Isometric Gaussian Process Latent Variable Model for Dissimilarity Data

Martin Jørgensen
marjor@dtu.dk
Technical University of Denmark

Søren Hauberg
sohau@dtu.dk
Technical University of Denmark

Abstract

We propose a fully generative model where the latent variable respects both the distances and the topology of the modeled data. The model leverages the Riemannian geometry of the generated manifold to endow the latent space with a well-defined stochastic distance measure, which is modeled as Nakagami distributions. These stochastic distances are sought to be as similar as possible to observed distances along a neighborhood graph through a censoring process. The model is inferred by variational inference and is therefore fully generative. We demonstrate how the new model can encode invariances in the learned manifolds.

1 Introduction

Dimensionality reduction aims to compress data to a lower dimensional representation while preserving the underlying signal and suppressing noise. Contemporary nonlinear methods mostly call upon the *manifold assumption* [Bengio et al., 2013] stating that the observed data is distributed near a low-dimensional manifold embedded in the observation space. Beyond this unifying assumption, methods often differ by focusing on one of three key properties (Table 1).

Topology preservation. A *topological space* is a set of points whose *connectivity* is invariant to continuous deformations. For finite data, connectivity is commonly interpreted as a clustering structure, such that topology preserving methods do not form new clusters or break apart existing ones. For visualization purposes, the *uniform manifold approximation projection (UMAP)* [McInnes et al., 2018] appears to be the current state-of-the-art within this domain.

	Generative	Topology	Distance
PCA	(✓)	✗	(✓)
MDS	✗	✗	✓
IsoMap	✗	(✗)	✓
t-SNE	✗	(✓)	✓
UMAP	✗	✓	✓
GPLVM	✓	✗	✗
Iso-GPLVM (our)	✓	✓	✓

Table 1: A list of common dimensionality reduction methods and coarse overview of their features.

Distance preservation. Methods designed to find low-dimensional representation with pairwise distances that are similar to those of the observed data may generally be viewed as a variant of *multi-dimensional scaling (MDS)* [Ripley, 2007]. Usually, this is achieved by a direct minimization of the *stress* defined as

$$\text{stress} = \sum_{i<j\leq N} (d_{ij} - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|)^2, \tag{1}$$

where d_{ij} are the *dissimilarity* (or *distance*) of two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ denote the low-dimensional representation in $\mathbb{R}^q$.

More advanced methods have been built on top of this idea. In particular, *IsoMap* [Tenenbaum et al., 2000] computes d_{ij} along a neighborhood graph using Dijkstra's algorithm. This bears some

Preprint. Under review.

resemblance to *t-SNE* [Maaten and Hinton, 2008] that uses the Kullback-Leibler divergence to match distribution in low-dimensional Euclidean spaces with the data in high dimensions.

Generative models. A common trait for the mentioned methods is that they learn features in a mapping from high-dimensions to low, but not the reverse. This makes the methods mostly useful for visualization. *Generative models* [Kingma and Welling, 2014, Rezende et al., 2014, Lawrence, 2005, Goodfellow et al., 2014, Rezende and Mohamed, 2015] allow us to make new samples in high-dimensional space. Of particular relevance to us, is the *Gaussian process latent variable model (GP-LVM)* [Lawrence, 2005, Titsias and Lawrence, 2010] that learns a stochastic mapping $f : \mathbb{R}^q \to \mathbb{R}^D$ jointly with the latent representations $\boldsymbol{z}$. This is achieved by marginalizing the mapping under a Gaussian process prior [Rasmussen and Williams, 2006]. The generative approach allows the methods to extend beyond visualization to e.g. missing data imputation, data augmentation and semi-supervised tasks [Mattei and Frellsen, 2019, Urtasun and Darrell, 2007].

In this paper we learn a Riemannian manifold using Gaussian processes on which distances on the manifold match the *local* distances as is implied by the Riemannian assumption. Assuming the observed data lies on a Riemannian q-submanifold of $\mathbb{R}^D$ with infinite injectivity radius, then our approach can learn a q-dimensional representation that is isometric to the original manifold. Similar statements only hold true for traditional manifold learning methods that embed into $\mathbb{R}^q$ if the original manifold is flat. We learn global and local structure through a common technique from survival analysis, combined with a likelihood model based on the theory of Gaussian process arc-lengths. Lastly, we show how the GP approach allow us to marginalize the latent representation and produce a fully Bayesian non-parametric generative model. We envision how learning generative models by pairwise dissimilarities easily allow for encoding invariances.

