf{w}\rangle + \lambda L(\mathbf{g}_1) \qquad (10)\\ s.t.\ &\mathbf{w} \geq 0\\ &\mathbb{1}_{n_1}^t \cdot \mathbf{w} \cdot \mathbb{1}_{n_2} = 1\\ &\mathbf{q}_1 - \mathbf{g}_1 \leq \mathbf{w} \cdot \mathbb{1}_{n_2} \leq \mathbf{q}_1 + \mathbf{g}_1\\ &\mathbf{w}^t \cdot \mathbb{1}_{n_1} = \mathbf{q}_2 \end{aligned}$$

We call Eq. (10) the *fixed target matching* problem. The marginal constraints on the target distribution is strict, but not so on the source.

In NN matching, each support point in the target is matched with the closest point in the source according to the definition of the cost $\mathbf{c}$, e.g., $c_{i,j} = \|\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)}\|^2$ with $\|\cdot\|$ being the L_2 norm. Suppose $\mathbf{x}_j^{(2)}$ is matched to $\mathbf{x}_{\nu^*(j)}^{(1)}$ in the source by NN. We define the matching weight matrix of NN, denoted by $\mathbf{w}^{NN}$, as follows. For the (i, j)th element, let

$$w_{i,j}^{NN} = \begin{cases} q_j^{(2)} & i = \nu^*(j) \\ 0 & i \neq \nu^*(j). \end{cases}.$$

Clearly $\mathbf{w}^{NN}$ is a joint distribution on $\mathbf{X}_1 \times \mathbf{X}_2$. Thus the NN induced distribution on the support of the source distribution is given by

$$\tilde{q}_i^{(1),NN} = \sum_{j:\nu^*(j)=i} q_j^{(2)}, \quad i = 1, \ldots, n_1 .$$

If a support point in the source is not matched with any point in the target, its induced probability is 0, otherwise, its induced probability is the sum of the probabilities on all its matched points in the target.

Theorem 2: The fixed target matching problem (10) at $\lambda = 0$ *is solved by NN matching weight matrix* $\mathbf{w}^{NN}$.

Proof: Let the matching weight matrix of the fixed target problem at $\lambda = 0$ be $\mathbf{w}^* = (w_{i,j}^*)_{i=1,\ldots,n_1, j=1,\ldots,n_2}$. Let the induced distribution by $\mathbf{w}^*$ on the source be $\tilde{\mathbf{q}}^{(1)}$.

The objective function of problem (10) becomes

$$\begin{aligned} R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) &= \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} w_{i,j}^* c_{i,j}\\ &\geq \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} w_{i,j}^* \cdot \min_{i'} c_{i',j}\\ &= \sum_{j=1}^{n_2} \min_{i'} c_{i',j} \cdot \sum_{i=1}^{n_1} w_{i,j}^*\\ &= \sum_{j=1}^{n_2} c_{\nu^*(j),j} \cdot q_j^{(2)} \qquad (11) \end{aligned}$$

The last equality comes from the strict marginal constraints on the target. If we set $\mathbf{w}^* = \mathbf{w}^{NN}$, then

$$R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2) = \sum_{j=1}^{n_2} c_{\nu^*(j),j} \cdot q_j^{(2)} .$$

That is, the lower bound on $R(\mathbf{c}, \mathbf{q}_1, \mathbf{q}_2)$ in Eq. (11) is achieved by $\mathbf{w}^{NN}$. If we define $\mathbf{g}_1 = \left|\mathbf{w}^{NN} \cdot \mathbb{1}_{n_2} - \mathbf{q}_1\right|$, obviously, $\mathbf{w}^{NN}$ and $\mathbf{g}_1$ satisfy all the constraints in problem (10). Thus $\mathbf{w}^{NN}$ solves the fixed target matching problem at $\lambda = 0$. □

