Unsupervised Labeling by Geometric and Spatially Regularized Self-Assignment

Matthias Zisler[1], Artjom Zern[1], Stefania Petra[2], and Christoph Schnörr[1]

[1] Image and Pattern Analysis Group, Heidelberg University, Germany
[2] Mathematical Imaging Group, Heidelberg University, Germany

Abstract. We introduce and study the *unsupervised self-assignment flow* for labeling image data (euclidean or manifold-valued) without specifying any class prototypes (labels) beforehand, and without alternating between data assignment and prototype evolution, which is common in unsupervised learning. Rather, a *single* smooth flow evolving on an elementary statistical manifold is geometrically integrated which assigns given data to *itself*. Specifying the *scale* of spatial regularization by geometric averaging suffices to induce a *low-rank* data representation, the emergence of prototypes together with their number, and the data labeling. Connections to the literature on low-rank matrix factorization and on data representations based on discrete optimal mass transport are discussed.

1 Introduction

The assignment flow introduced by [3] is a smooth dynamical system for image labeling. Its state is given by discrete distributions (assignment vectors) assigned to each pixel, like in established discrete graphical models [13]. These states evolve on probability simplices equipped with the Fisher-Rao metric in order to minimize locally a distance to class prototypes, commonly called *labels*. Label assignments emerge gradually as the flow evolves and are spatially regularized by geometric averaging over local neighborhoods. Convergence analysis for a particular multiplicative update scheme was reported in [4].

Due to its *smoothness* and the wide range of *sparse* numerical updates that can be derived by geometrically integrating the assignment flow [22], this approach provides an attractive alternative to established discrete graphical models for image labeling [13], that rely on convex relaxations for large-scale problems, which do not scale well with increasing problem size or increasing numbers of labels. The evaluation of graphical models using the assignment flow has been demonstrated recently [12].

A common problem of *supervised* image labeling concerns the specification of labels beforehand, that is to determine prototypical features that properly represent different data categories into which given image data should be classified in a spatially coherent way. As a remedy, the *adaption* of labels during the assignment process was proposed recently [23]: Starting with an 'overcomplete' dictionary of prototypes (labels) that is efficiently computed beforehand using

conventional metric clustering, prototypes evolved on a corresponding feature manifold while being coupled to the assignment flow.

While this approach largely compensates the lack of prior knowledge of adequate labels, it does not directly address the fundamental question: How may prototypes emerge *directly* from the data during the assignment process *without any* prior coding using conventional clustering? This paper describes our first step towards a *completely unsupervised* approach to image labeling.

Contribution. Our approach is to apply the supervised assignment flow (Section 2) to the *self-assignment* of the given data (Section 3). This gives rise to a *factorized* affinity matrix that is parametrized by the assignments and converges to a *low-rank representation* of the data, *solely* induced by the *scale* of the spatial regularization performed by the assignment flow. This process also defines the formation of feature prototypes (labels) and a proper number of classes.

Related work. Our work utilizes information geometry and relates to the current dynamically evolving literature on clustering using low-rank matrix factorizations, and on data representation using discrete optimal mass transport. These relations are discussed in *Section 4* after presenting our novel approach.

We conclude with experiments in Section 5. In order to illustrate one-to-one the content of the preceding sections, our implementation does *not* involve any further changes of the assignment flow. In particular, we did not change and adapt the numerics in order to exploit the low-rank structure for large problem sizes right from the beginning. We leave such aspects for future work and used small and medium problem sizes for our experiments, to be able to apply the assignment flow *directly* to the self-assignment of given data, and to study how labels emerge in a completely unsupervised way.

2 Assignment Flow: Supervised Labeling

We summarize the assignment flow for *supervised* image labeling introduced by [3]. Let $G = (I, E)$ be a given undirected graph with vertices $i \in I$ indexing data $\mathcal{F}_I = \{f_i \colon i \in I\} \subset \mathcal{F}$ given in a metric space $(\mathcal{F}, d)$. The edge set E specifies neigborhoods $\mathcal{N}_i = \{k \in I \colon ik = ki \in E\} \cup \{i\}$ for every pixel $i \in I$, together with positive weight vectors $w_i \in \operatorname{rint} \Delta_{|\mathcal{N}_i|}$, where $\Delta_n \subset \mathbb{R}^n$ denotes the probability simplex.

