An Information Geometry of Statistical Manifold Learning

Sun, Ke; Marchand-Maillet, Stéphane

An Information Geometry of Statistical Manifold Learning

Ke Sun KE.SUN@UNIGE.CH
Stéphane Marchand-Maillet STEPHANE.MARCHAND-MAILLET@UNIGE.CH
Viper Group, Computer Vision & Multimedia Laboratory, University of Geneva, Switzerland

Abstract

Manifold learning seeks low-dimensional representations of high-dimensional data. The main tactics have been exploring the geometry in an input data space and an output embedding space. We develop a manifold learning theory in a hypothesis space consisting of models. A model means a specific instance of a collection of points, e.g., the input data collectively or the output embedding collectively. The semi-Riemannian metric of this hypothesis space is uniquely derived in closed form based on the information geometry of probability distributions. There, manifold learning is interpreted as a trajectory of intermediate models. The volume of a continuous region reveals an amount of information. It can be measured to define model complexity and embedding quality. This provides deep unified perspectives of manifold learning theory.

NOTATIONS

$\boxed{\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{Z}_n}$ — respectively, a generic model in $\Re^{\mathfrak{D}}$, an MAL input model in $\Re^D$, and an MAL output model in $\Re^d$. The subscripts denote the sample size and can be omitted; $\boxed{\mathcal{M}(\boldsymbol{X}),\ \widetilde{\mathcal{M}},\ \mathcal{M}_z,\ \mathcal{O}^n_{\boldsymbol{X},\mathfrak{k}},\ \mathcal{H}^{2n}_{\boldsymbol{X},\boldsymbol{Y},\mathfrak{k}},\ \bar{\mathcal{H}}^2_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k},\kappa}}$ — different model manifolds. The superscripts denote the dimension. The parameters in parentheses denote the coordinate system. Both can be omitted; $\boxed{\mathcal{S}^{n-1}}$ — the $(n-1)$-dimensional statistical simplex; $\boxed{g, \mathfrak{G}}$ — semi-Riemannian metric of a model family and Fisher information metric; $\boxed{s^{\boldsymbol{X}}_{ij}}$ — pairwise dissimilarities of a model $\boldsymbol{X}$. The superscript can be omitted; $\boxed{p_{j|i}, p_{ij}}$ — neighbourhood probabilities; $\boxed{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)}$ — canonical parameters of probability distributions; $\boxed{|\mathcal{M}|}$ — volume or scale of $\mathcal{M}$; $\boxed{\mathfrak{k}\mathtt{NN}_i}$ — indexes of the $\mathfrak{k}$-nearest-neighbours of the i'th sample; $\boxed{\mathtt{det}(\cdot)}$ — determinant; $\boxed{\mathtt{diag}(x_1, \ldots, x_\ell)}$ — a diagonal matrix with $x_1, \ldots, x_\ell$ on its main diagonal.

Manifold learning (MAL), or non-linear dimensionality reduction, assumes that some given high-dimensional observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \Re^D$ lie around a low-dimensional sub-manifold $\{\Gamma(\boldsymbol{z}) : \boldsymbol{z} \in \Re^d\}$ induced by a smooth mapping $\Gamma : \Re^d \to \Re^D$ $(d \ll D)$. While it is possible to learn a parametric form of Γ (Hinton & Salakhutdinov, 2006), the majority of manifold learners are *non-parametric*. They learn directly a set of low-dimensional coordinates $\{\boldsymbol{z}_i\}_{i=1}^n$ to preserve certain information in $\{\boldsymbol{y}_i\}_{i=1}^n$.

Depending on the choice of information to be preserved, at least two families of MAL methods thrived in the last decade. The spectral methods (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003) and semi-definite embeddings (Weinberger et al., 2004; Sha & Saul, 2005) represent the family with natural convex formulations. They only preserve encodings of local information on a weighted $\mathfrak{k}$-nearest-neighbour ($\mathfrak{k}\mathtt{NN}$) graph of $\{\boldsymbol{y}_i\}$. Stochastic Neighbour Embedding (SNE) (Hinton & Roweis, 2003) and its extensions (Cook et al., 2007; Venna & Kaski, 2007; van der Maaten & Hinton, 2008) represent the non-convex family. They encode the input and output as probability distributions and optimize the embedding in a maximum-likelihood framework. By sacrificing convexity, non-local information can be preserved as well. The latter SNE-based family shows robustness to parameter configuration and favorable performance in data visualization. It is being actively developed (Carreira-Perpiñán, 2010; Vladymyrov & Carreira-Perpiñán, 2012; Sun et al., 2012; Yang et al., 2013) and stands as state-of-the-art MAL.

Despite such a diversity, several critical problems in the field of MAL remain unclear. Practically, no standard exists in gauging the data complexity and the embedding

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

quality. The performance is often empirically assessed via visualization or indirectly evaluated via classification. Theoretically, an intrinsic MAL theory with deep connections to classical statistical learning theory (Akaike, 1974; Schwarz, 1978; Amari, 1995; Vapnik, 1998) is not established. MAL emphasizes local information encoded into sample-wise structures. How to describe and measure such preservation of local information is unknown.

We attack these problems with a geometry, not in an observation space $\Re^D$ or an embedding space $\Re^d$, but in a very high-dimensional hypothesis space made of models.

Definition 1. *A model $\boldsymbol{X}_n$ is a specific instance of a set of vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.*

Remark 1.1. *By default, a model denoted by $\boldsymbol{X}_n$ is a coordinate matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. Alternatively, it can be implicitly specified by a $n \times n$ matrix of pairwise measurements, e.g., distances or (dis-)similarities.*

Definition 2. *A model family $\mathcal{M}$ is a smooth manifold consisting of continuous models.*

For example in MAL, the input $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ or the output $\boldsymbol{Z}_n = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$ is one single model. The model family $\mathcal{M}_z = \{\boldsymbol{Z}_n = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T : \sum_i \boldsymbol{z}_i = \boldsymbol{0};\ \forall i, \boldsymbol{z}_i \in \Re^d\}$ includes all possible embeddings centered at $\boldsymbol{0}$. Then, MAL can be described as a projection $\boldsymbol{Y} \to \boldsymbol{Z}^\star(\boldsymbol{Y}) \in \mathcal{M}_z$ through convex optimization, or a path $\boldsymbol{Z}^0(\boldsymbol{Y}), \boldsymbol{Z}^1(\boldsymbol{Y}), \ldots, \boldsymbol{Z}^\star(\boldsymbol{Y}) \in \mathcal{M}_z$ along the gradient of some non-convex objective function.

1. Preliminaries

1.1. Manifold Learning

Given a model family $\mathcal{M}$, any model $\boldsymbol{X}_n \in \mathcal{M}$, representing a collection of coordinates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, can be encoded into n distributions over $\{1, 2, \ldots, n\}$:

$$p_{j|i}(\boldsymbol{X}) = \frac{\exp\left(-s_{ij}^{\boldsymbol{X}}\right)}{\sum_{j \neq i} \exp\left(-s_{ij}^{\boldsymbol{X}}\right)}\ (\forall j \neq i),\ p_{i|i} = 0\ (\forall i) \tag{1}$$

or one single distribution over $\{1, 2, \ldots, n\}^2$:

$$p_{ij}(\boldsymbol{X}) = \frac{\exp\left(-s_{ij}^{\boldsymbol{X}}\right)}{\sum_{\imath,\jmath:\imath\neq\jmath} \exp\left(-s_{\imath\jmath}^{\boldsymbol{X}}\right)}\ (\forall j \neq i),\ p_{ii} = 0\ (\forall i). \tag{2}$$

In either case, $s_{ij}^{\boldsymbol{X}}$ is a possibly non-symmetric difference measure between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, e.g., square distance. After normalization, p represents the probability of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ being similar. The subscript "$j|i$" of p in eq. (1) signifies a conditional probability; the subscript "ij" in eq. (2) signifies a joint probability.

