Lecture 4. Kernel Density Estimation

Juno Kim

Department of Mathematics & Statistics
Seoul National University

Manifold Learning, Spring 2022

Table of Contents

Existence

Lemma

Let ϕ be an isometry between Riemannian manifolds $(\mathcal{M}, g)$ and $(\mathcal{M}', g')$. Then ϕ preserves the Levi-Civita connection, i.e.

$$\nabla'_{d\phi(X)} d\phi(Y) = d\phi \nabla_X Y$$

In particular, ϕ preserves the Riemann curvature tensor.

Thus, curvature acts as a *local* obstruction for the existence of isometries. Due to this, many distance-based algorithms cannot learn intrinsically curved manifolds.

Exercise. Prove the lemma by showing that the equation defines a compatible and torsionfree connection on $\mathcal{M}'$.

Theorem (Moser)

Let $\mathcal{M}$, $\mathcal{M}'$ be diffeomorphic closed, connected, orientable smooth manifolds with volume forms τ, τ'. Suppose $\int_{\mathcal{M}} \tau = \int_{\mathcal{M}'} \tau'$. Then there exists a diffeomorphism $\phi : \mathcal{M} \to \mathcal{M}'$ so that $\tau = \phi^* \tau'$.

Thus, the only obstruction to the existence of volume-preserving maps is the *global* invariant – total volume.

Corollary

In the setting above, let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ taking values on $(\mathcal{M}, \tau)$ with density f, that is, $dP_*(X) = fd\tau$. Then the pushforward measure on $(\mathcal{M}', \tau')$ has density $f \circ \phi^{-1}$.

Proof (Moser's trick)

- Let ψ be any diffeomorphism $\mathcal{M} \to \mathcal{M}'$
- τ, $\psi^*\tau'$ represent the same cohomology class in $H^n(\mathcal{M}, \mathbb{R})$. Let $\psi^*\tau' = \tau + d\eta$ and $\tau_t = \tau + t \cdot d\eta$
- Since τ_t is a volume form, $\mathcal{X}(\mathcal{M}) \to \Omega^{n-1}(\mathcal{M}) : X \mapsto \iota_X \tau_t$ is an isomorphism, so $\exists\, X_t$ solving $\iota_{X_t}\tau_t + \eta = 0$
- Let $\dot{\phi}_t = X_t \circ \phi_t$ be the flow on $\mathcal{M}$ generated by X_t, then:

$$\frac{d}{dt}\phi_t^*\tau_t = \phi_t^*\left(\mathcal{L}_{X_t}\tau_t + \frac{d}{dt}\tau_t\right) = \phi_t^* d(\iota_{X_t}\tau_t + \eta) = 0$$

- $\tau = \phi_0^*\tau_0 = \phi_1^*\psi^*\tau'$. Set $\phi = \psi \circ \phi_1$. ■

Exercise. Show that any two symplectic manifolds are locally symplectomorphic (Darboux theorem). Thus there are no local invariants in SG.

Remark

Moser's theorem extends to noncompact manifolds where $\mathrm{vol}_\tau(\mathcal{M}) = \mathrm{vol}_{\tau'}(\mathcal{M}') \leq \infty$ and each end of $\mathcal{M}$ has finite τ-volume iff it has finite τ'-volume.

- The end condition is necessary: let

$$\mathcal{M} = S^1 \times \mathbb{R} = (S^1 \times \mathbb{R}_{\geq 0}) \cup_S (S^1 \times \mathbb{R}_{\leq 0}) = C_+ \cup_S C_-$$

- Find volume forms τ, τ' such that $\mathrm{vol}_\tau(C_+) = \mathrm{vol}_\tau(C_-) = \mathrm{vol}_{\tau'}(C_+) = \infty$ but $\mathrm{vol}_{\tau'}(C_-) < \infty$
- For any $\phi \in \mathit{Diff}(\mathcal{M})$, $\phi(S)$ is homotopic to S. The 2 components of $\mathcal{M} \setminus \phi(S)$ must have unbounded volume. $\Rightarrow\Leftarrow$

The theorem holds for manifolds with boundary, and for nonvanishing odd forms on nonorientable manifolds.

Nonuniqueness

For the isometry case, the following holds:

Theorem (Myers-Steenrod)

The isometry group of any Riemannian manifold is a finite-dimensional Lie group.

In contrast, the space of volume-preserving maps $SDiff(\mathcal{M})$ on manifolds of dimension ≥ 2 is always infinite-dimensional. This fact is of importance in e.g. fluid dynamics or gauge theories.

Later on, we will outline an optimization process for choosing "good" mappings in this space.

