Unsupervised Riemannian Metric Learning for Histograms Using Aitchison Transformations

Tam Le TAM.LE@IIP.IST.I.KYOTO-U.AC.JP
Graduate School of Informatics, Kyoto University, Japan

Marco Cuturi MCUTURI@I.KYOTO-U.AC.JP
Graduate School of Informatics, Kyoto University, Japan

Abstract

Many applications in machine learning handle bags of features or histograms rather than simple vectors. In that context, defining a proper geometry to compare histograms can be crucial for many machine learning algorithms. While one might be tempted to use a default metric such as the Euclidean metric, empirical evidence shows this may not be the best choice when dealing with observations that lie in the probability simplex. Additionally, it might be desirable to choose a metric adaptively based on data. We consider in this paper the problem of learning a Riemannian metric on the simplex given *unlabeled* histogram data. We follow the approach of Lebanon (2006), who proposed to estimate such a metric within a parametric family by maximizing the inverse volume of a given data set of points under that metric. The metrics we consider on the multinomial simplex are pull-back metrics of the Fisher information parameterized by operations within the simplex known as Aitchison (1982) transformations. We propose an algorithmic approach to maximize inverse volumes using sampling and contrastive divergences. We provide experimental evidence that the metric obtained under our proposal outperforms alternative approaches.

1. Introduction

Learning distances to compare objects is an important topic in machine learning. Many approaches have been proposed to tackle this problem, notably by making the most of Mahalanobis distances in a supervised setting (Xing et al., 2002; Schultz & Joachims, 2003; Goldberger et al., 2004; Shalev-Shwartz et al., 2004; Globerson & Roweis, 2005; Weinberger et al., 2006; Davis et al., 2007; Weinberger & Saul, 2009).

Among such objects of interest, histograms – the normalized representation for bags of features – are popular in many applications, notably computer vision (Julesz, 1981; Sivic & Zisserman, 2003; Vedaldi & Zisserman, 2012), natural language processing (Salton & McGill, 1983; Salton, 1989; Joachims, 2002; Blei et al., 2003; Blei & Lafferty, 2009) and speech processing (Doddington, 2001; Campbell & Richardson, 2007). Mahalanobis distances can be used as such on histograms, but are known to perform poorly because they do not take into account the inherent constraints that histograms have (non-negativity and normalization). Cuturi & Avis (2014) and Kedem et al. (2012) proposed recently two supervised metric learning approaches in the simplex. Kedem et al.'s contribution is particularly relevant to this work: they proposed to compare two histograms $\mathbf{r}$ and $\mathbf{c}$ by using the χ^2 distance, $\chi^2(L\mathbf{r}, L\mathbf{c})$ between $L\mathbf{r}$ and $L\mathbf{c}$, where L is a linear map from and onto the simplex. This map L is learned by using labeled data and the Large Margin Nearest Neighbor framework (Weinberger et al., 2006; Weinberger & Saul, 2009). Our approaches also build on the idea of learning a map from and onto the simplex to parameterize a family of distances.

An even stronger influence on this paper lies in the work of Lebanon (2002; 2006) who proposed to learn a Riemannian metric for histograms using unlabeled data. The family of Riemannian metrics considered in these works can be seen as the standard Fisher information metric (instead of the χ_2 distance) using a particular family of transformations in the simplex. Cuturi & Avis (2014, §5.3) noticed that these transformations were defined in earlier references by Aitchison (1982; 1986; 2003) who called them simplicial pertubations.

Our contribution in this paper is two-fold: (1) we ex-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

tend Lebanon (2006)'s original approach to more general Aitchison transformations in the simplex; (2) we propose a new approach to solve a key step in Lebanon's procedure, namely the maximization of the inverse volume of a Riemannian metric.

This paper is organized as follows: after providing short reminders of Aitchison's tools and Riemannian geometry in Section 2, we proceed with the description of Fisher's information metric for histograms and show how all these elements can be used to form a parameterized family of Riemannian metrics in the simplex in Section 3. In Section 4, we propose a new algorithm to learn such metrics in an unsupervised way. In Section 5, we propose to use locally sensitive hashing to approximate k-nearest neighbors for our metrics to apply for large datasets. We study connections of this work with related approaches in Section 6, before providing experimental evidence in Section 7, and concluding this paper in Section 8.

2. Preliminary

We provide in this section a self-contained review of Aitchison's geometry as well as elements of Riemannian geometry that will be useful to define our methods.

2.1. Aitchison Geometry

We consider the n-simplex $\mathbb{P}_n$, defined by

$$\mathbb{P}_n \stackrel{\text{def}}{=} \left\{ \mathbf{x} \in \mathbb{R}^{n+1} \;\middle|\; \forall i, \mathbf{x}_i \geq 0 \text{ and } \sum_{i=1}^{n+1} \mathbf{x}_i = 1 \right\},$$

and write $\text{int}\mathbb{P}_n$ for its interior. Aitchison (1982; 1986; 2003) claims that the information reflected in histograms lies in the relative values of their coordinates rather than on their absolute value. Therefore, Aitchison proposes dedicated binary operations to combine two elements $\mathbf{x}$ and $\mathbf{z}$ in the interior of the simplex. Given $\gamma \in \mathbb{R}$, the pertubation and powering operations, denoted by $\oplus$ and $\otimes$, are respectively defined as

$$\mathbf{x} \oplus \mathbf{z} \stackrel{\text{def}}{=} C\left[\mathbf{x}_i \mathbf{z}_i\right]_{1 \leq i \leq n+1} \in \text{int}\mathbb{P}_n,$$
$$\gamma \otimes \mathbf{z} \stackrel{\text{def}}{=} C\left[\mathbf{z}_i^{\gamma}\right]_{1 \leq i \leq n+1} \in \text{int}\mathbb{P}_n,$$

where $C[x_1, x_2, \cdots, x_{n+1}] = \left[\frac{\mathbf{x}_i}{\sum_{j=1}^{n+1} \mathbf{x}_j}\right]_{1 \leq i \leq n+1}$ is the closure or normalization operator. A definition for the difference between $\mathbf{x}$ and $\mathbf{z}$ is naturally defined as:

$$\mathbf{x} \ominus \mathbf{z} = \mathbf{x} \oplus (-1 \otimes \mathbf{z}) = C\left[{}^{\mathbf{x}_i}/_{\mathbf{z}_i}\right]_{1 \leq i \leq n+1} \in \text{int}\mathbb{P}_n.$$

Note that the difference of two elements in the simplex with these operations remains in the simplex, unlike the results obtained in with the usual Euclidean geometry.