It is not arbitrary but natural to employ eqs. (1) and (2) for statistical MAL, because they encode *distributed local information*. The information in p is distributed in a sample-wise manner. Each sample $\boldsymbol{x}_i$ has limited knowledge encoded into $p_{\cdot|i}$ mostly regarding its neighbours.

Equations (1) and (2) are general enough to cover SNE (Hinton & Roweis, 2003), symmetric SNE (Cook et al., 2007), t-SNE (van der Maaten & Hinton, 2008), and a spectrum of extensions. For example, SNE applies $s_{ij}^{\boldsymbol{X}} = \tau_i ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$ to eq. (1) for encoding the input and output, where $\tau_i > 0$ is a scalar; t-SNE applies $s_{ij}^{\boldsymbol{X}} = \log(||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 + 1)$ to eq. (2) for encoding the output. From a kernel view (Ham et al., 2004), any MAL technique that encodes into kernels naturally extends to such probabilities.

Despite a model $\boldsymbol{X}$ can have various forms, after the encoding $p(\boldsymbol{X})$ is in a unified space. In eq. (1), $(p_{j|i}(\boldsymbol{X}))$ is a point on the product manifold $(\mathcal{S}^{n-1})^n$, where $\mathcal{S}^{n-1} = \{(p_1, \ldots, p_n) : \forall i, p_i > 0;\ \sum_{i=1}^n p_i = 1\}$ is a *statistical simplex* consisting of all distributions over $\{1, 2, \ldots, n\}$. In eq. (2), $(p_{ij}(\boldsymbol{X}))$ is a point on $\mathcal{S}^{n^2-1}$. Such a unified representation makes it possible to measure the difference between two models $\boldsymbol{Y}$ and $\boldsymbol{Z}$ with different original forms. It motivates us to develop an MAL theory on the statistical simplex regardless of the original representations.

1.2. Information Geometry

We introduce the Riemannian geometry of $\mathcal{S}^{n-1}$, the $(n-1)$-dimensional statistical simplex formed by all distributions in the form $(p_1, \ldots, p_n)$. The readers are referred to (Jost, 2008; Amari & Nagaoka, 2000) for a thorough view.

Any $(p_1, \ldots, p_n) \in \mathcal{S}^{n-1}$ uniquely corresponds to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ via the invertible mapping $\theta_i = \log(p_i/p_r)$, $\forall i$, where p_r $(1 \leq r \leq n)$ is a reference probability. These *canonical parameters* $\boldsymbol{\theta}$ serve as a global coordinate system of $\mathcal{S}^{n-1}$. Around $\forall \boldsymbol{\theta} \in \mathcal{S}^{n-1}$, the partial derivative operators $\{\partial/\partial\theta_1, \cdots, \partial/\partial\theta_{r-1}, \partial/\partial\theta_{r+1}, \cdots, \partial/\partial\theta_n\}$ represent the *velocities* passing through $\boldsymbol{\theta}$ along the coordinate curves. An infinitesimal patch of $\mathcal{S}^{n-1}$ around $\boldsymbol{\theta}$ can be studied as a linear space $\mathcal{T}_{\boldsymbol{\theta}}\mathcal{S}^{n-1} = \{\sum_{i:i\neq r}(\alpha_i \cdot \partial/\partial\theta_i) : \forall i, \alpha_i \in \Re\}$ called the *tangent space*. A *Riemannian metric* $\mathfrak{G}$ defines a local inner product $\langle \partial/\partial\theta_i, \partial/\partial\theta_j \rangle_{\mathfrak{G}(\boldsymbol{\theta})}$ on each tangent space $\mathcal{T}_{\boldsymbol{\theta}}\mathcal{S}^{n-1}$ and varies smoothly across different $\boldsymbol{\theta}$. Locally, it is given by the positive definite (p.d.) matrix $\mathfrak{G}_{ij}(\boldsymbol{\theta}) = \langle \partial/\partial\theta_i, \partial/\partial\theta_j \rangle_{\mathfrak{G}(\boldsymbol{\theta})}$. Under certain conditions (Čencov, 1982), the Riemannian metric of statistical manifolds, e.g. $\mathcal{S}^{n-1}$, is uniquely given by the Fisher information metric (FIM) (Rao, 1945) $\mathfrak{G}_{ij}(\boldsymbol{\theta}) = \sum_{k=1}^n p_k(\boldsymbol{\theta}) \left(\partial \log p_k(\boldsymbol{\theta})/\partial\theta_i\right) \left(\partial \log p_k(\boldsymbol{\theta})/\partial\theta_j\right)$.

Lemma 3. *On $\mathcal{S}^{n-1}$, $\mathfrak{G}_{ij}(\boldsymbol{\theta}) = p_i(\boldsymbol{\theta})\delta_{ij} - p_i(\boldsymbol{\theta})p_j(\boldsymbol{\theta})$.*

FIM grants us the power to measure information. With respect to a coordinate system, e.g. the canonical pa-

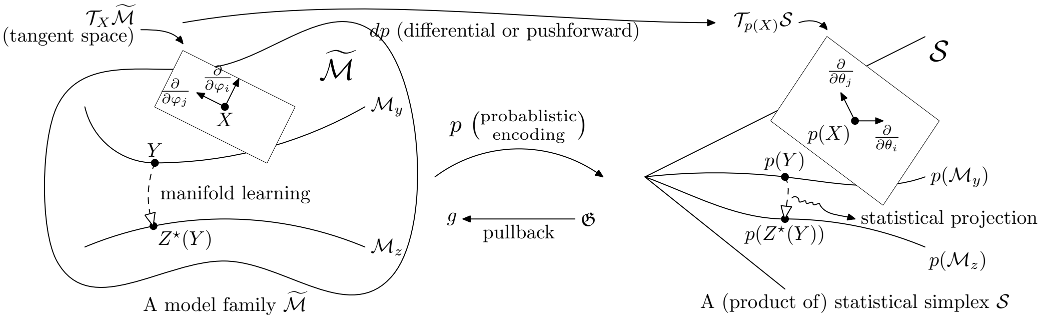

$\mathcal{T}_X\widetilde{\mathcal{M}}$
(tangent space)
dp (differential or pushforward)
$\mathcal{T}_{p(X)}\mathcal{S}$
$\mathcal{S}$
$\frac{\partial}{\partial\varphi_j}$
$\frac{\partial}{\partial\varphi_i}$
X
$\widetilde{\mathcal{M}}$
$\mathcal{M}_y$
Y
manifold learning
$Z^\star(Y)$
$\mathcal{M}_z$
A model family $\widetilde{\mathcal{M}}$
p $\binom{\text{probablistic}}{\text{encoding}}$
g
pullback
$\mathfrak{G}$
$\frac{\partial}{\partial\theta_j}$
$p(X)$
$\frac{\partial}{\partial\theta_i}$
$p(Y)$
$p(\mathcal{M}_y)$
statistical projection
$p(Z^\star(Y))$
$p(\mathcal{M}_z)$
A (product of) statistical simplex $\mathcal{S}$