Recall that D_W is the Wasserstein distance solved by OT between the fixed target distribution and the truncated conditional distribution of the source restricted to the selected pixels. We denote the conditional distribution by $\check{\mathcal{P}}_1$: $\{(\mathbf{x}_i^{(1)}, \check{q}_i^{(1)}), i = 1, \ldots, n_1\}$. Let the corresponding optimal matching weight matrix solved by OT be $\mathbf{w}^*$. For the p-Wasserstein distance, the cost is the pth power of a norm in $\mathbb{R}^d$, $c_{i,j} = c(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}) = \|\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)}\|^{\mathrm{p}}$, and $D_W^{\mathrm{p}} = \langle \mathbf{c}, \mathbf{w}^*\rangle$. We define a lower bound D_{LB} based on NN matching for distance D_W as follows:

$$D_{LB}^{\mathrm{p}} = \sum_{j=1}^{n_2} c_{\nu^*(j),j} \cdot q_j^{(2)} .$$

Following the proof for (11), it is straightforward to see that

$$D_W^{\mathrm{p}} = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} w_{i,j}^* c_{i,j} \geq \sum_{j=1}^{n_2} c_{\nu^*(j),j} \cdot q_j^{(2)} = D_{LB}^{\mathrm{p}} .$$

D_{LB} can be trivially solved without a numerical algorithm, and is used as a lower bound on D_W.

APPENDIX B ADDITIONAL RESULTS

The effect of translation on the solution of OT-RMC is demonstrated in Figure 11. The target pattern is the circle image shown on the top left corner of Figure 1, and the source image is the one to its right. The target pattern is shifted to 25 locations across the source image, and the corresponding pattern extracted (shown as black pixels) from the source is shown together with the target pattern (shown in light gray). Although a considerable portion of the embedded

