Uncertainty quantification for manifold valued models

Anton Mallasto, Søren Hauberg, **Aasa Feragen**

Section for Image Analysis & Computer Graphics, DTU Compute

afhar@dtu.dk

DALI/ELLIS workshop on Geometric Deep Learning

San Sebastian 05.09.2019

Danmarks
Tekniske
Universitet

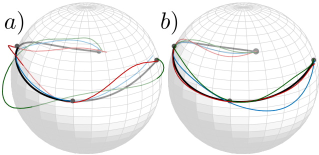
Most of the work in this talk is based on Anton Mallasto's upcoming PhD thesis

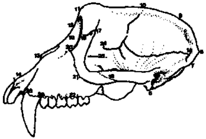
Ingrid (3 months) also helped!

Manifold valued models: motivation

Manifolds are everywhere:

- Implicitly defined via constraints
- Implicitly defined via wanted invariances
- Explicitly defined via a change of metric (learned or known)





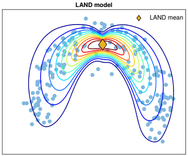
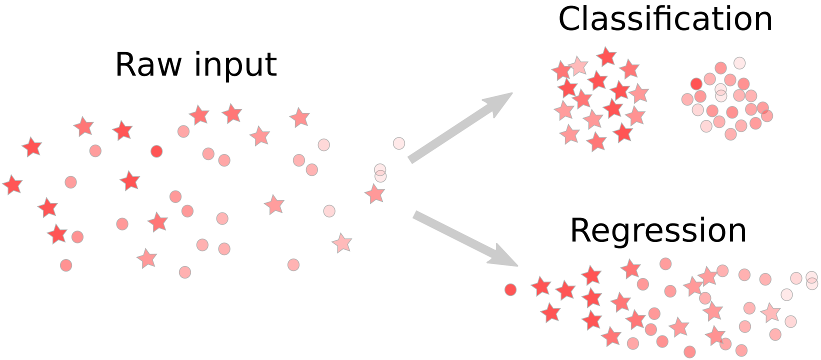


Figure sources: Dryden and Mardia (middle); Arvanitidis et al (right)

Manifold valued models: motivation

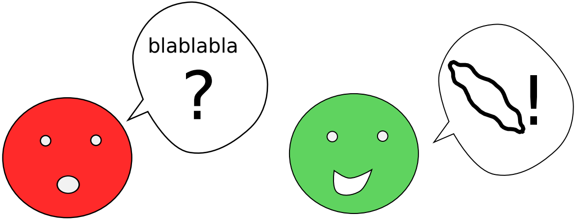
Manifolds as input is "easy": Map to feature space; "only" need to retain some level of order



Manifold valued models: motivation

Manifold-valued as output is often more difficult – mapping to feature space is often out of the question

- Manifold-valued regression
- Manifold-valued generative models
- Interpolation for manifold-valued data
- Interpretation



This talk

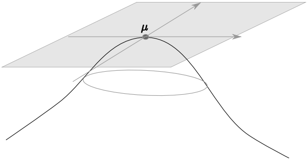
- Basic notation and definitions
- Generalizing GPs: Wrapped Gaussian Processes (WGPs)
- Manifold valued regression with UQ: WGP regression
- Uncertain submanifold learning: WGPLVM

Basic notation and definitions

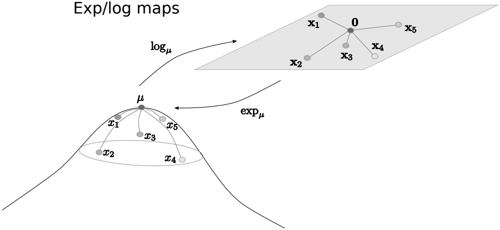
Riemannian manifolds

- *Riemannian manifold* = smooth manifold M with smoothly varying inner product (*Riemannian metric*) $g_p(\cdot,\cdot)$, aka $\langle\cdot,\cdot\rangle_p$ on tangent space T_pM
- Induces a distance function d and geodesics γ (locally distance minimizing) on M

Tangent space

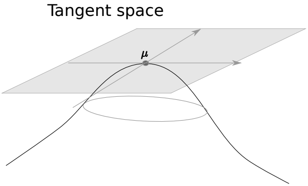


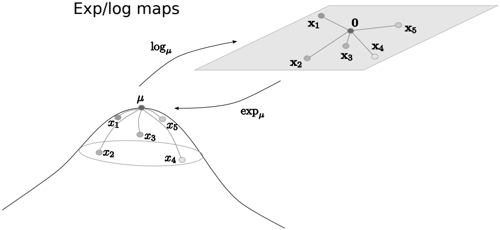
Exp/log maps



Riemannian manifolds

- *Riemannian manifold* = smooth manifold M with smoothly varying inner product (*Riemannian metric*) $g_p(\cdot,\cdot)$, aka $\langle\cdot,\cdot\rangle_p$ on tangent space T_pM
- Induces a distance function d and geodesics γ (locally distance minimizing) on M
- *Logarithmic* and *exponential maps* $\mathrm{Log}\colon M \to TM$, $\mathrm{Exp}\colon TM \to M$ locally linearize the manifold