2 Background material

2.1 Gaussian Processes

A Gaussian process (GP) [Rasmussen and Williams, 2006] is a distribution over functions, $f : \mathbb{R}^q \to \mathbb{R}$, which satisfy that for any finite set of points $\{\boldsymbol{z}_i\}_{i=1}^N$, in the domain $\mathbb{R}^q$, the output $\boldsymbol{f} = \big(f(\boldsymbol{z}_1), \dots, f(\boldsymbol{z}_N)\big)$ have a joint Gaussian distribution. This Gaussian is fully determined by a mean function $\mu : \mathbb{R}^q \to \mathbb{R}$ and a covariance function $k : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, such that

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}), \tag{2}$$

where $\boldsymbol{\mu} = \big(\mu(\boldsymbol{z}_1, \dots, \mu(\boldsymbol{z}_N)\big)$ and $\boldsymbol{K}$ is the $N \times N$-matrix with (i, j)-th entry $k(\boldsymbol{z}_i, \boldsymbol{z}_j)$.

GPs are well-suited for Bayesian non-parametric regression, since if we condition on data $\mathcal{D} = \{\boldsymbol{z}, x\}$, where x denote the labels, then the posterior of $f(\boldsymbol{z}^*)$, at a test location $\boldsymbol{z}^*$, is given as

$$p\big(f(\boldsymbol{z}^*)|\mathcal{D}\big) = \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{K}^*), \quad \begin{cases} \mu(\boldsymbol{z}^*) + \boldsymbol{\mu}^* = k(\boldsymbol{z}^*, \boldsymbol{z})^\top k(\boldsymbol{z}, \boldsymbol{z})^{-1} x, \\ \boldsymbol{K}^* = k(\boldsymbol{z}^*, \boldsymbol{z}^*) - k(\boldsymbol{z}^*, \boldsymbol{z})^\top k(\boldsymbol{z}, \boldsymbol{z})^{-1} k(\boldsymbol{z}^*, \boldsymbol{z}). \end{cases} \tag{3}$$

We see that this posterior computation involves inversion of the $N \times N$-matrix $\boldsymbol{K}$, which has complexity $\mathcal{O}(N^3)$. To overcome this computational burden in inference we consider variational sparse GP regression, which introduces M auxiliary points $\boldsymbol{u}$, that approximate the posterior of f with a variational distribution q. For a review of variational GP methods, we refer to Titsias [2009].

2.2 Riemannian Geometry

A *manifold* is a topological space, for which each point on it has a neighborhood that is homeomorphic to Euclidean space; that is, manifolds are locally linear spaces. Such manifolds can be embedded into spaces of higher dimension than the dimensionality of the associated Euclidean space; the manifold *itself* has the same dimension as the local Euclidean space. A q-dimensional manifold $\mathcal{M}$ can, for our purposes thus, be seen as a surface embedded in $\mathbb{R}^D$. In order to make quantitative statements along the manifold we require it to be *Riemannian*.

Definition 1. *A Riemannian manifold $\mathcal{M}$ is a smooth q-manifold equipped with an inner product*

$$\langle \cdot, \cdot \rangle_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \times \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \to \mathbb{R}, \qquad \boldsymbol{x} \in \mathcal{M}, \tag{4}$$

that is smooth in $\boldsymbol{x}$. Here $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ evaluated at $\boldsymbol{x}$.

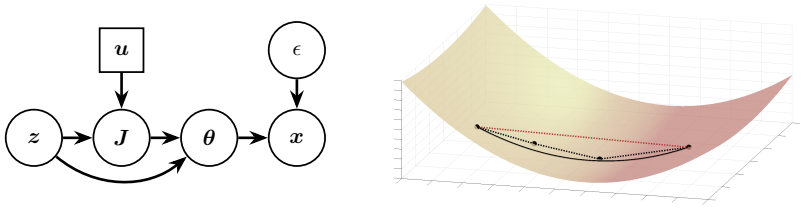

Figure 1: *Left:* A graphical representation of the model: $\boldsymbol{x}$ is the observational input, $\boldsymbol{J}$ is the Gaussian process manifold and $\boldsymbol{\theta}$ are the parameters it yields based on latent embedding $\boldsymbol{z}$. ϵ is a hyperparameter for the neighbor-graph embedding and $\boldsymbol{u}$ are variational parameters. *Right:* Illustration of the task: the dashed lines are Euclidean distances in three dimensions. The black ones are *neighbors* and their distance along the two-dimensional manifold should *match* the 3d-Euclidean distance. The red is not a neighbor-pair and the manifold distance should not match it.

The length of a curve is easily defined from the Riemannian inner product. If $\boldsymbol{c} : [0, 1] \to \mathcal{M}$ is a smooth curve, its length is given by $s = \int_0^1 \|\dot{\boldsymbol{c}}(t)\| \mathrm{d}t$. On an embedded manifold $f(\mathcal{M})$ this becomes

$$s = \int_0^1 \|\dot{f}(\boldsymbol{c}(t))\dot{\boldsymbol{c}}(t)\| \mathrm{d}t. \tag{5}$$

A metric on $\mathcal{M}$ can then be defined as

$$d_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y}) = \inf_{\boldsymbol{c} \in C^1(\mathcal{M})} \big\{ s | \boldsymbol{c}(0) = \boldsymbol{x} \text{ and } \boldsymbol{c}(1) = \boldsymbol{y} \big\}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}. \tag{6}$$

2.3 The Nakagami distribution

We consider random manifolds immersed by a GP. The length of a curve (5) on such a manifold is necessarily random as well. Fortunately, since this manifold is a Gaussian field, then curve lengths are well-approximated with the Nakagami m-distribution [Bewsher et al., 2017].