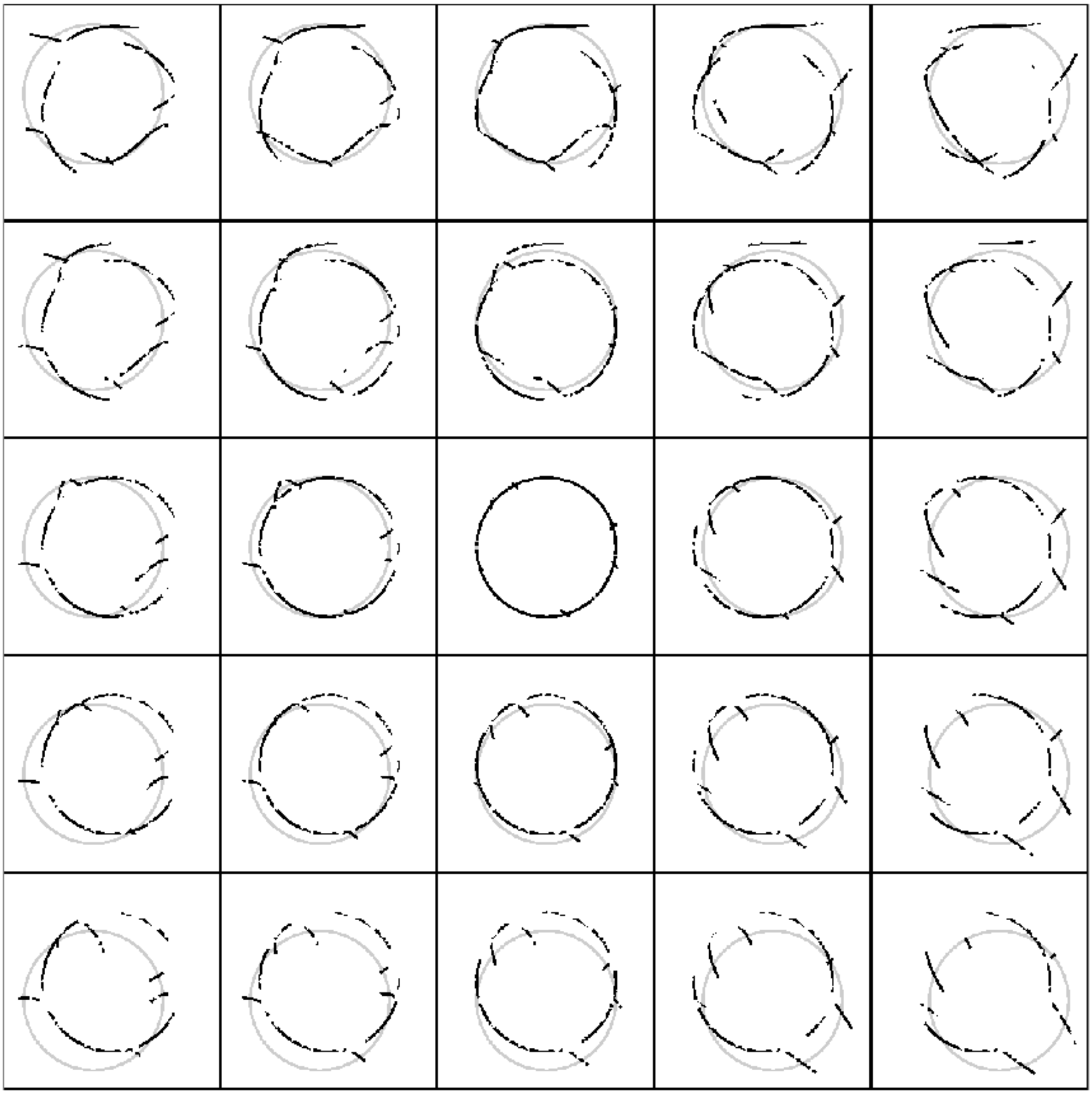


FIGURE 11. The extracted patterns from the source image when the target pattern is shifted to different locations. The light gray circle in each image is the shifted target pattern, and the black pixels are extracted from the source.

circle has been extracted even when the target circle is displaced from the embedded one, we cannot expect OT-RMC to fully reverse the effect of translation. In practice, we can achieve robustness against translation by centering both the source and target distributions or as we did in this experiment, by testing a collection of translated target distributions. Which approach is suitable depends on the application and available computational power.

APPENDIX C
DISTANCE DEFINITIONS AND IMAGE PATCH SELECTION FOR PIXEL PATTERN EXTRACTION

We hereby define the Translation-Rotation Adjusted Wasserstein distance D_{TRA} and the angle distance D_{ANG}, assuming the cost between support points is the square of the Euclidean distance. These distances are used for the pixel pattern matching problem, motivated by the fact that our perception of a pixel pattern should not be affected by translation and small or moderate amount of rotation. The Wasserstein distance D_W based on OT is not invariant to translation or rotation. We remark that for other applications such as color scheme matching, translation or rotation should matter for comparing similarity, and thus we simply use D_W.

Consider two distributions with support points specified by the data matrix $X_l \in \mathbb{R}^{n_l \times d}$, $l = 1, 2$, where n_l is the support size and d is the dimension of the data. Every row of the data matrix corresponds to one point. For pixel pattern matching, $d = 2$. Specifically, X_1 contains the vertical and horizontal coordinates of black pixels in the source image and X_2 contains those in the target pattern. Let the ith row, $i = 1, \ldots, n_l$ of X_l be $\mathbf{x}_i^{(l)}$. Let the probabilities on the support points be $\mathbf{q}_l = (q_1^{(l)}, q_2^{(l)}, \ldots, q_{n_l}^{(l)})^t$, $l = 1, 2$. Let the matching weight matrix solved by OT be $\mathbf{w}^* = (w_{i,j})_{i=1,\ldots,n_1, j=1,\ldots,n_2}$. In our setup, $\mathbf{q}_2$ is the target distribution, and $\mathbf{q}_1$ is the source distribution on the selected pixels by OT-RMC or NN. Although $\mathbf{q}_1$ and $\mathbf{q}_2$ are uniform in our experiments, we discuss for general distributions below.

To remove the effect of translation, we simply center the support points in both X_l, $l = 1, 2$. Specifically, let the expected data vector of X_l be $\bar{\mathbf{x}}_l = \sum_{i=1}^{n_l} q_i^{(l)} \mathbf{x}_i^{(l)}$. Subtract each row in X_l by $\bar{\mathbf{x}}_l$. We then solve OT using the centered support points. For brevity of notation, we will use X_l to denote the centered data matrix in the discussion below. From now on, we assume the operations presented below are all on centered data matrices.

