DENSITY RIDGE MANIFOLD TRAVERSAL

Jonas N. Myhre, Michael Kampffmeyer and Robert Jenssen

Machine Learning Group: site.uit.no/ml
Department of Physics and Technology, UiT - The Arctic University of Norway

ABSTRACT

The density ridge framework for estimating principal curves and surfaces has in a number of recent works been shown to capture manifold structure in data in an intuitive and effective manner. However, to date there exists no efficient way to traverse these manifolds as defined by density ridges. This is unfortunate, as manifold traversal is an important problem for example for shape estimation in medical imaging, or in general for being able to characterize and understand state transitions or local variability over the data manifold. In this paper, we remedy this situation by introducing a novel manifold traversal algorithm based on geodesics within the density ridge approach. The traversal is executed in a subspace capturing the intrinsic dimensionality of the data using dimensionality reduction techniques such as principal component analysis or kernel entropy component analysis. A mapping back to the ambient space is obtained by training a neural network. We compare against maximum mean discrepancy traversal, a recent approach, and obtain promising results.

***Index Terms*—** Manifold Learning, Dimensionality Reduction, Density Ridges, Neural Network

1. INTRODUCTION

Density ridges are estimates of principal manifolds, d-dimensional smooth curves or surfaces that pass through the middle of the data distribution. They play an important role in the field of manifold learning and signal processing [1, 2, 3] and have been used for tasks such as for example manifold unwrapping [4], similarity clustering [5], and contour representation [6]. However, there exists no efficient method for manifold traversal in the density ridge framework.

Manifold traversal is an important technique for modelling of shape priors for tasks such as image segmentation [7, 8] of Magnetic Resonance Images. Also, physical processes often inhabit high dimensional spaces, where state transitions or local variation can be seen as traversing along a low dimensional manifold. One such example, in a health analytics setting, would be to have a set of patients represented by various measurements. Moving from a healthy to a sick patient along the manifold could be interpreted as such a state transition. Traversing the manifold can then help to understand how the patient's measurements change as the patients status changes from healthy to sick.

Density ridge estimation is based on a non-parametric kernel density estimate, which makes it a flexible and model free method [1, 9]. Consequently, this also renders density ridge based methods impractical in very high dimensional spaces due to the inherent problems of density estimation in high dimensions. This makes it necessary to use dimensionality reduction techniques to learn the embedding function from the high dimensional ambient space to a low dimensional subspace. Of great benefit here is the fact that the so-called manifold assumption, where high dimensional data exhibits low dimensional intrinsic manifold structure, has gathered significant empirical evidence in recent years [10, 11, 12, 7, 13].

In this work, inspired by the connection between density ridges and differential geometry [4, 14], we propose a novel algorithm for manifold traversal along a density ridge manifold estimate. This is done by calculating smooth geodesics along the estimated manifold. We use dimensionality reduction techniques to learn the embedding function from the high dimensional ambient space to a low dimensional subspace and traverse along geodesics in the intrinsic space. Exploiting the power of neural networks [15] we learn a backprojection from the low dimensional space back to the ambient space, which allows us to illustrate effects of the manifold traversal algorithm on the input data space.

2. METHOD

Our proposed workflow consists of three stages: Dimensionality reduction in the form of principal component analysis (PCA) or kernel entropy component analysis (KECA) to capture the lower dimensional subspace (linear in the case of PCA) containing the data, followed by estimating the underlying smooth manifold by density ridge estimation. Given a smooth manifold estimate, we can perform operations inspired by differential geometry on the manifold such as, but not limited to, estimating geodesic distances or isometric unfolding. In the final stage the low dimensional data represen-

We gratefully acknowledge the support of NVIDIA Corporation. This work was partially funded by the Norwegian Research Council FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines*.

tation is mapped back to the input space. We learn the inverse mapping function using a two-layer neural network.