Exercise. Prove the assertion. Hint: use hyperspherical coordinates.

Table of Contents

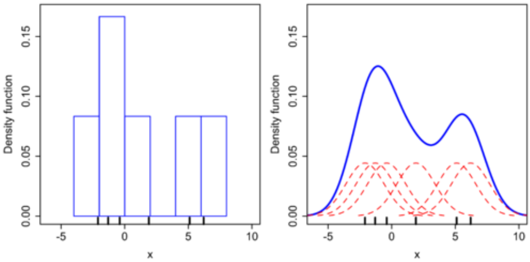

Figure: Histogram versus KDE with Gaussian kernel.

Kernel density estimation is a principal nonparametric method of estimating probability distributions. Local kernels around each data point are summed to yield the smoothed empirical density.

- A kernel is a bounded, integrable function $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\lim_{x\to\infty} xK(x) = 0$ and $\int_{\mathbb{R}} K(|x|)dx = 1$.
- Given $\mathbb{R}$-valued data $x_1, \cdots, x_n \overset{i.i.d.}{\sim} f$, the p.d.f. estimator is

$$\hat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{|x - x_i|}{h_n}\right)$$

where the bandwidth $h_n \to 0$ as $n \to \infty$.

Theorem (Parzen)

At all points of continuity of f, the following hold.
(1) $\hat{f}_n$ is pointwise asymptotically unbiased: $\lim_n \mathbb{E}\hat{f}_n(x) = f(x)$
(2) $\lim_n nh_n \cdot \mathrm{Var}(\hat{f}_n(x)) = f(x)\int_{\mathbb{R}} K^2(y)dy$

- Now assume $nh_n \to \infty$. By the MSE decomposition:

$$\mathbb{E}\left[\hat{f}_n(x) - f(x)\right]^2 = \mathrm{Var}\,\hat{f}_n(x) + (\mathrm{Bias}\,\hat{f}_n(x))^2 \to 0$$

- Thus $\hat{f}_n(x)$ is consistent. It is also asymptotically normal.
- The following results are classical:

Theorem (Parzen)

If f is uniformly continuous and $nh_n^2 \to 0$, then $\hat{f}_n$ is uniformly consistent. If f possesses a unique mode θ, the sample mode $\hat{\theta}_n := \arg\max \hat{f}_n$ is a consistent estimator of θ.

An estimator $\hat{\theta}_n$ of θ is *consistent* if $\hat{\theta}_n \xrightarrow{p} \theta$, i.e. $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \ \forall \epsilon > 0$.

- For r-dimensional data and normalized kernel $\int_{\mathbb{R}^r} K_r(\|x\|)d^r x = 1$, the estimator is:

$$\hat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n^r} K_r\left(\frac{\|x - x_i\|}{h_n}\right)$$

- Assume $h_n \to 0$ and $nh_n^r \to \infty$. That is, the # of data points contributing to the local density estimate $\to \infty$. Then:

$$MSE(\hat{f}_n) = O\left(h_n^4 + \frac{1}{nh_n^r}\right) \geq O(n^{-4/(r+4)})$$

- The 1st and 2nd terms are due to bias and variance, resp.
- The exceedingly slow convergence speed for high-dimensional feature spaces is called the curse of dimensionality.

On Submanifolds

- Now assume the data is drawn from a density f supported on an unknown submanifold $\mathcal{M}^t$ of $\mathbb{R}^r$, $t < r$
- KDE on $\mathbb{R}^r$ fails to converge to the correct density on $\mathcal{M}$ since it is not absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}^r$

Example. Let $\mathcal{M}$ be the unit t-cube and f be uniform on $\mathcal{M}$. Using the indicator kernel $K_r(x) \propto I(x \leq 1)$:

$$\hat{f}_n(x) \begin{cases} = \dfrac{V_t}{V_r} \dfrac{1}{h_n^{r-t}}(1 + o(1)) \to \infty \text{ if } x \in \mathcal{M} \\ \to 0 \text{ if } x \notin \mathcal{M} \end{cases}$$

where $V_r = \frac{\pi^{r/2}}{\Gamma(r/2+1)}$ is the volume of the unit r-ball.

- **Idea:** at small bandwidths, only local points contribute to $\hat{f}$ where the geometry is nearly flat, $d^{\mathcal{M}} \sim \|\cdot\|_{\mathbb{R}^r}$
- Instead apply t-dimensional KDE to $\{x_1, \cdots, x_n\} \subset \mathcal{M}$ *as if they were in* $\mathbb{R}^t$ to obtain density estimates at each x_j:

$$\hat{f}_n(x_j) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n^t} K_t \left(\frac{\|x_i - x_j\|_{\mathbb{R}^r}}{h_n} \right)$$

- Thus we obtain an estimator which does not depend on prior knowledge of $\mathcal{M}$ (except t) and has the usual KDE properties.
- In particular, $MSE \sim n^{-4/(t+4)}$ requiring much smaller samples for low *intrinsic* dimension.