Along with $\mathcal{F}_I$, *prototypical data (labels)* $\mathcal{G}_J = \{g_j \in \mathcal{F} \colon j \in J\}$ are given representing classes $j = 1, \ldots, |J|$. *Supervised image labeling* denotes the task to assign precisely one prototype g_j to each datum f_i in a spatially coherent way. These assignments are represented at each pixel i by probability vectors

$$W_i \in \mathcal{S} := \operatorname{rint} \Delta_{|J|}, \quad i \in I \tag{1}$$

on the relative interior of the simplex $\Delta_{|J|}$, that together with the Fisher-Rao metric g_{FR} becomes a Riemannian manifold denoted by $\mathcal{S}$. Collecting all assignment vectors into a strictly positive, row-stochastic matrix

$$W = (W_1, \ldots, W_{|I|})^\top \in \mathcal{W} = \mathcal{S} \times \cdots \times \mathcal{S} \subset \mathbb{R}^{|I| \times |J|} \tag{2}$$

defines a point on the *assignment manifold* $\mathcal{W}$. Image labeling is accomplished by geometrically integrating the *assignment flow* (the r.h.s. is defined below)

$$\dot{W} = \Pi_W\big(S(W)\big), \qquad W(0) = \mathbb{1}_{\mathcal{W}}, \tag{3}$$

that evolves from the barycenter $W(0)$ towards *pure* assignment vectors, i.e. each vector W_i approaches the ε-neighborhood of some unit vector at some vertex of $\mathcal{S}$ and hence a *labeling* after trivial rounding.

In order to explain the rationale behind (3), we need the following maps based on the affine e-connection of information geometry [2] in place of the Levi-Civita connection on the tangent bundle of the manifolds $\mathcal{S}$ and $\mathcal{W}$: With tangent space $T_0 = T_p\mathcal{S}$ independent of the base point $p \in \mathcal{S}$, we define

$$\mathbb{R}^{|J|} \ni z \mapsto \Pi_p(z) = \big(\operatorname{Diag}(p) - pp^\top\big)z \in T_0, \tag{4a}$$

$$\mathcal{S} \times T_0 \ni (p, v) \mapsto \operatorname{Exp}_p(v) = \frac{e^{\frac{v}{p}}}{\langle p, e^{\frac{v}{p}}\rangle} p \in \mathcal{S}, \tag{4b}$$

$$\mathcal{S} \times \mathcal{S} \ni (p, q) \mapsto \operatorname{Exp}_p^{-1}(q) = \Pi_p \log \frac{q}{p} \in T_0, \tag{4c}$$

$$\mathcal{S} \times \mathbb{R}^{|J|} \ni (p, z) \mapsto \exp_p(z) = \operatorname{Exp}_p \circ \Pi_p(z) = \frac{pe^z}{\langle p, e^z\rangle} \in \mathcal{S}, \tag{4d}$$

where multiplication, subdivision and the exponential function $e^{(\cdot)}$ apply *componentwise* to strictly positive vectors in $\mathcal{S}$. Corresponding maps $\Pi_W, \operatorname{Exp}_W, \exp_W$ in connection with the product manifold (2) are defined analogously.

The vector field defining the assignment flow on the right-hand side of (3) is defined as follows. Given the metric d, data $\mathcal{F}_I$ and labels $\mathcal{G}_J$, distance vectors $D_i = \big(d(f_i, g_1), \ldots, d(f_i, g_{|J|})\big)^\top$ are defined at each pixel $i \in I$ and mapped to the assignment manifold by

$$L(W) = \exp_W\big(-\tfrac{1}{\rho}D\big) \in \mathcal{W}, \quad L_i(W_i) = \exp_{W_i}\big(-\tfrac{1}{\rho}D_i\big) = \frac{W_i e^{-\frac{1}{\rho}D_i}}{\langle W_i, e^{-\frac{1}{\rho}D_i}\rangle}, \tag{5}$$

where $\rho > 0$ is a user parameter for normalizing the scale of the data. These *likelihood vectors* represent 'data terms' in conventional variational approaches, and they are *spatially regularized* in a way conforming to the geometry of $\mathcal{S}$, to obtain

$$S(W) = \mathcal{R}^w(L(W)) \in \mathcal{W}, \quad \mathcal{R}_i^w(W) := \operatorname{Exp}_{W_i}\Big(\sum_{k \in \mathcal{N}_i} w_{ik} \operatorname{Exp}_{W_i}^{-1}(W_k)\Big). \tag{6}$$

The assignment flow (3) is well-defined based on (6). In addition, following [3], it *may* also be interpreted from a variational perspective as *approximate* Riemannian gradient ascent flow $\dot{W} = \nabla_{\mathcal{W}} J(W)$ with respect to the correlation functional $J(W)$,

$$\nabla_{\mathcal{W}} J(W) = \Pi_W(\nabla J(W)), \qquad J(W) = \langle W, S(W)\rangle, \tag{7}$$

based on the approximation of the Euclidean gradient $\nabla J(W) \approx S(W)$, which is justified by the slow dynamics of $S(W(t))$ due to averaging (6), relative to the fast dynamics of $W(t)$.