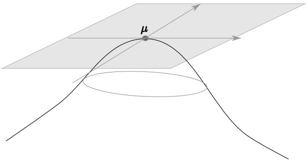




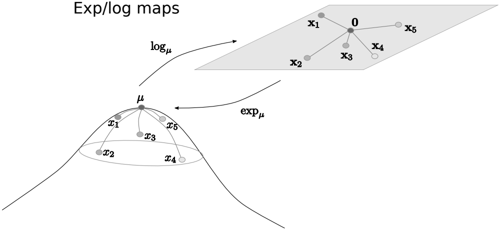
Riemannian manifolds

- *Riemannian manifold* = smooth manifold M with smoothly varying inner product (*Riemannian metric*) $g_p(\cdot,\cdot)$, aka $\langle\cdot,\cdot\rangle_p$ on tangent space T_pM
- Induces a distance function d and geodesics γ (locally distance minimizing) on M
- *Logarithmic* and *exponential maps* $\mathrm{Log}\colon M \to TM$, $\mathrm{Exp}\colon TM \to M$ locally linearize the manifold
- Exp_p is a diffeomorphism between a neighborhood $0 \in U \subset T_pM$ and neighbourhood $p \in V \subset M$, chosen maximally. V = *area of injectivity*.

Tangent space



Exp/log maps

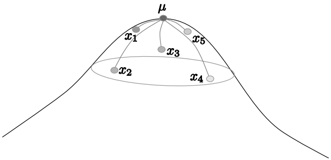


Product manifolds

- (M_i, g_i) Riemannian manifolds with, exponential maps Exp^i, logarithmic maps Log^i, $i = 1, 2$.
- $M = M_1 \times M_2$ is a Riemannian manifold with
  - metric $g = g_1 + g_2$,
  - component-wise computed exponential map $\mathrm{Exp}_{(p_1,p_2)}((v_1, v_2)) = \big(\mathrm{Exp}^1_{p_1}(v_1), \mathrm{Exp}^2_{p_2}(v_2)\big)$
  - component-wise log map as well

Expectations and means on Riemannian manifolds

- For a random point $X \in M$, its expectation, or set of *Fréchet means* is
$$\mathbb{E}[X] := \arg\min_{q \in M} (\mathbb{E}[d(q, X)^2]).$$
Can be multivalued!
- For a dataset $\boldsymbol{p} = \{p_i \in M\}_{i=1}^N$, the *empirical Fréchet mean* is the minimizer
$$\min_{q \in M} \sum_{i=1}^N d(q, p_i)^2.$$



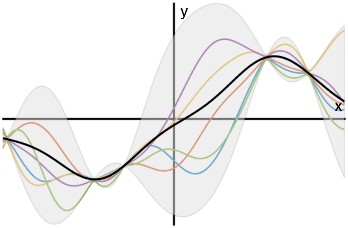
Gaussian Processes (GPs)

- *Gaussian process* (GP) = collection f of random variables s.t. any finite subcollection $(f(\omega_i))_{i=1}^N$ has a joint Gaussian distribution, where $\omega_i \in \Omega$ for the *index set* Ω.
- Entirely characterized by the *mean function* m and *covariance function* k:

$$m(\omega) = \mathbb{E}\left[f(\omega)\right], \tag{1}$$

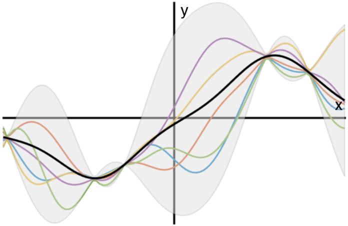
$$k(\omega, \omega') = \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \tag{2}$$

- Notation: $f \sim \mathcal{GP}(m, k)$.



What do we need to obtain manifold valued GPs?

- Joint “GDs”



Euclidean GP regression

- Training data: $\boldsymbol{D} = \{(x_i, y_i) \mid x_i \in \boldsymbol{x} \subset \mathbb{R}^l,\ y_i \in \boldsymbol{y} \subset \mathbb{R}^n\}$
- The GP predictive distribution at outputs $\boldsymbol{y}_*$ at test inputs $\boldsymbol{x}_*$:

$$p(\boldsymbol{y}_*|\boldsymbol{D}, \boldsymbol{x}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*), \tag{3}$$

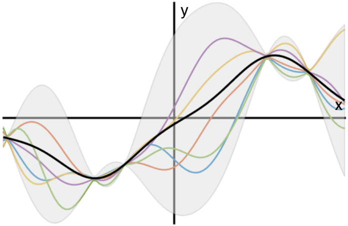
$$\boldsymbol{\mu}_* = \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{y}, \tag{4}$$

$$\Sigma_* = \boldsymbol{k}_{**} - \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{k}_*, \tag{5}$$

where, given a kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we use the notation $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}, \boldsymbol{x}_*)$, $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ and K_{err} is the measurement error variance.

What do we need to obtain manifold valued GPs?