2.1. Dimensionality reduction

High-dimensional data often exhibits low dimensional structure that is nonlinear. Separating dimensionality reduction and manifold estimation allows us to use established techniques that capture the overall variation in the data as a first stage. Take for example the swiss roll dataset [16], which is a two-dimensional plane rolled up into a swiss roll shape, and consider sampling from a noisy version of this shape in a high dimensional space. Given a low level of noise, i.e. points are not sampled too far from the manifold, we could reduce the dimension of this set to three without interfering with the nonlinear structure.

In principle any dimensionality reduction approach can be used to reduce the dimension of the ambient space of the manifold. We focus on the linear principal component analysis (PCA) [17] and the non-linear kernel entropy component analysis (KECA) [18] approach in this work. PCA is a widely used *variance* preserving linear dimensionality reduction method. KECA is an *entropy* preserving dimensionality reduction method, which yields an embedding where the projected subspace captures the largest amount of Renyi entropy as represented by a Parzen window estimate.

2.2. Manifold estimation using density ridges

Once the ambient dimension is reduced, the structure of the manifold can be estimated. In this work we use the density ridge framework of Ozertem and Erdogmus [1], which can be estimated by solving an initial value problem [19]. Density ridges are estimates of principal manifolds, d-dimensional hypersurfaces that pass through the middle of the data. Genovese et al. [9] showed that given noisy data sampled from a manifold, the density ridges are close to the underlying manifold bounded by Hausdorff distance and are thus good candidates for estimating manifolds.

Given data points $X \in \mathbb{R}^D$ sampled with additive noise $\Phi_\sigma = N(0, \sigma I)$ from some distribution W supported along a manifold M, we can define the probability density function [9]:

$$p(x) = W * \Phi_\sigma, \tag{1}$$

where $*$ denotes convolution [9]. The density ridges are then defined through the gradient $g(\mathbf{x}) \doteq \nabla^T p(\mathbf{x})$ and the Hessian matrix $H(\mathbf{x}) \doteq \nabla^T \nabla p(\mathbf{x})$ of the probability density $p(\mathbf{x})$:

Definition 1 (Ozertem 2011) *A point $\mathbf{x}$ is on the d-dimensional ridge, R, of its probability density function, when the gradient $g(\mathbf{x})$ is orthogonal to at least $D - d$ eigenvectors of $H(\mathbf{x})$ and the corresponding $D - d$ eigenvalues are all negative.*

We express the spectral decomposition of H as $H(\mathbf{x}) = Q(\mathbf{x})\Lambda(\mathbf{x})Q(\mathbf{x})^T$, where $Q(\mathbf{x})$ is the matrix of eigenvectors sorted according to the size of the eigenvalue. Furthermore $Q(\mathbf{x})$ can be decomposed into $\left[Q_\perp(\mathbf{x})\; Q_\parallel(\mathbf{x})\right]$, where $Q_\perp$ are the d first eigenvectors of $Q(\mathbf{x})$, and $Q_\parallel$ are the $D - d$ smallest. The latter is referred to as the *orthogonal subspace* due to the fact that when at a ridge point, all eigenvectors in $Q_\perp$ will be orthogonal to $g(\mathbf{x})$.

This motivates the following initial value problem for projecting points onto a density ridge:

$$\frac{\mathrm{d}\mathbf{y}_t}{\mathrm{d}t} = V_t V_t^T g(\mathbf{y}_t), \tag{2}$$

where $V_t = Q_\perp(\mathbf{x}(t))$ at $\mathbf{y}_t = \mathbf{y}(t)$, and $\mathbf{y}(0) = \mathbf{x}$. We denote the set of $\mathbf{y}$'s that satisfy equation (2), calculated via the kernel density estimator $\hat{f}(\mathbf{x})$, as the d-dimensional *ridge estimator* $\hat{M}$.

Finally, given an estimate, $\hat{M}$, of the underlying manifold, we can introduce concepts from differential geometry that allow smooth operations on the manifold.

