Kernel Density Estimation via Diffusion

April 30, 2020

In this note we look at the paper [1] titled 'Kernel Density Estimation via Diffusion' which highlights a nice connection between the class of kernel density estimators and diffusions/PDEs.

1 Diffusions

Suppose that the stochastic process $(X_t)_{t\geq 0} \subseteq \mathbb{R}$ satisfies the SDE

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t,$$

also called a diffusion, where B is a Brownian Motion. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a sufficiently regular function. We define the semigroup $(P_t)_{t\geq 0}$ associated with the process X by its action on functions:

$$(P_t\psi)(x) = \mathbb{E}\left[\psi(X_t)|\, X_0 = x\right].$$

The generator of the Markov process X is

$$\mathcal{L} := \frac{\mathrm{d}}{\mathrm{d}t}P_t\Big|_{t=0}$$

or in other words

$$\mathcal{L}\psi := \lim_{t\to 0}\frac{P_t\psi - \psi}{t} \qquad \forall\,\psi$$

where the limit is with respect to the $\|\cdot\|_\infty$-norm. On sufficiently smooth functions the generator $\mathcal{L}$ is a second order elliptic differential operator of the form

$$(\mathcal{L}\psi)(x) = a(x)\psi'(x) + \frac{\sigma(x)^2}{2}\psi''(x).$$

Crucially $(P_t)_{t\geq 0}$ and $\mathcal{L}$ satisfy Kolmogorov's Forward (a.k.a. Fokker-Planck) and Backwards equations. These take the form

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t = P_t\mathcal{L} \qquad \text{(KFE)}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}P_t = \mathcal{L}P_t \qquad \text{(KBE)}.$$

These equations can be used to derive two PDEs that characterize the transition kernel κ of the process X. Let $\kappa(x, y; t)$ denote the density of the particle X_t when started from point x. We have then

$$\int \frac{\partial}{\partial t}\kappa(x,y;t)\psi(y)\,\mathrm{d}y = \frac{\partial}{\partial t}P_t\psi(x) \overset{\text{(KFE)}}{=} \underbrace{(P_t\mathcal{L}\psi)(x)}_{(*)} \overset{\text{(KBE)}}{=} \underbrace{(\mathcal{L}P_t\psi)(x)}_{(**)}.$$

Looking at the two terms individually we find that

$$(*) = \int \kappa(x,y;t)(\mathcal{L}\psi)(y)\,\mathrm{d}y = \int \mathcal{L}_y^*\kappa(x,y;t)\psi(y)\,\mathrm{d}y$$

and

$$(**) = \mathcal{L}_x \int \kappa(x,y;t)\psi(y)\,\mathrm{d}y = \int \mathcal{L}_x \kappa(x,y;t)\psi(y)\,\mathrm{d}y.$$

Since the above holds for all $\psi \in \mathcal{C}_c^\infty$, we get that the transition kernel of X satisfies the following PDEs:

$$\frac{\partial}{\partial t}\kappa(x,y;t) = \mathcal{L}_x \kappa(x,y;t) = \mathcal{L}_y^* \kappa(x,y;t) \qquad \forall\, x,y \in \mathbb{R},\, t \geq 0$$

with the initial condition $\kappa(x,y;0) = \delta(x-y)$.

2 Kernel Density Estimators (KDEs)

Suppose that we have an i.i.d. sample $X_1, ..., X_n \in \mathbb{R}$ froma distribution with density f. A natural objective is to construct an estimator $\hat{f}$ for the density based on the sample. A popular method to do so are Kernel Density Estimators (KDEs). Given a bandwidth $h > 0$ and a kernel $K : \mathbb{R} \to \mathbb{R}_+$ they are usually defined as