- Joint "GDs"
- Conditioning of the "joint GD"



Generalizing GPs: Wrapped Gaussian Processes (WGPs)

Mallasto, **F**, CVPR'18

Wrapped Gaussian Distributions (WGDs)[2]

- n-dimensional Riemannian manifold (M, d)
- Stochastic variable $X \in M$ follows a *wrapped Gaussian distribution* (WGD) if for some $\mu \in M$ and SPD matrix $K \in \mathbb{R}^{n \times n}$,

$$X \sim \left(\mathrm{Exp}_\mu\right)_\# \left(\mathcal{N}(0, K)\right),$$

- Notation: $X \sim \mathcal{N}_M(\mu, K)$.
- The *basepoint* and *tangent space covariance* of X are

$$\mu_{\mathcal{N}_M}(X) := \mu, \ \mathrm{Cov}_{\mathcal{N}_M}(X) := K.$$

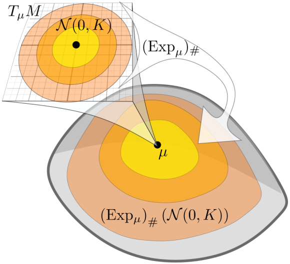


Figure: WGD defined as a Gaussian $\mathcal{N}(0, K)$ in the tangent space $T_\mu M$, pushed forward by Exp_μ to M.

[2]Mardia and Jupp, *Directional Statistics*, 2009

Needed for wrapped GPs: Jointly WGD stochastic variables

- Random points $X_i \sim \mathcal{N}_{M_i}(\mu_i, K_i)$, $i = 1, 2$, are *jointly WGD*, if the random point (X_1, X_2) is WGD on $M_1 \times M_2$:

$$(X_1, X_2) \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}K_1 & K_{12} \\ K_{21} & K_2\end{pmatrix}\right),$$

for some matrix $K_{12} = K_{21}^T$.

Needed for wrapped GPs: Conditioning

Theorem

Assume X_1, X_2 are jointly WGD as in (16), then we have the conditional distribution

$$X_1|(X_2 = p_2) \sim (\mathrm{Exp}_{\mu_1})_{\#} \left(\sum_{v \in A} \lambda_v \mathcal{N}(\mu_v, K_v) \right),$$

where

$$\begin{aligned} \mu_v &= K_{12} K_2^{-1} v, \\ K_v &= K_1 - K_{12} K_2^{-1} K_{12}^T, \\ \lambda_v &= \frac{\mathcal{N}(v|\mathbf{0}, K_2)}{\mathbb{P}\{A\}}, \\ A &= \{v \in T_{\mu_2} M \mid \mathrm{Exp}_{\mu_2}(v) = p_2\}, \\ \mathbb{P}\{A\} &= \sum_{v \in A} \mathcal{N}(v|\mathbf{0}, K_2). \end{aligned}$$

Special case: Infinite injectivity radius

- When the Exp and Log maps are globally 1-1
  - Manifolds of non-positive curvature
  - Wasserstein geometry on normal distributions
  - Typical Riemannian geometries on SPD matrices
- In this case, $\mu_{\mathcal{N}_M}(X) \in \mathbb{E}[X]$ (not generally)
- In this case,

$$X_1|(X_2 = p_2) \sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#} \left(\mathcal{N}\left(\mu_{\mathrm{Log}_{\mu_2}(p_2)}, K_{\mathrm{Log}_{\mu_2}(p_2)}\right)\right),$$

- **In practice:** Assume probability mass on the area of injectivity large $\rightsquigarrow$ this is a reasonable approximation, i.e. the Gaussian mixture in the tangent space is well approximated by a single Gaussian.

The Wrapped Gaussian Process (WGP)[3]

- A collection f of random points on a manifold M indexed over a set Ω is a *wrapped Gaussian process* (WGP), if every finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly WGD on M^N.
- We define

$$\begin{aligned} m(\omega) \quad &:= \mu_{\mathcal{N}_M}(f(\omega)) \\ k(\omega, \omega') \quad &:= \mathrm{Cov}_{\mathcal{N}_M}(f(\omega), f(\omega')) \end{aligned}$$

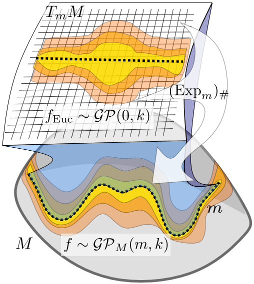
  called the *basepoint function* (BPF) and *tangent space covariance function* (TSCF) of f, respectively.
- Entirely characterized by the pair (m, k), similar to the Euclidean case.
- Notation: $f \sim \mathcal{GP}_M(m, k)$.

[3] Mallasto, **F**, CVPR'18

Remark: Viewed via an infinite product manifold

- A WGP f can be viewed as a WGD on the possibly infinite-dimensional product manifold $M^{|\Omega|}$.
- f defines a GP f_{Euc} in the tangent spaces $T_m M \subset M$ over the basepoint function, pushing each marginal $f(x_i)$ forward onto M by $(\mathrm{Exp}_{m(x_i)})_{\#}(f(x_i))$.
- Formally:

$$f \sim (\mathrm{Exp}_m)_{\#}(\mathcal{GP}(0, k)).$$



Manifold valued regression with UQ:

Wrapped Gaussian Process Regression on Riemannian Manifolds

Anton Mallasto, Aasa Feragen

CVPR 2018

Setting

- Infinite injectivity radius (or using the unimodal approximation)
- Noise-free training data (later with noise)

$$\boldsymbol{D}_M = \{(x_i, p_i) \mid x_i \in \mathbb{R}^l,\ p_i \in M,\ i = 1, ..., N\}.$$

- Denote $\boldsymbol{x} = (x_i)_{i=1}^N$ and $\boldsymbol{p} = (p_i)_{i=1}^N$; moreover $\boldsymbol{x}_*$ is used for test inputs, and $\boldsymbol{p}_*$ for test outputs.