The Nakagami distribution [Nakagami, 1960] describes the length of an isotropic Gaussian vector, but Bewsher et al. [2017] have meticulously demonstrated that this also provides a good approximation to the arc length of a GP. The Nakagami has density function

$$g(s) = \frac{2m^m}{\Gamma(m)\Omega^m} s^{2m-1} \exp\Big(-\frac{m}{\Omega} s^2 \Big), \qquad s \geq 0, \tag{7}$$

and it is parametrised by $m \geq 1/2$ and $\Omega > 0$; here Γ denotes the Gamma function. The parameters are interpretable by the equations

$$\Omega = \mathbb{E}[s^2] \quad \text{and} \quad m = \frac{\Omega^2}{\mathrm{Var}(s^2)}, \tag{8}$$

which can be used to infer the parameters through samples, although it does involve a fourth moment.

3 Model and variational inference

With prerequisites settled, we now set up a Gaussian process latent variable model that is *locally* distance preserving and *globally* topology preserving. Notation-wise we let $\mathcal{Z}$ denote the latent representation of a dataset $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x_i} \in \mathbb{R}^D$, and let $f : \boldsymbol{z} \mapsto \boldsymbol{x}$ be the generative mapping.

3.1 Distance and topology preservation

The *manifold assumption* hypothesizes that high-dimensional data in $\mathbb{R}^D$ lie near a manifold with small intrinsic dimension. A manifold suggests that, a neighborhood around any point is approximately homeomorphic to a linear space. So nearby points are approximately linear, but non-nearby points have distances *greater* than the linear approximation suggests.

We shall build a Gaussian process latent variable model (GP-LVM) [Lawrence, 2005] that is explicitly designed for distance and topology preservation. The vanilla GP-LVM takes on the Gaussian likelihood where observations $\mathcal{X}$ are assumed i.i.d. when conditioned on a Gaussian process f. That is, $p(\mathcal{X}|f) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|f(\boldsymbol{z}_i))$ and $p(\boldsymbol{x}_i|f(\boldsymbol{z}_i)) = \mathcal{N}(\boldsymbol{x}_i|f(\boldsymbol{z}_i), \sigma^2)$. In contrast, we consider a likelihood over pairwise distances between observations.

Neighborhood graph. To model locality, we condition our model on a graph embedding of the observed data $\mathcal{X}$. The graph is the *ϵ-nearest neighbor* embedded graph; that is, the undirected graph with vertices $V = \mathcal{X}$ and edges $E = \{e_{ij}\}$, where e_{ij} is in E, only if $d(\boldsymbol{x}_i, \boldsymbol{x}_j) < \epsilon$, for some metric d. Equivalently, $G = (V, E)$ can be represented by its adjacency matrix A_G with entries

$$a_{ij} = \mathbf{1}_{d(\boldsymbol{x}_i, \boldsymbol{x}_j) < \epsilon}. \tag{9}$$

In Sec. 3.5 we discuss how to choose ϵ informedly, but for now we view it as a hyperparameter.

Manifold distances. To arrive at a likelihood over pairwise distances, we first recall that the linear interpolation between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ in the latent space has curve length

$$s_{ij} = \int_0^1 \|\boldsymbol{J}(\boldsymbol{c}(t))\dot{\boldsymbol{c}}(t)\| \mathrm{d}t, \quad \boldsymbol{c}(t) = \boldsymbol{z}_i(1-t) + \boldsymbol{z}_j t. \tag{10}$$

As the manifold distance $d_{\mathcal{M}}$ is the length of the shortest connecting curve, then s_{ij} is by definition an upper bound on $d_{\mathcal{M}}$. However, as the manifold is locally homeomorphic to a Euclidean space, then we can expect s_{ij} to be a good approximation of the distance to nearby points, i.e.

$$d_{\mathcal{M}}(\boldsymbol{z}_i, \boldsymbol{z}_j) \approx s_{ij} \qquad \text{for } \|\boldsymbol{x}_i - \boldsymbol{x}_j\| < \epsilon \tag{11}$$

$$d_{\mathcal{M}}(\boldsymbol{z}_i, \boldsymbol{z}_j) \leq s_{ij} \qquad \text{otherwise.} \tag{12}$$

The behavior we seek is that local interpolation in latent space should mimic local interpolation in data space only if the points are close in data space. If they are far apart, they should *repel* each other in the sense that the linear interpolation in latent space should have *large* curve length.

Censoring. To encode this behavior in the likelihood, we introduce *censoring* [Lee and Wang, 2003] into our objective function. This method is usually applied to missing data in survival analysis, when the event of something happening is known to occur later than some time point.