To remove the rotation effect (reflection included), we take an iterative approach. In each iteration, OT solves the matching weight matrix $\mathbf{w}^*$, based on which a rotation matrix is solved and applied to the second distribution. Equivalently, we can apply rotation to the first distribution, while the problem is essentially the same. Once the second distribution has been updated by rotation, OT can be applied again to yield a new $\mathbf{w}^*$, and rotation can be applied to the second distribution again, so on and so forth. We find empirically that the rotation angle is not large, and after one round of rotation, the results vary negligibly. Hence in our experiments, only one iteration is applied. Denote the rotation matrix to be applied to the second data matrix by A. We solve A by the following optimization problem:

$$\underset{A}{argmin} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{i,j} \| \mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)} \cdot A^t \|^2 , \qquad (12)$$

where $\| \cdot \|$ is the L_2 norm. This is essentially the weighted orthogonal Procrustes problem. Here we have $n_1 \times n_2$ pairs of points, $\mathbf{x}_i^{(1)}$ versus $\mathbf{x}_j^{(2)}$ with weight $w_{i,j}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. We use the algorithm of [44] to solve the rotation matrix. Let $S \in \mathbb{R}^{d \times d}$ be

$$S = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{i,j} \mathbf{x}_j^{(2)^t} \cdot \mathbf{x}_i^{(1)} ,$$

and the singular value decomposition (SVD) of S be $S = U^t \Sigma_S V$. Then

$$A = V \cdot U^t . \qquad (13)$$

Once A is solved, OT is applied to data matrices X_1 and $X_2 \cdot A^t$, and the resulting D_W is defined as D_{TRA}. We summarize the process in Algorithm 1.

When comparing D_{TRA} or D_{LB} (see Appendix A for the definition of D_{LB}) across target patterns that are scaled versions of each other (a multiplicative factor applied to the data matrix), we would like to account for the scaling effect. In particular, we apply the inverse scaling to D_{TRA} and D_{LB}. The rationale is that D_{TRA} and D_{LB} will reduce proportionally if both the source and the target pattern are scaled down although we prefer to consider that the patterns are essentially the same unless the scaling is extreme, e.g., shrinking every pattern to a tiny dot.

```
Algorithm 1 Compute $D_{TRA}$
Input: Data matrices $X_1$, $X_2$, and the two probability vectors
    $\mathbf{q}_1$ and $\mathbf{q}_2$.
Output: $D_{TRA}$
1: Center $X_1$ and $X_2$ by subtracting the mean vector from
   each row.
2: Repeat the following steps $k$ times.
     1) Solve the matching weight matrix $\mathbf{w}^*$ by OT based
        on $X_l$ and $\mathbf{q}_l$, $l = 1, 2$.
     2) Solve the rotation matrix $A$ by Eq. (13).
     3) Let $X_2 \cdot A^t \rightarrow X_2$.
3: Based on the updated $X_1$ and $X_2$, compute the cost matrix
   $\mathbf{c}$. Solve the matching weight matrix $\mathbf{w}^*$ by OT using $\mathbf{c}$,
   $\mathbf{q}_1$, and $\mathbf{q}_2$.
4: Compute $D_{TRA} = \sqrt{\langle \mathbf{c}, \mathbf{w}^* \rangle}$.
```