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

The kernel K is usually taken to be a symmetric probability density so that $\hat{f}_h$ integrates to 1. We see that h controls the degree of smoothing where large h corresponding to flatter/smoother estimates while small h correspond to high-probability spikes at the sample points $X_1, ..., X_n$. Kernel Density Estimators have been studied a lot over the past 40 years with much effort going into deriving rules to select the kernel K and the bandwidth $h = h(n)$. A popular choice of kernel is the Gaussian/Heat kernel

$$\phi(x,y;t) := \frac{1}{\sqrt{2\pi t}} e^{-\frac{|x-y|^2}{2t}}.$$

The corresponding Gaussian KDE is

$$\hat{f}(x,t) := \frac{1}{n}\sum_{i=1}^n \phi(X_i, x; t). \tag{2.1}$$

The above is just the density of a Gaussian mixture with means at the sample points $X_1, ..., X_n$ and variances t. Another way to look at it is that it describes the density of a Brownian Motion after time t when started from a point randomly chosen from the set $\{X_1, ..., X_n\}$. To make this argument rigorous notice that the Gaussian KDE (2.1) satisfies the heat equation

$$\frac{\partial}{\partial t}\hat{f} = \frac{1}{2}\Delta \hat{f} \tag{2.2}$$

with initial condition $\hat{f}(x;0) := (1/n)\sum_{i=1}^n \delta_{X_i}$. This follows from the linearity of the heat equation and the fact that the Gaussian/Heat kernel ϕ is a solution. Thus the Gaussian KDE can be characterized as the solution to (2.2). The connection to diffusions is obvious once we note that the heat equation is just the KFE/KBE for Brownian Motion since the generator of Brownian Motion is $\mathcal{L} = \frac{1}{2}\Delta$ where Δ is self-adjoint.

The above PDE perspective to the Gaussian KDE has multiple advantages. Suppose for example that we know that the true density f is supported on the set $[0,1]$. There is no simple modification to the usual Kernel Density Estimator that incorporates this. In contrast, we can modify (2.2) by adding the Neumann boundary conditions

$$\frac{\partial}{\partial x}\hat{f}\Big|_{x=1} = \frac{\partial}{\partial x}\hat{f}\Big|_{x=0} = 0.$$

This ensures that $\int \hat{f}\,\mathrm{d}x = 1$ for all $t \geq 0$. Indeed, we have

$$\begin{aligned}\frac{\partial}{\partial t}\int_0^1 \hat{f}(x;t)\,\mathrm{d}x &= \int_0^1 \frac{\partial}{\partial t}\hat{f}(x;t)\,\mathrm{d}x \\ &= \frac{1}{2}\int_0^1 \Delta\hat{f}(x;t)\,\mathrm{d}x \\ &= \frac{1}{2}\left[\left(\frac{\partial}{\partial x}\hat{f}\right)(1;t) - \left(\frac{\partial}{\partial x}\hat{f}\right)(0;t)\right] \\ &= 0\end{aligned}$$

so that $1 = \int \hat{f}(x;0)\,\mathrm{d}x = \int \hat{f}(x;t)$ for all $t \geq 0$.

3 A new estimator

3.1 Description

Motivated by our observations in the previous section there is a natural generalization we can make. We saw that the Gaussian KDE is just

$$\hat{f}(x;t) = \frac{1}{n}\sum_{i=1}^{n}\kappa(X_i, x; t)$$

where the transition kernel κ solves the KBE and KFE of Brownian Motion

$$\frac{\partial}{\partial t}\kappa(x,y;t) = \frac{1}{2}\Delta_x\kappa(x,y;t) = \frac{1}{2}\Delta_y^*\kappa(x,y;t) \tag{3.1}$$

with initial condition $\kappa(x,y;0) = \delta(x-y)$. Now, we can replace the generator of Brownian Motion $\mathcal{L} = \frac{1}{2}\Delta$ in equation (3.1) with the generator of any other diffusion. Let us define

$$\mathcal{L} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left(a(x)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\cdot}{p(x)}\right)\right) \tag{3.2}$$

for functions $a, p \geq 0$. Onc can check that the adjoint operator is

$$\mathcal{L}^* = \frac{1}{2p(y)}\frac{\mathrm{d}}{\mathrm{d}y}\left(a(y)\frac{\partial}{\partial y}(\cdot)\right).$$

The above form is not fully general, it is chosen so that the corresponding transition kernel κ satisfies detailed balance with respect to p i.e. X's stationary distribution is proportional to p.