2.2. Riemannian Manifold

A Riemannian metric g on a manifold M is a function which assigns to each point $\mathbf{x} \in M$ an inner product $g_{\mathbf{x}}$ on the corresponding tangent space $T_{\mathbf{x}}M$. Consequently, we can measure the length of a tangent vector $\mathbf{v} \in T_{\mathbf{x}}M$ as $\|\mathbf{v}\|_{\mathbf{x}} = \sqrt{g_{\mathbf{x}}(\mathbf{v}, \mathbf{v})}$. Let $c : [a, b] \mapsto M$ be a curve in M. Its length is defined as $L(c) = \int_a^b \sqrt{g_{c(t)}(c'(t), c'(t))} \mathrm{d}t$, where $c'(t)$ belongs to $T_{c(t)}M$. The geodesic distance $d(\mathbf{x}, \mathbf{z})$ between two points $\mathbf{x}$ and $\mathbf{z}$ in the manifold M is defined as the length of the shortest curve connecting $\mathbf{x}$ and $\mathbf{z}$.

One way to specify a Riemannian metric on M is by using pull-back metrics. Let $F : M \mapsto N$ be a diffeomorphism that maps the manifold M onto the manifold N, and write h for a Riemannian metric on N. Let $T_{\mathbf{x}}M$, $T_{\mathbf{z}}N$ be the tangent spaces on the manifold M and N at $\mathbf{x}$ and $\mathbf{z}$ respectively. We can define a pull-back metric F^*h on M as follows:

$$F^*h_{\mathbf{x}}(\mathbf{u}, \mathbf{v}) = h_{F(\mathbf{x})}(F_*\mathbf{u}, F_*\mathbf{v}),$$

where F_* is the push-forward map which transforms a tangent vector $\mathbf{v} \in T_xM$ to a tangent vector $F_*\mathbf{v} \in T_{F(\mathbf{x})}N$. Thus, F is an isometric mapping between the manifold M and N:

$$d_{F^*h}(\mathbf{x}, \mathbf{z}) = d_h(F(\mathbf{x}), F(\mathbf{z})).$$

3. Fisher Information Metric for Histograms

In information geometry, the Fisher information metric is a particular Riemannian metric, defined on the simplex. It is well-known that the Fisher information metric can be described as a pull-back metric from the positive orthant of the sphere $\mathbb{S}_n^+$,

$$\mathbb{S}_n^+ \stackrel{\text{def}}{=} \left\{ \mathbf{x} \in \mathbb{R}^{n+1} \;\middle|\; \forall i, \mathbf{x}_i \geq 0 \text{ and } \sum_{i=1}^{n+1} \mathbf{x}_i^2 = 1 \right\}.$$

The diffeomorphism mapping $H : \mathbb{P}_n \mapsto \mathbb{S}_n^+$ is defined as the Hellinger mapping,

$$H(\mathbf{x}) \stackrel{\text{def}}{=} \sqrt{\mathbf{x}},$$

where the square root is an element-wise function. The mapping H pulls-back the Euclidean metric on the positive sphere $\mathbb{S}_n^+$ to the Fisher information metric on the simplex $\mathbb{P}_n$. Thus, the geodesic distance $d(\mathbf{x}, \mathbf{z})$ between two histograms $\mathbf{x}, \mathbf{z}$ in the simplex $\mathbb{P}_n$ under the Fisher information metric is equivalent to the length of the shortest curve on the positive sphere $\mathbb{S}_n^+$ between $H(\mathbf{x})$ and $H(\mathbf{z})$,

$$d(\mathbf{x}, \mathbf{z}) = \arccos\left(\langle H(\mathbf{x}), H(\mathbf{z}) \rangle\right) = \arccos\left(\sum_{i=1}^{n+1} \sqrt{\mathbf{x}_i \mathbf{z}_i}\right), \quad (1)$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product.

Let $G : \text{int}\mathbb{P}_n \mapsto \text{int}\mathbb{P}_n$ be a transformation inside the simplex. The Fisher information metric under the transformation G on the simplex $\mathbb{P}_n$, denoted as J, is a pull-back metric of the Euclidean metric on the positive sphere $\mathbb{S}_n^+$ through a transformation $F = H \circ G$. The geodesic distance that results by using J between $\mathbf{x}, \mathbf{z} \in \mathbb{P}_n$ is thus

$$d_J(\mathbf{x}, \mathbf{z}) = \arccos\left(\langle F(\mathbf{x}), F(\mathbf{z}) \rangle\right).$$

Therefore, we have a family of pull-back metrics J on the simplex $\mathbb{P}_n$, parameterized by the transformation G inside the simplex $\mathbb{P}_n$. In the next section, we will present a way to learn a suitable pull-back metric J based on a family of transformations G using only unlabeled data.

4. Unsupervised Riemannian Metric Learning for Histograms

4.1. Aitchison Transformation

We consider a family of transformations G on the simplex that can be defined using Aitchison elementary perturbation and powering operations presented in Section 2.1. The transformation we consider is parameterized by a vector $\boldsymbol{\alpha}$ in the strictly positive orthant $\mathbb{R}_+^{n+1}$, and by $\boldsymbol{\lambda} \in \text{int}\mathbb{P}_n$:

$$G(\mathbf{x}) = \boldsymbol{\alpha} \otimes \mathbf{x} \oplus \boldsymbol{\lambda} \in \text{int}\mathbb{P}_n. \tag{2}$$

Here, we generalize the powering operation for a histogram and a vector, so that we can have exponents that can vary for each coordinate:

$$\boldsymbol{\alpha} \otimes \mathbf{x} \stackrel{\text{def}}{=} C\left[\mathbf{x}_i^{\boldsymbol{\alpha}_i}\right]_{1 \le i \le n+1} \in \text{int}\mathbb{P}_n$$

Consequently, we have

$$G(\mathbf{x}) = C\left[\mathbf{x}_i^{\boldsymbol{\alpha}_i} \boldsymbol{\lambda}_i\right]_{1 \le i \le n+1} \in \text{int}\mathbb{P}_n.$$

We note that $\boldsymbol{\alpha} \otimes (\mathbf{x} \oplus \boldsymbol{\lambda}) = (\boldsymbol{\alpha} \otimes \mathbf{x}) \oplus (\boldsymbol{\alpha} \otimes \boldsymbol{\lambda})$. So, for the transformation $G(\mathbf{x})$, we can interpret that vector $\boldsymbol{\lambda}$ under operator $\oplus$ may be considered as a translation, and vector $\boldsymbol{\alpha}$ under operator $\otimes$ has a role as a linear mapping for a histogram $\mathbf{x}$ in the simplex.