2.3. Manifold traversal

Intuitively, manifold traversal can be performed in two different ways:

- By piece-wise traversal along a tangent vector followed by a mapping onto the manifold (exponential map[1]).
- If the end point of the traversal is known, calculating the *geodesic path* between the start and end point and following it corresponds to traversal along the manifold.

Both concepts are closely related, but in this work we adopt the latter framework and calculate smooth geodesics between points on the manifold as estimated by the density ridge.

We start by recalling that a geodesic is informally defined as the shortest distance between two points on a manifold given that the path between the points is constrained to lie completely on the manifold. A complete formal definition can be found in Lee [14]. Hauberg et al. [12] showed that given a metric tensor at each point of the manifold solving the Euler Lagrange equation admits a computationally tractable scheme for finding geodesics through solving a system of differential equation. As the density ridge framework does not explicitly provides a metric tensor, we instead use an iterative scheme taking advantage of the following connections between kernel density ridges and topics from differential geometry (all concepts can be found in complete detail in [14]): (1) The top d Hessian eigenvectors, $Q_\parallel$, span the tangent space of $\hat{M}$ at $\mathbf{x}$. This gives an estimate of the tangent bundle of M through $\hat{M}$. (2) The density ridge projections correspond, under some constraints, to an approximate exponential map if sufficiently close to some local mode.

[1]The exponential map maps a vector in the tangent space of a manifold to the endpoint of the geodesic of the same length starting in the same point.

These properties allows alternating between shortening the path between two points and projecting the solution back onto the density ridge. Formally, given a sequence of points $\{\gamma_i\}_{i=1}^n$ between two points, $\mathbf{x}, \mathbf{y} \in M$ the problem of finding a shortest path constrained to the manifold is formulated as follows [11]:

$$\begin{aligned} \underset{\gamma}{\text{minimize}} \quad & \sum_{l=2}^{n} ||\gamma_l - \gamma_{l-1}||^2 \\ \text{subject to} \quad & \gamma_1 = \mathbf{x},\ \gamma_n = \mathbf{y},\ \gamma \in M. \end{aligned} \tag{3}$$

To minimize (3) the path γ is initialized using Dijkstra's algorithm and further discretized with linear interpolation between the n given points. Then minimization is performed by alternating between gradient descent to minimize distance and density ridge projection to project points back onto the manifold. Combined with the back-projection scheme presented in Section 2.4 the geodesics gives smooth interpolation in the ambient space of the manifold.

2.4. Reconstruction

Finally, to map the traversal along the low dimensional density ridge estimate back to the ambient data space, we use a two layer neural network with 50 hidden units, which is trained to reconstruct the input data from the low dimensional vectors. We note that PCA has an inverse mapping, which would avoid the need for training. However, due to the radical reduction of dimensions, our results illustrate that the more powerful neural network is generally more capable of approximating the inverse mapping. For the KECA dimensionality reduction no inverse mapping exists, such that learning of a mapping is required.

The size and number of hidden units in the network were chosen by manual tuning. In our experiments we found that the number of hidden units used in the reconstruction stage was quite robust, however, for future research we will investigate more precise and methodical ways of tuning this parameter.

2.5. Summary of workflow

A summary of the workflow is presented in Figure 1, where interpolation is performed between a smiling and a frowning image from the Frey face dataset. The top and bottom image in the final stage correspond to the original images, while the middle images are results from geodesic interpolation and backprojection through the neural network.

3. EXPERIMENTAL RESULTS

We evaluate our method on two real-world datasets, the ones in the MNIST dataset [20] and the Frey Faces dataset[2]. The pixel space of these datasets represents high-dimensional data, which has a low dimensional manifold structure.

[2]Obtained with kind permission from Brandon Frey, University of Toronto.

