Topological Constraints and Kernel-Based Density Estimation

Florian T. Pokorny **Carl Henrik Ek** **Hedvig Kjellström** **Danica Kragic** *
School of Computer Science and Communication
KTH Royal Institute of Technology, Stockholm, Sweden
{fpokorny, chek, hedvig, danik}@csc.kth.se

Abstract

This extended abstract[1] explores the question of how to estimate a probability distribution from a finite number of samples when information about the *topology* of the support region of an underlying density is known. This workshop contribution is a continuation of our recent work [1] combining persistent homology and kernel-based density estimation for the first time and in which we explored an approach capable of incorporating topological constraints in bandwidth selection. We report on some recent experiments with high-dimensional motion capture data which show that our method is applicable even in high dimensions and develop our ideas for potential future applications of this framework.

1 Introduction

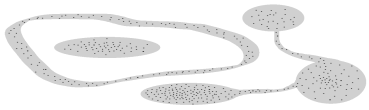


Figure 1: A topologically interesting point-cloud $S \subset \mathbb{R}^2$.

In recent years, novel topological techniques such as persistent homology have been developed to extract homological information from point-cloud data. This opens up a whole new range of possibilities for exploiting topological methods in a data-driven paradigm. Exciting theoretical results, providing guarantees on the reconstruction of homological information from point-cloud data for large sample sizes have been obtained [2] and a growing community of researchers has recently found interesting applications of the persistence algorithm, such as the analysis of computer vision and medical data in [3]. However, much of the motivation for this research has been coming from a theoretical point of view or has been focussed only on trying to *extract topological information from point-cloud data* without attempting to build a probabilistic model for the data. In the work presented here, we ask the question of how else persistent homology might be used as a tool for machine learning and suggest motion capture data as a potential domain of application. We present some preliminary experiments showing that our topological density estimation framework [1], which combines classical kernel-based density estimation [4] with persistent homology, can be applied even to high-dimensional motion capture data. Instead of trying to *recover* topological information from a finite number of samples, we are interested in the case where we are given *prior* topological information about the support region of some otherwise unknown probability distribution and where we would then like to utilize this information together with the persistence algorithm to select an optimal bandwidth parameter.

*This work was supported by the EU projects FLEXBOT (FP7-ERC-279933) and TOMSY (IST-FP7-270436) and the Swedish Foundation for Strategic Research

[1] Appearing in: Workshop on Algebraic Topology and Machine Learning, Advances in Neural Information Processing Systems (NIPS) 25, 2012

2 Background

Many probabilistic methods are based on Gaussian densities. Examples include Gaussian Mixture Models (GMM) [5], Gaussian Processes (GP) [6] and Gaussian Mixture Regression (GMR). While these methods are highly popular in application fields such as robotics, speech recognition and computer vision [5, 7, 8], a crucial and often overlooked property of such models is that they all generate densities on $\mathbb{R}^d$ for which $\operatorname{supp} f = \mathbb{R}^d$; i.e. *such models will assign a non-zero probability to every subset of* $\mathbb{R}^d$ *with non-zero volume*. This can cause problems in several real-world applications. Consider for example a probabilistic model of human pose positions in joint space. The two human hands, for example, are often in close contact during manipulation and grasping, yet we should assign zero probability to configurations corresponding to collisions with the environment or to self-intersections of the various body parts. Such constraints are impossible to enforce directly using Gaussian densities. A better approach is to base such a model on densities with bounded support which can be achieved using *kernel-based density estimation* [9, 10, 11]. There, one tries to reconstruct a probability density $f : \mathbb{R}^d \to \mathbb{R}$ of which only an i.i.d. sample $S = \{X_1, \cdots, X_n\}$ is known using an estimator $\hat{f}_{\varepsilon,n}$ defined by $\hat{f}_{\varepsilon,n}(x) = \frac{1}{n\varepsilon^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\varepsilon}\right)$, and where the kernel function $K : \mathbb{R}^d \to \mathbb{R}$ is a suitably chosen probability density which is concentrated near the origin. We will in particular focus on spherical kernels that are symmetric functions of the norm $\|x\|$ of their input variable $x \in \mathbb{R}^d$ and which satisfy $\operatorname{supp} K = \mathbb{B}_1(0)$, where $\mathbb{B}_\varepsilon(p) = \{x \in \mathbb{R}^d : \|x - p\| \leqslant \varepsilon\}$. Then $\operatorname{supp} \hat{f}_{\varepsilon,n} = Y_\varepsilon(S)$, where $Y_\varepsilon(S) = \bigcup_{i=1}^{n} \mathbb{B}_\varepsilon(X_i)$. By choosing ε small enough, we can hence 'design' a support region that is as concentrated as we want around the data points and which does not suffer from the Gaussians' trivial support region which we alluded to earlier. In kernel-based density estimation ε is called the *bandwidth* [4]. Theorem 2.1 provides reassurance that, for large sample sizes, $\hat{f}_{\varepsilon,n} \to f$ pointwise if the bandwidth is chosen correctly:

Theorem 2.1 (Theorem 3.1 and 3.2 [11])**.** *Suppose that K is a Borel measurable function on $\mathbb{R}^d$ such that $\sup_{x \in \mathbb{R}^d} |K(x)| < \infty$, $\int_{\mathbb{R}^d} |K(x)| \, \mathrm{d}x < \infty$, $\int_{\mathbb{R}^d} K(x) \mathrm{d}x = 1$ and $\lim_{\|x\| \to \infty} \|x\|^d K(x) = 0$. Suppose that $\{\varepsilon_n\}_{n=1}^{\infty}$ is a sequence of positive numbers such that $\lim_{n \to \infty} \varepsilon_n = 0$. Then $\lim_{n \to \infty} \mathbb{E}[\hat{f}_{\varepsilon_n,n}(x)] = f(x)$ at every point of continuity of f. If furthermore $\lim_{n \to \infty} n\varepsilon_n^d = \infty$, then $\lim_{n \to \infty} \mathbb{E}[(\hat{f}_{\varepsilon_n,n}(x) - f(x))^2] = 0$ at every point of continuity of f.*

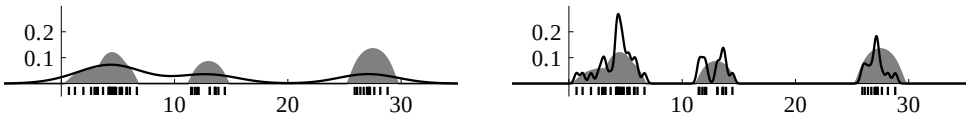


Figure 2: Graph of a density function f (shaded, grey), a selection of 50 samples (displayed below the x-axis) and a traditional Gaussian kernel-based estimate ($\sigma^2 = \frac{1}{4}$, black line) with a bandwidth that is too large ($\varepsilon = 5$, on the left) and too small ($\varepsilon = 0.4$, on the right) respectively.

The question of optimal bandwidth selection is an ongoing area of research [12] and optimality is traditionally studied in an asymptotic L^1 or L^2-error context [4, 13]. Given knowledge of the underlying density f and a spherically symmetric kernel K, the asymptotic mean integrated squared error is given by:

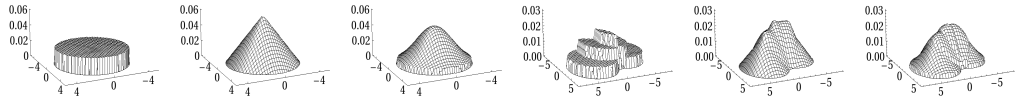
$$AMISE(\varepsilon_n) = \frac{1}{n\varepsilon_n^d} \int K(x)^2 \, \mathrm{d}x + \frac{\varepsilon_n^4}{4} \mu_2(K)^2 \int \{\operatorname{tr}(\operatorname{Hess} f(x))\}^2 \, \mathrm{d}x,$$