Additionally, we can express the transformation $F(\mathbf{x})$ as the element-wise square root for $G(\mathbf{x})$:

$$F(\mathbf{x}) = H \circ G(\mathbf{x}) = \left[\sqrt{\frac{\mathbf{x}_i^{\boldsymbol{\alpha}_i} \boldsymbol{\lambda}_i}{\sum_{j=1}^{n+1} \mathbf{x}_j^{\boldsymbol{\alpha}_j} \boldsymbol{\lambda}_j}} \right]_{1 \le i \le n+1} \in \mathbb{S}_n^+.$$

Hence, we have a closed form for the geodesic distance under Riemannian metric J – the pull-back metric of the Euclidean metric on the positive sphere $\mathbb{S}_n^+$ through a transformation $F = H \circ G$,

$$d_J(\mathbf{x}, \mathbf{z}) = \arccos\left(\sum_{i=1}^{n+1} \sqrt{\frac{\mathbf{x}_i^{\boldsymbol{\alpha}_i} \boldsymbol{\lambda}_i}{\sum_{j=1}^{n+1} \mathbf{x}_j^{\boldsymbol{\alpha}_j} \boldsymbol{\lambda}_j} \frac{\mathbf{z}_i^{\boldsymbol{\alpha}_i} \boldsymbol{\lambda}_i}{\sum_{\ell=1}^{n+1} \mathbf{z}_\ell^{\boldsymbol{\alpha}_\ell} \boldsymbol{\lambda}_\ell}} \right). \tag{3}$$

4.2. Criterion

Let $D = \{\mathbf{x_i}, 1 \le i \le m\}$ be a dataset of unlabeled histograms in the interior of the simplex. We will learn a Riemannian metric from a family of pull-back metrics J on the simplex as described in Section 3. Since J is parameterized by Aitchison transformation G, defined in Equation (2), we equivalently learn an Aitchison transformation on the simplex.

The volume element of the Riemanian metric J at point $\mathbf{x}$ is defined as:

$$\text{dvol}J(\mathbf{x}) \stackrel{\text{def}}{=} \sqrt{\det \mathcal{G}(\mathbf{x})},$$

where $\mathcal{G}(\mathbf{x})$ is the Gram matrix, whose components $[\mathcal{G}]_{ij} = J(\partial_i, \partial_j)$, where $\{\partial_i\}_{1 \le i \le n}$ is a basis of a tangent space $T_{\mathbf{x}}\mathbb{P}_n$ of the simplex $\mathbb{P}_n$ at point $\mathbf{x}$. Intuitively, the volume element $\text{dvol}J(\mathbf{x})$ summaries the size of metric J at $\mathbf{x}$ in a scalar. Paths over areas with smaller volume will tend to be shorter than similar paths over areas with higher volume. Lebanon (2002; 2006) propose to maximize inverse volume to obtain shorter curves across densely populated regions of the simplex $\mathbb{P}_n$. Therefore, the geodesic distances will also tend to pass densely populated regions. It matches with an intuition about distance which should be measured on the lower dimensional data submanifold to capture intrinsic geometrical structure of data. We note that volume element $\text{dvol}J(\mathbf{x})$ is a homogeneous function, normalization for inverse volume is necessary to bound its quantity in optimization.

Following these intuitions, we consider a metric learning problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\lambda}} \quad & \mathcal{F} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \log \frac{\text{dvol}J^{-1}(\mathbf{x_i})}{\int_{\mathbb{P}_n} \text{dvol}J^{-1}(\mathbf{x}) d\mathbf{x}} - \frac{\mu}{2} \left\| \log \boldsymbol{\alpha} \right\|_2^2 \\ \text{s. t.} \quad & \boldsymbol{\lambda} \in \text{int}\mathbb{P}_n, \quad \boldsymbol{\alpha} \in \mathbb{R}_+^{n+1}, \end{aligned} \tag{4}$$

where $\log \boldsymbol{\alpha}$ is an element-wise function and $\mu > 0$ is a regularization parameter. We apply the logarithm function to the normalized inverse volume element in the criterion to simplify our learning procedure. We regularize this objective by the ℓ_2-norm of the element-wise logarithm $\boldsymbol{\alpha}$, that tends to avoid 0 values for our exponents. We do not regularize $\boldsymbol{\lambda}$ since $\boldsymbol{\lambda} \in \text{int}\mathbb{P}_n$ (or $\|\boldsymbol{\lambda}\|_1 = 1$).

4.3. Volume Element

We recall that the volume element of the Riemannian metric J at a point $\mathbf{x}$ is defined as $\text{dvol}J(\mathbf{x}) = \sqrt{\det \mathcal{G}(\mathbf{x})}$, and $[\mathcal{G}]_{ij} = J(\partial_i, \partial_j)$ where $\{\partial_i\}_{1 \le i \le n}$ is a basis of a tangent space of the simplex $T_{\mathbf{x}}\mathbb{P}_n$, described as rows of the matrix

$$U = \begin{bmatrix} 1 & \cdots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{n \times (n+1)}.$$

Algorithm 1 Gradient Ascent using Contrastive Divergence

Input: data $(\mathbf{x_i})_{1\leq i\leq m}$, gradient step size t_0^{α} and t_0^{λ}, initial vectors $\boldsymbol{\alpha}_0$, $\boldsymbol{\lambda}_0$ and a tolerance ϵ.
Set $t \leftarrow 1$.
Set $\alpha_{\mathbf{t}} \leftarrow \alpha_{\mathbf{0}}$.
Set $\lambda_{\mathbf{t}} \leftarrow \lambda_{\mathbf{0}}$.
repeat
  Use Metropolis-Hasting sampling algorithm where its proposal distribution is logistic normal distribution to transform training data $(\mathbf{x_i})_{1\leq i\leq m}$ into data drawn from $p(\mathbf{x})$.
  Compute gradient of the objective function with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ using Proposition 3.
  Update $\alpha_{\mathbf{t+1}} \leftarrow \Pi\Big(\alpha_{\mathbf{t}} + \frac{t_0^{\alpha}}{\sqrt{t}}\frac{\partial \mathcal{F}}{\partial \boldsymbol{\alpha}}\Big)$.
  Update $\lambda_{\mathbf{t+1}} \leftarrow C\Big[\lambda_{\mathbf{t}} \bullet \exp\Big(\frac{t_0^{\lambda}}{\sqrt{t}}\frac{\partial \mathcal{F}}{\partial \boldsymbol{\lambda}}\Big)\Big]$.
  Set $t \leftarrow t+1$.
until $(t > t_{\max})$ or $(\|\alpha_{\mathbf{t}} - \alpha_{\mathbf{t-1}}\| < \epsilon)$ or $(\|\lambda_{\mathbf{t}} - \lambda_{\mathbf{t-1}}\| < \epsilon)$.
Output: vectors $\alpha_{\mathbf{t}}$ and $\lambda_{\mathbf{t}}$.