3 Approach: Label Learning through Self-Assignment

In this section we generalize the assignment flow to completely *unsupervised* scenarios. Specifically, we do *not* assume a set of prototypes $\mathcal{G}_J$ to be given. Rather, we initially set $\mathcal{G}_J = \mathcal{F}_I$ and consider each datum *both* as data point $f_i \in \mathcal{F}_I$ and (its copy) as label $f_i \in \mathcal{G}_J$. Consequently, the distance matrix of (5) is now defined as

$$D = \big(d(f_i, f_k)\big)_{i,k \in I}. \tag{8}$$

Integrating the assignment flow then performs a *spatially regularized self-assignment* of the data, based on which the set $\mathcal{G}_J$ evolves and forms *prototypes* in an unbiased and unsupervised way.

We regard these prototypes as *latent* variables denoted by $g_j \in \mathcal{G}_J$, to be distinguished from $f_i \in \mathcal{F}_I$ which are both *data points* and *labels*.

3.1 Rationale

Due to initially setting $\mathcal{G}_J = \mathcal{F}_I$, we have $J = I$ and the row-stochastic assignment matrix (2) is quadratic: $W \in \mathcal{W} \subset \mathbb{R}^{|I| \times |I|}$. Adopting from [3] the interpretation of the entry W_{ij} as posterior probability of assigning label f_j conditioned on the observation f_i,

$$W_{ij} = P(j|i), \quad j \in J,\ i \in I, \qquad\qquad P(i) = \frac{1}{|I|}, \quad i \in I \tag{9}$$

together with uniform prior probabilities $P(i)$ due to the absence of any supervision, Bayes' rule yields the probability of observing datum f_i conditioned on the label f_j,

$$P(i|j) = \frac{P(j|i)P(i)}{P(j)} = \frac{P(j|i)P(i)}{\sum_{l \in I} P(j|l)P(l)} \overset{(9)}{=} \frac{P(j|i)}{\sum_{l \in I} P(j|l)} \overset{(9)}{=} \frac{W_{ij}}{\sum_{l \in I} W_{lj}} \tag{10a}$$

$$= \big(WC(W)^{-1}\big)_{ij} \quad \text{with} \quad C(W) := \operatorname{Diag}(W^\top \mathbb{1}_{|I|}), \tag{10b}$$

that is by *normalizing* the *columns* of W. Since the *rows* of W are normalized by definition, this *symmetry* reflects our ansatz to form prototypes from the *entire* given data set $\mathcal{F}_I$.

Next we introduce and compute the *probabilities of self-assignments* $f_i \leftrightarrow f_k$ by marginalizing over the labels $f_j,\ j \in J$,

$$A_{ki}(W) := \sum_{j \in J} P(k|j)P(j|i) \overset{(9),(10)}{=} \sum_{j \in J} (WC(W)^{-1})_{kj} W_{ij} = (WC(W)^{-1}W^\top)_{ki}. \tag{11}$$

The resulting *(self)-affinity matrix*

$$A(W) \in \mathbb{R}_+^{|I| \times |I|}, \qquad A(W) = A(W)^\top, \quad A(W)\mathbb{1}_{|I|} = A(W)^\top \mathbb{1}_{|I|} = \mathbb{1}_{|I|} \tag{12}$$

is nonnegative, symmetric and doubly stochastic. It represents the mutual influence of the features at all pixels, as a function of the assignment matrix W.

As a consequence, we propose to replace the objective (7) used in the *supervised* case which maximizes the correlation of assignments and a spatially regularized representation of the affinity between data and prototypes, by the objective function

$$\min_{W \in \mathcal{W}} E(W), \qquad E(W) = \langle D, A(W) \rangle, \tag{13}$$

which in the present *unsupervised* scenario minimizes the correlation between the data (distance) matrix D (8) and the self-affinity matrix $A(W)$: whenever the feature distance between pixel i and k is *large*, the affinity probability $A_{ik}(W)$ between this pair of pixels should *decrease*, subject to mass conservation (12).

The *latent* (hidden) prototypes $\mathcal{G}_J$ that emerge from the data $\mathcal{F}_I$ are *implicitly* determined by the column-normalized assignment matrix (10) that minimizes (13): entries $(P(i|j))_{i \in I}$ signal the relative contribution of each data point i to forming the prototype g_j. How g_j is actually computed depends on the nature of the feature space $\mathcal{F}$ whose properties only matter at this point: a corresponding weighted average has to be well-defined. In the simplest case, the space $\mathcal{F}$ is Euclidean and prototype g_j, $j \in J$ is defined as the convex combination of all data points f_i, $i \in I$ with the probabilities $P(i|j)$, $i \in I$ as coefficients, i.e.

$$g_j = \sum_{i \in I} \big(WC(W)^{-1}\big)_{ij} f_i. \tag{14}$$

Of particular interest is the capability of this process to represent given data by *few* prototypes. Our approach accomplishes this in a natural way, solely depending on the *scale* at which spatial regularity is enforced in terms of the neighborhood size $|\mathcal{N}_i|$ for geometric averaging (6).

