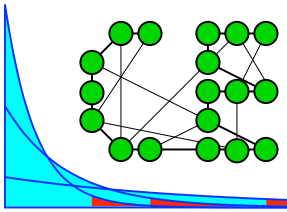
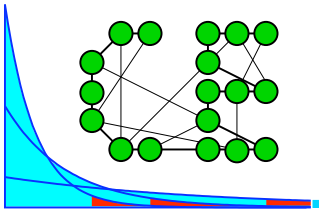


Density Estimation by Diffusion

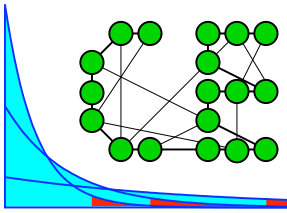
Dirk Kroese*, Zdravko Botev, Joe Grotowski

Department of Mathematics
The University of Queensland



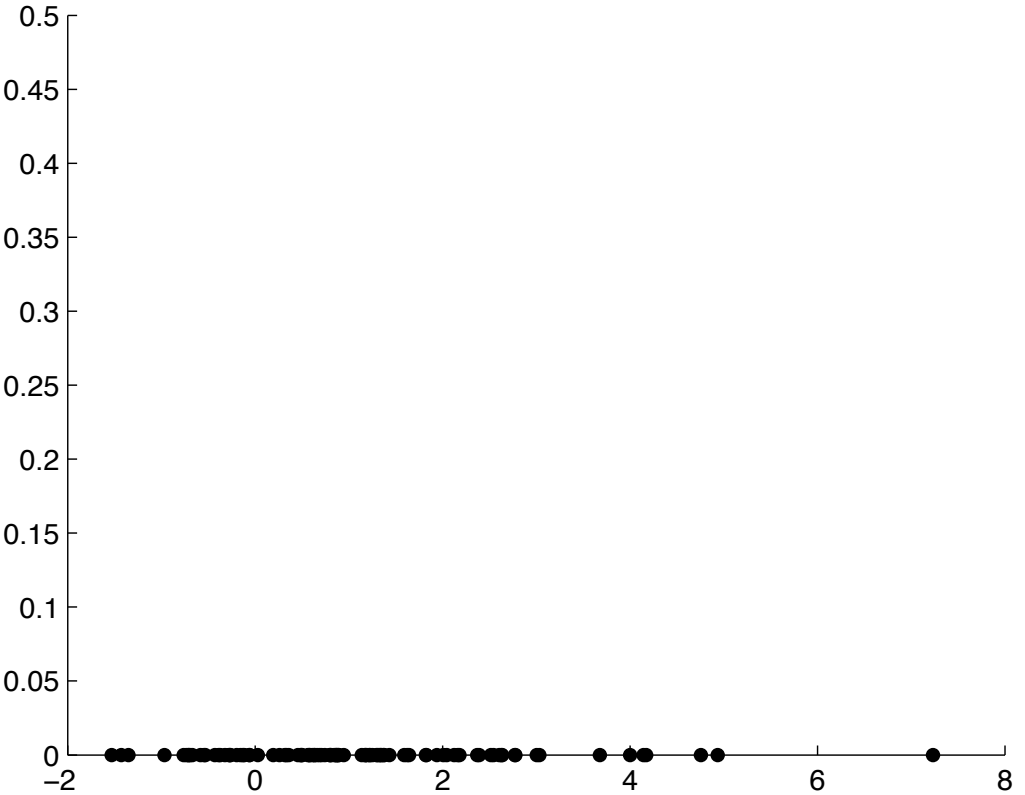
Outline

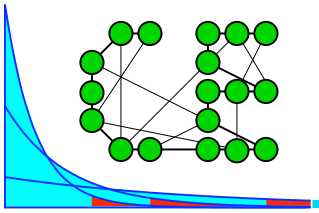
- Background on Kernel Density Estimation
- The Wiener Process and the Heat Equation
- Main Idea: Formulating Gaussian KDEs via the Heat Equation
- Extensions
  - Using General Diffusion Processes
  - Higher Dimensions



