Maximum likelihood estimation of Riemannian metrics from Euclidean data

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg

Section for Cognitive Systems, Technical University of Denmark,
Richard Petersens Plads, B321, 2800 Kgs. Lyngby, Denmark.
{gear, lkai, sohau}@dtu.dk

Abstract. Euclidean data often exhibit a nonlinear behavior, which may be modeled by assuming the data is distributed near a nonlinear submanifold in the data space. One approach to find such a manifold is to estimate a Riemannian metric that locally models the given data. Data distributions with respect to this metric will then tend to follow the nonlinear structure of the data. In practice, the learned metric rely on parameters that are hand-tuned for a given task. We propose to estimate such parameters by maximizing the data likelihood under the assumed distribution. This is complicated by two issues: (1) a change of parameters imply a change of measure such that different likelihoods are incomparable; (2) some choice of parameters renders the numerical calculation of distances and geodesics unstable such that likelihoods cannot be evaluated. As a practical solution, we propose to (1) re-normalize likelihoods with respect to the usual Lebesgue measure of the data space, and (2) to bound the likelihood when its exact value is unattainable. We provide practical algorithms for these ideas and illustrate their use on synthetic data, images of digits and faces, as well as signals extracted from EEG scalp measurements.

Keywords: manifold learning, metric learning, statistics on manifolds.

1 Introduction

The *"manifold assumption"* is often applied in machine learning research to express that data is believed to lie near a (nonlinear) submanifold embedded in the data space. Such an assumption finds uses e.g. in dynamical or periodic systems, and in many problems with a smooth behavior. When the manifold structure is known a priori it can be incorporated into the problem specification, but unfortunately such structure is often not known. In these cases it is necessary to estimate the manifold structure from the observed data, a process known as *manifold learning*. In this work, we approach manifold learning *geometrically* by estimating a Riemannian metric that captures local behavior of the data, and *probabilistically* by estimating unknown parameters of the metric using maximum likelihood. First we set the stage with background information on manifold learning (Sec. 1.1) and geometry (Sec. 1.2), followed by an exposition of our model (Sec. 2) and the proposed maximum likelihood scheme (Sec. 3). Finally results are presented (Sec. 4) and discussed (Sec. 5).

1.1 Background and related work

Given observations $\mathbf{x}_{1:N} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^D$, the key task in manifold learning is to estimate a data representation that reflect the nonlinear structure of the original data. The intuition behind most methods for this was phrased by Saul & Roweis [18] as *"Think Globally, Fit Locally"*, practically meaning that locally linear models are fitted to all points in data space and these then are merged to a global representation (details depend on the method).

The *Isomap* method [20] famously replace Euclidean distances with geodesic distances defined on a neighborhood graph and then embed the data in a lower dimensional space where Euclidean distances approximate the geodesic counterparts. While this approach is popular, its discrete nature only describes the observed data points and consequently cannot be used to develop probabilistic generative models. Similar comments hold for other graph-based methods [18,2].

As a smooth alternative, Lawrence [16] proposed a probabilistic extension of standard surface models by assuming that each dimension of the data is described as $x_d = f_d(\mathbf{z})$, where $\mathbf{z}$ is a low-dimensional latent variable and f_d is a Gaussian process. The latent variables then provide a low-dimensional parametrization that capture the manifold strucure. Tosi et al. [21] give this a geometric interpretation by deriving the distribution of the induced Riemannian pull-back metric and show how geodesics can be computed under this uncertain metric.

Often manifold learning is viewed as a form of dimensionality reduction, but this need not be the case. Hauberg et al. [12] suggest to model the local behavior of the data manifold via a locally-defined Riemannian metric, which is constructed by interpolating a set of pre-trained metric tensors at a few select points in data space. Once a Riemannian metric is available existing tools can be used for dimensionality reduction [22,11,8], mixture modeling [1,19], tracking [13,14], hypothesis testing [17], transfer learning [9] and more. Our approach follow this line of work.

1.2 The basics of Riemannian geometry

For completeness we start with an informal review of Riemannian manifolds, but refer the reader to standard text books [5] for a more detailed exposition.