The angle distance D_{ANG} is inspired by approaches to characterize the similarity between curves on a plane. The local similarity between two points on the two curves often depends on both the positions of the points and the orientations of the curves at those points. Consider two sets of points $\mathbf{Z}_l = \{\mathbf{z}_i^{(l)}, i = 1, \ldots, n\}$, $\mathbf{z}_i^{(l)} = (z_{i,1}^{(l)}, z_{i,2}^{(l)})$, $l = 1, 2$, where a one-to-one correspondence between the points is established. Without loss of generality, assume the correspondence $\mathbf{z}_i^{(1)} \leftrightarrow \mathbf{z}_i^{(2)}$. Take $\mathbf{Z}_2$ as the reference set. For each point $\mathbf{z}_i^{(2)}$, find its two nearest neighbors according to the Euclidean distance from $\mathbf{Z}_2$. Suppose the indices for the two nearest neighbors of the ith point are $\nu_{i,1}$ and $\nu_{i,2}$. For example, if the points in a set line up on a continuous and non-self-intersecting curve and we trace the points along the curve, the two nearest neighbors of a point are the two points immediately proceeding and following it. The angle of the vector $\mathbf{z}_i^{(2)} - \mathbf{z}_{\nu_{i,1}}^{(2)}$ is computed by $\theta_{i,1} = \arccos \frac{z_{i,1}^{(2)} - z_{\nu_{i,1},1}^{(2)}}{\|\mathbf{z}_i^{(2)} - \mathbf{z}_{\nu_{i,1}}^{(2)}\|}$. Similarly we can define the angle $\theta_{i,2}$ for vector $\mathbf{z}_i^{(2)} - \mathbf{z}_{\nu_{i,2}}^{(2)}$. If the two sets $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the same or similar, the nearest neighbors of a point in $\mathbf{Z}_2$ should carry over to its corresponding point in $\mathbf{Z}_1$, and the angles formed between the point and each of its neighbor should be retained. That is, if the two sets capture the same pattern, the neighbors of $\mathbf{z}_i^{(1)}$ in $\mathbf{Z}_1$ should be $\mathbf{z}_{\nu_{i,1}}^{(1)}$ and $\mathbf{z}_{\nu_{i,2}}^{(1)}$, and the angles $\theta'_{i,j} = \arccos \frac{z_{i,1}^{(1)} - z_{\nu_{i,j},1}^{(1)}}{\|\mathbf{z}_i^{(1)} - \mathbf{z}_{\nu_{i,j}}^{(1)}\|}$ should equal $\theta_{i,j}$ for $j = 1, 2$ respectively. The average difference $|\theta_{i,j} - \theta'_{i,j}|$ across $i = 1, \ldots, n$ and $j = 1, 2$ can be used to quantify the local orientation difference between the two sets.

We treat the support points in the target distribution (the second distribution) as $\mathbf{Z}_2$. Here, we assume that the support sets of the source and target distributions have been centered and the latter has been subject to the rotation described above. That is, $\mathbf{z}_i^{(2)} = \mathbf{x}_i^{(2)}$, $i = 1, \ldots, n_2$. For each $\mathbf{x}_i^{(2)}$, $i = 1, \ldots, n_2$, its corresponding point in $\mathbf{Z}_1$ is the weighted average of its mapped point in the first distribution:

$$\mathbf{z}_i^{(1)} = \frac{\sum_{j=1}^{n_1} w_{j,i}\mathbf{x}_j^{(1)}}{\sum_{j=1}^{n_1} w_{j,i}}.$$

Then we define the angle distance D_{ANG} between the source and target distributions by

$$D_{ANG} = \sum_{i=1}^{n_2} q_i^{(2)} \cdot \frac{|\theta_{i,1} - \theta'_{i,1}| + |\theta_{i,2} - \theta'_{i,2}|}{2}.$$

We selected at most one patch from each image to conduct detailed experiments described in Subsection IV-A. We experimented with three scales of the target distributions (scale 1.0, 0.85, 0.65) and the reflected patterns, thus total of 6 versions of the same pattern. For each version of the pattern, it is shifted across the image, and the lower bound distance D_{LB} is computed. If the lower bound D_{LB} is above a given threshold, the patch will be marked as a poor match, and the exact D_W based on the selected pixels by OT-RMC will not be computed. If D_{LB} is no greater than the threshold, OT-RMC with $\lambda = 20$ is applied to select pixels from the patch. Then D_{ANG} is computed using the selected pixels. For any image in which D_{ANG} is computed for multiple patches, we choose the patch yielding the minimum D_{ANG}.

REFERENCES

[1] C. Villani, *Topics in Optimal Transportation*. Providence, RI, USA: American Mathematical Society, 2003.

[2] R. M. Gray. (2013). *Transportation Distance, Shannon Information, and Source Coding*. [Online]. Available: https://ee.stanford.edu/~gray/gretsi.pdf

[3] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[4] X. Zheng, J. Ye, J. Z. Wang, and J. Li, "SCOTT: Shape-location combined tracking with optimal transport," *SIAM J. Math. Data Sci.*, vol. 2, no. 2, pp. 284–308, Jan. 2020.

[5] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander, "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming," *Cell*, vol. 176, no. 4, pp. 928–943, 2019.

[6] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM J. Math. Anal.*, vol. 43, no. 2, pp. 904–924, Jan. 2011.