Figure 1. A geometry of statistical manifold learning.

rameters $\boldsymbol{\theta}$, the *information density* of a statistical model $\boldsymbol{\theta} \in \mathcal{M}$ is given by the Riemannian volume element $\sqrt{\mathtt{det}\,(\mathfrak{G}(\boldsymbol{\theta}))}$ (Jost, 2008). It means the amount of information a single observation possesses with respect to $\boldsymbol{\theta}$. A small $\sqrt{\mathtt{det}\,(\mathfrak{G}(\boldsymbol{\theta}))}$ means that $\boldsymbol{\theta}$ contains much uncertainty and requires many observations to learn. Then, the *information capacity* of a statistical model family $\mathcal{M} \subset \mathcal{S}^{n-1}$ is given by its volume $|\mathcal{M}| = \int_{\boldsymbol{\theta}\in\mathcal{M}} \sqrt{\mathtt{det}\,(\mathfrak{G}(\boldsymbol{\theta}))}d\boldsymbol{\theta}$. It means the total amount of information, or the "number" of distinct models, contained in $\mathcal{M}$ (Myung et al., 2000). This volume is an intrinsic measure invariant to the choice of coordinate system.

To extend FIM to a general non-statistical manifold, e.g., a model family $\mathcal{M}^r$ parametrized by $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_r)$, one need to construct a mapping from $\mathcal{M}^r$ to $\mathcal{S}^{n-1}$. Following such a mapping, any tangent vector $\partial/\partial\varphi_i$ of $\mathcal{M}^r$ can be *pushed forward* to $\mathcal{T}_{\boldsymbol{\theta}}\mathcal{S}^{n-1}$ as $\partial/\partial\varphi_i = \sum_{k=1}^n (\partial\theta_k/\partial\varphi_i \cdot \partial/\partial\theta_k)$. In this way, the inner product $\langle\partial/\partial\varphi_i, \partial/\partial\varphi_j\rangle$ can be measured with FIM and used to define the Riemannian metric of $\mathcal{M}^r$. Such a strategy is called *pullback* (Jost, 2008). This is intuitively shown in fig. 1. $\mathcal{M}^r$ (left) as in definition 2 mirrors the information geometry (right) through a probabilistic encoding p.

2. An Intrinsic Geometry of MAL

The central result of this paper is summarized as follows.

Theorem 4. *Consider a model family $\mathcal{M}^r = \{s_{ij}(\boldsymbol{\varphi}) : \boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_r)^T \in \Phi\}$, where $\Phi \subset \Re^r$. The pullback information metric with respect to Eq.(1) is given by*

$$g(\boldsymbol{\varphi}) = \sum_{k=1}^n \left[\sum_{l=1}^n p_{l|k} \left(\frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right) \left(\frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right)^T - \left(\sum_{l=1}^n p_{l|k} \frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right) \left(\sum_{l=1}^n p_{l|k} \frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right)^T \right];$$

the pullback metric with respect to Eq.(2) is

$$g(\boldsymbol{\varphi}) = \sum_{k=1}^n \sum_{l=1}^n p_{kl} \left(\frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right) \left(\frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right)^T - \left(\sum_{k=1}^n \sum_{l=1}^n p_{kl} \frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right) \left(\sum_{k=1}^n \sum_{l=1}^n p_{kl} \frac{\partial s_{kl}}{\partial\boldsymbol{\varphi}}\right)^T,$$

where s_{kl}, $p_{l|k}$ and p_{kl} vary with $\boldsymbol{\varphi}$ as in eqs. (1) *and* (2).

Remark 4.1. *$g(\boldsymbol{\varphi})$ is a meta-metric. Its exact form depends on $s_{ij}(\boldsymbol{\varphi})$ and the choice between eqs.* (1) *and* (2). *Different encodings lead to different geometries of information.*

Remark 4.2. *We leave the reader to verify $g(\boldsymbol{\varphi}) \succeq 0$. Therefore $g(\boldsymbol{\varphi})$ is called a semi-Riemannian metric. "Semi" means $g(\boldsymbol{\varphi})$ is positive semi-definite (p.s.d.) rather than p.d* [1]. *A model $\boldsymbol{X}$ moving rigidly forms a null space (Jost, 2008), a model family with zero volume, meaning that such movements contribute zero information.*

From theorem 4, the pullback metrics induced by eq. (1) and eq. (2) are in similar forms. Both are some covariance of $\partial s_{kl}/\partial\boldsymbol{\varphi}$. Such similarity agrees with previous studies where SNE and symmetric SNE show similar performance (Cook et al., 2007). Because of space limitation, the following results are derived based on the metric induced by eq. (1) only. The subtle difference between these two kinds of normalizations is left to a longer version.

A natural question arises as to what kind of geometry is endowed to the *ambient model family* $\widetilde{\mathcal{M}}^{n\mathfrak{D}} = \{\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T : \forall i, \boldsymbol{x}_i \in \Re^{\mathfrak{D}}\}$. This is meaningful because a model family $\mathcal{M}$ of interest is often a sub-manifold of $\widetilde{\mathcal{M}}$. With respect to the Euclidean coordinates, its semi-Riemannian metric $\tilde{g}$ defines for any $\boldsymbol{X} \in \widetilde{\mathcal{M}}$ an $(n\mathfrak{D} \times n\mathfrak{D})$ p.s.d. matrix $\tilde{g}(\boldsymbol{X})$. We investigate its $\mathfrak{D} \times \mathfrak{D}$ sub-block $\langle\partial/\partial\boldsymbol{x}_i, \partial/\partial\boldsymbol{x}_i\rangle_{\tilde{g}(\boldsymbol{X})}$, which can be used to measure the infinitesimal length $\|d\boldsymbol{x}_i\|$ when $\boldsymbol{x}_i$ is shifted to

[1] In other contexts, a semi-Riemannian metric is also defined to be non-degenerate with full rank (Jost, 2008)

$\boldsymbol{x}_i + d\boldsymbol{x}_i$ while the other samples stay, or the intrinsic difference caused by such an increment $d\boldsymbol{x}_i$.