Definition 1. *A smooth manifold $\mathcal{M}$ together with a Riemannian metric $\mathbf{M} : \mathcal{M} \to \mathbb{R}^{D \times D}$ and $\mathbf{M} \succ 0$ is called a Riemannian manifold. The Riemannian metric $\mathbf{M}$ encodes a smoothly changing inner product $\langle \mathbf{u}, \mathbf{M}(\mathbf{x})\mathbf{v} \rangle$ on the tangent space $\mathbf{u}, \mathbf{v} \in \mathcal{T}_{\mathbf{x}}\mathcal{M}$ of each point $\mathbf{x} \in \mathcal{M}$.*

Since the Riemannian metric $\mathbf{M}(\mathbf{x})$ acts on tangent vectors it may be interpreted as a standard Mahalanobis metric restricted to an infinitesimal region around $\mathbf{x}$. This local inner product is a suitable model for capturing local behavior of data, i.e. *manifold learning.* Shortest paths (*geodesics*) are then length-minimizing curves connecting two points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, i.e.

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \int_0^1 \sqrt{\langle \boldsymbol{\gamma}'(t), \mathbf{M}(\boldsymbol{\gamma}(t))\boldsymbol{\gamma}'(t) \rangle} \mathrm{d}t, \quad \text{s.t.} \quad \boldsymbol{\gamma}(0) = \mathbf{x},\ \boldsymbol{\gamma}(1) = \mathbf{y}. \qquad (1)$$

Here $\mathbf{M}(\boldsymbol{\gamma}(t))$ is the metric tensor at $\boldsymbol{\gamma}(t)$, and the tangent vector $\boldsymbol{\gamma}'$ denotes the derivative (velocity) of $\boldsymbol{\gamma}$. The distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as the length of the geodesic. Geodesic can be found as the solution to a system of 2^{nd} order ordinary differential equations (ODEs):

$$\boldsymbol{\gamma}''(t) = -\frac{1}{2}\mathbf{M}^{-1}(\boldsymbol{\gamma}(t))\left[\frac{\partial \mathrm{vec}[\mathbf{M}(\boldsymbol{\gamma}(t))]}{\partial \boldsymbol{\gamma}(t)}\right]^{\intercal}(\boldsymbol{\gamma}'(t)\otimes\boldsymbol{\gamma}'(t)) \tag{2}$$

subject to $\boldsymbol{\gamma}(0) = \mathbf{x},\ \boldsymbol{\gamma}(1) = \mathbf{y}$. Here $\mathrm{vec}[\cdot]$ stacks the columns of a matrix into a vector and $\otimes$ is the Kronecker product.

This differential equation allows us to define basic operations on the manifold. The *exponential map* at a point $\mathbf{x}$ takes a tangent vector $\mathbf{v} \in \mathcal{T}_{\mathbf{x}}\mathcal{M}$ to $\mathbf{y} = \mathrm{Exp}_{\mathbf{x}}(\mathbf{v}) \in \mathcal{M}$ such that the curve $\boldsymbol{\gamma}(t) = \mathrm{Exp}_{\mathbf{x}}(t \cdot \mathbf{v})$ is a geodesic originating at $\mathbf{x}$ with initial velocity $\mathbf{v}$ and length $\|\mathbf{v}\|$. The inverse mapping, which takes $\mathbf{y}$ to $\mathcal{T}_{\mathbf{x}}\mathcal{M}$ is known as the *logarithm map* and is denoted $\mathrm{Log}_{\mathbf{x}}(\mathbf{y})$. By definition $\|\mathrm{Log}_{\mathbf{x}}(\mathbf{y})\|$ corresponds to the geodesic distance from $\mathbf{x}$ to $\mathbf{y}$. The exponential and the logarithmic map can be computed by solving Eq. 2 numerically, as an *initial value problem* or a *boundary value problem* respectively.

2 A locally adaptive normal distribution

We have previously provided a simple nonparametric manifold learning scheme that conceptually mimics a local principal component analysis [1]. At each point $\mathbf{x} \in \mathbb{R}^D$ a local covariance matrix is computed and its inverse then specify a local metric. For computational efficiency and to prevent overfitting we restrict ourselves to diagonal covariances

$$M_{dd}(\mathbf{x}) = \left(\sum_{n=1}^{N} w_n(\mathbf{x})(x_{nd} - x_d)^2 + \rho\right)^{-1}, \tag{3}$$

$$w_n(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}\|_2^2}{2\sigma^2}\right). \tag{4}$$

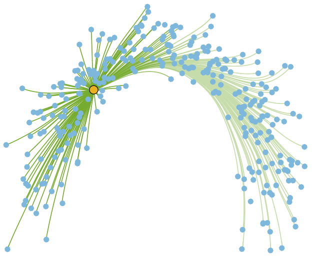


Fig. 1: Example geodesics.