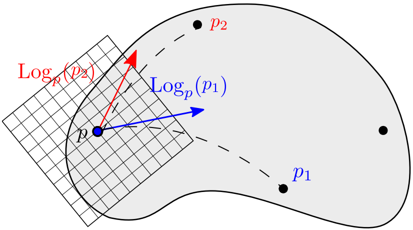
GP regression on manifolds: A naïve benchmark

- Choose $p \in M$ (typically $p \in \mathbb{E}[\boldsymbol{p}]$; transfrom the training data $\boldsymbol{D}_M$ into $\boldsymbol{D}_{T_pM}$ by

$$\boldsymbol{D}_{T_pM} = (\boldsymbol{x}, \boldsymbol{y}) := \{(x_i, y_i) \mid y_i = \mathrm{Log}_p(p_i)\}.$$

- Apply GP regression $f_{euc} \sim \mathcal{GP}(m_{euc}, k_{euc})$ in the tangent space, giving a predictive distribution $\boldsymbol{y}_*|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)$.
- Map back to the manifold M, resulting in

$$\boldsymbol{p}_*|\boldsymbol{p} = \mathrm{Exp}_p(\boldsymbol{y}_*) \sim \left(\mathrm{Exp}_p\right)_{\#}\left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right).$$



WGP regression: Noise-free

- Assuming a WGP prior $f_{prior} \sim \mathcal{GP}_M(m, k)$, the joint distribution between the training outputs $\boldsymbol{p}$ and test outputs $\boldsymbol{p}_*$ at $\boldsymbol{x}_*$ is

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2} \left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} \end{pmatrix} \right),$$

  where $\boldsymbol{m} = m(\boldsymbol{x})$, $\boldsymbol{m}_* = m(\boldsymbol{x}_*)$, $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}_*, \boldsymbol{x})$, and $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$.
- Therefore (using the unimodal approximation if necessary):

$$\begin{aligned} \boldsymbol{p}_* | \boldsymbol{p} &\sim \left(\mathrm{Exp}_{\boldsymbol{m}_*}\right)_{\#} \left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right), \\ \boldsymbol{\mu}_* &= \boldsymbol{k}_* \boldsymbol{k}^{-1} \mathrm{Log}_{\boldsymbol{m}} \boldsymbol{p}, \\ \Sigma_* &= \boldsymbol{k}_{**} - \boldsymbol{k}_* \boldsymbol{k}^{-1} \boldsymbol{k}_*^T. \end{aligned}$$

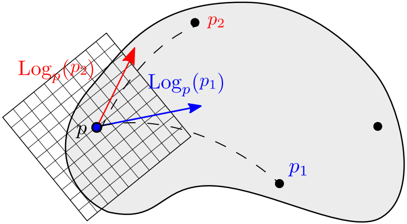
WGP regression: Noise-free

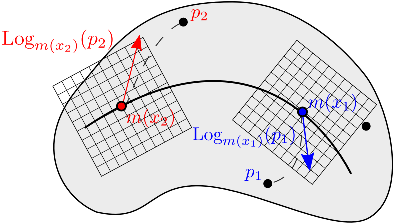
Remark

- The predictive distribution $\boldsymbol{p}_*|\boldsymbol{p}$ is not necessarily WGD, as $\boldsymbol{\mu}_*$ might be non-zero.
- The distribution can be sampled from, but computing quantities such as $\mathbb{E}[\boldsymbol{p}_*|\boldsymbol{p}]$ exactly is not trivial.
- $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ is not necessarily a Fréchet mean of $\boldsymbol{p}_*|\boldsymbol{p}$. However, it is the *maximum a posteriori* (MAP) estimate.

Choosing a prior

- An "informed" choice of prior base point function helps correctly localize the regressor
- We used (left) the Fréchet mean (constant function, giving naïve baseline) or (right) the output of geodesic regression or principal curves





WGP algorithm

Input Manifold-valued training data $\boldsymbol{D}_M = \{(x_i, p_i)\}_{i=1}^{n}$.

Output Predictive distribution for $\boldsymbol{p}_*|\boldsymbol{p}$ at $\boldsymbol{x}_*$.

i. Choose a prior BPF m.

ii. Transform $\boldsymbol{D}_{T_{\boldsymbol{m}}M} \leftarrow \{(x_i, \mathrm{Log}_{m(x_i)}(p_i))\}_{i=1}^{N}$.

iii. Choose a parametric prior TSCF k

iv. Using GP prior $\mathcal{GP}(0, k)$, carry out Euclidean GP regression for the transformed data $\boldsymbol{D}_{T_{\boldsymbol{m}}M}$, yielding the mean and covariance $(\boldsymbol{\mu}_*, \Sigma_*)$.

vi. End with the predictive distribution $\boldsymbol{p}_*|\boldsymbol{p} \sim (\mathrm{Exp}_{\boldsymbol{m}_*})_{\#}(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*))$

WGP regression: Noisy case

- The standard Euclidean noise model is $p_i = f(x_i) + \epsilon$, $\epsilon \sim \mathcal{N}(0, K_{\mathrm{err}})$
- We thus propose the error model $\mathrm{Log}_{m(x_i)}(p_i) = \mathrm{Log}_{m(x_i)}(f(x_i)) + \epsilon$. That is, the error lives in the tangent space of the prior mean at x_i.