Corollary 5.

$$\left\langle \frac{\partial}{\partial \boldsymbol{x}_i}, \frac{\partial}{\partial \boldsymbol{x}_i} \right\rangle_{\tilde{g}(\boldsymbol{X})} = 4 \sum_{j=1}^{n} p_{i|j}(1-p_{i|j})(\boldsymbol{x}_j - \boldsymbol{x}_i)(\boldsymbol{x}_j - \boldsymbol{x}_i)^T$$
$$+ 4 \sum_{j=1}^{n} p_{j|i} \left(\boldsymbol{x}_j - \sum_{j=1}^{n} p_{j|i}\boldsymbol{x}_j \right) \left(\boldsymbol{x}_j - \sum_{j=1}^{n} p_{j|i}\boldsymbol{x}_j \right)^T .$$

Corollary 5 reveals an interesting relationship between information geometry and data geometry. The right-hand side is like a local covariance matrix around $\boldsymbol{x}_i$ with locality defined by p. If $d\boldsymbol{x}_i$ is orthogonal to the data manifold, the resulting $\|d\boldsymbol{x}_i\|$ is small. This explains: while the data manifold is *unfolded* in the sense that the samples mainly move along the normal directions, the corresponding model goes along a path on $\widetilde{\mathcal{M}}$ with small information volume. Across different samples, Corollary 5 quantifies the potential information in each $\boldsymbol{x}_i$ with the Riemannian volume element $\sqrt{det(\langle \partial/\partial\boldsymbol{x}_i, \partial/\partial\boldsymbol{x}_i\rangle_{\tilde{g}(\boldsymbol{X})})}$. Such sample-wise information provides theoretical quantities to outlier identification or landmark selection in sub-sampling procedures (van der Maaten & Hinton, 2008). Corollary 5 gives a geometric interpretation of manifold kernel density estimation (Vincent & Bengio, 2003) as growing density to maximize the information variance on $\widetilde{\mathcal{M}}$.

The target of this paper is a learning theory centered on the above theorem 4 and supported by corollaries 5, 6 and 10 with illustrative simulations. We investigate two independent-yet-related model families, corresponding to information geometries of model complexity and model quality. We do not present a systematical comparison with other measurements. They are as many as manifold learners. *None escapes the fact that it is measured in the observation space or the embedding space* (Venna & Kaski, 2007; Zhang et al., 2011). In contrast, we regard all samples collectively as one point and measure information on a differentiable manifold of such models. In the history of statistical learning, similar information geometric theories contributed deep insights (Akaike, 1974; Amari & Nagaoka, 2000; Myung et al., 2000; Lebanon, 2003; Xu, 2004; Nock et al., 2011). We echo these previous works and adapt to recent developments of MAL. Given the uniqueness of FIM, the proposed measurements try to estimate the true information loss in MAL. This is more general than and fundamentally different from empirical measurements.

3. Locally Accumulated Information

We measure the complexity of a fixed model $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ given by a matrix $(s_{ij}^{\boldsymbol{X}})_{n\times n}$ of pairwise differences. We only study the case that $s_{ij}^{\boldsymbol{X}} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$. To generalize to other similar cases is trivial. Our strategy is to vary $\boldsymbol{X}$ in certain ways to form a model family. Such variation represents different perspectives to measure and perceive information. The scale of this model family exposes the total amount of information in $\boldsymbol{X}$ with respect to these variations.

We install on each sample $\boldsymbol{x}_i$ an isotropic *observer*. It perceives information encoded with eq. (1), where s_{ij} is parametrized as

$$s_{ij}(\boldsymbol{\tau}) = \begin{cases} \tau_i \cdot s_{ij}^{\boldsymbol{X}} & \text{if } j \in \mathfrak{k}\mathtt{NN}_i; \\ +\infty & \text{if otherwise,} \end{cases} \tag{3}$$

or

$$s_{ij}^{t}(\boldsymbol{\tau}) = \begin{cases} \log(\tau_i \cdot s_{ij}^{\boldsymbol{X}} + 1) & \text{if } j \in \mathfrak{k}\mathtt{NN}_i; \\ +\infty & \text{if otherwise.} \end{cases} \tag{4}$$

$\forall i$, $\tau_i > 0$ zooms other samples near or far from $\boldsymbol{x}_i$, so that information at different scales can be incorporated. $\mathfrak{k}$ ($2 \leq \mathfrak{k} \leq n-1$) specifies the visual range in the maximum number of samples that can be observed by any $\boldsymbol{x}_i$. The purpose of $\mathfrak{k}$ is to ignore distant relationships to make related computation scalable. Datasets of different size are measured on the same statistical manifold $\mathcal{S}^{\mathfrak{k}-1}$. Such measurements are therefore comparable. As compared to eq. (3), the distribution defined by eq. (4) has higher entropy. Even if $\tau_i \to \infty$, meaning zero sight, distant pairs still occupy some probability mass. Hence, eq. (4) puts more emphasis on non-local information.

Once $\boldsymbol{X}$ and $\mathfrak{k}$ are fixed, all possible configurations of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ forms a n-dimensional model manifold denoted as $\mathcal{O}^n_{\boldsymbol{X},\mathfrak{k}}(\boldsymbol{\tau})$. Its geometry is given as follows.

Corollary 6. *If the observers are characterized by eq.* (3)*, the Riemannian metric of $\mathcal{O}^n_{\boldsymbol{X},\mathfrak{k}}(\boldsymbol{\tau})$ is $g(\boldsymbol{\tau}) =$ $\mathtt{diag}\,(g_1(\tau_1), \cdots, g_n(\tau_n))$, where*

$$g_i(\tau_i) = -\frac{\partial}{\partial \tau_i} \left(\sum_{j\in\mathfrak{k}\mathtt{NN}_i} p_{j|i}(\tau_i)\, s_{ij}^{\boldsymbol{X}} \right)$$
$$= \sum_{j\in\mathfrak{k}\mathtt{NN}_i} p_{j|i}(\tau_i) \left(s_{ij}^{\boldsymbol{X}}\right)^2 - \left(\sum_{j\in\mathfrak{k}\mathtt{NN}_i} p_{j|i}(\tau_i) s_{ij}^{\boldsymbol{X}} \right)^2 ;$$

if the observers in eq. (4) *are used instead, the corresponding metric is $g^t(\boldsymbol{\tau}) = \mathtt{diag}\,(g^t_1(\tau_1), \cdots, g^t_n(\tau_n))$, where*

$$g^t_i(\tau_i) = \sum_{j\in\mathfrak{k}\mathtt{NN}_i} p_{j|i}(\tau_i) \left(\frac{s_{ij}^{\boldsymbol{X}}}{1+\tau_i s_{ij}^{\boldsymbol{X}}} \right)^2$$
$$- \left(\sum_{j\in\mathfrak{k}\mathtt{NN}_i} p_{j|i}(\tau_i) \frac{s_{ij}^{\boldsymbol{X}}}{1+\tau_i s_{ij}^{\boldsymbol{X}}} \right)^2 .$$

DATASET	n	$\mathfrak{D}$	TRUE d	MLE	LAI	t-LAI
Spiral	200	3	1	1.3 ± 0.23	1.2 ± 0.21	1.1 ± 0.06
Swiss roll	500	3	2	2.1 ± 0.13	2.0 ± 0.08	2.0 ± 0.05
Faces	698	$4k$	$\sim$3	3.8 ± 0.42	3.4 ± 0.27	3.2 ± 0.15
Hands	481	$245k$	$\sim$2	2.2 ± 0.37	1.7 ± 0.21	2.0 ± 0.10
MNIST	60k	784	unknown	10.1 ± 0.27	9.7 ± 0.55	9.8 ± 0.26

Table 1. Estimated intrinsic dimension ($avg. \pm std.$) for each $k \in \{5, 10, \ldots, 100\}$.

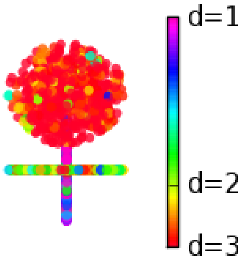


Figure 2. Local dimension estimation.