The Gram matrix $\mathcal{G}$ is provided by Proposition 1 while its determinant is studied in Proposition 2. The proofs for these two propositions are given in the Supplementary.

Proposition 1 *Let T be a $n \times (n+1)$ matrix whose rows are $\{F_*\partial_i\}_{1\leq i\leq n}$, I is an identity matrix in $\mathbb{R}^{(n+1)\times(n+1)}$, D is a diagonal matrix in $\mathbb{R}^{(n+1)\times(n+1)}$ where $[D]_{ii} = \frac{\mathbf{x}_i^{\frac{\boldsymbol{\alpha}_i}{2}-1}\boldsymbol{\alpha}_i\sqrt{\boldsymbol{\lambda}_i}}{2\sqrt{\sum_{\ell=1}^{n+1}\mathbf{x}_\ell^{\boldsymbol{\alpha}_\ell}\boldsymbol{\lambda}_\ell}}$, $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are column vectors in $\mathbb{R}^{n+1}$ where $\boldsymbol{\beta}_i = \mathbf{x}_i^{\boldsymbol{\alpha}_i-1}\boldsymbol{\alpha}_i\boldsymbol{\lambda}_i$ and $\boldsymbol{\eta}_i = \frac{\mathbf{x}_i}{\boldsymbol{\alpha}_i\sum_{\ell=1}^{n+1}\mathbf{x}_\ell^{\boldsymbol{\alpha}_\ell}\boldsymbol{\lambda}_\ell}$ for all $1 \leq i \leq (n+1)$. We have $T = U(I - \boldsymbol{\beta}\boldsymbol{\eta}^T)D$, and the Gram matrix is given by*

$$\mathcal{G} = TT^T = U(I - \boldsymbol{\beta}\boldsymbol{\eta}^T)D^2(I - \boldsymbol{\beta}\boldsymbol{\eta}^T)^TU^T.$$

Proposition 2 *The determinant of the Gram matrix $\mathcal{G}$ is*

$$det\mathcal{G} \propto \frac{\left(\sum_{i=1}^{n+1}\frac{\mathbf{x}_i}{\boldsymbol{\alpha}_i}\right)^2\left(\prod_{i=1}^{n+1}\mathbf{x}_i^{\boldsymbol{\alpha}_i-2}\right)}{\left(\sum_{i=1}^{n+1}\mathbf{x}_i^{\boldsymbol{\alpha}_i}\boldsymbol{\lambda}_i\right)^{n+1}}$$

4.4. Gradient Ascent using Contrastive Divergences

The main obstacle of our optimization problem is the normalization term of the inverse volume element since it is not known in closed form. However, we can bypass this factor to compute a partial derivative of the objective function $\mathcal{F}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ as given in Proposition 3. Its proof is given in the Supplementary.

Proposition 3 *Let $E(\cdot)_{\mathbf{X}}$ denote the expectation of $\cdot$ given the data distribution $\mathbf{X}$, and a distribution,*

$$p(\mathbf{x}) = \frac{dvolJ^{-1}(\mathbf{x})}{\int_{\mathbb{P}_n} dvolJ^{-1}(\mathbf{z})d\mathbf{z}}. \tag{5}$$

The partial derivative of the objective function $\mathcal{F}$ with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ in the optimization problem are:

$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{\alpha}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \log dvolJ^{-1}(\mathbf{x_i})}{\partial \boldsymbol{\alpha}} - E\left(\frac{\partial \log dvolJ^{-1}(\mathbf{x})}{\partial \boldsymbol{\alpha}}\right)_{p(\mathbf{x})} - \mu\sum_{j=1}^{n+1}\frac{\log \boldsymbol{\alpha}_j}{\boldsymbol{\alpha}_j}$$

where

$$\frac{\partial \log dvolJ^{-1}(\mathbf{x})}{\partial \boldsymbol{\alpha}} = \frac{n+1}{2\sum_{i=1}^{n+1}\mathbf{x}_i^{\boldsymbol{\alpha}_i}\boldsymbol{\lambda}_i}\left[\mathbf{x}_j^{\boldsymbol{\alpha}_j}\boldsymbol{\lambda}_j\log\mathbf{x}_j\right]_{1\leq j\leq n+1} + \frac{1}{\sum_{i=1}^{n+1}\frac{\mathbf{x}_i}{\boldsymbol{\alpha}_i}}\left[\frac{\mathbf{x}_j}{\boldsymbol{\alpha}_j^2}\right]_{1\leq j\leq n+1} - \frac{1}{2}\left[\log\mathbf{x}_j\right]_{1\leq j\leq n+1}$$

and

$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{\lambda}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \log dvolJ^{-1}(\mathbf{x_i})}{\partial \boldsymbol{\lambda}} - E\left(\frac{\partial \log dvolJ^{-1}(\mathbf{x})}{\partial \boldsymbol{\lambda}}\right)_{p(\mathbf{x})}$$

where

$$\frac{\partial \log dvolJ^{-1}(\mathbf{x})}{\partial \boldsymbol{\lambda}} = \frac{n+1}{2\sum_{i=1}^{n+1}\mathbf{x}_i^{\boldsymbol{\alpha}_i}\boldsymbol{\lambda}_i}\left[\mathbf{x}_j^{\boldsymbol{\alpha}_j}\right]_{1\leq j\leq n+1}$$

We propose to approximate the expectation $E(\cdot)_{p(\mathbf{x})}$ that appears in Proposition 3 by drawing samples from the distribution $p(\mathbf{x})$. Since the partition function $\int_{\mathbb{P}_n}\text{dvol}J^{-1}(\mathbf{z})\text{d}\mathbf{z}$ is not known in closed form, we can not draw samples directly from $p(\mathbf{x})$. However, we can use Markov Chain Monte Carlo (MCMC) sampling methods to draw such samples. Because we only need to compute the ratio of two probabilities, ${}^{p(\mathbf{x})}/_{p(\mathbf{z})}$ an approximation for the partition function itself is not required. Moreover, Hinton (2002) suggests that only a few cycles of MCMC can provide in certain settings a useful approximation. The intuition is that the data have moved from the target distribution – training data – towards the proposed distribution $p(\mathbf{x})$ after a few iterations.