Density Estimation: Motivation

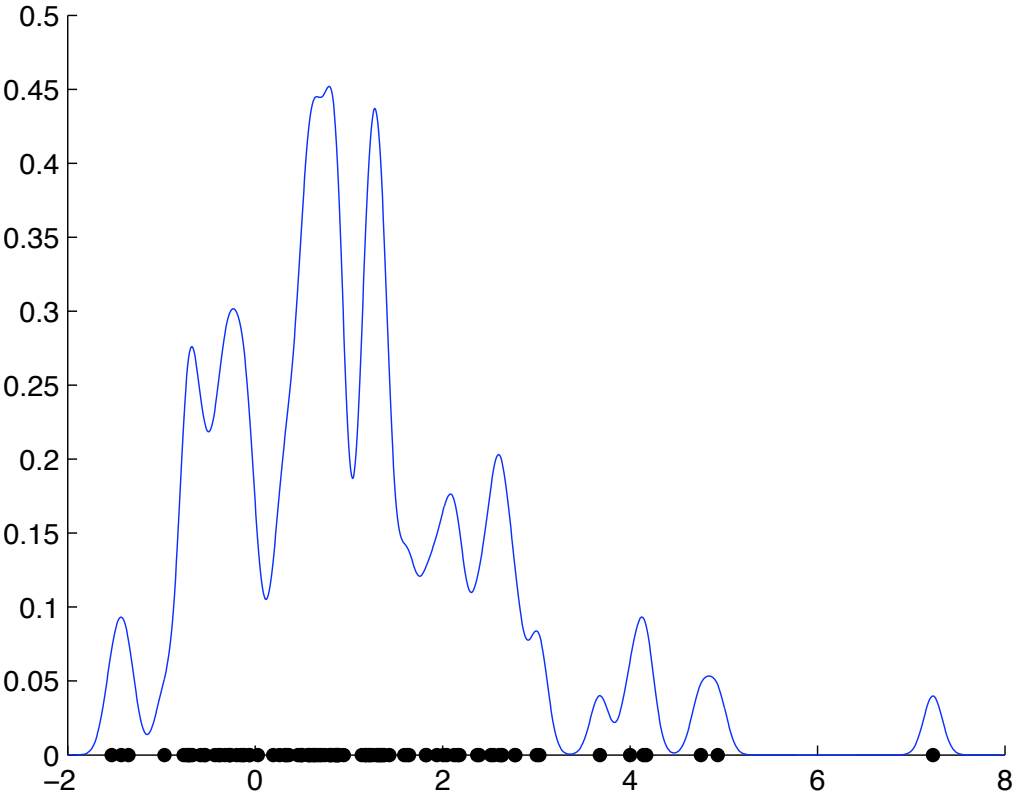
- Let $x_1, \ldots, x_N$ be independent draws from an unknown probability density function (pdf) f.
- How can we best estimate f?