3.2 Properties

One appeal of the new estimator is that it unifies a wide range of methods developed for kernel/bandwidth selection for KDEs. By varying the choice of the functions a and p in (3.2) we can recover previously studied estimators. For example:

- If we choose $a = p = 1$ we get the Gaussian KDE.
- If we choose $a = 1$ and $p = f_p$ for some pilot density estimate f_p, asymptotically we get a Gaussian KDE with adaptive bandwidth $\sqrt{t/f_p}$ which is precisely the adaptive bandwidth modification of Abramson [2] applied to the Gaussian KDE.

In Density estimation a common performance metric is the Mean Integrated Squared Error (MISE)

$$\mathrm{MISE}\{\hat{f}\}(t) = \int (\mathbb{E}_f\hat{f}(x;t) - f(x))^2\,\mathrm{d}x + \int \mathrm{Var}_f\hat{f}(x;t)\,\mathrm{d}x.$$

The authors analyze the Asymptotic ($n \to \infty$) MISE (AMISE) of the estimator and derive the following results:

Theorem 3.1 — *Asymptotically as $n \to \infty$ it holds that*

$$AMISE\{\hat{f}\}(t) = \underbrace{\frac{1}{4} t^2 \left\| a(f/p)' \right\|^2}_{bias} + \underbrace{\frac{\mathbb{E}_f \sigma^{-1}(X)}{2n\sqrt{\pi t}}}_{variance}.$$

Hence, the asymptotically optimal bandwidth is

$$t^* = \left(\frac{\mathbb{E}_f \sigma^{-1}(X)}{2n\sqrt{\pi} \left\| \mathcal{L} f \right\|^2} \right)^{2/5}$$

resulting in

$$AMISE \sim n^{-4/5}.$$

Remark 3.1. 1. The rate $n^{-4/5}$ is the same as for the Gaussian KDE, but the constant term can be made much smaller by a good pilot estimate p. The practical preformance of the new estimator is demonstrably better.

2. An important property of this new estimator is that in the case of known bounded support it doesn't suffer from boundary bias (if Neumann boundary conditions are enforced) as it is provably consistent at the boundary. This is not the case for regular KDEs.

Proof idea. Recall that the kernel κ solves KFE and KBE. In general these PDEs cannot be solved analytically, so the proof of the asymptotic results relies on calculating the asymptotic behaviour of the solution in the small $t > 0$ regime. To do so the authors use a well-known method (WKBJ) from Pertirbation Theory [3]. The main idea of the technique is to assume a functional series expansion form for κ, namely

$$\kappa(x, y; t) \sim e^{-\frac{s^2(x,y)}{2t}} \sum_{m=0}^{\infty} t^{m-1/2} C_m(x, y) \qquad \text{for small } t$$

for unknown functions s and C_m. This expression is then plugged into KFE and KBE and then the coefficients of the powers of t are matched. This results in PDEs for the functions s and C_m which can be solved explicitly. The authors go to some length to justify the validity of such a small t expansion by imposing conditions on the functions a, p. Once established, getting the asymptotic form of the AMISE is a simple calculation. □

Through a simulation study the authors demonstrate that the new estimator outperforms other state-of-the-art KDE algorithms. The pipeline for constructing the estimator is the following ([1, Algorithm 2])

1. Given the data $X_1, ..., X_n$ construct a pilot density estimate f_p using a Gaussian KDE with bandwidth chosen appropriately (authors suggest the Improved Sheather–Jones method).

2. Set $p = f_p$ and $a = p^\alpha$ for some $\alpha \in [0, 1]$, where α interpolates between Abramson's estimator ($\alpha = 0$) and the 'data-sharpening' [4] method ($\alpha = 1$).

3. Solve KFE and KBE numerically to get $\hat{f}$ with a, p as above and bandwidth chosen appropriately (authors propose a method to do so).

References

[1] Zdravko I Botev, Joseph F Grotowski, Dirk P Kroese, et al. Kernel density estimation via diffusion. *The annals of Statistics*, 38(5):2916–2957, 2010.

[2] Ian S Abramson. On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223, 1982.

[3] Yakar Kannai. Off diagonal short time asymptotics for fundamental solution of diffusion equation. *Communications in Partial Differential Equations*, 2(8):781–830, 1977.

[4] M Samiuddin and GM El-Sayyad. On nonparametric kernel density estimates. *Biometrika*, 77(4):865–874, 1990.