We only discuss $g(\boldsymbol{\tau})$ and leave $g^t(\boldsymbol{\tau})$ for future extentions. The scale of $\mathcal{O}^n_{\boldsymbol{X},\mathfrak{k}}(\boldsymbol{\tau})$ can be measured as follows.

Definition 7. *The locally accumulated information (LAI) of $\boldsymbol{X}$ given the visual range $\mathfrak{k}$ is defined as*

$$|\mathcal{O}_{\mathfrak{k}}|(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_0^{\infty}\sqrt{g_i(t)}\,dt\right). \tag{5}$$

Remark 7.1. *$\mathcal{O}^n_{\boldsymbol{X},\mathfrak{k}}(\boldsymbol{\tau})$ resembles an orthant $(0,\infty)^n$, and LAI measures its average side length.*

To understand LAI, one need to grasp the concept of information. Shannon's information, i.e. entropy, measures the absolute uncertainty of a single distribution. Fisher's information (Rao, 1945) measures the relative uncertainty within a continuous region of distributions (Myung et al., 2000). In eq. (5), each term in the sum is Fisher's information integrated along a statistical curve corresponding to a local observation process. LAI measures how much intrinsic difference, or how "many" distinct distributions, are observed during zooming the observation radius from 0 to ∞.

Proposition 8. $\forall\boldsymbol{X}, \forall\mathfrak{k}, \forall\lambda > 0$, $|\mathcal{O}_{\mathfrak{k}}|(\boldsymbol{X}) = |\mathcal{O}_{\mathfrak{k}}|(\lambda\boldsymbol{X})$.

Proposition 9. *If $\forall i$, $\boldsymbol{x}_i$'s 1-NN is unique, then $\forall\mathfrak{k}$,* $\arccos\left(1/\sqrt{\mathfrak{k}}\right) \le |\mathcal{O}_{\mathfrak{k}}|(\boldsymbol{X}) < \infty$.

LAI is invariant to scalling (proposition 8). LAI is always finite and has a lower bound (proposition 9). This lower bound can be approached by approximately placing $\boldsymbol{X}$ as a regular n-simplex. This is only possible when the dimension of $\{\boldsymbol{x}_i\}$ is large enough. In fact, LAI reflects the intrinsic dimensionality. In high dimension, the pairwise distances present less variance (Bellman, 1961). Correspondingly, the curve of distributions defined by eqs. (3) and (4) is straighter and shorter. In low dimension, the observerations at different scales are more different. Correspondingly, the curve of distributions is more bended. Consider a line of cities (London, Paris, Geneva, Rome). As an observer in London expands its sight, it discovers the other three cities one by one. On $\mathcal{S}^2$, a curve starts from the Paris vertex, then bends towards Geneva, then bends towards Rome. In this model, LAI ≈ 5.3. A rectangular model (Berlin, Paris, Vienna, Marseille) has "higher dimensionality" and therefore a smaller value of LAI (≈ 5.1).

LAI can be conveniently calculated with an off-the-shelf numerical integrator. The computation involves n integrations, which can be reduced by averaging over a random subset of samples in eq. (5). The cost of each integration scales with $\mathfrak{k}$ ($\mathfrak{k} \ll n$). Overall, the computation is scalable.

A Dictionary-based Intrinsic Dimensionality Estimator

We generate random artificial data with different dimension $\mathfrak{D}$ to build a dictionary of LAI values indexed by $\mathfrak{k}$ and $\mathfrak{D}$. This dictionary is used to map any input data to an intrinsic dimensionality. Table 1 shows its performance compared to the maximum likelihood estimator (MLE) (Levina & Bickel, 2005) on several benchmark datasets, including a spiral with one intrinsic dimension, a Swiss roll with two intrinsic dimensions, an artificial face dataset[2] rendered with different light directions and different orientations (three degrees of freedom), an image sequence[3] recording a hand holding a bowl and rotating (around two degrees of freedom), and MNIST hand-written digits [4]. In order to suppress the curse of dimensionality (Bellman, 1961), a dataset is projected to $\Re^{50}$ with principal component analysis (PCA) before going to the estimators. Overall, the LAI estimation is closer to the (estimated) ground truth with less variance. t-LAI is based on eq. (4) with similar definitions. It shows the best robustness to the choice of $\mathfrak{k}$, because s^t_{ij} is able to incorporate more non-local information. MNIST is closer to real-world datasets, where the ground truth is unknown and the intrinsic dimension varies from region to region. The large variance of (t-)LAI is because the intrinsic dimension changes with $\mathfrak{k}$. At a small scale ($\mathfrak{k} = 5$), the intrinsic dimension tends to be overestimated (~ 10) because of local high-dimensional noise. At a larger scale ($k = 100$), the intrinsic dimension is esti-

[2] http://isomap.stanford.edu/datasets.html

[3] http://vasc.ri.cmu.edu//idb/html/motion/hand/index.html

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

mated as $8 \sim 9$ as the data manifold shows up. This observation agrees with the characteristics of real-world data.

LAI can be used for local dimensionality estimation (Carter et al., 2010) by computing a local average of $\int_0^\infty \sqrt{g_i(t)}dt$ around any $\boldsymbol{x}_i$. Figure 2 shows the side view of a candy bar dataset with a 1D stick, a 2D disk, and a 3D head. Its local dimension showed by the colors is well-estimated.

4. A Gap between Two Models

A central problem in MAL is to define the difference between an input $\boldsymbol{Y}_n$ and an output $\boldsymbol{Z}_n$. Then, the embedding quality can be evaluated, and MAL can be implemented through optimization. According to section 3, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ individually extend to two families modeling their internal complexity. Our strategy is to continuously deform one family to the other along a bridging manifold. The volume of this bridge measures their intrinsic difference.

Fortunately, such a bridge exists for any given $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Consider the model family $\mathcal{H}^{2n}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k}}$ defined by

$$s_{ij}(\boldsymbol{c}) = \begin{cases} a_i s^{\boldsymbol{Y}}_{ij} + b_i s^{\boldsymbol{Z}}_{ij} & \text{if } j \in \mathfrak{k}\mathtt{NN}_i; \\ +\infty & \text{if otherwise,} \end{cases} \tag{6}$$

or

$$s^t_{ij}(\boldsymbol{c}) = \begin{cases} a_i s^{\boldsymbol{Y}}_{ij} + \log\left(b_i s^{\boldsymbol{Z}}_{ij} + 1\right) & \text{if } j \in \mathfrak{k}\mathtt{NN}_i; \\ +\infty & \text{if otherwise,} \end{cases} \tag{7}$$

where $\forall i$, $\mathfrak{k}\mathtt{NN}_i = \mathfrak{k}\mathtt{NN}_i(\boldsymbol{Y}) \cup \mathfrak{k}\mathtt{NN}_i(\boldsymbol{Z})$ are the input or output neighbours of i, $a_i > 0$, $b_i > 0$, and $\boldsymbol{c} = (a_1, b_1, \ldots, a_n, b_n)$ serves as a global coordinate system. The boundary $\boldsymbol{b} = \boldsymbol{0}$ deteriorates to a model family induced by $\boldsymbol{Y}$ (similar to $\mathcal{O}^n_{\boldsymbol{Y},\mathfrak{k}}$). The boundary $\boldsymbol{a} = \boldsymbol{0}$ deteriorates to a family induced by $\boldsymbol{Z}$ (similar to $\mathcal{O}^n_{\boldsymbol{Z},\mathfrak{k}}$). Among all possible interpolations, eqs. (6) and (7) are simple and natural. $\mathcal{H}^{2n}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k}}$ defined in eq. (6) is somehow flat (Amari & Nagaoka, 2000). Its geometry is given as follows.