We may think of censoring as modeling inequalities in data. The censored likelihood function for i.i.d. data t_i following distribution function G_θ, with density function g_θ, is defined as

$$L(\{t_i\}_{i=1}^N|\theta, T) = \prod_{t_i<T} g_\theta(t_i) \prod_{t_i \geq T} (1 - G_\theta(T)), \tag{13}$$

where θ are the parameters of the distribution G and T is some 'time point', where the experiment ended. Carreira-Perpiñan [2010] remark that most neighborhood-embedding methods have loss functions with two terms: one attracting close point and one scattering term for far away connections. Censoring provides a *likelihood* with similar such terms.

Local distance likelihood. From earlier, we know that if the manifold $f(\mathcal{M})$ is a Gaussian field, then distances (10) are approximately Nakagami distributed. Thus, we write our likelihood as

$$L(\left\{\{e_{ij}\}_{i<j}\right\}_{i=1}^{N-1}|\theta, \epsilon) = \prod_{e_{ij}<\epsilon} g_\theta(e_{ij}) \prod_{e_{ij} \geq \epsilon} (1 - G_\theta(\epsilon)), \tag{14}$$

where G_θ is the distribution function of a Nakagami with parameters $\theta = \{m, \Omega\}$. Hence, the log-likelihood we shall maximise is

$$\begin{aligned} l\left(\left\{\{e_{ij}\}_{i<j}\right\}_{i=1}^{N-1}\middle|\theta, \epsilon\right) = &-\sum_{e_{ij}<\epsilon}\left(\log\Gamma(m_{ij}) + m_{ij}\log\left(\frac{\Omega_{ij}}{m_{ij}}\right) - (2m_{ij}-1)\log(e_{ij}) + \frac{m_{ij}e_{ij}^2}{\Omega_{ij}}\right) \\ &-\sum_{e_{ij}\geq\epsilon}\left(\log\Gamma(m_{ij}) - \log\left(\Gamma(m_{ij}) - \gamma(m_{ij}, \frac{m_{ij}}{\Omega_{ij}}e_{ij}^2)\right)\right), \end{aligned} \tag{15}$$

where Γ and γ denotes the Gamma function and lower incomplete gamma function respectively and m_{ij} and Ω_{ij} are the Nakagami-parameters of Eq. 10.

Until now, we have introduced the log-likelihood based of an ϵ-NN graph, that preserves geometric features. Next we marginalize all other parameters to make a generative model.

3.2 Marginalizing the representation

We have a loss function (15) that matches distances e_{ij} with parameters $\theta_{ij} = \{m_{ij}, \Omega_{ij}\}$. We now seek to first fit these parameters and marginalize them to obtain a full generative approach. First, we will assume that conditioned on θ, we get the independent observations, i.e.

$$p(\mathcal{E}|\theta, \epsilon) = \prod_{1 \leq i < j \leq N} p(e_{ij}|\theta_{ij}, \epsilon) = L(\{\{e_{ij}\}_{i<j}\}_{i=1}^{N-1}|\theta, \epsilon), \tag{16}$$

as known from Eq. 14. We infer these parameters of the Nakagami by introducing a latent Gaussian field J and a latent representation $\boldsymbol{z}$. This allows us to define curve length (10), which we assume is also Nakagami distributed. In practice, we draw[1] m samples of s_{ij} from Eq. 10, and estimate the mean and variance of their second moment. This gives estimates of m_{ij} and Ω_{ij} via Eq. 8.

Essentially, we match distances on the manifold J with the observed distances $\mathcal{E}$. We marginalize this manifold

$$p(\mathcal{E}|\boldsymbol{z}) = \int p(\mathcal{E}|\theta) p(\theta|\boldsymbol{J}, \boldsymbol{z}) p(\boldsymbol{J}) \mathrm{d}\theta \mathrm{d}\boldsymbol{J}, \qquad \text{where} \tag{17}$$

$$p(\theta|\boldsymbol{J}, \boldsymbol{z}) := \int p(\theta|s) p(s|\boldsymbol{J}, \boldsymbol{z}) \mathrm{d}s, \quad \text{and} \quad p(\theta|s) = \begin{cases} \delta_{\mathbb{E}s^2}(\Omega) \\ \delta_{\Omega/\mathrm{Var}(s^2)}\big(m\big), \end{cases} \tag{18}$$

where δ denotes the Dirac probability measure and $p(s|\boldsymbol{J}, \boldsymbol{z})$ is the approximate Nakagami distribution (10). This means that s_{ij} and e_{ij} are both Nakagami variables that share the same parameters, which interpretively means the manifold distances s_{ij} match the embedding distances e_{ij}.