3.2 Computational Approach

We explain our approach as adaption of the *supervised* assignment flow (7) to *unsupervised* scenarios.

The vector field on the right-hand side of (7) involves the geometrically and spatially averaged likelihood vectors (5), which in turn result from mapping the feature distance matrix D to $L(W) \in \mathcal{W}$ on the assignment manifold. In view of the data terms of established variational segmentation approaches (see [6, 15] for the binary and non-binary case, respectively), we regard $D = \nabla_W \langle D, W \rangle$ as Euclidean gradient of such a basic data term.

A natural way to adapt the assignment flow to the present unsupervised setting is to replace this gradient by the Euclidean gradient of the objective (13), that is we redefine (5) as

$$L(W) = \exp_W \big(-\tfrac{1}{\rho} \nabla E(W)\big) \in \mathcal{W}, \quad L_i(W) = \frac{W_i e^{-\frac{1}{\rho} \nabla E(W)_i}}{\big\langle W_i, e^{-\frac{1}{\rho} \nabla E(W)_i} \big\rangle}, \quad \rho > 0 \tag{15}$$

with the gradient of (13) given by

$$\nabla E(W) = 2DWC(W)^{-1} - \mathbb{1}_{|I|} \operatorname{diag}\big(C(W)^{-1} W^\top DWC(W)^{-1}\big)^\top, \tag{16}$$

where diag$(\cdot)$ denotes the vector of diagonal elements of a matrix.

Besides this modification of (5), the remaining formulas (6), (7) do not change. The *unsupervised self-assignment flow*, therefore, reads

$$\dot{W}(t) = \Pi_{W(t)}\big(S(W(t))\big), \quad W(0) = \exp_{\mathbb{1}_{\mathcal{W}}}(-\varepsilon D) \in \mathcal{W}, \quad 0 < \varepsilon \ll 1, \tag{17}$$

where the initial point is a small perturbation of the barycenter $\mathbb{1}_{\mathcal{W}}$, in order to break the symmetry of the expression defining the gradient (16) that would result from choosing $W = \mathbb{1}_{\mathcal{W}}$ as initial point. For numerical schemes that properly integrate flows of the form (17) evolving on the manifold $\mathcal{W}$, we refer to [22].

We point out that the perturbed initialization (17) is not required in the supervised case in which distances are not averaged: compare $L(W)$ of (15) with the supervised version (5). Indeed, comparing again (5) and (15), we may view the gradient (16) as a *time-varying* distance matrix

$$D(t) = D(W(t)) := \nabla E(W(t)), \qquad D(0) = \nabla E\big(\exp_{\mathbb{1}_{\mathcal{W}}}(-\varepsilon D)\big), \tag{18a}$$

$$D(t)_{ij} \overset{(16)}{=} 2\langle D_i, W_C^j\rangle - \langle W_C^j, DW_C^j\rangle \quad \text{with} \quad W_C^j := (WC(W)^{-1})_{*j}, \tag{18b}$$

that emanates from (8) and takes into account the formation of the latent prototypes g_j, $j \in J$, caused by the self-assignment process (17). The prototypes are *implicitly* represented by the normalized column vectors W_C^j of the assignment matrix W, where each component $(W_C^j)_i = P(i|j)$ (see (10)) represents the support (affinity) of pixel $i \in I$. Accordingly, distances $D_{ij} = d(f_i, f_j)$ due to (8) are replaced in (18) by the *time-varying* averaged distances

$$D(t)_{ij} = 2\mathbb{E}\big[d(f_i, \mathcal{F}_I)\big|g_j\big] - \mathbb{E}\big[d(\mathcal{F}_I, \mathcal{F}_I)\big|g_j\big], \tag{19a}$$

$$\text{where} \quad \mathbb{E}[d(f_i, \mathcal{F}_I)\big|g_j\big] = \sum_{k\in I} P(k|j)d(f_i, f_k), \tag{19b}$$

of feature f_i to all features f_k supporting prototype g_j.

3.3 Spatially Regularized Optimal Transport

Problem (13) may also be regarded as *discrete optimal transport* problem [17], with D as cost matrix, with a transportation plan $A(W)$ *parametrized* by the assignment matrix W, and with marginal constraints (12) implied by the constraint $W \in \mathcal{W}$ of (13).