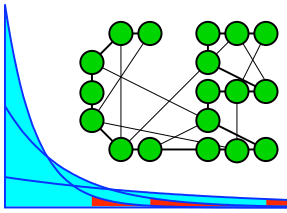




Density Estimation: Motivation

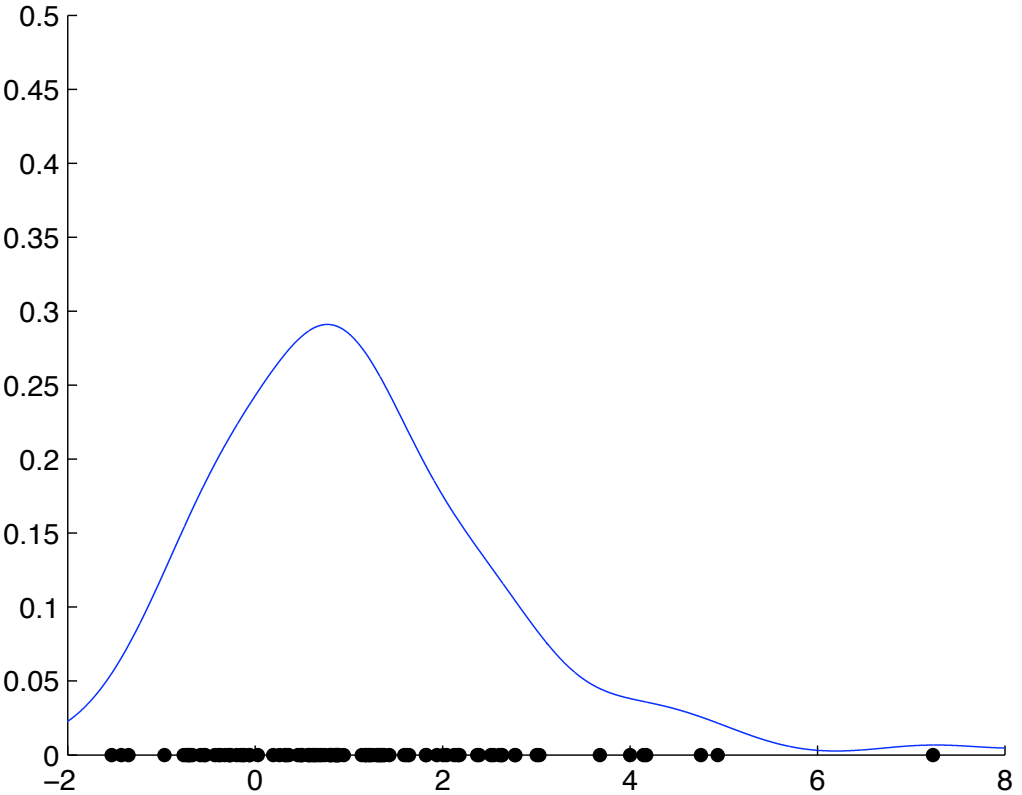
- Let $x_1, \ldots, x_N$ be independent draws from an unknown probability density function (pdf) f.
- How can we best estimate f?

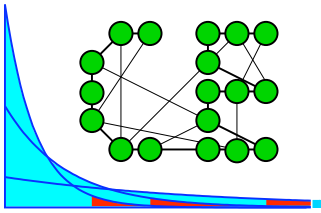




Density Estimation: Motivation

- Let $x_1, \ldots, x_N$ be independent draws from an unknown probability density function (pdf) f.
- How can we best estimate f?