Further, to make it generative, we can pose a prior on $\boldsymbol{z}$ and marginalize this in Eq. 17. We infer everything variationally [Blei et al., 2017], and choose a variational distribution over the marginalized variables. We approximate the posterior $p(\theta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{u}|\mathcal{E})$ with

$$q(\theta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{u}) := q(\theta|\boldsymbol{J}, \boldsymbol{z}) q(\boldsymbol{J}, \boldsymbol{u}) q(\boldsymbol{z}), \tag{19}$$

where $\boldsymbol{u}$ is an inducing variable [Titsias, 2009], and

$$q(\theta|\boldsymbol{J}, \boldsymbol{z}) = p(\theta|\boldsymbol{J}, \boldsymbol{z}), \quad q(\boldsymbol{J}, \boldsymbol{u}) = p(\boldsymbol{J}|\boldsymbol{u}) q(\boldsymbol{u}) \quad \text{and} \quad q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{A}_z), \tag{20}$$

where $\boldsymbol{\mu}_z$ is a vector of size N and $\boldsymbol{A}_z$ is a diagonal $N \times N$-matrix. Further $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{S})$ is a full M-dimensional Gaussian.

This allow us to bound the log-likelihood (15), with the evidence lower bound (ELBO)

$$\log p(\mathcal{E}) = \log \int \frac{p(\mathcal{E}, \theta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{u})}{q(\theta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{u})} q(\theta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{u}) \mathrm{d}\theta \mathrm{d}\boldsymbol{J} \mathrm{d}\boldsymbol{u} \mathrm{d}\boldsymbol{z} \tag{21}$$

$$\geq \mathbb{E}_\theta[l(\mathcal{E}|\theta)] - \mathrm{KL}\big(q(\boldsymbol{u})||p(\boldsymbol{u})\big) - \mathrm{KL}\big(q(\boldsymbol{z})||p(\boldsymbol{z})\big), \tag{22}$$

where both KL-terms are analytically tractable, but the first term has to be approximated using Monte Carlo. The right hand side here is readily optimized with gradient descent type algorithms.

In summary, we have a latent representation $\mathcal{Z}$ and a Riemannian manifold immersed as a GP $\boldsymbol{J}$. This implies that between any two points $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, we can compute s_{ij}, which is approximately Nakagami. With censoring we can match s_{ij} with observation e_{ij}, if $e_{ij} < \epsilon$; else we push s_{ij} to have all its mass on $[\epsilon, \infty)$. It is optimized with variational inference, by maximizing Eq. 22.

3.3 Generating new samples

All inference thus far has been done in a *coordinate-free* manner; in other words, we have yet to embed our manifold $f(\mathcal{M})$ in $\mathbb{R}^D$. We can do this embedding with Euclidean isometries, translation and rotation, and inspired by the fundamental theorem of analysis

$$f(\boldsymbol{z}_i) = f(\boldsymbol{z}_j) + \int_0^1 j(\boldsymbol{c}(t)) \dot{\boldsymbol{c}}(t) \mathrm{d}t, \quad \boldsymbol{c}(t) = \boldsymbol{z}_j(1-t) + \boldsymbol{z}_i t. \tag{23}$$

[1] We can approximate s by finely discretizing $\boldsymbol{c}$ and sum over the integrand.

In this view, the translation part can be done by the original points, as we assume $f(\boldsymbol{z}_j) \approx \boldsymbol{x}_j$ and we can, for a new point $\boldsymbol{z}^*$, define a generator as

$$x^* := f(\boldsymbol{z}_i) + \mathfrak{R}(\boldsymbol{z}_i) \int_0^1 \boldsymbol{J}(\boldsymbol{c}(t))\dot{\boldsymbol{c}}(t)\mathrm{d}t, \quad \boldsymbol{c}(t) = \boldsymbol{z}_i(1-t) + \boldsymbol{z}^* t, \tag{24}$$

where $\mathfrak{R}$ is a $D \times D$ rotation-matrix, that can be optimized to best fit with the original data $\mathcal{D}$ and $\boldsymbol{J}$ is the inferred Jacobian from Eq. 22. This is a rather naive way, since it needs many local embeddings and follows the intuition of Zhang and Zha [2003]. A more principled way would be to learn an isometry f by regression methods.

3.4 Invariance learning and geometric constraints

Why is it worth learning the manifold in a coordinate-free way, if we still need to fit values afterwards? Invariances are easily encoded via dissimilarity pairs by introducing equivalence classes in saying $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are in the same equivalence class. Popular choices of such equivalence classes are rotations, translations and scaling. Many constraints one could wish to impose on models can be formulated as geometric constraints. It holds true also for GPLVM-based models as seen in Urtasun et al. [2008], who wish to encode topological information, and Zhang et al. [2010], who highlight invariant models' usefulness in causal inference. Geometric constraints can alternatively be encoded with GPs that take their output directly on a Riemannian manifold [Mallasto et al., 2018].

The geometry of latent variable models in general is an active field of study [Arvanitidis et al., 2018, Tosi et al., 2014], and Simard et al. [2012] and Kumar et al. [2017] argues that the tangent (Jacobian) space serves a convenient way to encode invariances.