Example

Figure: A sample of MNIST database.

- MNIST dataset: 60,000 square 28×28 pixel grayscale images of handwritten single digits
- The subset of 2's are essentially parametrized by the upper arch and lower loop
- extrinsic dimension 28^2, error $\sim n^{-0.005}$
- intrinsic dimension 2, error $\sim n^{-0.67}$

Table of Contents

This section is devoted to proving the basic properties of the *submanifold kernel density estimator (SKDE)* discussed previously.

Theorem (Ozakin)

Let $\mathcal{M}^t$ be a complete embedded Riemannian submanifold of $\mathbb{R}^r$ with injectivity radius $r_{inj} > 0$. Let the kernel $K_t : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be C^1, supported on $[0,1]$, and normalized: $\int_{\mathbb{R}^t} K_t(\|x\|)d^t x = 1$.

Suppose f is a probability density on $\mathcal{M}$ and C^2 on a neighborhood of $p \in \mathcal{M}$. Let $h_n \to 0$ and $nh_n^t \to \infty$. Then the SKDE, defined as

$$\hat{f}_n(p) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n^t} K_t\left(\frac{\|p - x_j\|_{\mathbb{R}^r}}{h_n}\right)$$

has mean squared error bounded by $O(h_n^4 + 1/nh_n^t)$. Thus, $\hat{f}_n(p)$ is asymptotically unbiased and consistent.

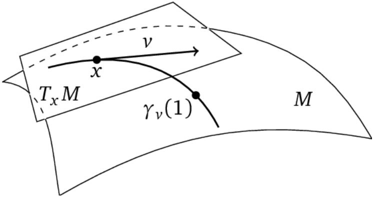

Figure: Geodesic normal coordinates.

- The *injectivity radius* at $p \in \mathcal{M}$ is defined as the largest radius $r = r_{inj}(p)$ for which $\exp_p : T_p\mathcal{M} \supset B(0, r) \to \mathcal{M}$ is a diffeomorphism onto its image
- The injectivity radius of $\mathcal{M}$ is $r_{inj} := \inf_{p \in \mathcal{M}} r_{inj}(p)$
- Define *geodesic normal coordinates* $y = (y^1, \cdots, y^r)$ around p by pushing forward orthonormal coordinates on $T_p\mathcal{M}$
- Then geodesics are of the form $y(t) = \gamma \cdot t$, thus $\Gamma^k_{ij}(p) = 0$

Lemma

In geodesic coordinates near p, the following identities hold.

- *(metric)* $g_{ij}(y) = \delta_{ij} - \frac{1}{3} R_{ikj\ell} y^k y^\ell + O(\|y\|^3)$
- *(volume element)* $\sqrt{\det(g_{ij})} = 1 - \frac{1}{6} \mathrm{Ric}_{k\ell}\, y^k y^\ell + O(\|y\|^3)$
- *For a 2-plane* $\Pi \subset T^p\mathcal{M}$ *and* $C_r = \partial B(0, r) \subset \Pi$,

  (sectional curvature) $K_p(\Pi) = \lim_{r \to 0} \frac{3}{\pi} \cdot \frac{2\pi r - \mathit{length}\,(\exp C_r)}{r^3}$
- *(scalar curvature)* $\frac{\mathrm{vol}_{\mathcal{M}}\, B(p, r)}{\mathrm{vol}_{\mathbb{R}^t}\, B(0, r)} = 1 - \frac{S(p)}{6(t+2)} r^2 + O(r^3)$

Exercise. Prove the Lemma.

The kernel behaves like a mollifier of the δ distribution:

Proposition

For any $\xi : \mathcal{M} \to \mathbb{R}$ C^2 near p and $0 < h \lesssim r_{inj}$,

$$\xi_h(p) := \frac{1}{h^t} \int_{\mathcal{M}} K_t\left(\frac{\|p-q\|_{\mathbb{R}^r}}{h}\right) \xi(q) \operatorname{vol}_{\mathcal{M}}(q)$$

satisfies $\xi_h(p) = \xi(p) + O(h^2)$.