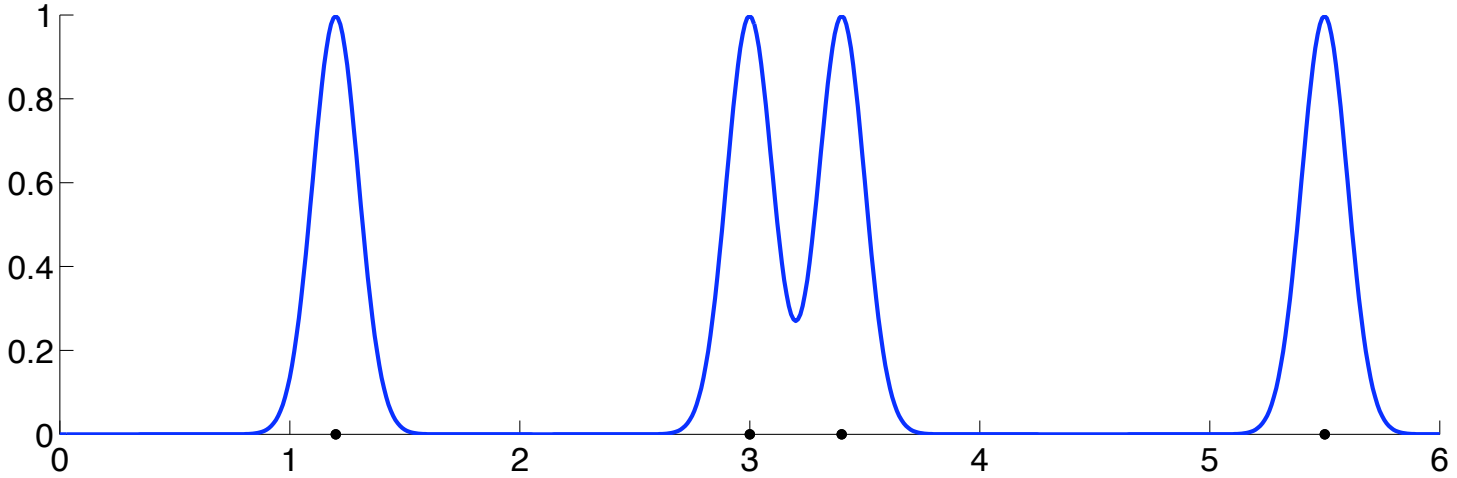


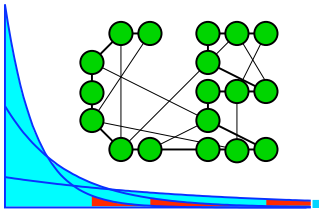
Kernel Density Estimation

General form of kernel density estimator (for data $x_1, \ldots, x_N$):

$$\widehat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa \left(\frac{x - x_i}{h} \right) \frac{1}{h} \ ,$$

where κ is a symmetric pdf on $\mathbb{R}$ and $h > 0$ is a "bandwidth" parameter. We will use the bandwidth parameter $t = h^2$.





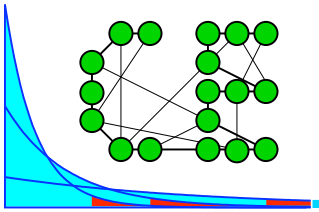
Questions

- What kernel κ should we use?
  - Most popular: Gaussian kernel:

$$\kappa(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R} .$$

- How should the bandwidth be chosen?
  - Small bandwidth: highly multimodal estimate.
  - Large bandwidth: irregularities are smoothed out.
- How do we measure the accuracy of the estimate?
  - Mean integrated squared error (MISE):

$$\mathbb{E}_f \int (\widehat{f}(x) - f(x))^2 \mathrm{d}x = \mathbb{E}_f \|\widehat{f} - f\|^2 .$$

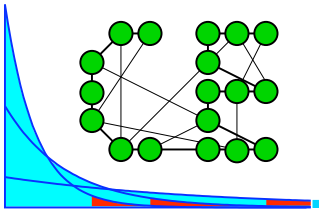


Problems

- Automatic bandwidth selection rules can suffer from a range of deficiencies:
  - boundary bias
  - assumption of normality
  - oversmoothing of well-separated modes

Literature abounds with partial solutions.

What is missing is a unified framework.



Gaussian Rule of Thumb

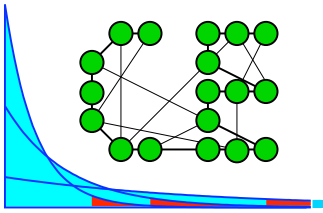
Consider a Gaussian kernel density estimator. It can be shown that asymptotically (for large N) the MISE is of the form

$$\frac{1}{4}\, t^2\, \|f''\|^2 + \frac{1}{2N\sqrt{\pi t}}\,, \qquad \text{(AMISE)}$$