3.5 Topological Data Analysis

The model is naturally affected by the hyperparameter ϵ. We argue that it can be chosen in a geometrically founded way using Topological Data Analysis [Carlsson, 2009]. By constructing a *Rips diagram* [Fasy et al., 2014] one can find ϵ such that the ϵ-NN graph captures the right topology of data. It is beyond this paper to summarize the techniques; we refer readers to Chazal and Michel [2017].

4 Experiments

We perform experiments first on a classical toy dataset and on the image dataset MNIST. We refer to the presented model as *Isometric Gaussian Process Latent Variable Model* (Iso-GPLVM). For some comparisons we evaluate other models also based on dissimilarity data. In all cases we initialize Iso-GPLVM with IsoMap, as it is known that GP-based methods are sensitive to initializations [Bitzer and Williams, 2010]. We use the Adam-optimizer [Kingma and Ba, 2014] with a learning rate of $3 \cdot 10^{-3}$ and optimize sequentially $q(\boldsymbol{z})$ and $q(\boldsymbol{u})$ separately. We use $m = 100$ inducing points for $q(\boldsymbol{u})$ and an ARD-kernel as covariance function for the GP.

4.1 Swiss roll

The 'swiss roll' was introduced by Tenenbaum et al. [2000] to highlight the difficulties of non-linear manifold learning. The point cloud resides on a 2-dimensional manifold embedded in $\mathbb{R}^3$ and can be thought as a paper rolled around itself (see Fig 2A).

We find a 2-dimensional latent embedding by four methods: classical MDS, t-SNE, IsoMap and Iso-GPLVM. From Fig. 2 we observe the linear MDS is unable to capture the highly non-linear manifold. t-SNE captures some local structure, but the global outlook is far from the ground truth. We tried with several tunings of the perplexity hyperparameter (60 in the plot), but none of them successfully captured the structure. It is known that t-SNE is prone to make clusters, even if clusters are not a natural part of a dataset [Amid and Warmuth, 2018].

Naturally, as the dataset was constructed for the 'geodesic' approach of IsoMap, this captures both global and local structure. On closer inspection, we see the linear interpolations, stemming from Dijkstra's algorithm, leaves some artificial 'holes' in the manifold. Hence, on a smaller scale it can be argued that he topology of the manifold is captured imperfectly. The plot suggests Iso-GPLVM closes these holes and approximates the topology of an unfolded paper. We used $\epsilon = 0.4$.

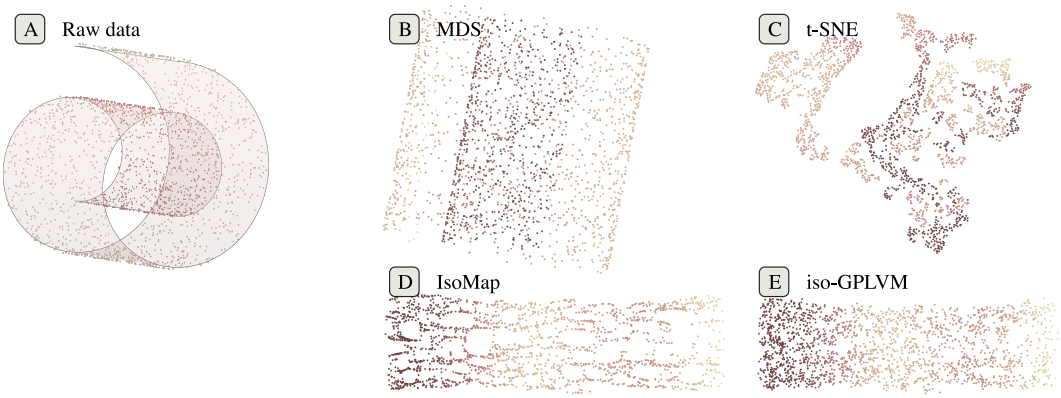

Figure 2: Data (A) and embeddings (B–E). All embeddings are shown with a unit aspect ratio to highlight that only IsoMap (D) and Iso-GPLVM (E) recover the elongated structure of the swiss roll.

4.2 MNIST

Metrics. We evaluate our model on 5000 images from MNIST, and we foremost wish to highlight how invariances can be encoded with dissimilarity data. In particular, we consider fitting our model to data under three different distance measures. We consider the classical Euclidean distance measure

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{25}$$

Further, we consider a metric that is invariant under image rotations

$$d_{\text{ROT}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \inf_{\theta \in [0, 2\pi)} \left\{ d\left(R_\theta(\boldsymbol{x}_i), \boldsymbol{x}_j\right) \right\}, \tag{26}$$

where R_θ rotates an image by θ radians. We note $d_{\text{ROT}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ always. Finally, we introduce a *lexicographic* metric [Rodriguez-Velazquez, 2018]

$$d_{\text{LEX}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} r, & \text{if } y_i \neq y_j, \\ \min\{2r,\ d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}, & \text{if } y_i = y_j, \end{cases} \tag{27}$$

which in the censoring phase enforce images carrying different labels to repel each other. This is a handy way to encode a topology or clustering based on discrete variables, when such are available. For all metrics, we have normalized the data and have set $r = 7$.