At first glance, this problem looks uninteresting since a uniform measure $\mathbb{1}_{|I|}$ assigned to all data is transported to another uniform measure (see (12)). This, however, reflects the fact that our approach is *completely* unsupervised. Moreover, since $D_{ii} = d(f_i, f_i) = 0$, $i \in I$, the trivial solution $A^* = I$ would attain the lower bound $0 \leq E(W) = \langle D, A\rangle$ and hence be optimal, *if* A were *not* constrained through the parametrization $A = A(W)$. And this trivial solution would correspond to the (useless) situation in which *each* data point is kept as prototype, which can only happen if assignment to pixels do not interact at all.

This trivial situation is ruled out through the *spatial interaction* of pixel assignments in terms of the geometric averaging map (6), which is a key property of the assignment flow (17). This interaction induces the formation of a (depending on the spatial scale: much) smaller subset of prototypes as latent variables and in turn a *low-rank factorization* of the affinity matrix (12), because many components j of the diagonal matrix $C(W)$ in (11) converge to 0. Rewriting the objective (13) as

$$\langle D, A(W)\rangle \overset{(11)}{=} \langle D, WC(W)^{-1}W^\top\rangle = \langle DWC^{-1}, W\rangle \overset{(18b)}{=} \langle DW_C, W\rangle, \quad (20)$$

with column-normalized, *asymmetric* assignment matrix W_C, admits the following interpretation complementing the discussion of (18) and takes into account that *spatial regularization* is 'built in' the unsupervised assignment flow (17):

Minimizing (13) through the flow (17) amounts to *spatially regularized* optimal transport of the uniform measure $\mathbb{1}_{|I|}$ assigned to all data points f_i, $i \in I$, to the positive measure $w(\mathcal{G}_J(t))$,

$$\mathbb{1}_{|I|} = W(t)\mathbb{1}_{|I|} \quad \to \quad W(t)^\top \mathbb{1}_{|I|} =: w(\mathcal{G}_J(t)), \quad (21)$$

where the latter concentrates on the *effective* support $J(w) \subset I$ of the emerging prototypes $g_j(t)$, $j \in J$. The corresponding cost matrix $DW_C = DW_C(t)$ of (20) for determining the transport plan $W(t)$ is given by the initial distance matrix (8) after averaging these distances with respect to the probability distributions that correspond to the normalized columns of $W(t)$,

$$\big(DW_C(t)\big)_{ij} \overset{(10)}{=} \sum_{i'\in I} P(i'|j)(t)d(f_i, f_{i'}) \overset{(19)}{=} \mathbb{E}\big[d(f_i, \mathcal{F}_I)\big|g_j\big], \quad j \in J. \quad (22)$$

At each point of time t, the assignment matrix $W(t)$ concentrates measure on those prototypes g_j, $j \in J$ (implicitly represented by the support of (21)) for which the average ('within-cluster') distances (22) are small.

4 Related Work

The literature on unsupervised learning and clustering is vast. We briefly comment on relations of our new approach to few closely related works from two general viewpoints: *nonnegative matrix factorization* and *discrete optimal transport* – see [7] and [17, 18] as general references.

4.1 Nonnegative Matrix Factorization (NMF)

NMF is concerned with representing a nonnegative input data matrix $F \in \mathbb{R}_+^{|I|\times d}$ (d is the feature dimension) in terms of a product of two nonnegative matrices $G, H \geq 0$. Non-negativity is key for interpreting the factors as weights H and as a dictionary G of prototypes. In addition, the rank constraint $\mathrm{rank}(HG^\top) \leq k$ has to be supplied as user parameter.

Archetypal Analysis [8] is an early work where the representation of data by prototypes was proposed. The factorization approach reads

$$F \approx HG^\top F \tag{23}$$

where both nonnegative factors $G^\top$ and H are constrained to be row-stochastic, so that $G^\top F$ are prototypes formed by convex combinations of data (feature) vectors that, in turn, are combined in a convex way using the weights H. The factorization is determined as local minimum by alternating minimization. Alternating updates of data assignment and prototype formation is common to most algorithms for unsupervised learning.

A major shortcoming of this early work is that the algorithm directly operates on the data rather than abstracting from the data space through a distance or a similarity function, as is common nowadays in machine learning. Likewise, equation (8) shows that we abstract from the data space through distance functions which could simply be induces by Euclidean norms or – when using more involved data models – by Riemannian distances of manifold-valued features.