where $\mu_2(K) = \int x_i^2 K(x) \mathrm{d}x$ is independent of $j \in \{1, \ldots, d\}$ by the spherical symmetry. $AMISE(\varepsilon_n)$ provides an asymptotic estimate for the natural mean integrated squared error $MISE(\varepsilon) = \mathbb{E}\left[\int (\hat{f}_{\varepsilon,n}(x) - f(x))^2 \mathrm{d}x\right]$ if $\lim_{n \to \infty} \varepsilon_n = 0$ and $\lim_{n \to \infty} n\varepsilon_n^d = \infty$ [4]. A popular branch of bandwidth selection methods attempts to determine bandwidths which minimize AMISE [12] instead of working with the more complicated MISE directly.

Considering the point cloud S displayed in Figure 1, a human observer – and especially anyone familiar with persistent homology – will immediately notice that S intuitively has three connected components, one of which has a 'hole'. We can reformulate the existence of the 'hole' in Figure 1 in a mathematically precise way using the machinery of persistent homology by inspecting the

corresponding barcode. We will assume here that the specialized audience of this workshop is familiar with the basic concepts from persistent homology ([3] provides an excellent introduction to the field). Observe that, in persistent homology, we attempt to reconstruct information about the Betti numbers of an underlying space X from the samples S by studying the space $Y_\varepsilon(S)$ as $\varepsilon > 0$ varies – just like we vary ε in the case of bandwidth selection with spherical kernels. In practice, the homology groups of $Y_\varepsilon(S)$ are then approximated by computing the homology groups of the Vietoris-Rips complex $\mathcal{V}_{2\varepsilon}(S)$.

We consider three kernels with $\operatorname{supp} K = \mathbb{B}_1(0) \subset \mathbb{R}^d$ [1]. All these kernels are zero outside $\mathbb{B}_1(0)$ and defined by $K_u = \operatorname{Vol}(\mathbb{B}_1(0))^{-1}$ (uniform), $K_c(x) = \frac{d(d+1)\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}}(1 - \|x\|)$ (conic) and $K_t(x) = (2\pi\sigma^2)^{-\frac{d}{2}}\left(1 - \frac{\Gamma\left(\frac{d}{2}, \frac{1}{2\sigma^2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^{-1} e^{-\frac{\|x\|^2}{2\sigma^2}}$ (truncated Gaussian) respectively for $x \in \mathbb{B}_1(0)$ (see Figure 3 a-c).



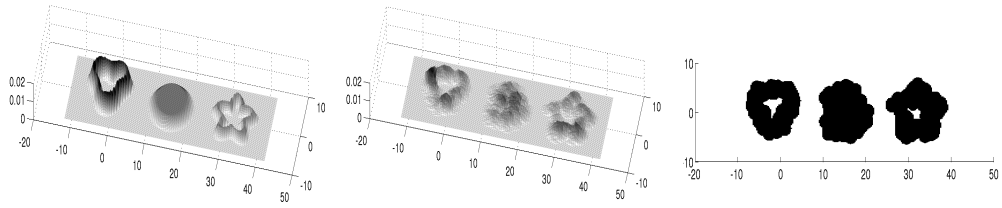
(a) K_u (b) K_c (c) K_t (d) $\hat{f}, K_u$ (e) $\hat{f}, K_c$ (f) $\hat{f}, K_u$

Figure 3: (a-c): $\frac{1}{\varepsilon^2}K(\frac{x}{\varepsilon})$, for $\varepsilon = 4$ and $K = K_u, K_c, K_t$ respectively. (d-e): Density approximation for $K = K_u, K_c, K_t$ in dimension 2 and with $\varepsilon = 4$ and for a few data points in the plane. For K_t, $\sigma^2 = \frac{1}{4}$.