Results. Figure 3(A—C) show the latent embeddings of the three metrics. The background color is the measure $\mathbb{E}\left[\sqrt{\det(\boldsymbol{J}^\top \boldsymbol{J})}\right]$, which provide a view of the Riemannian geometry of the latent space. Bishop et al. [1997] call this measure the *magnification factor*. Large values (light color) imply that trajectories moving in this area are longer and likely also more uncertain [Hauberg, 2018].

Panels A, D and E base their latent embedding on the Euclidean metric. We observe that IsoMap (D) and Iso-GPLVM (A) appear similar in shape, unsurprisingly as we initialize with IsoMap, but Iso-GPLVM finds a cleaner separation of the digits. Particularly, this is evident for the *six, three and eight digits*. The *fives* seem to group into several tighter cluster, and this behavior is found for t-SNE as well. Overall, from a clustering perspective, t-SNE visually is superior; but distances *between* clusters in (A) can be larger than the straight lines that connect them. This is evident from the lighter background color between cluster, say, *zeros* and *threes*. We note that IsoMap and t-SNE has no associated Riemannian metric and as such distances between any input cannot be computed.

The rotation invariant metric result in a latent embedding where different classes significantly overlap. Upon closer inspection we, however, note several interesting properties of the embedding. *Zero digits* are well separated from other classes as a rotated 0 does not resemble any other digits; the *one digits* form a cluster that is significantly more compact than other digits as there is limited variation left after rotations have been factored out; *two and five digits* significantly overlap, which is most likely due to 5 digits resembling 2 digits when rotated 180°; similar observations hold for the *four, nine and six digits*; and a partial overlap between *three and eight digits* as is often observed. The overall darker background is due to the rotational invariant metric being shorter than the Euclidean counterpart.

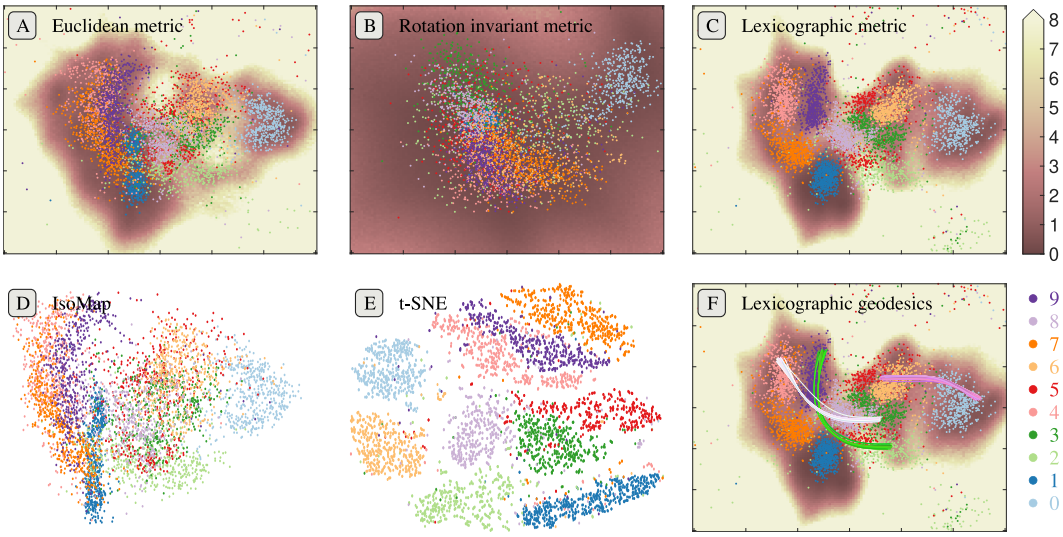

Figure 3: Embeddings of MNIST attained with our method under different metrics (A—C) and for baselines IsoMap (D) and t-SNE (E). The background color show the expected volume measure associated with the Riemannian metric $\mathbb{E}\left[\sqrt{\det(\boldsymbol{J}^\top \boldsymbol{J})}\right]$. A large measure generally indicate high uncertainty of the manifold. Panel F shows Riemannian geodesics under the lexicographic metric.

In terms of clustering the lexicographic approach outshines the other metrics. This is expected as the metric use label information, but neatly illustrate how domain-specific metrics can be developed from weak or partial information. Most classes are well-separated except for a region in the middle of the plot. Note how this region has high uncertainty.

The Riemannian geometry of the latent space imply that geodesics (shortest paths) can be computed in our model. Figure 3F shows example geodesics under the lexicographic metric. Their highly non-linear appearance emphasizes the curvature of the learned manifold. The green geodesics has one endpoint in a cluster of nine digits and move along this cluster avoiding the uncertain area of eights and fives, as opposed on linearly interpolating through them.