Zass and Shashua [21] studied the clustering problem in the form

$$\max_G \langle A(F), GG^\top \rangle \quad \text{subject to} \quad G \geq 0, \quad G^\top G = I, \quad GG^\top \mathbb{1} = \mathbb{1} \tag{24}$$

where the data F are represented by any positive-semidefinite, symmetric affinity matrix $A(F)$. In particular, they showed that all three constraints together imply *hard* clustering which is combinatorially difficult, and that the second orthogonality constraint which accounts for normalization as done with basic spectral relaxations [19, 16], is the weakest one. Accordingly, they proposed a two-step procedure after dropping the second constraint: compute the closest symmetric doubly-stochastic nonnegative approximation of the data matrix $A(F)$, followed by a second step for determining a completely positive factorization $GG^\top$, $G \geq 0$. The same set-up was proposed by [20] except for determining a locally optimal solution in a single iterative process using DC-programming. Likewise, [14] explored symmetric nonnegative factorizations but ignored the constraint enforcing that $GG^\top$ is doubly-stochastic which is crucial for cluster normalization.

These works are close to our approach (13) (the distances D_{ij} to be minimized are turned into affinities by (15)), in that $A(W)$ given by (11), (12) is a nonnegative, symmetric, doubly-stochastic low-rank factorization whose factors are not constrained to be orthogonal. Key differences are that we determine $A(W) = WC(W)^{-1}W^\top$ by a single *smooth continuous* process (17) that may be turned into a discrete iterative process performing numerical integration techniques [22], and that a numerical low rank is induced by *spatial regularization*.

4.2 Discrete Optimal Mass Transport (DOMT)

Authors of [5] adopted an interesting viewpoint on the clustering problem: associating a sample distribution q with the given data and a distribution p with an unknown subset of the data regarded as prototypes, the problem of determining the latter is defined as the task to minimize the transport costs between

q and p, subject to a cardinality (sparsity) constraint on p in order to obtain a *smaller* subset of representative prototypes. The combinatorially hard cardinality constraint is turned into a specific convex penalty, imposed on the transport map having p as marginal. Since the number of data points in our scenarios is already large, working in the 'lifted' space of transport mappings with quadratic dimension is computationally infeasible, however.

The recent work [11] provides a related and natural reformulation of the clustering problem based on discrete optimal transport. Prototypes are defined as Wasserstein barycenters [1, 9] of the assignment distributions, and the squared Wasserstein distance is decomposed into a sum of within-cluster and between-cluster distances, analogous to the classical decomposition of the total scatter matrix associated with patterns in a Euclidean feature space [10]. The authors promote low-rank transport maps not only to cope with the curse of dimensionality but also as an effective method to achieve stability under sampling noise.

The relations to our work are not as direct as the relations to work on NMF discussed in Section 4.1. On the one hand, the objectives (13), (20) together with the constraints admit an interpretation as DOMT, and our approach involving spatial regularization also leads to low-rank transport maps. On the other hand, averaging for prototype formation is based on the geometry of the Fisher-Rao metric rather than on the Wasserstein distance. We leave a more detailed discussion of these interesting aspects for future work.

5 Experiments

We demonstrate the proposed **unsupervised self-assignment flow (USAF)**, Eq. (17) by depicting the self-assignments of various scenes.

Implementation. The USAF was integrated numerically using the geometric Euler scheme [22] with step-size $h = 1.0$, together with the renormalization strategy from [3] with $\varepsilon = 10^{-10}$, termination criterion (average entropy $\leq 10^{-3}$) which ensures almost unique assignments. Default values are $\rho = 0.1$ (scale normalization), a $|\mathcal{N}| = 3 \times 3$ neighborhood with uniform weights w_i for spatial averaging and the ℓ_1-norm for the distance matrix of (5). We restricted the problem size to 64×64 pixels, since exploiting numerically the low-rank structure for larger problem sizes is beyond the scope of this paper. Finally, for Euclidean data, the self-assignments are $u_i = \sum_{j \in J} W_{ij} g_j$, $i \in I$ with g_j due to (14).

Parameter influence. Figure 1 illustrates the influence of the only two user parameters on the self-assignment (labeling) of prototypes the emerge from the RGB input data. We observe that increasing the spatial scale (averaging) or the selectivity parameter reduces the number of prototypes.