We propose to use a Metropolis-Hasting sampling method with a logistic normal distribution (Aitchison & Shen,

1980) proposal. We note that the logistic normal distribution is also a by-product of Aitchison's simplicial geometry. We apply contrastive divergences (Hinton, 2002) to compute approximations of the partial derivative of $\mathcal{F}$ as shown in the proof of the Proposition 3.

We propose to use a gradient ascent to optimize for the metric learning problem following the results in the Proposition 3. At iteration t, we can update $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ using preset step size $\frac{t_0^\alpha}{\sqrt{t}}$ and $\frac{t_0^\lambda}{\sqrt{t}}$ respectively, as follow

$$\alpha_{\mathbf{t+1}} = \Pi\left(\alpha_{\mathbf{t}} + \frac{t_0^\alpha}{\sqrt{t}} \frac{\partial \mathcal{F}}{\partial \boldsymbol{\alpha}}\right),$$
$$\lambda_{\mathbf{t+1}} = C\left[\lambda_{\mathbf{t}} \bullet \exp\left(\frac{t_0^\lambda}{\sqrt{t}} \frac{\partial \mathcal{F}}{\partial \boldsymbol{\lambda}}\right)\right],$$

where $\Pi(\mathbf{x})$ is the projection of $\mathbf{x}$ on the positive orthant offset by a small minimum threshold $\varepsilon = 10^{-20}$, namely the set of all vectors whose coordinates are larger or equal to 10^{-20}, and $\bullet$ is the Schur product between vectors or matrices, and the $\exp$ operator is here applied element-wise. Since we have a constraint $\boldsymbol{\lambda} \in \text{int}\mathbb{P}_n$ in the optimization problem (4), we use an exponentiated gradient update for $\boldsymbol{\lambda}$ (Kivinen & Warmuth, 1997).

We recall that computing an approximation of the normalization term for a specific transformation in (Lebanon, 2002; 2006) takes $O(n^2 \log n)$ by careful dynamic programming. So, our proposal is more efficient and general than Lebanon's approach. A pseudo-code for the projected gradient ascent algorithm is summarized in Algorithm 1.

We also note that the optimization problem (4) can be interpreted as maximizing log-likelihood for the probabilistic model on the simplex (Equation (5) and Proposition 2) which assigns probabilities propositional to the inverse Riemannian volume element, with a regularization.

5. Locally Sensitive Hashing to Approximate k-Nearest Neighbors Search

We recall that our proposed family of distances (Equation (3) in Section 4.1) is the pull-back metric of the Euclidean metric on the positive sphere through a composition transformation of Hellinger mapping and Aitchison transformation. Equivalently, it can be considered as measuring the angle between two mapped vectors from the composition transformation. So, we can apply the Locally Sensitive Hashing family proposed by Charikar (2002) to approximate k-nearest neighbors search.

For two histogram vectors $\mathbf{x}, \mathbf{z} \in \text{int}\mathbb{P}_n$, we have the corresponding mapped vector $\bar{\mathbf{x}} = F(\mathbf{x}), \bar{\mathbf{z}} = F(\mathbf{z}) \in \mathbb{S}_n^+$ via the composition transformation F. Charikar (2002) defines a hash function

$$h_{\mathbf{r}}(\bar{\mathbf{x}}) = \text{sign}(\mathbf{r}^T \bar{\mathbf{x}}),$$

where $\mathbf{r}$ is a random unit-length vector in $\mathbb{R}^{n+1}$. The hash function can be considered as a randomly chosen hyperplane to partition the space into two half-spaces. The probability of collision is as follow

$$\Pr\left[h_{\mathbf{r}}(\bar{\mathbf{x}}) = h_{\mathbf{r}}(\bar{\mathbf{z}})\right] = 1 - \frac{d_J(\mathbf{x}, \mathbf{z})}{\pi}.$$

For a random vector $\mathbf{r}$, we have a hash-bit $h_{\mathbf{r}}(\cdot)$ for each histogram $\mathbf{x}$ in a database. We use b random vectors for a total b hash functions to obtain hash keys (b hash bits) for each histogram. For a query histogram $\mathbf{z}$, we apply the same b hash functions, and then use the approximated similarity search method in (Charikar, 2002) which requires to search $O(m^{1/(1+\varepsilon)})$ histograms for $k = 1$ approximated nearest neighbor.

6. Related Work

Lebanon's use of Aitchison's perturbation operator provided the main inspiration for the metric learning approach advocated in this work (2002; 2006). We propose to extend this idea to other operations in the simplex. We also propose to adapt the contrastive divergence method for the purpose of computing a gradient to maximize inverse volumes, whereas Lebanon uses an approximation for the partition function which only applies to the pertubation transformation. We also show in the experimental section that our approach can also be used in Lebanon's original setting.

Recently, Le & Cuturi (2014) proposed generalized Aitchison embeddings to learn metrics for histograms. Rather than using Aitchison transformations, the authors focus on a different family of tools, Aitchison maps, that can map points in the simplex onto a Euclidean space $\mathbb{R}^d$. Le & Cuturi (2014) propose to learn simultaneously the parameters of such maps and the metric (a Mahalanobis metric) on $\mathbb{R}^d$ that will be used on such representations. This is related, although very different, from the approach we propose here that learns in an unsupervised way a map from and onto the simplex, to be used with Fisher's information metric.

7. Experiments

7.1. Clustering application with K-Medoids

7.1.1. Datasets and Experimental Setting

We use the K-medoids clustering algorithm seeded with different metrics and compute their clusters. We set the number of clusters K equal to the number of classes in corresponding datasets. To evaluate the adequacy of a metric for given data, we check that these clusters agree with a class typology provided for these points[1]. We test our

[1] In this setting for clustering application, we process with unlabelled data (for both learning the distance and applying to K-

Table 1. Properties of datasets and their corresponding experimental parameters.