5 Discussion

We introduced a model for non-linear dimensionality reduction from dissimilarity data. It is the first of its kind based on Gaussian processes. The non-linearity of the method stems both from the Gaussian processes, but also from the censoring in the likelihood. It unifies ideas from Gaussian processes, Riemannian geometry and neighborhood graph embeddings. Unlike traditional manifold learning methods that embed into $\mathbb{R}^q$, we embed into a q-dimensional Riemannian manifold through the learned metric. This allows us to learn latent representations that are isometric to the true underlying manifold.

The model does have limitations. The generation of new samples was only naively considered, and further research of how to isometrically embed a manifold $\mathcal{M}$ into $\mathbb{R}^D$ to fit with observation is warranted. Existence is ensured as the observed data manifold is one such embedding. The Nakagami distribution that approximates the arc lengths of Gaussian processes is prone to overestimate the variance [Bewsher et al., 2017] and better approximations would improve our method. Further, the model inherits problems of optimizing the latent variables and it has previously been noted that good performance in this regime is linked with good initialization [Bitzer and Williams, 2010].

Our experiments highlight that Iso-GPLVM can learn the geometry of data and geometric constraints are easier encoded by learning a manifold contra doing GP regression. The uncertainty quantification associated with GPs follow through and further highlights the connection between uncertainty, geometry and topology. To the best of our knowledge, our model is the first of its kind that, locally, can asses the quality of the manifold approximation through the associated Riemannian measure.

6 Broader Impact

We present a general methodology for learning low-dimensional representations from pairwise distances, such that the associated model is fully generative and both distance and topology preserving. The model is further suitable for encoding *a priori* known invariances through a choice of metric. The contribution is largely methodological.

We envision the model being applied for data where it is easier to express prior knowledge through the design of an appropriate distance function. For instance, in much biological data there is side-information regarding the underlying evolutionary structure, which can be used to develop suitable evolutionary metrics.

The flexibility of the approach does open the door for misuse. For instance misleading visualizations (of the latent variables) can be easily created by a malicious choice of metric. The lexicographic example illustrates this potential misuse as one can imagine forcing groups apart with this mechanism, even if the data disapproves such groupings.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757360). MJ and SH were supported in part by a research grant (15334) from VILLUM FONDEN.

References

E. Amid and M. K. Warmuth. A more globally accurate dimensionality reduction method using triplets. *arXiv:1803.00854 [cs]*, Mar. 2018.

G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: On the curvature of deep generative models. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013.

J. Bewsher, A. Tosi, M. Osborne, and S. Roberts. Distribution of gaussian process arc lengths. In *Artificial Intelligence and Statistics*, pages 1412–1420, 2017.

C. M. Bishop, M. Svens' en, and C. K. Williams. Magnification factors for the som and gtm algorithms. In *Proceedings 1997 Workshop on Self-Organizing Maps*, 1997.

S. Bitzer and C. K. Williams. Kick-starting gplvm optimization via a connection to metric mds. In *NIPS 2010 Workshop on Challenges of Data Visualization*, 2010.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

M. Á. Carreira-Perpiñan. The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 167–174, 2010.

F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists, 2017.

B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 12 2014. doi: 10.1214/14-AOS1252. URL `https://doi.org/10.1214/14-AOS1252`.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

S. Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems*, pages 5534–5544, 2017.

N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of machine learning research*, 6(Nov):1783–1816, 2005.

E. T. Lee and J. Wang. *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons, 2003.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

A. Mallasto, S. Hauberg, and A. Feragen. Probabilistic riemannian submanifold learning with wrapped gaussian process latent variable models. In *Proceedings of the 19th international Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

P.-A. Mattei and J. Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423, 2019.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

M. Nakagami. The m-distribution—a general formula of intensity distribution of rapid fading. In *Statistical Methods in Radio Wave Propagation*, pages 3–36. Elsevier, 1960.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2006.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

J. A. Rodriguez-Velazquez. Lexicographic metric spaces: Basic properties and the metric dimension, 2018.

P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri. Transformation invariance in pattern recognition–tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 235–269. Springer, 2012.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319, 2000.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

M. Titsias and N. D. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

A. Tosi, S. Hauberg, A. Vellido, and N. D. Lawrence. Metrics for Probabilistic Geometries. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2014.

R. Urtasun and T. Darrell. Discriminative gaussian process latent variable model for classification. In *Proceedings of the 24th International Conference on Machine Learning*, pages 927–934, 2007.

R. Urtasun, D. J. Fleet, A. Geiger, J. Popović, T. J. Darrell, and N. D. Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1080–1087, 2008.

K. Zhang, B. Schölkopf, and D. Janzing. Invariant gaussian process latent variable models and application in causal discovery. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 717–724, 2010.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Sci. Comput.*, 26, 01 2003. doi: 10.1137/S1064827502419154.