WGP regression: Noisy case

- The standard Euclidean noise model is $p_i = f(x_i) + \epsilon$, $\epsilon \sim \mathcal{N}(0, K_{\mathrm{err}})$
- We thus propose the error model $\mathrm{Log}_{m(x_i)}(p_i) = \mathrm{Log}_{m(x_i)}(f(x_i)) + \epsilon$. That is, the error lives in the tangent space of the prior mean at x_i.
- The joint distribution of $\boldsymbol{p}$ and $\boldsymbol{p}_*$ changes into

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} + K_{\mathrm{err}} \end{pmatrix} \right).$$

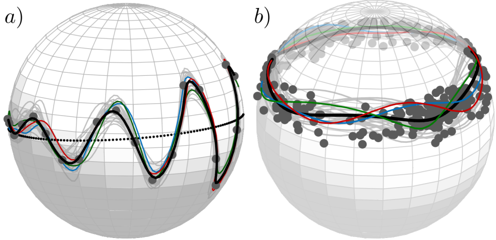
WGP regression: Noisy case

- The standard Euclidean noise model is $p_i = f(x_i) + \epsilon$, $\epsilon \sim \mathcal{N}(0, K_{\mathrm{err}})$
- We thus propose the error model $\mathrm{Log}_{m(x_i)}(p_i) = \mathrm{Log}_{m(x_i)}(f(x_i)) + \epsilon$. That is, the error lives in the tangent space of the prior mean at x_i.
- The joint distribution of $\boldsymbol{p}$ and $\boldsymbol{p}_*$ changes into

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} + K_{\mathrm{err}} \end{pmatrix} \right).$$

- The remaining computations are then carried out similarly, with the replacement of $\boldsymbol{k}$ with $\boldsymbol{k} + K_{\mathrm{err}}$ everywhere.

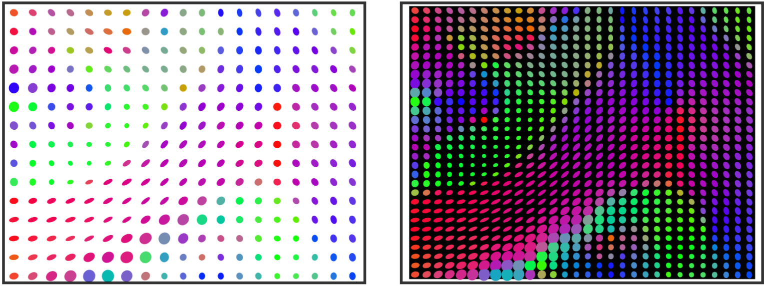
WGP regression in action on the sphere



- a) WGP regression using a prior BPF given by geodesic regression (dotted black) on a toy data set (grey dots) on S^2. The predictive distribution is visualized using the MAP estimate (black line), and 20 samples from the distribution (in gray) with three samples emphasized (in red, green and blue).
- b) A motion capture dataset of the orientation of the left *femur* of a walking person. The independent variables were estimated by *principal curve* analysis, and a WGP was fitted.

WGP regression in action on diffusion tensors

$a)$ Original Dataset $b)$ WGP Prediction

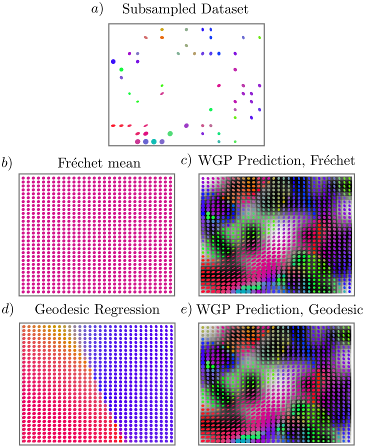


- Upsampling DTI tensor field by WGP regression.
- Colors depict the direction of the principal eigenvector of the respective tensor.
- Upsampling using the MAP estimate of the predictive distribution of WGP regression on the original data set with uncertainty visualized below (white = maximum relative error, black = no error).

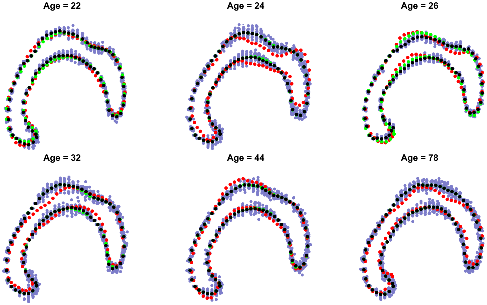
WGP regression in action on diffusion tensors

- Upsampling a subsampled DTI tensor field by WGP regression based on 20% of the original elements
- Regression using two different prior WGP BPFs:
  - b-c) the Fréchet mean
  - d-e) geodesic regression

  in both cases predicting via the MAP estimate
- The uncertainty fields in c) and e) have similar shapes, but the magnitudes differ.