- There exists $R_p(h)$, defined for $h \lesssim r_{inj}$, so that

$$\|p-q\|_{\mathbb{R}^r} < h \Rightarrow d^{\mathcal{M}}(p,q) < R_p(h)$$

  and $h \leq R_p(h) \leq h + O(h^3)$.
- If $\|y(q)\| > R_p(h)$, q does not contribute to the integral

$$
\begin{aligned}
\xi_h(p) - \xi(p) &= \frac{1}{h^t} \int_{\|y\| \le R_p(h)} K_t\left(\frac{\|p - y\|_{\mathbb{R}^r}}{h}\right) \xi(y) \sqrt{\det g(y)} d^t y \\
&\quad - \xi(0) \int_{\|z\| \le R_p(h)} K_t(\|z\|) d^t z \\
= \int_{\|z\| \le 1} & K_t\left(\frac{\|p - zh\|_{\mathbb{R}^r}}{h}\right) \xi(zh)(\sqrt{\det g(zh)} - 1) d^t z \\
+ \int_{\|z\| \le 1} & \xi(zh)\left(K_t\left(\frac{\|p - zh\|_{\mathbb{R}^r}}{h}\right) - K_t(\|z\|)\right) d^t z \\
+ \int_{\|z\| \le 1} & K_t(\|z\|)(\xi(zh) - \xi(0)) d^t z \\
+ \int_{1 \ge \|z\| \ge R_p(h)/h} & K_t\left(\frac{\|p - zh\|_{\mathbb{R}^r}}{h}\right) \xi(zh) \sqrt{\det g(zh)} d^t z
\end{aligned}
$$

Thus, $|\xi_h(p) - \xi(p)|$

$$\leq \|K_t\|_\infty \cdot \sup_{\|z\|\leq 1} |\xi(zh)| \cdot \sup_{\|z\|\leq 1} \left|\sqrt{\det g(zh)} - 1\right| \cdot V_t$$

$$+ \sup_{\|z\|\leq 1} |\xi(zh)| \cdot \sup_{\|z\|\leq 1} \left| K_t\left(\frac{\|p - zh\|_{\mathbb{R}^r}}{h}\right) - K_t(\|z\|)\right| \cdot V_t$$

$$+ \left| \int_{\|z\|\leq 1} K_t(\|z\|)(\xi(zh) - \xi(0))d^t z \right|$$

$$+ \|K_t\|_\infty \cdot \sup_{1\geq\|z\|\geq R_p(h)/h} \sqrt{\det g(zh)}|\xi(zh)| \cdot \int_{1\geq\|z\|\geq R_p(h)/h} d^t z$$

The black terms are bounded. We show the red terms are $O(h^2)$:

1. By the Lemma, $|\sqrt{\det g(zh)} - 1|$ is uniformly bounded by $O(h^2)$ and Ricci curvature
2. $\|p - zh\|_{\mathbb{R}^r} - \|zh\| = O(h^3)$ since $\|zh\|$ is the geodesic distance, and K_t is uniformly continuous due to C^1
3. By Taylor expansion, $\xi(zh) - \xi(0) = h \sum z^j \partial_j \xi|_0 + O(h^2)$ and 1st order terms vanish since $\int_{\|z\| \leq 1} zK_t(\|z\|)d^t z = 0$
4. The spherical shell $1 \leq \|z\| \leq 1 + \epsilon$ has volume $O(t\epsilon)$, so the term is $O(R_p(h)/h - 1) = O(h^2)$

Thus the Proposition is proved.

We conclude:

$$\operatorname{Bias} \hat{f}_n(p) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{x_i} \left[\frac{1}{h_n^t} K_t \left(\frac{\|p - x_i\|}{h_n} \right) \right] - f(p) = O(h_n^2)$$

and $\operatorname{Var} \hat{f}_n(p) = \frac{1}{n} \operatorname{Var}_{x_i} \left[\frac{1}{h_n^t} K_t \left(\frac{\|p - x_i\|}{h_n} \right) \right]$

$$= \frac{1}{n} \mathbb{E}_{x_i} \left[\frac{1}{h_n^{2t}} K_t^2 \left(\frac{\|p - x_i\|}{h_n} \right) \right] - \frac{1}{n} \left[\mathbb{E}_{x_i} \frac{1}{h_n^t} K_t \left(\frac{\|p - x_i\|}{h_n} \right) \right]^2$$

$$= \frac{1}{nh_n^t} \int K_t^2(\|z\|) d^t z \cdot (f(p) + O(h_n^2)) - O(n^{-1}) = O\left(\frac{1}{nh_n^t} \right)$$