Corollary 10. *With respect to eq. (6) and the global coordinate system $\boldsymbol{c} = (a_1, b_1, \ldots, a_n, b_n)$, the Riemannian metric of $\mathcal{H}^{2n}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k}}$ is in the block-diagonal form*

$$g(\boldsymbol{c}) = \begin{bmatrix} g^1_{aa} & g^1_{ab} & & & \\ g^1_{ba} & g^1_{bb} & & & \\ & & \ddots & & \\ & & & g^n_{aa} & g^n_{ab} \\ & & & g^n_{ba} & g^n_{bb} \end{bmatrix},$$

where $\forall i$,

$$g^i_{aa} = \sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i} \left(s^{\boldsymbol{Y}}_{ij}\right)^2 - \left(\sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i}\, s^{\boldsymbol{Y}}_{ij}\right)^2,$$

$$g^i_{bb} = \sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i} \left(s^{\boldsymbol{Z}}_{ij}\right)^2 - \left(\sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i}\, s^{\boldsymbol{Z}}_{ij}\right)^2,$$

$$g^i_{ab} = g^i_{ba} = \sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i} \left(s^{\boldsymbol{Y}}_{ij} \cdot s^{\boldsymbol{Z}}_{ij}\right) - \left(\sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i}\, s^{\boldsymbol{Y}}_{ij}\right)\left(\sum_{j \in \mathfrak{k}\mathtt{NN}_i} p_{j|i}\, s^{\boldsymbol{Z}}_{ij}\right).$$

The metric $g^t(\boldsymbol{c})$ with respect to eq. (7) is obtained by replacing $s^{\boldsymbol{Z}}_{ij}$ with $s^{\boldsymbol{Z}}_{ij} / \left(1 + b_i s^{\boldsymbol{Z}}_{ij}\right)$ in the above equations.

Due to space limitation, the following discussion is only based on the geometry induced by eq. (6).

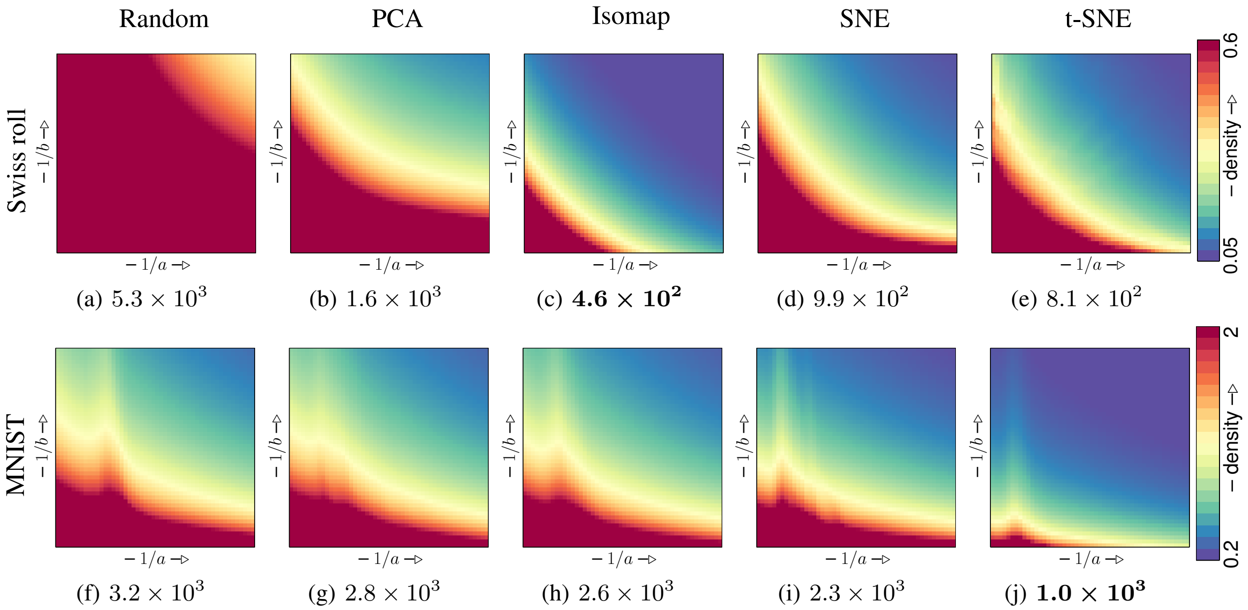
$\mathcal{H}^{2n}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k}}$ embeds all information regarding the intra- and inter-complexity of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Across this bridge, we construct a low dimensional *film*, whose volume can be easily computed to estimate the closeness between the input and output boundaries. Consider the 2D sub-manifold

$$\bar{\mathcal{H}}^2_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k},\kappa} = \{(a_1, b_1, \ldots, a_n, , b_n)\ :\ \forall i, a_i = a\tau_i(\boldsymbol{Y}, \kappa),\ b_i = b\tau_i(\boldsymbol{Z}, \kappa);\ a \in (0, \infty); b \in (0, \infty); \kappa < \mathfrak{k}\}$$

with a global coordinate system (a, b). All observers in $\boldsymbol{Y}$ (resp. $\boldsymbol{Z}$) simultaneously zoom according to one single parameter a (resp. b). A large value of a (resp. b) corresponds to high frequency local information; a small value of a (resp. b) corresponds to low frequency distant information. For each i, the scalars $\tau_i(\boldsymbol{Y}, \kappa)$ and $\tau_i(\boldsymbol{Z}, \kappa)$ are computed by binary search, so that the distribution $p_{\cdot|i}$ defined by eqs. (1), (6) and (7) has fixed entropy given by $\log\kappa$ and each sample has effectively the same number of neighbours given by κ (Hinton & Roweis, 2003). Such alignment is to derive two lines $a = 0$ and $b = 0$ as close as possible on each side of the gap. The volume of the film $\bar{\mathcal{H}}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k},\kappa}$ in between these lines approximates the minimal efforts needed to shift a continuous spectrum of information from one family to the other.

Proposition 11. *The volume (area) of some region Ω on $\bar{\mathcal{H}}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k},\kappa}$ is*

$$|\bar{\mathcal{H}}_{\mathfrak{k},\kappa,\Omega}|(\boldsymbol{Y}, \boldsymbol{Z}) = \iint_\Omega da\, db \sqrt{\left(\sum_{i=1}^n \tau_i^2(\boldsymbol{Y}, \kappa) g^i_{aa}\right) \times \left(\sum_{i=1}^n \tau_i^2(\boldsymbol{Z}, \kappa) g^i_{bb}\right) - \left(\sum_{i=1}^n \tau_i(\boldsymbol{Y}, \kappa)\tau_i(\boldsymbol{Z}, \kappa) g^i_{ab}\right)^2}. \tag{8}$$



(a) 5.3×10^3 (b) 1.6×10^3 (c) $\mathbf{4.6 \times 10^2}$ (d) 9.9×10^2 (e) 8.1×10^2

(f) 3.2×10^3 (g) 2.8×10^3 (h) 2.6×10^3 (i) 2.3×10^3 (j) $\mathbf{1.0 \times 10^3}$

Figure 3. Performance measurements of different embeddings. In each sub-figure, the color-map shows the local densities $vol(a,b)d\boldsymbol{\sigma}(a,b)$ over Ω. From left to right (resp. bottom to up), the input (resp. output) observation radius expands from 5 to 50. The x-axis (resp. y-axis) is linear in $1/a$ (resp. $1/b$). The number below each square shows the volume, i.e. the integration of local densities in the corresponding color-map. Note, the colors are different between the two datasets.