[7] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. NIPS*, 2013, vol. 2, no. 3, p. 4.

[8] G. Carlier, A. Oberman, and E. Oudet, "Numerical methods for matching for teams and Wasserstein barycenters," *ESAIM, Math. Model. Numer. Anal.*, vol. 49, no. 6, pp. 1621–1642, Nov. 2015.

[9] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative Bregman projections for regularized transportation problems," *SIAM J. Sci. Comput.*, vol. 37, no. 2, pp. A1111–A1138, Jan. 2015.

[10] E. Anderes, S. Borgwardt, and J. Miller, "Discrete Wasserstein barycenters: Optimal transport for discrete data," *Math. Methods Oper. Res.*, vol. 84, no. 2, pp. 389–409, Oct. 2016.

[11] S. Borgwardt and S. Patterson, "Improved linear programs for discrete barycenters," *INFORMS J. Optim.*, vol. 2, no. 1, pp. 14–33, Jan. 2020.

[12] L. Yang, J. Li, D. Sun, and K.-C. Toh, "A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters," *J. Mach. Learn. Res.*, vol. 22, no. 21, pp. 1–37, 2021. [Online]. Available: http://jmlr.org/papers/v22/19-629.html

[13] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.

[14] Y. Zhang, J. Z. Wang, and J. Li, "Parallel massive clustering of discrete distributions," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 4, pp. 1–24, Jun. 2015.

[15] J. Ye, P. Wu, J. Z. Wang, and J. Li, "Fast discrete distribution clustering using Wasserstein barycenter with sparse support," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2317–2332, May 2017.

[16] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–11, Jul. 2015.

[17] J. Solomon and A. Vaxman, "Optimal transport-based polar interpolation of directional fields," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–13, Jul. 2019.

[18] J. Li and F. Zhang, "Geometry-sensitive ensemble mean based on Wasserstein barycenters: Proof-of-concept on cloud simulations," *J. Comput. Graph. Statist.*, vol. 27, no. 4, pp. 785–797, Oct. 2018.

[19] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Optimal transport for Gaussian mixture models," *IEEE Access*, vol. 7, pp. 6269–6278, 2019.

[20] Y. Chen, J. Ye, and J. Li, "Aggregated Wasserstein distance and state registration for hidden Markov models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2133–2147, Sep. 2020.

[21] J. Delon and A. Desolneux, "A Wasserstein-type distance in the space of Gaussian mixture models," *SIAM J. Imag. Sci.*, vol. 13, no. 2, pp. 936–970, Jan. 2020.

[22] F. Luo and S. Mehrotra, "Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models," *Eur. J. Oper. Res.*, vol. 278, no. 1, pp. 20–35, Oct. 2019.

[23] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *J. Appl. Probab.*, vol. 56, no. 3, pp. 830–857, Sep. 2019.

[24] A. Levine and S. Feizi, "Wasserstein smoothing: Certified robustness against Wasserstein adversarial attacks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3938–3947.

[25] J. Li, B. Seo, and L. Lin, "Optimal transport, mean partition, and uncertainty assessment in cluster analysis," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 12, no. 5, pp. 359–377, Oct. 2019.

[26] L. Zhang, L. Lin, and J. Li, "CPS analysis: Self-contained validation of biomedical data clustering," *Bioinformatics*, vol. 36, no. 11, pp. 3516–3521, Jun. 2020.

[27] Wikepedia. *Color Scheme*. Accessed: Feb. 23, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Color_scheme

[28] J. Li, S. Ray, and B. G. Lindsay, "A nonparametric statistical approach to clustering via mode identification," *J. Mach. Learn. Res.*, vol. 8, no. 8, pp. 1687–1723, 2007.

[29] G. X. Y. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 14049.

[30] D. R. Burton, R. Ahmed, D. H. Barouch, S. T. Butera, S. Crotty, A. Godzik, D. E. Kaufmann, M. J. McElrath, M. C. Nussenzweig, B. Pulendran, C. N. Scanlan, W. R. Schief, G. Silvestri, H. Streeck, B. D. Walker, L. M. Walker, A. B. Ward, I. A. Wilson, and R. Wyatt, "A blueprint for HIV vaccine discovery," *Cell Host Microbe*, vol. 12, no. 4, pp. 396–407, Oct. 2012.