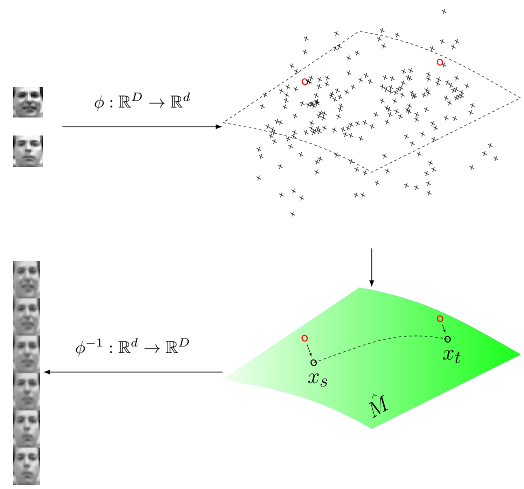
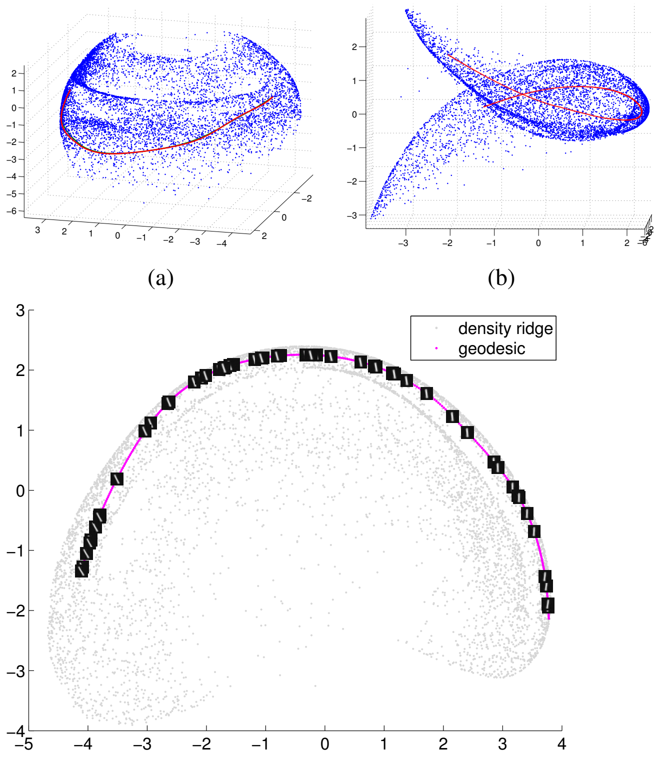


Fig. 1: Summary of workflow: ϕ is either PCA or KECA, ϕ^{-1} is the backprojecting neural network, and $\hat{M}$ is the density ridge estimate of the manifold. The images corresponds to the experiments in Section 3.2.

3.1. MNIST

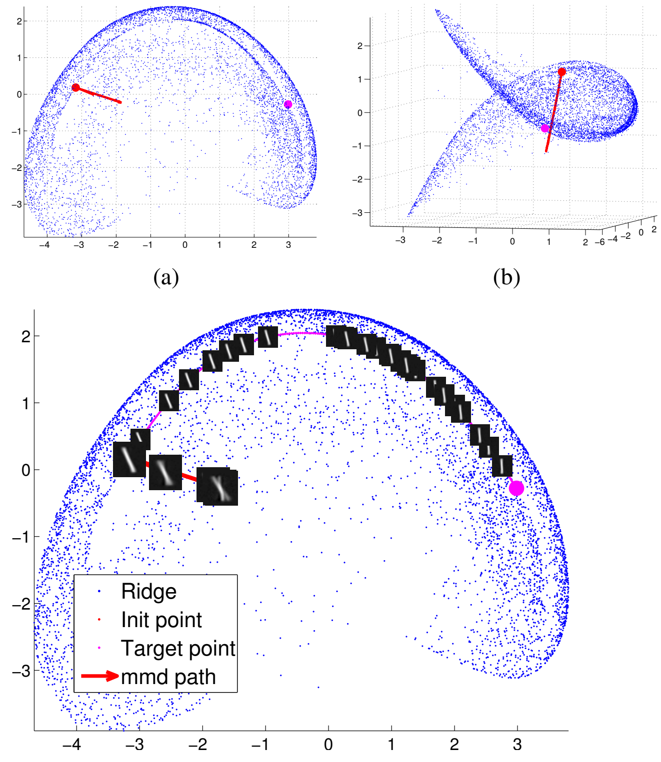
The intrinsic dimensionality of the MNIST dataset was estimated using the methods of Lombardi et al. [21] and Ceruti et al. [22], which indicate a ten dimensional manifold structure – also corroborated by the work of Sun and Marchand-Maillet [23]. When reducing the dimensionality to ten dimensions using PCA, we observe a smooth manifold structure (see Figure 2a for the first three PCA dimensions), however, we also note that the manifold is curved along several dimensions (see Figure 2b for an example). Figure 2c shows the result of our approach, where the leftmost and rightmost image are the start and final image of our traversal from the original dataset, respectively. All remaining images are generated by following the geodesic and projecting points back to the high dimensional ambient space. We note that the orientation of the one smoothly changes from leaning left to being straight.