WGP regression in action on Kendall shape space



- WGP regression predicting Corpus Callosum shape from age
- Red = data points from the test set, not used for training
- Black = the MAP estimates of the predictive distributions
- Green = values of the prior BPF (tangent space geodesic regression) at corresponding ages
- Blue = 20 samples from the predictive distribution

Uncertain submanifold learning: WGPLVM

Anton Mallasto, Søren Hauberg, Aasa Feragen

Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models

AISTATS 2019

Learning latent representations

In differential geometric terms: A latent variable or (sub)manifold learning model learns a (sometimes stochastic) *chart* for the manifold on which the data lies.

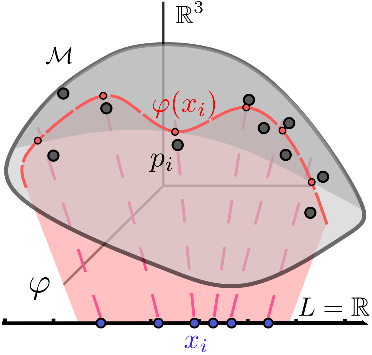


Figure: Submanifold learning

Gaussian Process Latent Variable Model (GPLVM)

- Aims to learn a probabilistic model relating elements in the low dimensional *latent space* $L \subseteq \mathbb{R}^{n'}$ to observed data $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^n$, with $n' < n$.
- In geometric terms, learns a latent space by optimizing over input variables for GP regression predicting the observed data.
- Computed by: Choosing a prior GP $f \sim \mathcal{GP}(m, k_\theta)$ with hyper-parameters $\theta \in \Theta$. The hyper-parameters are optimized with the *latent variables* $X = \{x_i\}_{i=1}^N \in L$ to maximize the log-likelihood

$$\begin{aligned}\log(\mathbb{P}(Y|X,\theta)) = &-\frac{nN}{2}\ln(2\pi) - \frac{n}{2}\ln|K_{X,\theta}| \\ &- \frac{1}{2}\mathrm{Tr}\left(K_{X,\theta}^{-1}YY^T\right),\end{aligned}$$

The Wrapped Gaussian Process Latent Variable Model (WGPLVM)

- $P = \{p_i\}_{i=1}^{N}$ on n-dim ambient Riemannian manifold $\mathcal{M}$.
- Consider a family of WGPs $f \sim \mathcal{GP}_{\mathcal{M}}(m, k_\theta)$, $f\colon L \to \mathcal{M}$ ($\theta \in \Theta$ hyperparameters)

The Wrapped Gaussian Process Latent Variable Model (WGPLVM)

- $P = \{p_i\}_{i=1}^N$ on n-dim ambient Riemannian manifold $\mathcal{M}$.
- Consider a family of WGPs $f \sim \mathcal{GP}_{\mathcal{M}}(m, k_\theta)$, $f : L \to \mathcal{M}$ ($\theta \in \Theta$ hyperparameters)
- **The likelihood** assigned by the prior f to a data point p with associated latent variable x is

$$\begin{aligned}\mathbb{P}\{p|x,\theta\} &= \sum_{v \in \mathrm{Exp}_{m(x)}^{-1}(p)} \mathcal{N}(v|\mathbf{0}, K_{x,\theta}) \\ &\approx \mathcal{N}\left(\mathrm{Log}_{m(x)}(p)|\mathbf{0}, K_{x,\theta}\right),\end{aligned}$$

where $(K_{x,\theta})_{ij} = k_\theta(x^i, x^j)$ and $x = (x^1, x^2, ..., x^n)$.

The Wrapped Gaussian Process Latent Variable Model (WGPLVM)

- $P = \{p_i\}_{i=1}^N$ on n-dim ambient Riemannian manifold $\mathcal{M}$.
- Consider a family of WGPs $f \sim \mathcal{GP}_{\mathcal{M}}(m, k_\theta)$, $f \colon L \to \mathcal{M}$ ($\theta \in \Theta$ hyperparameters)
- **The likelihood** assigned by the prior f to a data point p with associated latent variable x is

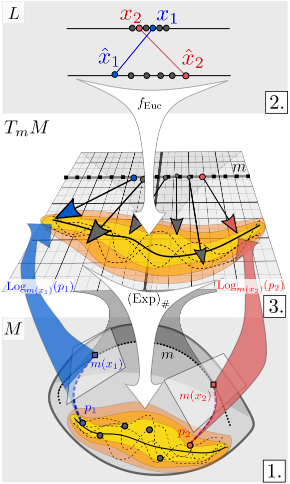
$$\begin{aligned}\mathbb{P}\{p|x,\theta\} &= \sum_{v \in \mathrm{Exp}_{m(x)}^{-1}(p)} \mathcal{N}(v|\mathbf{0}, K_{x,\theta}) \\ &\approx \mathcal{N}\left(\mathrm{Log}_{m(x)}(p)|\mathbf{0}, K_{x,\theta}\right),\end{aligned}$$

  where $(K_{x,\theta})_{ij} = k_\theta(x^i, x^j)$ and $x = (x^1, x^2, ..., x^n)$.
- **Maximize** the approximate log-likelihood

$$\begin{aligned}\ln\left(\mathbb{P}\{p|x,\theta\}\right) \approx &-\frac{nN}{2}\ln(2\pi) - \frac{n}{2}\ln|K_{x,\theta}| \\ &-\frac{1}{2}\mathrm{Log}_{m(x)}(p)^T K_{x,\theta}^{-1}\mathrm{Log}_{m(x)}(p),\end{aligned}$$

