Kernel Density Estimation via Diffusion

June 22, 2020

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

1 Diffusions

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

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

where B is a Brownian Motion. Such processes are often called (Itô) diffusions.

1.1 Semigroups and Generators

For a bounded function $\psi : \mathbb{R} \to \mathbb{R}$ we define the semigroup $(P_t)_{t\geq 0}$ associated with the process X by its action on functions:

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

$(P_t)_{t\geq 0}$ is called a semigroup because $P_{t+s} = P_tP_s$ for all $t, s \geq 0$. The semigroup property is equivalent to the process X having the Markov property. The generator of the Markov process X is

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

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 functions $\psi \in \mathcal{C}_c^2$ the generator $\mathcal{L}$ is a second order elliptic differential operator [2, Theorem 8.7] of the form

$$(\mathcal{L}\psi)(x) = \mu(x)\psi'(x) + \frac{\sigma(x)^2}{2}\psi''(x). \tag{1.1}$$

1.2 Kolmogorov's equations

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

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

The key to the proof is to use the semigroup property:

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t\psi = \lim_{h\downarrow 0}\frac{P_{t+h}\psi - P_t\psi}{h} \overset{\text{(KFE)}}{=} P_t\lim_{h\downarrow 0}\frac{P_h\psi - \psi}{h} \overset{\text{(KBE)}}{=} \lim_{h\downarrow 0}\frac{P_h(P_t\psi) - (P_t\psi)}{h}.$$

These equations can be used to derive two PDEs that characterize the transition kernel κ of the process X. Let $\kappa(x, \mathrm{d}y; t)$ 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 $\psi \in \mathcal{C}_c^2$, 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 $\kappa(x, y; 0) = \delta(x - y)$.

1.3 Stationary distributions

Suppose that we start the diffusion from the initial distribution $\rho(x)\,\mathrm{d}x$. Then the density after time t is given by

$$\rho_t(x) = \int \kappa(y, x; t)\rho(y)\,\mathrm{d}x.$$

Let $\psi \in \mathcal{C}_c^2$. We say that ρ is stationary if

$$\int \psi(x)\rho(x)\,\mathrm{d}x = \int \psi(x)\rho_t(x)\,\mathrm{d}x$$

for all ψ, i.e. if ρ_t is independent of t. Such measures are also called equilibrium measures because under suitable assumptions $\nu_t \to \rho$ as $t \to \infty$ for arbitrary starting measures ν (see e.g. [3, Theorem 2.18]). Differentiating both sides with respect to t we get

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \int \psi(x) \int \kappa(y, x; t)\rho(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \int (P_t\psi)(y)\rho(y)\,\mathrm{d}y \\
&= \int (\mathcal{L}\psi)(y)\rho(y)\,\mathrm{d}y \\
&= \int \psi(y)(\mathcal{L}^*\rho)(y)\,\mathrm{d}y.
\end{aligned}$$

Thus, the density ρ has to solve the equation $\mathcal{L}^*\rho = 0$. For an example, consider the Ornstein–Uhlenbeck process which solves the SDE

$$\mathrm{d}X_t = bX_t\,\mathrm{d}t + \sigma\,\mathrm{d}B_t.$$

By (1.1) we get

$$(\mathcal{L}\psi)(x) = bx\psi'(x) + \frac{\sigma^2}{2}\psi''(x) \qquad \text{and} \qquad (\mathcal{L}^*\psi)(y) = -b\frac{\mathrm{d}}{\mathrm{d}y}(y\psi(y)) + \frac{\sigma^2}{2}\psi''(y).$$

Solving the equation $\mathcal{L}^*\rho$ we obtain

$$0 = -b\frac{\mathrm{d}}{\mathrm{d}y}(y\rho(y)) + \frac{\sigma^2}{2}\rho''(y)$$

$$\implies \quad \text{constant} = -by\rho(y) + \frac{\sigma^2}{2}\rho'(y) \tag{1.2}$$

In the above equation consider taking $|y| \to \infty$. Since ρ is a probability density it decays to zero so (1.2) implies that ρ' converges to a constant. To have ρ nonnegative and normalizable this constant needs to be zero. We obtain

$$\rho(y) \propto \exp\left(\frac{by^2}{\sigma^2}\right).$$

This shows that the Ornstein–Uhlenbeck process has a mean-zero Gaussian stationary distribution with variance $\frac{\sigma^2}{2|b|}$, provided that $b < 0$.