Dataset	#Samples	#Class	Feature	Rep	#Dim	#Run
MIT Scene	1600	8	SIFT	BoF	200	100
UIUC Scene	3000	15	SIFT	BoF	200	100
OXFORD Flower	1360	17	SIFT	BoF	200	100
CALTECH-101	3060	102	SIFT	BoF	200	100
20 News Group	10000	20	BoW	LDA	200	100
Reuters	2500	10	BoW	LDA	200	100
MNIST-60K	60000	10	Normalized Intensity		784	4
CIFAR-10	60000	10	BoW	SIFT	200	4

method on 6 benchmark datasets. Table 1 displays their properties and parameters. These datasets include different kinds of data such as scene images in MIT Scene[2] and UIUC Scene[3] datasets, flower images in Oxford Flower[4] dataset, object images in CALTECH-101[5] dataset and texts in Reuters[6] and 20 News Group[7] datasets.

7.1.2. Implementation Notes

For image datasets, we compute dense SIFT features by operating a SIFT descriptor of 16×16 patches computed over each pixel of an image. We also convert images into gray scale ones before computing dense SIFT to improve robustness. We use the LabelMe toolbox[8] for computing dense SIFT features. Then, we use bag-of-features (BoF) to represent for each image as a histogram, the size of dictionary for visual words is set 200.

For text datasets, we calculate bag of words (BoW) for each document, and then compute topic modelling to reduce the dimension of histograms using the *gensim* toolbox[9]. Each document can be thus described as a histogram of topics (Blei et al., 2003; Blei & Lafferty, 2009).

We use the PMTK3 toolbox[10] implementation of the K-medoids algorithm. For each metric, we performs K-medoids algorithm 100 times with different random initializations, resulting in box-plots for our error statistics.

We may use $\alpha_0 = [1, 1, \cdots, 1]$ and $\lambda_0 = C[\alpha_0]$ for initialization since our proposed distance (Equation (3) in Section 4.1) is equivalent to the Fisher Information Metric (Equation (1) in Section 3) at these values for α and λ. We also propose to use an internal criterion - Davies-Bouldin index (Davies & Bouldin, 1979) to select parameters via applying K-medoids clustering algorithm. We choose gradient step size t_0^α and t_0^λ from the sets $\frac{1}{\left\|\frac{\partial \mathcal{F}}{\partial \alpha}(\alpha_0,\lambda_0)\right\|_2}\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and $\frac{1}{\left\|\frac{\partial \mathcal{F}}{\partial \lambda}(\alpha_0,\lambda_0)\right\|_2}\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ respectively and μ from $\{0.1, 1, 10\}$. We set maximum iterations t_{max} = 10000 and a tolerance $\epsilon = 10^{-5}$. We also set 5 cycles for Metropolis-Hasting sampling algorithm to transform training data into data drawn from $p(\mathbf{x})$. The logistic normal distribution is used as the proposal distribution for Metropolis-Hasting algorithm where its mean is training data point, and its covariance is set $0.01I$ where I is an identity matrix.

7.1.3. Metrics and Metric Learning Baseline Method

We use usual metrics on the simplex such as the Euclidean, the total variation, χ^2 and Hellinger distances. We recall that the Hellinger distance between two histograms $\mathbf{x}$ and $\mathbf{z}$ in the simplex $\mathbb{P}_n$ is $d_{\text{Hellinger}}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n+1} \left(\sqrt{\mathbf{x}_i} - \sqrt{\mathbf{z}_i}\right)^2$. We also consider the cosine similarity as suggested in (Lebanon, 2002; 2006) and the most popular of Aitchison mappings, known as isometric log-ratio (**ilr**) (Egozcue et al., 2003; Le & Cuturi, 2014). Additionally, we compare our proposal with the work of (Lebanon, 2002; 2006) implemented using our algorithm to maximize inverse volumes, denoted as `pFIM`.

7.1.4. $\mathbf{F}_\beta$ measure

We use the $\mathbf{F}_\beta$ measure to compare results of K-medoids clustering with different metrics (Manning et al., 2008). The intuition is that a pair of histograms is assigned to the same cluster if and only if they are in the same class and otherwise[11]. So, a true positive (TP) decision assigns a pair

medoids clustering method). Labels are only used to evaluate the clustering results. We use K-medoids clustering algorithm instead of a traditional K-means since it is not trivial to compute a mean with respect to a specific distance (i.e our proposed distance).

[2] http://people.csail.mit.edu/torralba/code/spatialenve-lope/

[3] http://www.cs.illinois.edu/homes/slazebni/research/

[4] http://www.robots.ox.ac.uk/∼vgg/data/flowers/17/

[5] http://www.vision.caltech.edu/Image_Datasets/Cal-tech101/

[6] http://archive.ics.uci.edu/ml/datasets/Reuters–21578+Text+Categorization+Collection

[7] http://qwone.com/∼jason/20Newsgroups/

[8] http://new-labelme.csail.mit.edu/Release3.0/

[9] http://radimrehurek.com/gensim/

[10] https://github.com/probml/pmtk3

[11] We note that the class label y_i corresponding for a histogram $\mathbf{x_i}$, for all $1 \leq i \leq m$, is only used for evaluation. In training