The WGPLVM pipeline

1. The data $p_i \in \mathcal{M}$ (blue and red dots) is transformed to the tangent bundle by $p_i \mapsto \mathrm{Log}_{m(x_i)}(p_i) \in T_{m(x_i)}\mathcal{M} \subset T_m\mathcal{M}$ along the prior basepoint function m (dotted black line) at initial latent variables x_i.
2. A GPLVM is learned, yielding the latent variables $\hat{x}_i \in L$ and the GP f_{Euc} from L to the tangent bundle.
3. The GP f_{Euc} is then pushed forward onto $\mathcal{M}$ by $(\mathrm{Exp})_\#(f_{\mathrm{Euc}})$, resulting in the predicted data submanifold.

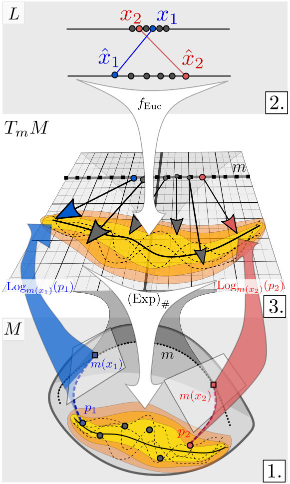


Interpretation

- Basepoint function m can delocalize the learning process in order to avoid distortions of the metric caused by linearization of the curved $\mathcal{M}$.
- Kernel k_θ governs interaction between observations in different tangent spaces

Predictions

- **Approximate submanifold** can be predicted at arbitrary latent variables X_{Pred}, by conditioning $\hat{f} \sim \mathcal{GP}_{\mathcal{M}}(m, k_{\hat{\theta}})$ on the data P with the associated latent variables $\hat{X}$.
- The conditional distribution will then be a non-centered GP $f_{\mathrm{Euc}} \sim \mathcal{GP}(m_{\mathrm{Euc}}, k_{\mathrm{Euc}})$ defined on $T_m\mathcal{M}$ pushed forward by the exponential map, resulting in the predictive distribution $\varphi_{\mathrm{pred}} \sim (\mathrm{Exp}_{m(x)})_{\#}(f_{\mathrm{Euc}})$.
- The *mean prediction* is given by $\bar{\varphi}_{\mathrm{pred}}(x) = (\mathrm{Exp}_{m(x)})_{\#}(m_{\mathrm{Euc}})(x))$.

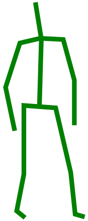


Optimization and computation

- **The initial latent variables** $X = \{x_i\}_{i=1}^N$ can be chosen strategically to aid optimization. We use *principal geodesic analysis* (for geodesic trend) and *principal curves* (otherwise)
- **The basepoint function** was set to the Fréchet mean, but could in principle be optimized over, in particular for very spread-out data
- **Computational complexity** is $\mathcal{O}(NL + N^3)$, where L is the cost of computing the Riemannian logarithm.

WGPLVM in action: Datasets and manifolds used

Femur dataset on S^2. A set of directions $P = \{p_i\}_{i=1}^N \in S^2$ of the left *femur* bone of a person walking in a circular pattern is measured at $N = 338$ time points.



WGPLVM in action: Datasets and manifolds used

Diatom shapes in Kendall's shape space. Diatoms are unicellular algae, whose species are related to their shapes. In Kendall's shape space M_K we analyze a set of outline shapes of 780 *diatoms* from 37 different species.

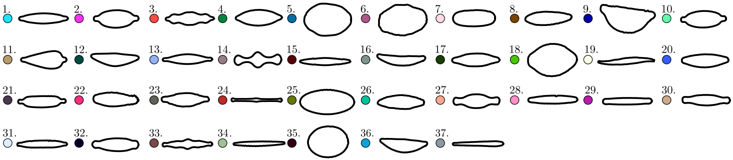


Figure: Representatives of each of the 37 diatom classes.

WGPLVM in action: Datasets and manifolds used

Diffusion tensors in $SPD(3)$ and Crypto-tensors in $SPD(10)$, Log-Euclidean metric.

- $SPD(3)$: Collect a set of 750 diffusion tensors from a diffusion MRI dataset, sampled with approximately uniform fractional anisotropy values.
- $SPD(10)$: Collect price of 10 popular crypto-currencies in the time 2.12.2014-15.5.2018; encode the crypto-currency intra-relationship at a given time in the covariance matrix between the prices in the past 20 days. Include every 7th day in the period, resulting in 126 10×10 covariance matrices.

WGPLVM in action: Visualization

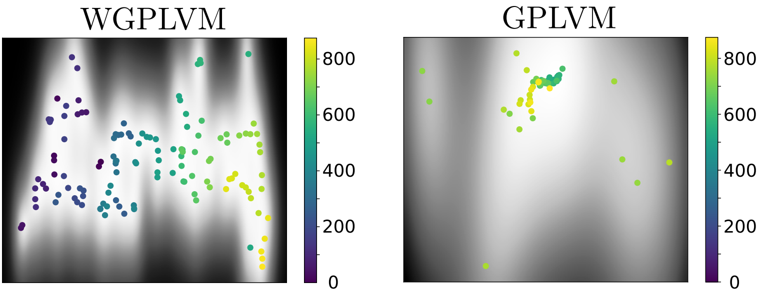


Figure: The latent space for the crypto-tensor dataset, with days visualized by color. Note that for GPLVM, the dark blue points corresponding to early times are hidden underneath the green points.