Here the subscript d is the dimension, n corresponds to the given data, and ρ is a regularization parameter to avoid singular covariances. The weight-function $w_n(\mathbf{x})$ changes smoothly such that the resulting metric is Riemannian. It is easy to see that if $\mathbf{x}$ is outside of the support of the data, then the metric tensor is large. Thus, geodesics are "pulled" towards the data where the metric is small (see Fig. 1).

The weight-function $w_n(\mathbf{x})$ depends on a parameter σ that effectively determine the size of the neighborhood used to define the data manifold. Small values of σ gives a manifold with high curvature, while a large σ gives an almost flat manifold. The main contribution of this paper is a systematic approach to determine this parameter.

For a given metric (and hence σ), we can estimate data distributions with respect to this metric. We consider Riemannian normal distributions [17]

$$p_{\mathcal{M}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\mathcal{C}} \exp\left(-\frac{1}{2} d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathcal{M} \tag{5}$$

and mixtures thereof. Here $\mathcal{M}$ denote the manifold induced by the learned metric, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance, and $d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = \langle \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}), \boldsymbol{\Sigma}^{-\mathbf{1}} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}) \rangle$. The normalization constant $\mathcal{C}$ is by definition

$$\mathcal{C}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathcal{M}} \exp\left(-\frac{1}{2} d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu})\right) \mathrm{d}\mathcal{M}(\mathbf{x}), \tag{6}$$

where $\mathrm{d}\mathcal{M}(\mathbf{x})$ denotes the measure induced by the Riemannian metric. Note that this measure depends in σ. Figure 2 show an example of the resulting distribution under the proposed metric. As the distribution adapts locally to the data we coin it a *locally adaptive normal distribution (LAND)*.

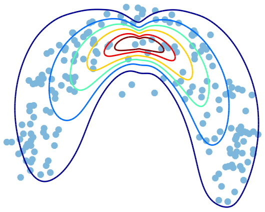


Fig. 2: Example of the *locally adaptive normal distribution (LAND)*.

Assuming that the data are generated from a distribution $q_{\mathcal{M}}$ then commonly the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ are estimated with intrinsic least squares (ILS)

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu} \in \mathcal{M}}{\mathrm{argmin}} \int_{\mathcal{M}} d^2(\boldsymbol{\mu}, \mathbf{x}) q_{\mathcal{M}}(\mathbf{x}) \mathrm{d}\mathcal{M}(\mathbf{x}), \tag{7}$$

$$\hat{\boldsymbol{\Sigma}} = \int_{\mathcal{T}_{\hat{\boldsymbol{\mu}}}\mathcal{M}} \mathrm{Log}_{\hat{\boldsymbol{\mu}}}(\mathbf{x}) \mathrm{Log}_{\hat{\boldsymbol{\mu}}}(\mathbf{x})^{\intercal} p_{\mathcal{M}}(\mathbf{x}) \mathrm{d}\mathcal{M}(\mathbf{x}), \tag{8}$$

where $d^2(\cdot, \cdot)$ denotes the squared geodesic distance. These parameter estimates naturally generalize their Euclidean counterparts, and they can be further shown to have maximal likelihood when the manifold is also a symmetric space [7]. For more general manifolds, like the ones under consideration in this paper, these estimates do not attain maximal likelihood. Figure 3 show both the ILS estimate of $\boldsymbol{\mu}$ and the maximum likelihood (ML) estimate. Since the ILS estimate falls outside the support of the data, a significantly larger covariance matrix is needed to explain the data, which gives a poor likelihood. To find the maximum likelihood parameters of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we perform steepest descent directly on the data log-likelihood using an efficient Monte Carlo estimator of the normalization constant $\mathcal{C}$ [1].

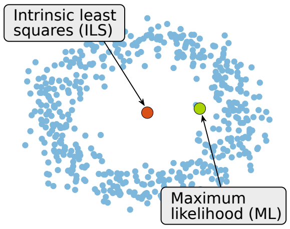


Fig. 3: ML and ILS means.