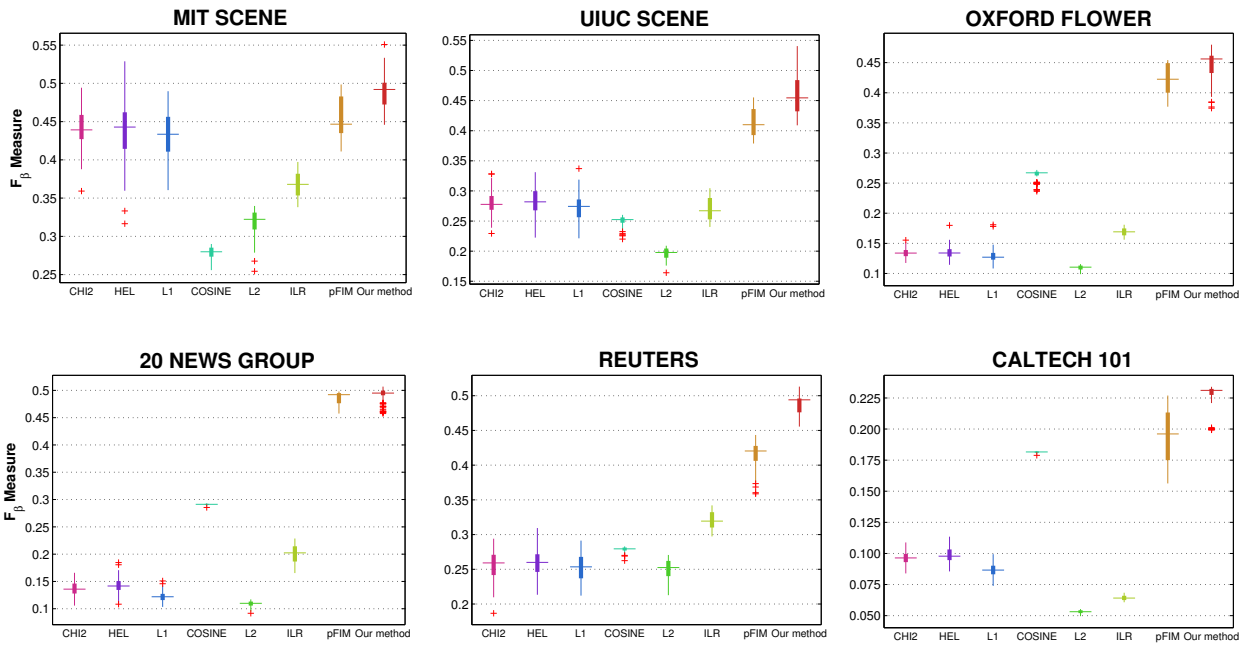

Figure 1. $\mathbf{F}_\beta$ measure for K-medoids clustering on MIT Scene, UIUC Scene, Oxford Flower, 20 News Group, Reuters, and CALTECH 101 datasets where we denote CHI2 for χ^2 distance, HEL for Hellinger distance, L1 for total variation distance, COSINE for cosine similarity, L2 for Euclidean distance, ILR for isometric log-ratio mapping - the most popular Aitchison mapping and pFIM for Fisher information metric pameterized by a pertubation transformation (Lebanon, 2002; 2006).

of histograms in the same class to the same cluster while a true negative (TN) one assigns a pair of histograms in the different classes to the different clusters. We have two types of errors. A false positive (FP) decision assigns a pair of histograms of different classes to the same cluster, and a false negative (FN) one assigns a pair of histograms of the same class to different clusters. Therefore, we can measure the precision $\mathbf{P} = \frac{TP}{TP+FP}$ and recall $\mathbf{R} = \frac{TP}{TP+FN}$. Since we have more pairs of histograms in different classes than in the same class, we need to penalize false negative more strongly than false positives. $\mathbf{F}_\beta$ measure can take into account of that idea through a scalar $\beta > 1$ as $\mathbf{F}_\beta = \frac{(\beta^2+1)\mathbf{PR}}{\beta^2\mathbf{P}+\mathbf{R}}$. By replacing $\mathbf{P}$ and $\mathbf{R}$ into $\mathbf{F}_\beta$, we note that $\mathbf{F}_\beta$ penalizes false negative β^2 times more than false positives. So, let $\mathcal{D}$ and $\mathcal{S}$ be sets of pairs of histograms in different and same classes of a dataset respectively, we can set $\beta = \sqrt{\frac{|\mathcal{D}|}{|\mathcal{S}|}}$ where $|\cdot|$ denotes a cardinality of a set.

7.1.5. Results

Figure 1 illustrates $\mathbf{F}_\beta$ measure for K-medoids clustering on 6 benchmark datasets. It shows that the Euclidean distance, which fails to incorporate the geometrical constraints in the simplex, does not work well for histogram data. Some popular distances for histograms such as total variation distance, χ^2 distance and Hellinger distance as well as the Aitchison mapping - **ilr** give better results than the simple Euclidean distance. Cosine similarity (or angular distance) has a better or comparative performance to these popular distances for histograms, except on MIT Scene and UIUC Scene datasets. The performances of Riemannian metric learning using Aitchison transformations is significantly better, notably on the UIUC Scene, Oxford Flower, 20 News Group and Reuters datasets.

7.2. k-Nearest Neighbors Classification with Locally Sensitive Hashing

We also carry out k-nearest neighbors classification with locally sensitive hashing. We use 2 large datasets MNIST-60K[12] and CIFAR-10[13]. Each dataset consists of 60000 images, we randomly choose 50000 images as a database and use the rest 10000 images for queries. Table 1 displays their properties and parameters.

procedure, only histograms (without labels) are available.