WGPLVM in action: Visualization

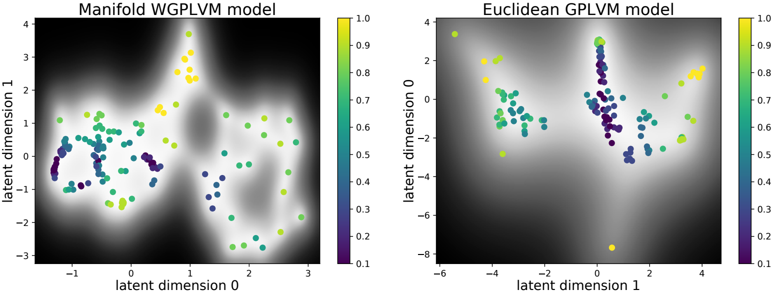


Figure: The latent spaces for the diffusion-tensor dataset learned using the WGPLVM and GPLVM models. The colors indicate the FA of the given tensor.

WGPLVM in action: Visualization

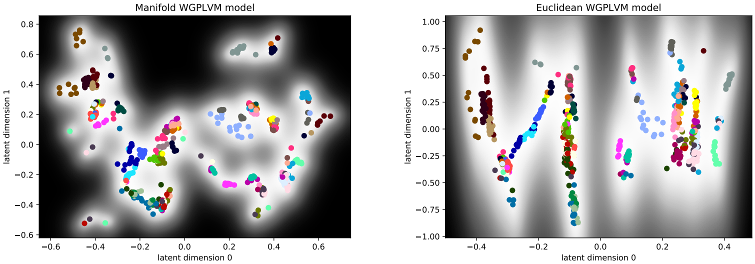
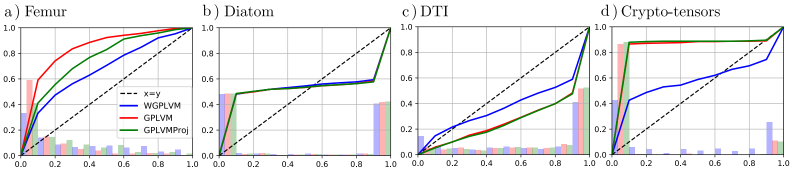


Figure: The latent spaces for the diatom dataset learned using the WGPLVM and GPLVM models. The colors indicate the species of the diatom corresponding to the latent variable.

WGPLVM in action: Uncertainty quantification



- Uncertainty estimates given by the WGPLVM, GPLVM and projected GPLVM models for the four datasets.
- Bars represent the frequency of occurrences, where the fraction of samples, given by the x-value, lies closer to the mean prediction than a test point.
- Continuous curves represent the cumulative distributions.
- If the cumulative distribution lies above $x = y$, we are overestimating the corresponding quantile, and vice versa.
- “Close to diagonal” = “good model fit”

WGPLVM in action: Encoding

Riemannian	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVMProj	$(9.22 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	0.582 ± 0.025	21.91 ± 2.26
WGPLVM	$(\mathbf{9.20} \pm \mathbf{0.53}) \times \mathbf{10^{-2}}$	$(\mathbf{2.39} \pm \mathbf{0.15}) \times \mathbf{10^{-2}}$	$\mathbf{0.391} \pm \mathbf{0.035}$	$\mathbf{3.04} \pm \mathbf{0.26}$

Euclidean	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVM	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	$(\mathbf{6.03} \pm \mathbf{0.34}) \times \mathbf{10^{-2}}$	$(7.36 \pm 5.27) \times 10^{5}$
GPLVMProj	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25 \times 10^{-2}$	$(\mathbf{6.03} \pm \mathbf{0.34}) \times \mathbf{10^{-2}}$	$(5.49 \pm 3.17) \times 10^{5}$
WGPLVM	$(\mathbf{9.19} \pm \mathbf{0.53}) \times \mathbf{10^{-2}}$	$(\mathbf{2.39} \pm \mathbf{0.15}) \times \mathbf{10^{-2}}$	$(7.54 \pm 0.36) \times 10^{-2}$	$(8.69 \pm 7.12) \times 10^{5}$

Figure: Mean reconstruction errors (top = intrinsic distance, bottom = Euclidean distance)

Discussion – what did we see?

Summary:

- WGPs: Generalization of GPs that takes values (as opposed to input) on a manifold
- Applications in WGP regression and WGPLVM
- Clearly improved uncertainty quantification over the Euclidean models

Discussion:

- These datasets were not particularly big, but even in the Euclidean models, the mean function learned the manifold anyway!
- However, in the Euclidean models, the covariance function does *not* learn the manifold on its own

Explanation:

- The uncertainty covers up a poor model fit of the parameterized covariance
- As a result, the Euclidean model assigns positive probability mass to impossible points.

Outlook

- GPs are rather restrictive – more flexible models of uncertainty?
- In particular (and in view of the name of the workshop) – deep WGPs?
- Closely related: Deep learning with manifold valued *output*?