3 Maximum likelihood metric learning

Determining the optimal metric (parametrized by σ) is an open question. Since the LAND is a parametric probabilistic model it is natural to perform this model selection using maximum likelihood. The complete data log-likelihood is

$$\mathcal{L}(\sigma) = -\frac{1}{2} \sum_{n=1}^{N} d^2_{\boldsymbol{\Sigma}}(\mathbf{x}_n, \boldsymbol{\mu}) - N \log \mathcal{C}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{9}$$

It is tempting to evaluate $\mathcal{L}(\sigma)$ for several values of σ and pick the one with maximal likelihood. This, however, is both theoretically and practically flawed.

The first issue is that the measure $\mathrm{d}\mathcal{M}(\cdot)$ used to define the LAND depends on σ. This imply that $\mathcal{L}(\sigma)$ cannot be compared for different values of σ as they do not rely on the same measure. The second issue is that $\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)$ must be evaluated numerically, which can become unstable when $\mathcal{M}$ has high curvature. This imply that $\mathcal{L}(\sigma)$ can often not be fully evaluated when σ is small.

3.1 Likelihood bounds to cope with numerical instabilities

When numerical instabilities prevent us from evaluating $\mathcal{L}(\sigma)$ we instead rely on an easy-to-evaluate lower bound $\overline{\mathcal{L}(\sigma)}$. To derive this, let $\mathbf{v}_n = \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}_n)$. Then $\|\mathbf{v}_n\|$ is the geodesic distance between $\boldsymbol{\mu}$ and $\mathbf{x}_n$, while ${}^{\mathbf{v}_n}/_{\|\mathbf{v}_n\|}$ is the initial direction of the connecting geodesic. It is easy to provide an upper bound on the geodesic distance by taking the length of a non-geodesic connecting curve, here chosen as the straight line connecting $\boldsymbol{\mu}$ and $\mathbf{x}_n$. The bound then becomes

$$\|\mathbf{v}_n\| \leq \tilde{d}_n = \int_0^1 \sqrt{\langle(\mathbf{x}_n - \boldsymbol{\mu}), \mathbf{M}(t\mathbf{x}_n + (1-t)\boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})\rangle}\mathrm{d}t. \tag{10}$$

The initial orientation ${}^{\mathbf{v}_n}/_{\|\mathbf{v}_n\|}$ influence the log-likelihood as the covariance $\boldsymbol{\Sigma}$ is generally anisotropic. This is, however, easily bounded by picking the initial direction as the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the smallest eigenvalue $\lambda_{\min}$. This then gives the final lower bound

$$\overline{\mathcal{L}(\sigma)} = -\frac{1}{2}\sum_{n=1}^{N} \frac{\tilde{d}_n^2}{\lambda_{\min}} - N \log \mathcal{C}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{11}$$

In practice, we only use the bound for data points $\mathbf{x}$ where the logarithm map cannot be evaluated, and otherwise use the correct log-likelihood.

3.2 Comparing likelihoods

Since the measure $\mathrm{d}\mathcal{M}(\cdot)$ changes with σ we cannot directly compare $\mathcal{L}(\sigma)$ across inputs. In order to make this comparison feasible, we propose to re-normalize the LAND with respect to the usual Lebesgue measure of the data space $\mathbb{R}^D$. This amount to changing the applied measure in Eq. 6. As we lack closed-form expressions, we perform this re-normalization using importance sampling [4]