In [1], we introduced the idea of *topological constraints* to bandwidth selection. i.e., given a density f such that some constraints on the Betti numbers of $\operatorname{supp} f$ and an i.i.d. sample S are known, we compute the barcode for S and search for barcode intervals such that these constraints are satisfied. We then chose the first such interval $I = [\varepsilon_{min}, \varepsilon_{max}]$ (which also coincided with the largest such interval in our application) and defined $\varepsilon_{top}(n) = \varepsilon_{min}(n) + \frac{\varepsilon_{max}(n) - \varepsilon_{min}(n)}{2} n^{-\frac{1}{4+d}}$ and were able to show in examples that the empirical MISE for this bandwidth choice converges quickly to zero as $n \to \infty$ for K_u, K_t, K_c. We believe that this novel type of *topological bandwidth selection*, i.e. choosing a bandwidth ε_n such that topological constraints are satisfied *and such that the expected squared error is minimized* for a given sample size n is a very interesting topic, which we believe should be explored further both from an experimental and a theoretical perspective.

3 Experiments

To gain an intuition for our approach, let us consider first a new 2D synthetic example of our topological density estimation framework. We consider the density f displayed Figure 4 a). Observe that $\Omega = \operatorname{supp} f$ has three connected components ($b_0(\Omega, \mathbb{Z}_2)) = 3$) and that two of these components have a 2-dimensional hole each ($b_1(\Omega, \mathbb{Z}_2) = 2$). Taking this observation as our prior topological knowledge about f during reconstruction, we are now able to compute a density estimator as displayed in 4 b) from only 300 samples and using ε_{top}. We can also clearly see in that figure and in 4 c) that the reconstructed density has the correct topological features. In this example, $\varepsilon_{min} \approxeq 1.36$, $\varepsilon_{max} \approxeq 2.23$ and $\varepsilon_{top} \approxeq 1.53$. The reconstruction L^1 and L^2 errors for ε_{top} were approximately 0.55 and 0.04 respectively. A very interesting topic that remains to be explored more fully and which we look forward to discussing during this workshop is the question of how else one could choose $\varepsilon_{top} \in [\varepsilon_{min}, \varepsilon_{max}]$ in order to minimize L^1 or L^2 errors. In [1], we determined an ε_{top} which ensures point-wise asymptotic convergence under some assumptions on $\varepsilon_{min}, \varepsilon_{max}$. In many application domains, one might however prefer to choose ε_{top} as small as possible in order not to 'generalize' the data too much. As outlined in the introduction, the probabilistic modelling of motion is an interesting area of application for our density estimation approach which we would like to begin to explore in this work. For this purpose, we generated a point-cloud $S = \{X_1, \ldots, X_{2276}\} \subset \mathbb{R}^{60}$, where each vector X_i is a concatenation of 20 vectors in $\mathbb{R}^3$. Each of these 20 vectors describes the (x, y, z) positions of one of twenty key/joint positions of a person at a fixed point in time (see the stick-person in Figure 5 for a visualization of one such X_i). We start with a short 35 frame motion-capture sequence of a person walking along a straight line taken from [14]. We then repeat these frames and translate and rotate each frame to obtain semi-synthetic data of a person walking

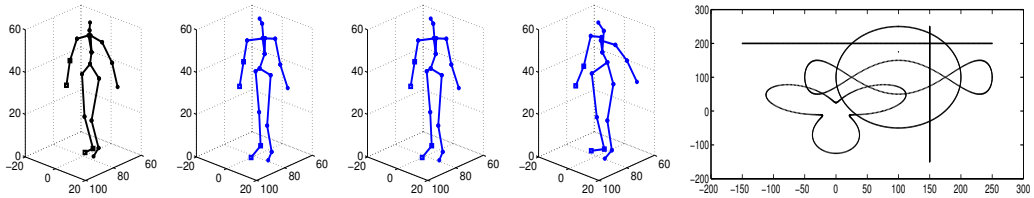


(a) (b) (c)

Figure 4: a) a $2D$ density. b) a uniform kernel estimate based on 300 samples and with $\varepsilon_{top} = 1.53$. c) the support region of the kernel estimate in b) with the correct number of 'holes' and components.