2 Kernel Density Estimators

Suppose that we have an i.i.d. sample $X_1, ..., X_n \in \mathbb{R}$ from a 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 $h = h(n)$.

2.1 Gaussian KDE

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 see it is as the convolution of the empirical measure of the X_i and the density of a centered Gaussian with variance t. Yet another way to look at it is that it equals 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 connection 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 because the heat kernel is the fundamental solution of the heat equation. 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.

2.2 The Diffusion Estimator

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

$$\frac{\partial}{\partial t}\kappa(x,y;t) = \frac{1}{2}\Delta_x \kappa(x,y;t) = \frac{1}{2}\Delta_y^* \kappa(x,y;t) \tag{2.3}$$

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 (2.3) with the generator of any other diffusion. Let us define

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

for functions $a, p > 0$. One can check that the adjoint operator is

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

Note that the above form the most general for which p is a stationary distribution.

2.3 Properties of the Diffusion Estimator

The authors highlight three issues with regular Kernel Density Estimators:

1. Popular bandwidth selection algorithms (plug-in method) rely on a preliminary normal model of the data, thereby diminishing the 'nonparametric' quality of KDEs.
2. Lack of local adaptivity.
3. Boundary bias.

Sophisticated methods have been introduced to tackle each of these issues, but these efforts are somewhat disjointed and unsatisfying (for example many estimators don't integrate to 1). The main appeal of the Diffusion Estimator is that it unifies a wide range of these methods and in particular tackles all three points above. For example, suppose that we know that the true density f is supported on the set $[0,1]$. We can simply modify (2.2) by adding the Neumann boundary conditions

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

This ensures that $\int_0^1 \hat{f}\,\mathrm{d}x = 1$ 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 $t \geq 0$.
By varying the choice of the functions a and p in (2.2) 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 [4] applied to the Gaussian KDE.

Through a simulation study the authors demonstrate that the Diffusion Estimator along with their proposed bandwidth selection algorithm (for 1 and 2-dimensional data) outperforms other state-of-the-art KDE algorithms. The pipeline for constructing the estimator is as follows ([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' [5] 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).

In the Kernel Density Estimation literature the Mean Integrated Squared Error (MISE) is a common performance metric

$$\begin{aligned}
\mathrm{MISE}\{\hat{f}\}(t) &:= \int \mathbb{E}_f(\hat{f} - f)^2\,\mathrm{d}x \\
&= \int \mathbb{E}_f(\hat{f} - \mathbb{E}_f\hat{f} + \mathbb{E}_f\hat{f} - f)^2\,\mathrm{d}f \\
&= \int \underbrace{(\mathbb{E}_f\hat{f} - f)^2}_{\text{bias}^2}\,\mathrm{d}x + \int \underbrace{\mathrm{Var}_f\hat{f}}_{\text{variance}}\,\mathrm{d}x.
\end{aligned}$$

Theorem 2.1 — *Let the support of f be $\mathbb{R}$ and let $t = t(n)$ be such that $t_n \to 0$ and $n\sqrt{t_n} \to \infty$ as $n \to \infty$. Then*

(i) *bias* $= \mathbb{E}_f\hat{f}(x;t) - f(x) = t(\mathcal{L}^*f)(x) + \mathcal{O}(t^2)$ *as* $n \to \infty$

(ii) *variance* $= \mathrm{Var}_f\,\hat{f}(x;t) \sim \frac{f(x)}{2N\sqrt{\pi t}}\sqrt{\frac{p(x)}{a(x)}}$ *as* $n \to \infty$.

Proof. For the bias we have

$$\begin{aligned}
\mathbb{E}_f\left[\hat{f}(x;t)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f\kappa(X_i, x; t) \\
&= \int_{\mathbb{R}}\kappa(y, x; t)f(y)\,\mathrm{d}y.
\end{aligned}$$