[31] L. Lin *et al.*, "COMPASS identifies T-cell subsets correlated with clinical outcomes," *Nature Biotechnol.*, vol. 33, no. 6, pp. 610–616, Jun. 2015.

[32] A. C. Carrano, F. Mulas, C. Zeng, and M. Sander, "Interrogating islets in health and disease with single-cell technologies," *Mol. Metabolism*, vol. 6, no. 9, pp. 991–1001, Sep. 2017.

[33] A.-C. Villani *et al.*, "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors," *Science*, vol. 356, no. 6335, Apr. 2017, Art. no. eaah4573.

[34] A. K. Shalek and M. Benson, "Single-cell analyses to tailor treatments," *Sci. Transl. Med.*, vol. 9, no. 408, Sep. 2017, Art. no. eaan4730.

[35] L. Dobnikar, A. L. Taylor, J. Chappell, P. Oldach, J. L. Harman, E. Oerton, E. Dzierzak, M. R. Bennett, M. Spivakov, and H. F. Jørgensen, "Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels," *Nature Commun.*, vol. 9, no. 1, p. 4567, 2018.

[36] D. A. Lawson, K. Kessenbrock, R. T. Davis, N. Pervolarakis, and Z. Werb, "Tumour heterogeneity and metastasis at single-cell resolution," *Nature Cell Biol.*, vol. 20, no. 12, p. 1349, 2018.

[37] D. R. Gawel *et al.*, "A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases," *Genome Med.*, vol. 11, no. 1, p. 47, Dec. 2019.

[38] H. T. Maecker, J. P. McCoy, and R. Nussenblatt, "Standardizing immunophenotyping for the human immunology project," *Nature Rev. Immunol.*, vol. 12, no. 3, pp. 191–200, Mar. 2012.

[39] G. Finak *et al.*, "Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 020686.

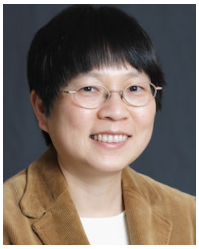
[40] L. Lin and B. P. Hejblum, "Bayesian mixture models for cytometry data analysis," *WIREs Comput. Statist.*, vol. 12, p. e1535, Oct. 2020, doi: 10.1002/wics.1535.

[41] M. Baron, A. Veres, S. Wolock, A. Faust, R. Gaujoux, A. Vetere, J. Ryu, B. Wagner, S. Shen-Orr, A. Klein, D. Melton, and I. Yanai, "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure," *Cell Syst.*, vol. 3, no. 4, pp. 346.e4–360.e4, Oct. 2016, doi: 10.1016/j.cels.2016.08.011.

[42] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Nov. 2008.

[43] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature Biotechnol.*, vol. 36, no. 5, pp. 411–420, May 2018.

[44] O. Sorkine-Hornung and M. Rabinovich, "Least-squares rigid motion using SVD," *Comput.*, vol. 1, no. 1, pp. 1–5, 2017.



JIA LI (Fellow, IEEE) received the M.S. degree in electrical engineering, the M.S. degree in statistics, and the Ph.D. degree in electrical engineering from Stanford University, in 1995, 1998, and 1999, respectively. She worked as a Program Director with the Division of Mathematical Sciences, National Science Foundation, from 2011 to 2013; a Visiting Scientist with the Google Labs, Pittsburgh, from 2007 to 2008; a Researcher with the Xerox Palo Alto Research Center, from 1999 to 2000; and a Research Associate with the Computer Science Department, Stanford University, in 1999. She is currently a Professor of statistics with The Pennsylvania State University. Her research interests include statistical/machine learning, image analysis, bioinformatics, and artificial intelligence.

LIN LIN received the M.S. degree in statistics from the National University of Singapore, in 2008, and the Ph.D. degree in statistics from Duke University, in 2012. She worked as a Postdoctoral Researcher with the Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, from 2012 to 2015. She is currently an Assistant Professor of statistics with The Pennsylvania State University. Her research interests include Bayesian analysis, machine learning, and bioinformatics.