[12] http://yann.lecun.com/exdb/mnist/

[13] http://www.cs.toronto.edu/~kriz/cifar.html

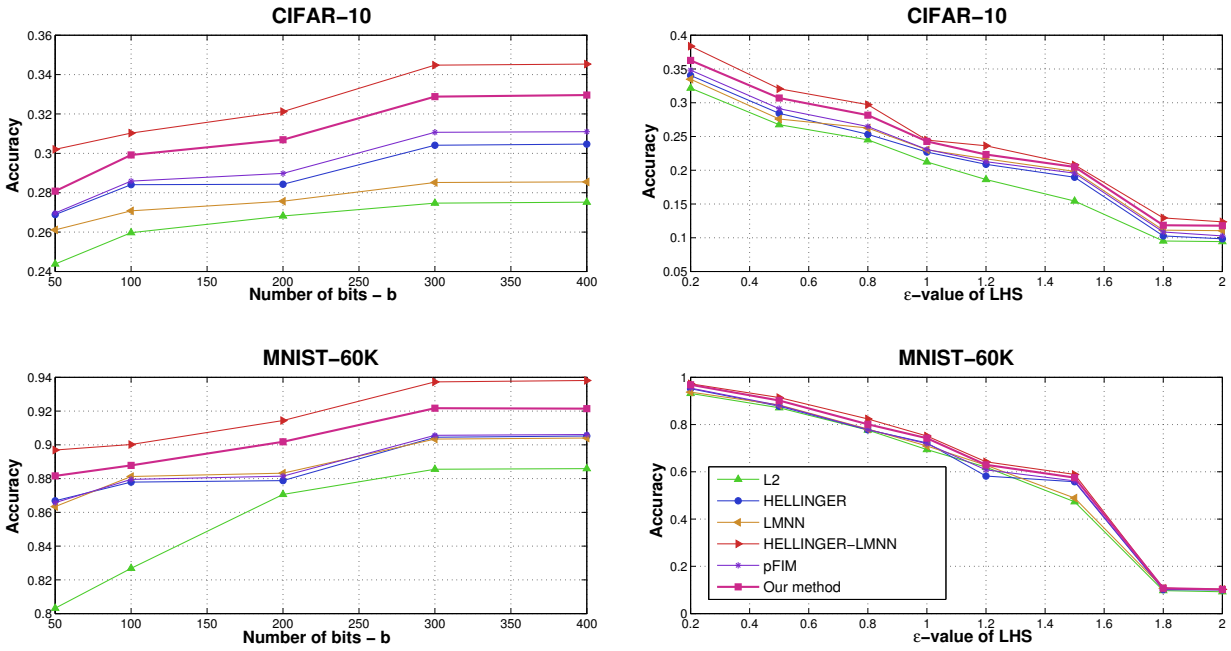

Figure 2. Performances of k-Nearest neighbors with locally sensitive hashing on CIFAR-10 and MNIST-60K datasets, averaged over 4 repetitions where we denote L2 for Euclidean distance, HELLINGER for Hellinger distance, LMNN for Mahalanobis distance learned by Large Margin Nearest Neighbor Weinberger et al. (2006); Weinberger & Saul (2009) algorithm, HELLINGER-LMNN for LMNN learned from data mapped by Hellinger transformation and pFIM for Fisher information metric pameterized by a pertubation transformation (Lebanon, 2002; 2006). For figure Accuracy vs Number of bits - b, we set ε=0.5. For figure Accuracy vs ε-value, we set b=200. All figures are reported with k=7, since in our experiments, the relative performance of these classifiers does not vary with k.

To handle large datasets, we propose a variance of Algorithm 1 by using a mini-batch stochastic gradient (Bengio, 2007). Instead of using the whole samples at each iteration to compute gradients, we randomly choose a small subset of the order of 10 samples as suggested in (Bengio, 2007) to speed up the learning procedure.

As baselines, we consider the Euclidean, a Mahalanobis distance learned by using Large Margin Nearest Neighbors (`LMNN`) Weinberger et al. (2006); Weinberger & Saul (2009) algorithm. We also consider Hellinger distance and Hellinger mapping with a Mahalanobis distance learned by using `LMNN`, denoted as `HELLINGER-LMNN`, as well as the approach of (Lebanon, 2002; 2006) as mentioned in Section 7.1.3.

Figure 2 illustrates our results on MNIST-60K and CIFAR-10 datasets. Our approach outperforms other alternative distances except `HELLINGER-LMNN` which should be expected, given that it is a state of the art *supervised* metric learning approach for histograms. Figure 2 also shows that Euclidean distance and a straightforward application of `LMNN` do not work well for histogram data. We insist that `HELLINGER-LMNN` uses labels to learn a Mahalanobis matrix while our approach do *not* consider them.

8. Conclusion

We propose a new unsupervised metric learning approach for histograms that leverages Aitchison transformations for histograms in the simplex. These transformations are learned with the maximum inverse volume framework of Lebanon (2006). We provide a new algorithm to carry out such a maximization using contrastive divergences which solves the key obstacle - the partition function for a general case. We show empirically that our proposal can learn effectively histogram metrics for unlabeled data. It outperforms alternative popular metrics for histograms such as χ^2, Hellinger, total variation, Euclidean distance, cosine similarity and an Aitchison map (**ilr**) in clustering problem on many benchmark datasets. Additionally, it also improves the performances of k-nearest neighbors classification with locally sensitive hashing for large datasets.

Acknowledgments

We thank anonymous reviewers for their comments. TL acknowledges the support of the MEXT scholarship 123353. MC acknowledges the support of the Japanese Society for the Promotion of Science grant 25540100.

References

Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society*, 44:139–177, 1982.

Aitchison, J. *The statistical analysis of compositional data*. Chapman and Hall, Ltd., 1986.

Aitchison, J. A concise guide to compositional data analysis. In *Compositional Data Analysis Workshop*, 2003.

Aitchison, J. and Shen, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika*, pp. 261–272, 1980.

Bengio, Y. Speeding up stochastic gradient descent. In *Workshop on Efficient Machine Learning, Neural Information Processing Systems*, 2007.

Blei, D. and Lafferty, J. *Topic models*. Text Mining: Classification, Clustering, and Applications, 2009.

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Campbell, W. M. and Richardson, F. S. Discriminative keyword selection using support vector machines. In *Advances in Neural Information Processing Systems*, 2007.

Charikar, M. S. Similarity estimation techniques from rounding algorithms. In *ACM symposium on Theory of computing*, pp. 380–388. ACM, 2002.

Cuturi, M. and Avis, D. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227, 1979.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic metric learning. In *International Conference on Machine Learning*, pp. 209–216, 2007.

Doddington, G. Speaker recognition based on idiolectal differences between speakers. In *Eurospeech*, pp. 2521–2524, 2001.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcel-Vidal, C. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

Globerson, A. and Roweis, S. T. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems*, pp. 451–458, 2005.

Goldberger, J., Roweis, S. T., Hinton, G. E., and Salakhutdinov, R. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, 2004.

Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800, 2002.

Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Springer, 2002.

Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature*, 1981.

Kedem, D., Tyree, S., Weinberger, K. Q., Sha, F., and Lanckriet, G. Nonlinear metric learning. In *Advances in Neural Information Processing Systems*, pp. 2582–2590, 2012.

Kivinen, J. and Warmuth, M.K. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Le, T. and Cuturi, M. Adaptive euclidean maps for histograms: generalized aitchison embeddings. *Machine Learning*, pp. 1–19, 2014.

Lebanon, G. Learning riemannian metrics. In *Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann Publishers Inc., 2002.

Lebanon, G. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508, 2006.

Manning, C.D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Salton, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of information by Computer*. Addison-Wesley, 1989.

Salton, G. and McGill, M. J. *Introduction to Moderm Information Retrieval*. McGraw-Hill, 1983.

Schultz, M. and Joachims, T. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems*, volume 16, pp. 41, 2003.

Shalev-Shwartz, S., Singer, Y., and Ng, A. Y. Online and batch learning of pseudo-metrics. In *International Conference on Machine Learning*, pp. 94, 2004.

Sivic, J. and Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, 2003.

Vedaldi, A. and Zisserman, A. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

Weinberger, K.Q. and Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

Weinberger, K.Q., Blitzer, J., and Saul, L. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 1473–1480, 2006.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pp. 1473–1480, 2002.