Differentiating with respect to time the KBE gives us

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}_f\left[\hat{f}(x;t)\right] &= \int_{\mathbb{R}}\mathcal{L}_y\kappa(y, x; t)f(y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}}\kappa(y, x; t)\mathcal{L}_y^*f(y)\,\mathrm{d}t
\end{aligned}$$

where the boundary terms vanish due to technical assumptions on a, p. As $\mathbb{E}_f \hat{f}$ is a convolution it is sufficiently differentiable, and a second order Taylor expansion yields

$$\begin{aligned}\mathbb{E}_f\left[\hat{f}(x;t)\right] &= \mathbb{E}_f \hat{f}(x,0) + t\frac{\partial}{\partial t}\mathbb{E}_f \hat{f}(x;t)\Big|_{t=0} + \mathcal{O}(t^2) \\ &= f(x) + t(\mathcal{L}^* f)(x) + \mathcal{O}(t^2)\end{aligned}$$

where we used the initial condition $\kappa(x,y;0) = \delta(x-y)$. The expression for the variance also follows by a direct calculation using the small t asymptotic found in Lemma 2.2. □

Lemma 2.2 ([1, Lemma 1]) —

$$\kappa(x,y;t) \sim \frac{p(x)}{\sqrt{2\pi t}\left[a(x)a(y)p(x)p(y)\right]^{1/4}} \times \exp\left(-\frac{1}{2t}\left[\int_x^y \sqrt{\frac{p(s)}{a(s)}}\,\mathrm{d}s\right]^2\right) \qquad \text{as } t \downarrow 0.$$

Proof outline. 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 the WKBJ method from Perturbation Theory [6]. The idea of the technique is to assume a functional series expansion form for κ, namely

$$\kappa(x,y;t) \sim \tilde{\kappa}(x,y;t) := e^{-\frac{s^2(x,y)}{2t}} \sum_{m=0}^{\infty} t^{m-1/2} C_m(x,y) \qquad \text{for small } t \tag{2.4}$$

for unknown functions s and C_m. This expression is then plugged into KFE and KBE upon which 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. □

Proof Sketch. Assume the expansion (2.4) for κ and recall the KFE

$$\frac{\partial}{\partial t}\kappa(x,y;t) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}y}\left(a(y)\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\kappa(x,y;t)}{p(y)}\right)\right)$$

for all $x, y \in \mathbb{R}$ and $t \geq 0$ with $\kappa(x,y;0) = \delta(x-y)$. Let us match the lowest power of t, namely $t^{-5/2}$. After some inspection (and cancelling the exponential term) we get

$$\frac{s^2(x,y)}{2} C_0(x,y) = \frac{a(y)}{2p(y)} s^2(x,y)\partial_y s(x,y)^2 C_0(x,y).$$

Rearranging again we arrive at

$$a(y)\left(\frac{\mathrm{d}}{\mathrm{d}y}s(x,y)\right)^2 - p(y) = 0.$$

We can freely impose $s(x,x) = 0$ as it is necessary to satisfy $\kappa(x,y;0) = \delta(x-y)$. We obtain

$$s(x,y) = \int_x^y \sqrt{\frac{p(s)}{a(s)}}\,\mathrm{d}s.$$

To further simplify the calculation we reuquire that κ be reversible with respect to p i.e. it satisfy detailed balance

$$p(x)\tilde{\kappa}(x,y;t) = p(y)\tilde{\kappa}(y,x;t) \qquad \text{(DB)}.$$

Under this simplification $\tilde{\kappa}$ automatically satisfies the KBE. Indeed,

$$\begin{aligned}
\mathcal{L}_x\tilde{\kappa}(x,y;t) &\overset{(\mathrm{DB})}{=} \mathcal{L}_x\left\{\frac{p(y)}{p(x)}\tilde{\kappa}(y,x;t)\right\}\\
&= \frac{p(y)}{p(x)}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left(a(x)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\tilde{\kappa}(y,x;t)}{p(x)}\right)\right)\\
&= \frac{p(y)}{p(x)}\mathcal{L}_x^*\tilde{\kappa}(y,x;t)\\
&\overset{(\mathrm{KFE})}{=} \frac{p(y)}{p(x)}\frac{\partial}{\partial t}\tilde{\kappa}(y,x;t)\\
&\overset{(\mathrm{DB})}{=} \frac{\partial}{\partial t}\tilde{\kappa}(x,y;t).
\end{aligned}$$

Armed with (2.3) and (2.3) we can continue and match coefficients of $t^{-3/2}$. We omit the rest of the details, and just state the result of this:

$$C_0(x,y) = \frac{p(x)}{\sqrt{2\pi}\left[a(x)a(y)p(x)p(y)\right]^{1/4}}.$$

□

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] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.

[3] Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

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

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

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