
Supplementary material for Learning from uncertain curves: The 2-Wasserstein metric for Gaussian Processes

Anonymous Author(s)

Affiliation

Address

email

1 Content

The supplementary material consists of the movie file `climate.avi` as well as the appendices listed below.

A Details of data processing

A.1 The uncertain white-matter tract trajectories

The white-matter tract example is carried out on subject 103414 from the Q3 release of the Human Connectome Project [2–4]. The pre-processed HCP diffusion data contains 270 diffusion directions distributed equally over 3 shells with b -values = 1000, 2000 and 3000 s/mm². We perform probabilistic shortest-path tractography [5], which returns uncertain tracts in the form of GPs by estimating uncertain geodesics in an uncertain Riemannian manifold as described in [5]. The produced GPs are indexed over $X = [0, 1]$, have continuous covariance functions and L^2 integrable means, thus allowing us to apply the developed framework.

A.2 The GP temperature curves

The temperature curve GPs were estimated from a publicly available dataset of temperature measurements from Russian meteorological stations [6, 7]. We used data from the years 1940 – 2009, and 30 stations were randomly sampled among those that covered all these years. For each year, and each station, a GP was fitted to the measurements of minimal daily temperature, using a Gaussian covariance function. The maximum likelihood scale and bandwidth parameters of the covariance function were estimated using grid search. Years where more than 20 daily measurements were missing, were discarded from the dataset.

B Details of kernel regression

Given two metric spaces (X, d_X) and (Y, d_Y) , where the barycenter

$$\operatorname{argmin}_{y \in Y} \sum_{n=1}^N \xi_n d_Y^2(y, y_n)$$

is well defined, we can define a nonparametric *kernel regression* model as follows:

Given a sample of pairs $(x_1, y_1), \dots, (x_N, y_N) \in X \times Y$ of free variables $x \in X$ and dependent variables $y \in Y$, we predict Y -valued observations $y_0 = h(x_0)$ given new observations x_0 of x , as

26 follows:

$$y_0 = \operatorname{argmin}_{y \in Y} \sum_{n=1}^N k(x_n, x_0) d_Y^2(y, y_n)$$

27 where k is some kernel function. In our case, we use the Gaussian kernel

$$k(x_n, x_0) = C \cdot e^{-\frac{1}{2\sigma^2} d_X^2(x_n, x_0)},$$

28 where C is set to ensure that $\sum_{n=1}^N k(x_n, x_0) = 1$.

29 C Details of permutation test

30 Given a metric space (X, d_X) where the barycenter

$$\mu = \operatorname{argmin}_{y \in Y} \sum_{n=1}^N \xi_n d_Y^2(y, y_n)$$

31 is well defined, we can also define nonparametric hypothesis testing through permutation tests.
 32 Assume that $A = \{a_i\}_{i=1}^{N_1}$ and $B = \{b_j\}_{j=1}^{N_2}$ are two finite, sampled datasets in X ; now we define
 33 the test statistic $T(A, B) = d(\mu_A, \mu_B)$, where μ_A and μ_B are barycenters of the sets A and B ,
 34 respectively. Under the null hypothesis the samples A and B are drawn from the same distribution on
 35 \mathcal{T} , and randomly permuting the elements of A and B should not affect the value of T .

36 Join the sets in a larger dataset $S = A \cup B$ and consider partitions of X into datasets of size N_1 and
 37 N_2 . Due to the size of X we cannot usually check all possible permutations, but compute the test
 38 statistics $T_m = d(\mu_{A_m}, \mu_{B_m})$, $m = 1, \dots, M$, for M different random partitions $X = A_m \cup B_m$,
 39 with $|A_m| = N_1$ and $|B_m| = N_2$. Comparing the T_m to the original statistic value $T_0 = d(\hat{\mu}_A, \hat{\mu}_B)$
 40 we obtain a p -value approximating the probability of observing T_0 under the null hypothesis:

$$p = \frac{1 + \sum_{T_m \geq T_0, m \in \{1, \dots, M\}} 1}{M + 1},$$

41 where the additional 1 is added to avoid $p = 0$, which is impossible in the limit where all permutations
 42 are tested [1].

43 References

- 44 [1] Knijnenburg, T. A. and Wessels, L. F. A. and Reinders, M. J. T. and Shmulevich, I. Fewer
 45 permutations, more accurate P-values. In *Bioinformatics* (2009) Vol. 25, No. 12, pp. i161-i168.
- 46 [2] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., &
 47 WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: an overview.
 48 *Neuroimage*, 80, 62-79.
- 49 [3] Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., ...
 50 & Van Essen, D. C. (2013). The minimal preprocessing pipelines for the Human Connectome
 51 Project. *Neuroimage*, 80, 105-124.
- 52 [4] Sotiropoulos, S. N., Moeller, S., Jbabdi, S., Xu, J., Andersson, J. L., Auerbach, E. J., ... &
 53 Behrens, T. E. J. (2013). Effects of image reconstruction on fiber orientation mapping from
 54 multichannel diffusion MRI: reducing the noise floor using SENSE. *Magnetic resonance in*
 55 *medicine*, 70(6), 1682-1689.
- 56 [5] Hauberg, S., Schober, M., Liptrot, M., Hennig, P., & Feragen, A. (2015). A random riemannian
 57 metric for probabilistic shortest-path tractography. In *International Conference on Medical*
 58 *Image Computing and Computer-Assisted Intervention* (pp. 597-604). Springer International
 59 Publishing.
- 60 [6] Tatusko, R. L. (1990) Cooperation in climate research: An evaluation of the activities conducted
 61 under the US-USSR agreement for environmental protection since 1974. In *National Oceanic*
 62 *and Atmospheric Administration*, Washington, D.C.
- 63 [7] Bulygina, O. N. and Razuvaev, V. N. (2012). Daily Temperature and Precipitation Data
 64 for 518 Russian Meteorological Stations. In *Carbon Dioxide Information Analysis Center*,
 65 Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee. doi:
 66 10.3334/CDIAC/cli.100.