Proposition 12. $\forall \boldsymbol{Y}, \forall \boldsymbol{Z}, \forall \lambda_y > 0, \forall \lambda_z > 0,$

$$|\bar{\mathcal{H}}_{\mathfrak{k},\kappa,\Omega}|(\boldsymbol{Y},\boldsymbol{Z}) = |\bar{\mathcal{H}}_{\mathfrak{k},\kappa,\Omega}|(\lambda_y \boldsymbol{Y}, \lambda_z \boldsymbol{Z}).$$

The region Ω means interested information in MAL. It can be chosen empirically as a rectangle to exclude low frequency information above the radius κ and high frequency information below a minimal radius k_s. Given $\boldsymbol{Y}$ and $\boldsymbol{Z}$, the volume $|\bar{\mathcal{H}}_{\mathfrak{k},\kappa,\Omega}|(\boldsymbol{Y},\boldsymbol{Z})$, shortly denoted as $|\bar{\mathcal{H}}_{\Omega}|$, can be efficiently computed with a Monte Carlo integrator. It forms a theoretical objective of MAL that is scale-invariant (proposition 12).

We re-write eq. (8) as $|\bar{\mathcal{H}}_{\Omega}| = \iint_{\Omega} vol(a,b)\ d\boldsymbol{\sigma}(a,b)$. By noting that g^i_{aa}, g^i_{bb}, g^i_{ab} and g^i_{ba} in corollary 10 are all in the form of (co-)variances, the normalized scalar

$$vol(a,b) = \sqrt{1 - \frac{\left(\sum_{i=1}^n \tau_i(\boldsymbol{Y},\kappa)\tau_i(\boldsymbol{Z},\kappa)g^i_{ab}\right)^2}{\sum_{i=1}^n \tau_i^2(\boldsymbol{Y},\kappa)g^i_{aa}\sum_{i=1}^n \tau_i^2(\boldsymbol{Z},\kappa)g^i_{bb}}} \qquad (9)$$

ranges in $[0,1]$ and measures the overall linear correlation between $s^{\boldsymbol{Y}}_{ij}$ and $s^{\boldsymbol{Z}}_{ij}$ for the same i. The more linearly correlated $s^{\boldsymbol{Y}}_{ij}$ and $s^{\boldsymbol{Z}}_{ij}$, the smaller the value of $vol(a,b)$.

Proposition 13. $|\bar{\mathcal{H}}_{\Omega}| = 0$ *iff* $\exists a,b \in \Omega$, *s.t.* $vol(a,b) = 0$.

$vol(a,b)$ is usually positive (proposition 13) and tells the input-output agreement at a specific frequency. The term

$$d\boldsymbol{\sigma}(a,b) = \left(\sqrt{\sum_{i=1}^n \tau_i^2(\boldsymbol{Y},\kappa)g^i_{aa}da}\right)\left(\sqrt{\sum_{i=1}^n \tau_i^2(\boldsymbol{Z},\kappa)g^i_{bb}db}\right)$$

measures the information density or interestingness at (a,b) based on separate observations on $\boldsymbol{Y}$ and $\boldsymbol{Z}$ (see section 3). Therefore, $|\bar{\mathcal{H}}_{\Omega}|$ can be understood as the linear agreement at different scales weighted by information density.

Two classical datasets in MAL, Swiss roll ($n = 10^3$) and MNIST ($n = 5 \times 10^3$; five classes), are embedded into $\Re^2$ by PCA, Isomap, SNE and t-SNE with typical configurations. The parameters $\mathfrak{k}$, κ and k_s are empirically set to 100, 50 and 5, respectively. The gap on MNIST is computed by randomly sampling 10^3 input-output pairs to be comparable with Swiss roll. In fig. 3, the color-maps show $vol(a,b)d\boldsymbol{\sigma}(a,b)$ over Ω. Their appearances depend on the coordinate system of $\bar{\mathcal{H}}_{\boldsymbol{Y},\boldsymbol{Z},\mathfrak{k},\kappa}$. Here, the axises are linear in $1/a$ and $1/b$ and represent the observation radii. For example, the upper-left corner means that the input (resp. output) information is examined at a radius of 5 (resp. 50) samples. The gap volume $|\bar{\mathcal{H}}_{\Omega}|$ below each color-map is independent of the coordinate system. The best method for each dataset (Isomap for Swiss roll; t-SNE for MNIST) is identified by the bluest square with the smallest volume. The patterns on the color-maps tell more detailed information. The redness on the lower-right corner (e.g. fig. 3(b)) indicates that the original neighbours are heavily invaded in the embedding. Apparently, t-SNE outperforms SNE in this region. This is related to an information retrieval perspective (Venna & Kaski, 2007). By comparing fig. 3(f) with fig. 3(a), MNIST with high dimensional noise is closer to random data. It is more difficult to improve over random embeddings in this dataset. The spiky pattern in fig. 3(f-j) shows that some structural information that distinguishes MNIST from random data is forced to a thin band due to the high dimensionality (Bellman, 1961).

Most MAL techniques preserve a single frequency or scale of a specific type of local information. The result strongly favors this frequency and this type of information. The proposed gap yields a family of criteria that are less-biased towards such choices. By mapping onto the statistical manifold, some redundant information is factored out. By aligning and seeking a minimal gap, an intrinsic difference is exposed. The integration over a spectrum gives accurate estimation of the true information loss. In practice, to compute the gap always faces the choice of a statistical encoding and associated parameters, e.g., $\mathfrak{k}$, κ, and k_s. However, the relative order of the gap volumes should be robust to such choices. Despite that the results are developed based on the SNE encodings, the gap volume is fundamentally different from SNE's objective and does not necessarily favor SNE.

5. Related Works and Discussion

Information geometry (Rao, 1945; Čencov, 1982; Nagaoka & Amari, 1982) plays a vital role in statistical learning theory. MAL has been developed along a statistical approach. It is a natural and meaningful step to bridge the profound information geometry. Efforts (Weinberger et al., 2007; Carreira-Perpiñán, 2010; Vladymyrov & Carreira-Perpiñán, 2012) in seeking efficient MAL implicitly used such a geometry. There, a common technique is to bend the gradient $\bigtriangledown(\boldsymbol{Z})$ of a cost function with $M^{-1}(\boldsymbol{Z})\bigtriangledown(\boldsymbol{Z})$, where $M(\boldsymbol{Z})\succ 0$. This is equivalent to compute the gradient with respect to a Riemannian metric $M(\boldsymbol{Z})$ on the solution space. Such a metric, however, has not been explicitly mentioned or formally studied.