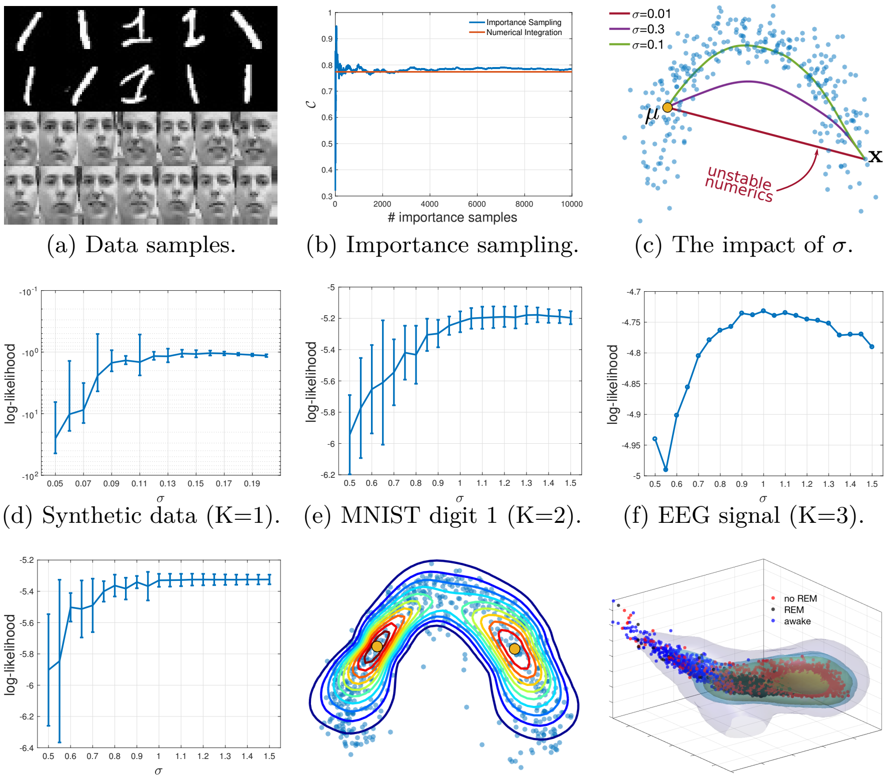
$$\tilde{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^D} \exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}, \mathbf{x})\right)\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^D} \frac{\exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}, \mathbf{x})\right)}{q(\mathbf{x})} q(\mathbf{x})\mathrm{d}\mathbf{x} \tag{12}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} w_s \exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}, \mathbf{x}_s)\right), \quad \mathbf{x}_s \sim q(\mathbf{x}),\ w_s = \frac{1}{q(\mathbf{x}_s)}, \tag{13}$$

where $q(\mathbf{x})$ is the proposal distribution from which we draw S samples. In our experiments we choose $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the linear mean and covariance of the data. Thus, we ensure that the support of the proposal captures the data manifold, but any other distribution with the desired properties can be used.

4 Results

Experimental setup: We evaluate the proposed method on both synthetic and real data. The two-dimensional synthetic data is drawn from an arc-shaped distribution (see Fig. 4c) [1]. We further consider features extracted from EEG measurements during human sleep [1]; the digit "1" from MNIST; and the "Frey faces"[1]. Both image modalities are projected onto their first two principal components, and are separated into 10 and 5 folds respectively. To each data modality, we fit a mixture of LANDs with K components.



(a) Data samples. (b) Importance sampling. (c) The impact of σ.

(d) Synthetic data (K=1). (e) MNIST digit 1 (K=2). (f) EEG signal (K=3).

(g) Frey faces (K=2). (h) Digit 1 density. (i) EEG sleep density.

Fig. 4: Experimental results on various data sets; see text for details.

Verification: First, we validate the importance sampling scheme in Fig. 4b where we compare with an expensive numerical integration scheme on a predefined grid. It is evident that importance sampling quickly gives a good approximation to to the true normalization constant. However, choosing the correct proposal

[1] http://www.cs.nyu.edu/~roweis/data.html

distribution is usually crucial for the success of the approximation [4]. Then, in Fig. 4c we show the impact of σ on the geodesic solution. When σ is small (0.01) the true geodesic cannot be computed numerically and a straight line is used to bound the likelihood (Sec. 3). For larger values of σ the geodesic can be computed. Note that the geodesic becomes increasingly "straight" for large values of σ.

Model selection: Figures 4d-4g show the log-likelihood bound proposed in Sec. 3 for all data sets. In particular, we can distinguish three different regions for the σ parameter. (1) For small values of σ the manifold has high curvature and some geodesics cannot be computed, such that the bound penalizes the data log-likelihood. (2) There is a range of σ values where the construction of the manifold captures the actual underlying data structure, and in those cases we achieve the best log-likelihood. (3) For larger values of σ the manifold becomes flat, and even if we are able compute all the geodesics the likelihood is reduced. The reason is that when the manifolds becomes flat, significant probability mass is assigned to regions outside of the data support, while in the other case all the probability mass is concentrated near the data resulting to higher likelihood.