Comparison with supervised assignment flows (SAF) [3] and unsupervised assignment flows (UAF) [23]. We adopted the implementation from [3, 23] and determined first the *effective* number k of protoypes using USAF, given by $\operatorname{rank}(W)$, and then used k-means to determine also k prototypes for both (SAF: prototypes are fixed) and (UAF: prototypes may evolve). The comparison (Figure 2) shows that *decoupling* prototype formation and spatial infer-

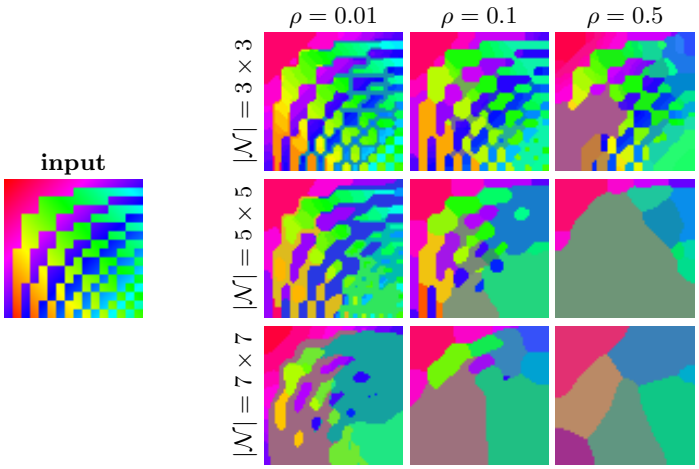

Fig. 1. Unsupervised image labeling through self-assignment of RGB data, depending on $|\mathcal{N}|$ and ρ. Increasing either value decreases the number of prototypes.

ence (SAF and UAF) leads to labelings that *mix spatial scales* in a way that is difficult to control. The self-assignment returned by (USAF: keeping the default value $\rho = 0.1$ fixed), on the other hand, clearly demonstrates that prototype formation is *solely determined by spatial scale*, i.e. by a single parameter.

Manifold valued data. Figure 3 shows $\mathcal{S}^1$-valued orientation data (panel 'angular') extracted from the fingerprint image using the structure tensor. The USAF returns a natural partition in terms of *prototypical orientations* extracted from the data itself, just based on the spatial scale at which the USAF operates. This happens *without extra costs*, since USAF *separates* data from the assignment manifold and hence manifold-specific operations are *not* needed. Prototypes *on* the manifold *may* be computed, of course, as weighted Riemannian means using the probabilities of (10).

6 Conclusion

We presented a novel geometric flow for completely unsupervised image labeling. A clear probabilistic interpretable stochastic factorization of the self-affinity matrix constitutes a low-rank representation of the input data, whereas the complexity (number of effective prototypes) is exclusively induced by spatial regularization, which is performed in an geometric and unbiased way. Experiments demonstrated the approach on Euclidean and manifold valued data.

In future work, we plan to exploit the low-rank structure by globally restricting the complexity of the solution, which immediately enables handling large problem instances.

Acknowledgement. Support from the German Science Foundation, grant GRK 1653, is gratefully acknowledged.

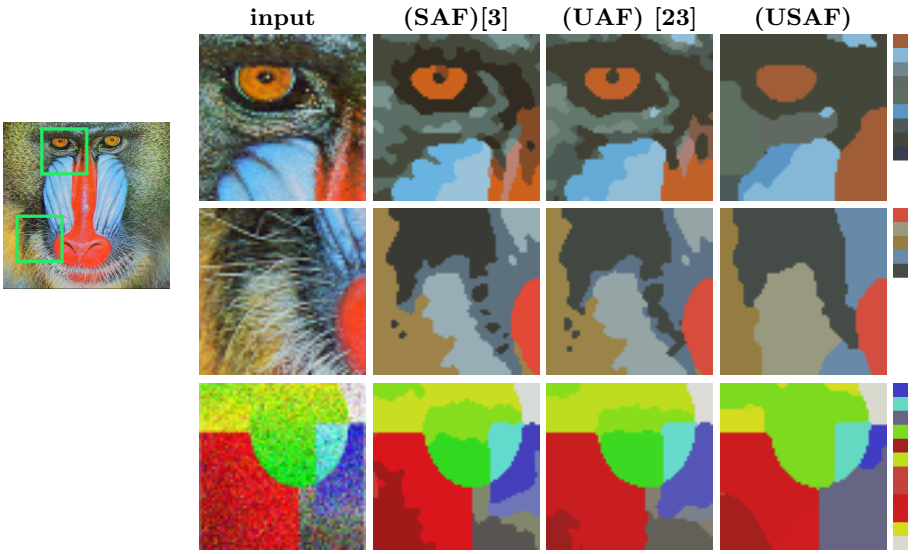

Fig. 2. Comparison of supervised (SAF), unsupervised (UAF) and self-assignment (USAF) flows. The right-most column depicts the prototypes $\mathcal{G}_J$ returned by the USAF, solely determined by the spatial scale, as the corresponding labeling (partition) reflects. By contrast, both SAF and UAF mix spatial scales due to prespecified prototypes.