We compare our approach of traversing manifolds along the geodesic with the alternative method of using the maximum mean discrepancy (MMD) [24] trajectory, as presented in Gardner et al. [13], to move between two data points. Figure 3 illustrates the above experiment using MMD. Figures 3a and 3b show the first steps of the MMD trajectory in the first three and second to fourth PCA dimension, respectively. We observe that the MMD trajectory clearly does not follow the manifold structure and therefore results in a considerably worse reconstruction (Figure 3c). Note that using the difference between means (MMD) to describe a trajectory along the manifold, is clearly not sufficient.



(a) (b) (c)

Fig. 2: Manifold traversal on MNIST. **(a)** PCA dimensions 1 to 3. **(b)** PCA dimensions 2 to 4. **(c)** The estimated ridge and the generated images when traversing the geodesic path (red line). The illustrated images are generated via the backprojection and are not in the original dataset.

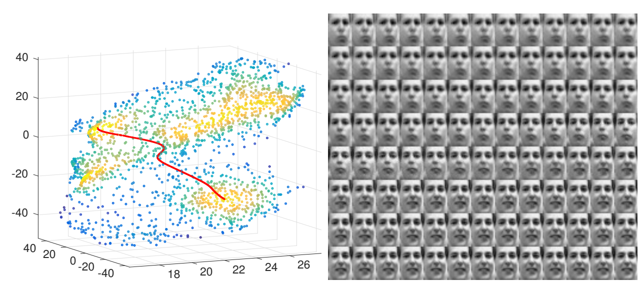


(a) (b) (c)

Fig. 3: Manifold traversal on MNIST using MMD. **(a)** PCA dimensions 1 to 3. **(b)** PCA dimensions 2 to 4. **(c)** The estimated ridge and the generated images following the MMD trajectory (red line).

3.2. Frey Faces

Figure 4a shows the density ridge estimate for the Frey Face dataset after dimensionality reduction using KECA. For visualization purposes we reduce the number of dimensions to three in the dimensionality reduction step. Similar to the MNIST dataset we observe a smooth manifold structure. Figure 4 illustrates the results of traversing from a frowning face along the geodesic (red line in Figure 4a) to a smiling image. The first image (top left) is the original start image, and the bottom right is the original end image, whereas images in-between do not exist in the original dataset and are generated using our proposed method. Initially, the main change in the images appears to be the rotation of the face, however, as we move closer to the smiling image the facial expression of the generated images clearly changes as well.



(a) (b)

Fig. 4: Interpolating between frowning and smiling images using KECA for dimensionality reduction. **(a)** The density ridge estimate for the top three entropy components of the Frey Faces dataset. The red line indicates the geodesic between a smiling and a frowning face. **(b)** The reconstructed images along the geodesic.