5 Discussion

Probability density estimation in non-linear spaces is essential in data analysis [6]. With the current work, we have proposed practical tools for model selection of the metric underlying the *locally adaptive normal distribution (LAND)* [1]. The basic idea amounts to picking the metric that maximize the data likelihood. A *theoretical* concern is that different metrics gives different measures implying that likelihoods are not comparable. We have proposed to solve this by re-normalizing according to the Lebesgue measure associated with the data space. *Practically* our idea face numerical challenges when the metric has high curvature as geodesics then become unstable to compute. Here we have proposed an easy-to-compute bound on the data likelihood, which has the added benefit that metrics giving rise to numerical instabilities are penalized. Experimental results on diverse data sets indicate that the approach is suitable for model selection.

In this paper we have considered maximum likelihood estimation on the training data, which can potentially overfit [10]. While we did not observe such behavior in our experiments it is still worth investigating model selection on a held-out test set or to put a prior on σ and pick the value that maximize the posterior probability. Both choices are straight-forward.

An interesting alternative to bounding the likelihood appears when considering probabilistic solvers [15] for the geodesic equations (2). These represent the numerical estimate of geodesics with a Gaussian process whose uncertainty captures numerical approximation errors. Efficient algorithms then exist for estimating the distribution of the geodesic arc length [3]. With these solvers, hard-to-estimate geodesics will be associated with high variance, such that the now-stochastic data log-likelihood also has high variance. Model selection should then take this variance into account.

Acknowledgements. SH was supported by a research grant (15334) from VILLUM FONDEN. LKH was supported by the Novo Nordisk Foundation Interdisciplinary Synergy Program 2014, 'Biophysically adjusted state-informed cortex stimulation (BASICS)'.

References

1. Arvanitidis, G., Hansen, L.K., Hauberg, S.: A locally adaptive normal distribution. In: Advances in Neural Information Processing Systems (NIPS) (2016)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation 15(6), 1373–1396 (2003)
3. Bewsher, J., Tosi, A., Osborne, M., Roberts, S.: Distribution of Gaussian Process Arc Lengths. In: AISTATS (2017)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
5. do Carmo, M.: Riemannian Geometry. Birkhäuser (1992)
6. Chevallier, E., Barbaresco, F., Angulo, J.: Probability Density Estimation on the Hyperbolic Space Applied to Radar Processing. In: GSI. pp. 753–761 (2015)
7. Fletcher, P.T.: Geodesic regression and the theory of least squares on Riemannian manifolds. IJCV 105(2), 171–185 (2013)
8. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE TMI 23(8), 995–1005 (2004)
9. Freifeld, O., Hauberg, S., Black, M.J.: Model transport: Towards scalable transfer learning on manifolds. In: CVPR (2014)
10. Hansen, L.K., Larsen, J.: Unsupervised learning and generalization. In: Neural Networks, 1996., IEEE International Conference on. vol. 1, pp. 25–30. IEEE (1996)
11. Hauberg, S.: Principal Curves on Riemannian Manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2016)
12. Hauberg, S., Freifeld, O., Black, M.J.: A Geometric Take on Metric Learning. In: Advances in Neural Information Processing Systems (NIPS). pp. 2033–2041 (2012)
13. Hauberg, S., Lauze, F., Pedersen, K.S.: Unscented Kalman Filtering on Riemannian manifolds. Journal of Mathematical Imaging and Vision 46(1), 103–120 (May 2013)
14. Hauberg, S., Pedersen, K.S.: Stick it! articulated tracking using spatial rigid object priors. In: ACCV. LNCS, vol. 6494, pp. 758–769. Springer (2010)
15. Hennig, P., Hauberg, S.: Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics. In: AISTATS. vol. 33 (2014)
16. Lawrence, N.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J. of Machine Learning Research 6, 1783–1816 (2005)
17. Pennec, X.: Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. Journal of Mathematical Imaging and Vision 25(1), 127–154 (2006)
18. Saul, L.K., Roweis, S.T.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research 4, 119–155 (2003)
19. Straub, J., Chang, J., Freifeld, O., Fisher III, J.W.: A Dirichlet Process Mixture Model for Spherical Data. In: AISTATS (2015)
20. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319 (2000)
21. Tosi, A., Hauberg, S., Vellido, A., Lawrence, N.D.: Metrics for Probabilistic Geometries. In: The Conference on Uncertainty in Artificial Intelligence (UAI) (2014)
22. Zhang, M., Fletcher, P.T.: Probabilistic Principal Geodesic Analysis. In: Advances in Neural Information Processing Systems (NIPS) 26. pp. 1178–1186 (2013)