Lebanon (2003; 2005) parametrized the Riemannian metric of a statistical simplex and proposed a metric learning objective to maximize the inverse volume element. Carter et al. (2009; 2011) studied MAL on a collections of probability density functions. In these works, the subject is still a data geometry, where the observed data is assumed to lie on a statistical manifold. This is different from the picture shown here, which views all input or output information jointly as one point and studies its dynamics.

We formally introduce a semi-Riemannian geometry of a model manifold. It broadens our horizons so that MAL appears as a curve (figs. 1 and 4) and different manifold learners are viewed from a unified perspective. An intriguing aspect is that any volume corresponds to an amount of information. It can be measured to define intrinsic quantities. On two specific model manifolds, we demonstrate how to apply the theoretical results to measure the complexity and quality of models. These measurements are only briefly sketched here to testify the learning theory. They can be further unfold into meaningful theories. This work is summarized in fig. 4. A fundamental trade-off of MAL is to minimize the volume of the gap (lost information; see section 4) and to maximize the volume of the output (remained information; see section 3). To unify and combine LAI and the gap volume into one criterion and to seek parameter-free invariants are worthy of future work.

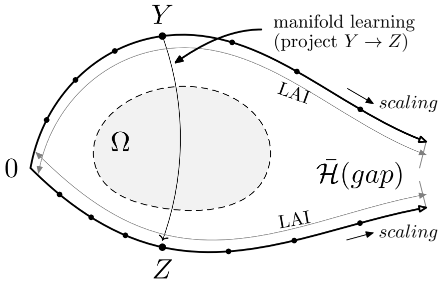


Figure 4. Internal complexity of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ and their gap.

The gap volume in proposition 11 as a theoretical objective is hard to optimize directly. This is expected and fits in a usual two-stage learning scheme (Akaike, 1974; Schwarz, 1978; Xu, 2010). In the parameter learning stage, a simple objective function is optimized for each candidate model. In the model selection stage, a sophisticated criterion that better approximates the generalization error is computed to select the best model. We seek to derive simple approximations of the gap volume and develop related MAL algorithms.

Several possible extensions are discussed at the end of section 2. A problem that fits in the recent advancements (Vladymyrov & Carreira-Perpiñán, 2012; Yang et al., 2013) of MAL is to find efficient optimization based on generalizations of Amari's natural gradient (Amari & Nagaoka, 2000; Nock et al., 2011). A theoretical problem is to explore the relationship with graph Laplacian regularization (Belkin & Niyogi, 2003; Weinberger et al., 2007).

Acknowledgments

This work is jointly supported by the Swiss National Science Foundation (SNSF) and the European COST Action on Multilingual and Multifaceted Interactive Information Access (MUMIA) via the Swiss State Secretariat for Education and Research (SER project CC11.0043).

References

Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, 1974.

Amari, S. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

Amari, S. and Nagaoka, H. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. AMS and OUP, 2000. (Published in Japanese in 1993).

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6): 1373–1396, 2003.

Bellman, R. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

Carreira-Perpiñán, M. Á. The elastic embedding algorithm for dimensionality reduction. In *ICML*, pp. 167–174, 2010.

Carter, K. M., Raich, R., Finn, W. G., and Hero, A. O. FINE: Fisher Information Nonparametric Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):2093–2098, 2009.

Carter, K. M., Raich, R., and Hero, A. O. On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Processing*, 58(2):650–663, 2010.

Carter, K. M., Raich, R., Finn, W. G., and Hero, A. O. Information-geometric dimensionality reduction. *IEEE Signal Process. Mag.*, 28(2):89–99, 2011.

Čencov, N. N. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. AMS, 1982. (Published in Russian in 1972).

Cook, J., Sutskever, I., Mnih, A., and Hinton, G. E. Visualizing similarity data with a mixture of maps. In *AISTATS, JMLR: W&CP 2*, pp. 67–74, 2007.

Ham, J., Lee, D. D., Mika, S., and Schölkopf, B. A kernel view of the dimensionality reduction of manifolds. In *ICML*, pp. 47–54, 2004.

Hinton, G. E. and Roweis, S. T. Stochastic Neighbor Embedding. In *NIPS 15*, pp. 833–840. 2003.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Jost, J. *Riemannian Geometry and Geometric Analysis*. Springer, 5th edition, 2008.

Lebanon, G. Learning Riemannian metrics. In *UAI*, pp. 362–369, 2003.

Lebanon, G. *Riemannian geometry and statistical machine learning*. PhD thesis, CMU, 2005.

Levina, E. and Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In *NIPS 17*, pp. 777–784. 2005.

Myung, J., Balasubramanian, V., and Pitt, M. A. Counting probability distributions: differential geometry and model selection. *PNAS*, 97(21):11170–11175, 2000.

Nagaoka, H. and Amari, S. Differential geometry of smooth families of probability distributions. Technical Report METR 82-7, Univ. of Tokyo, 1982.

Nock, R., Magdalou, B., Briys, E., and Nielsen, F. On tracking portfolios with certainty equivalents on a generalization of markowitz model: the fool, the wise and the adaptive. In *ICML*, pp. 73–80, 2011.

Rao, C. R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Cal. Math. Soc.*, 37(3):81–91, 1945.

Roweis, Sam T. and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000.

Schwarz, G. Estimating the dimension of a model. *Ann. Stat.*, 6 (2):461–464, 1978.

Sha, F. and Saul, L.K. Analysis and extension of spectral methods for nonlinear dimensionality reduction. In *ICML*, pp. 784–791, 2005.

Sun, K., Bruno, E., and Marchand-Maillet, S. Stochastic unfolding. In *MLSP*, pp. 1–6, 2012.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

van der Maaten, L. J. P. and Hinton, G. E. Visualizing data using t-SNE. *JMLR*, 9(Nov):2579–2605, 2008.

Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Venna, J. and Kaski, S. Nonlinear dimensionality reduction as information retrieval. In *AISTATS, JMLR: W&CP 2*, pp. 572–579, 2007.

Vincent, P. and Bengio, Y. Manifold Parzen windows. In *NIPS 15*, pp. 825–832. 2003.

Vladymyrov, M. and Carreira-Perpiñán, M. Á. Partial-Hessian strategies for fast learning of nonlinear embeddings. In *ICML*, pp. 345–352, 2012.

Weinberger, K. Q., Sha, F., and Saul, L. K. Learning a kernel matrix for nonlinear dimensionality reduction. In *ICML*, pp. 839–846, 2004.

Weinberger, K. Q., Sha, F., Zhu, Q., and Saul, L. K. Graph Laplacian regularization for large-scale semidefinite programming. In *NIPS 19*, pp. 1489–1496. 2007.

Xu, L. Advances on BYY harmony learning: Information theoretic perspective, generalized projection geometry, and independent factor auto-determination. *IEEE Trans. Neural Networks*, 15(4):885–902, 2004.

Xu, L. Machine learning problems from optimization perspective. *J. Global Optim*, 47(3):369–401, 2010.

Yang, Z., Peltonen, J., and Kaski, S. Scalable optimization of neighbor embedding for visualization. In *ICML, JMLR: W&CP 28.2*, pp. 127–135, 2013.

Zhang, J., Wang, Q., He, L., and Zhou, Z. H. Quantitative analysis of nonlinear embedding. *IEEE Trans. Neural Networks*, 22 (12):1987–1998, 2011.