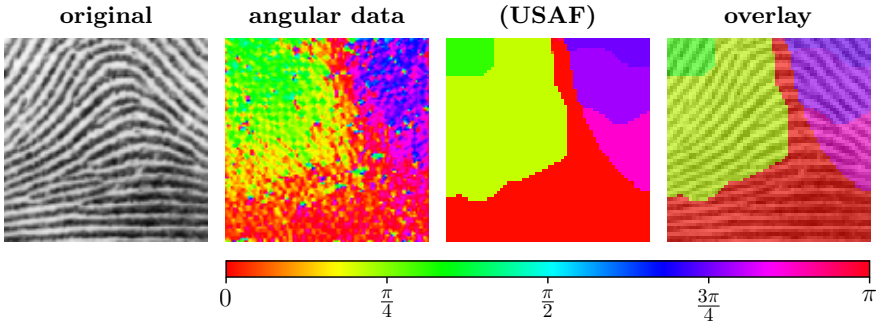

Fig. 3. Manifold valued data. $\mathcal{S}^1$-values angular data are extracted as input data from a fingerprint image. The USAF, operating with 5×5 neighborhoods, returns $|\mathcal{G}_J| = 6$ prototypes and a natural labeling (partition) on $\mathcal{W}$, *without* any operation *on* the data manifold $\mathcal{F}$, due to the separation of $\mathcal{F}$ and $\mathcal{W}$. Nevertheless, the partition *does* reflect the geometry of $\mathcal{F}$: orientations 0 and π are identified, for example.

References

1. Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM J. Math. Anal. **43**(2), 904–924 (2011)
2. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Amer. Math. Soc. and Oxford Univ. Press (2000)
3. Åström, F., Petra, S., Schmitzer, B., Schnörr, C.: Image Labeling by Assignment. Journal of Mathematical Imaging and Vision **58**(2), 211–238 (2017)
4. Bergmann, R., Fitschen, J.H., Persch, J., Steidl, G.: Iterative multiplicative filters for data labeling. International Journal of Computer Vision **123**(3), 435–453 (2017)
5. Carli, F.P., Ning, L., Georgiou, T.T.: Convex clustering via optimal mass transport. arXiv preprint arXiv:1307.5459 (2013)
6. Chan, T.F., Esedoglu, S., Nikolova, M.: Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models. SIAM J. Appl. Math. **66**(5) (2006)
7. Cichocki, A., Zdunek, A., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. John Wiley & Sons (2009)
8. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**(4), 338–347 (1994)
9. Cuturi, M., Peyré, G.: Semidual Regularized Optimal Transport. SIAM Review **60**(4), 941–965 (2018)
10. Devyver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall (1982)
11. Forrow, A., Hütter, J.C., Nitzan, M., Schiebinger, G., Rigollet, P., Weed, J.: Statistical optimal transport via geodesic hubs. arXiv preprint arXiv:1806.07348 (2018)
12. Hühnerbein, R., Savarino, F., Åström, F., Schnörr, C.: Image Labeling Based on Graphical Models Using Wasserstein Messages and Geometric Assignment. SIAM J. Imaging Science **11**(2), 1317–1362 (2018)
13. Kappes, J., Andres, B., Hamprecht, F., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., Rother, C.: A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems. Int. J. Computer Vision **115**(2), 155–184 (2015)
14. Kuang, D., Yun, S., Park, H.: SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. Journal of Global Optimization **62**(3), 545–574 (2015)
15. Lellmann, J., Schnörr, C.: Continuous Multiclass Labeling Approaches and Algorithms. SIAM J. Imag. Sci. **4**(4), 1049–1096 (2011)
16. von Luxburg, U.: A Tutorial on Spectral Clustering. Statistics and Computing **17**(4), 395–416 (2007)
17. Peyré, G.and Cuturi, M.: Computational Optimal Transport. CNRS (2018)
18. Santambrogio, F.: Optimal Transport for Applied Mathematicians. Birkhäuser (2015)
19. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Patt. Anal. Mach. Intell. **22**, 888–905 (2000)
20. Yang, Z., Corander, J., Oja, E.: Low-Rank Doubly Stochastic Matrix Decomposition for Cluster Analysis. J. Mach. Learning Res. **17**(1–25) (2016)
21. Zass, R., Shashua, A.: A unifying approach to hard and probabilistic clustering. In: Proc. ICCV (2005)
22. Zeilmann, A., Savarino, F., Petra, S., Schnörr, C.: Geometric Numerical Integration of the Assignment Flow. CoRR abs/1810.06970 (2018)
23. Zern, A., Zisler, M., Åström, F., Petra, S., Schnörr, C.: Unsupervised Label Learning on Manifolds by Spatially Regularized Geometric Assignment. In: Proc. GCPR (2018)