along the patterns in the plane displayed in Figure 5. Three of these paths are periodic closed-curve walking patterns, two are straight line walking segments and the last is a single isolated static posture. As given topological data, we assume the knowledge that $b_0(\Omega, \mathbb{Z}_2) = 6$ (*i.e.* there are 6 underlying motion patterns) and that $b_1(\Omega, \mathbb{Z}_2) = 3$ (*i.e.* there are 3 'cyclic' motions). This example is typical in that we do not have knowledge of any underlying probability density and can hence not evaluate the error incurred by our estimator. We do know however that a probabilistic model should not 'mix up' the various motion patterns by creating a density with too few connected components in its support region. We use JavaPlex [15] to determine the smallest bandwidth ε_{min} such that $b_0(\mathcal{V}_{2\varepsilon_{min}}(S), \mathbb{Z}_2) = 6$ and $b_1(\mathcal{V}_{2\varepsilon_{min}}(S), \mathbb{Z}_2) = 3$. In our experiment $\varepsilon_{min} \approxeq 12$. In this experiment, we choose $\varepsilon_{top} = \varepsilon_{min}$ as a conservative estimate for the bandwidth. This yields a density $\hat{f}$ with support equal to $Y_{12}(S)$ from which we can generate samples by selecting a data point with uniform probability and then sampling from the rescaled kernel K at that data point. Here, we chose $K = K_c$. Note that we can now continuously move from one data point to the next inside $Y_{12}(S)$ in each of the 6 connected components (*e.g.* to potentially obtain an interpolation of the whole walking sequence from the samples). We observe also that, unlike the projection of the paths in Figure 5, the paths in the full 60 dimensional space do not intersect.

Figure 5 shows a data point from S in 5 a) and 3 random samples from a conical kernel scaled by the bandwidth $\varepsilon_{top} = 12$ and centred at that data point 5 b-d). We note that, even though our dataset is very sparse in 60 dimensions, our framework enables us to *generalize from the data* while *satisfying the given topological constraints*.



(a) (b) (c) (d) (e)

Figure 5: Stick-person from the data (a) and three samples from a conical kernel with bandwidth 12 centred at that data-point (b-d). (e) displays the movement patterns in the xy-plane along which we generated our data points.

4 Conclusion and Future Work

In this work, we have begun to explore applications of our novel topological density estimation framework [1] to a new synthetic data set in $2D$ and to motion capture. From our evaluation of the estimation errors in [1] and the encouraging results with real motion data, we believe that these novel uses of persistent homology in kernel-based density estimation can be a valuable tool in real world applications. We are intending to investigate various aspects of our approach in the coming months. Incorporating a time-dependence in the motion-capture data is an obvious starting point for example.

References

[1] F. T. Pokorny, C. H. Ek, H. Kjellström, and D. Kragic, "Persistent homology for learning densities with bounded support," in *Advances in Neural Information Processing Systems 25*, pp. 1826–1834, 2012.

[2] P. Niyogi, S. Smale, and S. Weinberger, "A topological view of unsupervised learning from noisy data," *SIAM Journal of Computing*, vol. 40, no. 3, pp. 646–663, 2011.

[3] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009.

[4] M. P. Wand and M. C. Jones, *Kernel Smoothing*. Chapman and Hall/CRC, 1 ed., Dec. 1994.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[6] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[7] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2007.

[8] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *PAMI*, vol. 27, no. 5, pp. 827–832, 2005.

[9] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[10] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[11] T. Cacoullos, "Estimation of a multivariate density," *Annals of the Institute of Statistical Mathematics*, vol. 18, pp. 179–189, 1966.

[12] B. A. Turlach, "Bandwidth Selection in Kernel Density Estimation: A Review," in *CORE and Institut de Statistique*, pp. 23–493, 1993.

[13] L. Devroye, *Nonparametric Density Estimation: The L1 View*. Wiley, 1985.

[14] A. Agarwal and B. Triggs, "3d human pose from silhouettes by relevance vector regression," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 882–888, 2004.

[15] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A software package for computing persistent topological invariants." Software, 2011.