4. CONCLUSION

This work solves the problem of manifold traversal in the density ridge framework based on a novel method to estimate geodesics. We have shown that our low dimensional manifold approaches can model realistic complex shape variation and smooth transitions (or interpolations) can be achieved for real-world datasets. We compare our method to an established manifold traversal method, showing promising results.

5. REFERENCES

[1] Umut Ozertem and Deniz Erdogmus, "Locally defined principal curves and surfaces," *The Journal of Machine Learning Research*, vol. 12, pp. 1249–1286, 2011.

[2] Umut Ozertem and Deniz Erdogmus, "Signal denoising using principal curves: application to timewarping," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3709–3712.

[3] Deniz Erdogmus and Umut Ozertem, "Self-consistent locally defined principal surfaces," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 2, pp. II–549.

[4] Jonas Nordhaug Myhre, Matineh Shaker, Devrim Kaba, Robert Jenssen, and Deniz Erdogmus, "Manifold unwrapping using density ridges," *arXiv preprint arXiv:1604.01602*, 2016.

[5] Umut Ozertem, Deniz Erdogmus, and Miguel A Carreira-Perpinán, "Density geodesics for similarity clustering," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1977–1980.

[6] Esra Ataer-Cansizoglu, Erhan Bas, Jayashree Kalpathy-Cramer, Greg C Sharp, and Deniz Erdogmus, "Contour-based shape representation using principal curves," *Pattern Recognition*, vol. 46, no. 4, pp. 1140–1150, 2013.

[7] Arturo Mendoza Quispe and Caroline Petitjean, "Shape prior based image segmentation using manifold learning," in *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*. IEEE, 2015, pp. 137–142.

[8] Oliver Moolan-Feroze, Majid Mirmehdi, Mark Hamilton, and Chiara Bucciarelli-Ducci, "Segmentation of the right ventricle using diffusion maps and markov random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 682–689.

[9] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, Larry Wasserman, et al., "Nonparametric ridge estimation," *The Annals of Statistics*, vol. 42, no. 4, pp. 1511–1545, 2014.

[10] Joshua B Tenenbaum, Vin De Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[11] Piotr Dollár, Vincent Rabaud, and Serge Belongie, "Non-isometric manifold learning: Analysis and an algorithm," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 241–248.

[12] Søren Hauberg, Oren Freifeld, and Michael J Black, "A geometric take on metric learning," in *Advances in Neural Information Processing Systems*, 2012, pp. 2024–2032.

[13] Jacob R Gardner, Matt J Kusner, Yixuan Li, Paul Upchurch, Kilian Q Weinberger, and John E Hopcroft, "Deep manifold traversal: Changing labels with convolutional features," *arXiv preprint arXiv:1511.06421*, 2015.

[14] John M Lee, *Riemannian manifolds: an introduction to curvature*, vol. 176, Springer Science & Business Media, 2006.

[15] A Cochocki and Rolf Unbehauen, *Neural networks for optimization and signal processing*, John Wiley & Sons, Inc., 1993.

[16] Samuel Gerber and Ross Whitaker, "Regularization-free principal curve estimation," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1285–1302, 2013.

[17] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[18] Robert Jenssen, "Kernel entropy component analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 847–860, 2010.

[19] Matineh Shaker, Jonas N Myhre, M Devrim Kaba, and Deniz Erdogmus, "Invertible nonlinear cluster unwrapping," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*. IEEE, 2014, pp. 1–6.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] Gabriele Lombardi, Alessandro Rozza, Claudio Ceruti, Elena Casiraghi, and Paola Campadelli, "Minimum neighbor distance estimators of intrinsic dimension," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 374–389.

[22] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli, "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern recognition*, vol. 47, no. 8, pp. 2569–2581, 2014.

[23] Ke Sun and Stéphane Marchand-Maillet, "An information geometry of statistical manifold learning.," in *ICML*, 2014, pp. 1–9.

[24] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.