where we have applied the Proposition to normalized K^2. ■

Table of Contents

- The SKDE is defined on all points $p \in \mathcal{M}$. However if $\mathcal{M}$ is unknown, we can only calculate $\hat{f}_n(x_i)$ of each data point.
- Assumption: $\mathcal{M}^t$ is diffeomorphic to a region in $\mathbb{R}^t$
- Goal: find a density-preserving map $\phi : \mathcal{M} \to \mathbb{R}^t$ which (1) best preserves the SKDE estimates and (2) is optimal in some sense in $SDiff(\mathcal{M})$.
- In particular, find $\{y_1, \cdots, y_n\} \subset \mathbb{R}^t$ such that ordinary KDE:

$$\hat{f}_n(y_i) := \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n^t} K_t \left(\frac{\|y_i - y_j\|}{h_n} \right) \simeq \hat{f}_n(x_i)$$

  with scale constraints on y_i's.
- In general, this approach is nonconvex.

SDP

Semidefinite programming (SDP) generalizes linear programming (maximizing a linear objective over a polytope) to multidimensional variables satisfying semidefiniteness constraints.

Let $Sym(n)$ be the space of $n \times n$ real symmetric matrices with the inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr} \mathbf{A}^T \mathbf{B} = \sum_{i,j} A_{ij} B_{ij}$.

Form 1	**Form 2**
find vectors $x_1, \cdots, x_n$	find Gram matrix $\mathbf{G}$
$\max\limits_{x_i \in \mathbb{R}^n} \sum c_{ij}(x_i^T x_j)$	$\max\limits_{\mathbf{G} \in Sym(n)} \langle \mathbf{C}, \mathbf{G} \rangle$
subject to	subject to
$\sum a_{ij}^{(k)}(x_i^T x_j) \leq b^{(k)}$	$\langle \mathbf{A}^{(k)}, \mathbf{G} \rangle \leq b^{(k)}$ and $\mathbf{G} \succeq 0$

Parameters

- In practice, we implement *variable bandwidth* to compensate for inhomogeneous data: $h_n(x_i)$ depends on x_i
- Here we set $h_n(x_i) :=$ the distance of the Kth nearest data point, so only the K nearest x_j, $j \in N_i$ contribute
- K may be chosen using e.g. leave-one-out cross-validation with log-likelihood score:

$$K = \operatorname{argmin} \sum_i \log \hat{f}_{n-1}^{K,(-i)}(x_i)$$

  where $\hat{f}_{n-1}^{K,(-i)}$ is the SKDE for deleted x_i
- We use the asymptotically optimal Epanechnikov kernel:

$$E_t(\|x_i - x_j\|) = e_t \cdot (1 - \|x_i - x_j\|^2)_+$$

The objective function is $\operatorname{tr}\mathbf{G}$ as in *maximum variance unfolding* methods (e.g. LLE). The conditions are:

$$d_{ij}^2 := \mathbf{G}_{ii} - 2\mathbf{G}_{ij} + \mathbf{G}_{jj} \leq h_n(x_i)^2 \quad \text{for} \quad j \in N_i$$

$$\hat{f}_n(x_i) = \frac{e_t}{h_n(x_i)^t} \sum_{j \in N_i} \left(1 - \frac{d_{ij}^2}{h_n(x_i)^2}\right) \quad \forall i$$

$$\mathbf{G} \succeq 0 \quad \text{and} \quad \sum_{i,j=1}^{n} \mathbf{G}_{ij} = 0 \quad \text{(centering)}$$

Using the original $h_n(x_i)$ in RHS of line 2 allows the $e_t/h_n(x_i)^t$ term to cancel out, so t need not be predetermined. Rather, t may be determined along with y_i by eigenanalysis of $\mathbf{G}$.

This is justified by the 1st condition (neighborhood stability).

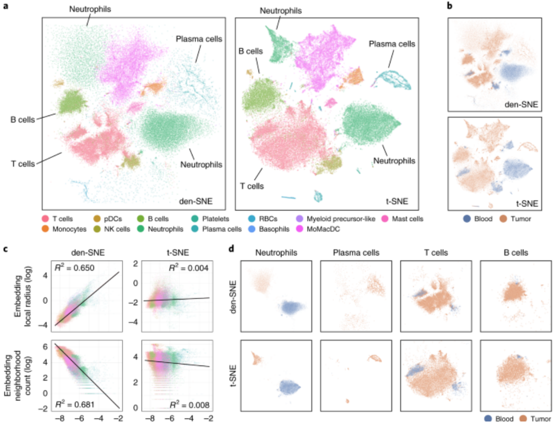

Figure: Assessing RNA transcriptional diversity through density-preserving data visualization (Nature).