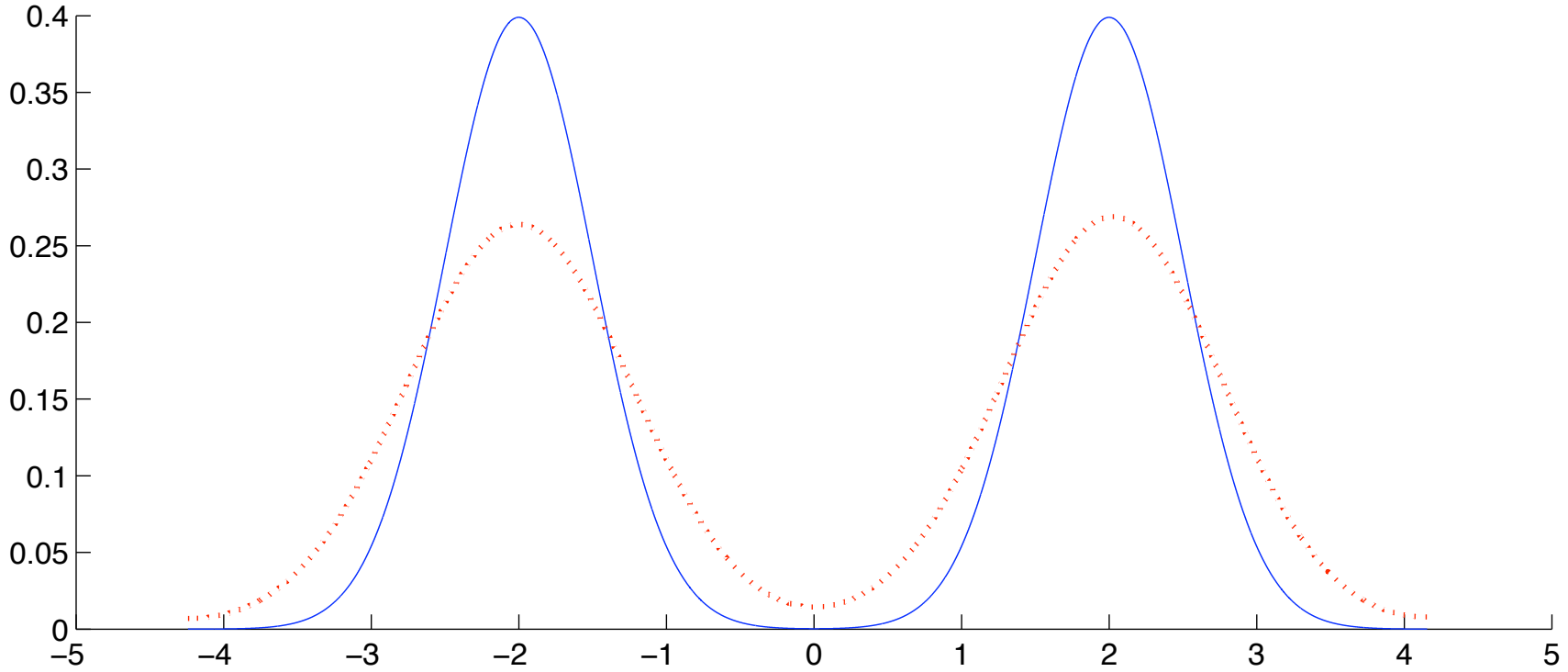
where $\|f''\|^2 = \int (f''(x))^2\, \mathrm{d}x$. The corresponding optimal value of t is $t^* = \left(\frac{1}{2N\sqrt{\pi}\,\|f''\|^2}\right)^{2/5}$.

The Gaussian rule of thumb is to assume that f is the density of the $\mathsf{N}(\text{sample mean}, \text{sample variance} = \widehat{\sigma}^2)$ distribution. In this case $\|f''\|^2 = \widehat{\sigma}^{-5}\pi^{-1/2}3/8$ and

$$t_{\text{rot}} = \left(\frac{4\,\widehat{\sigma}^5}{3\,N}\right)^{2/5} \approx 1.12\,\widehat{\sigma}^2\, N^{-2/5}\,.$$

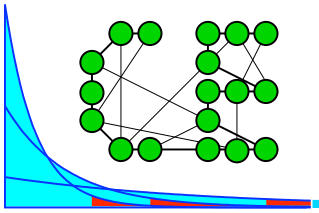


Gaussian RoT: Separated Modes

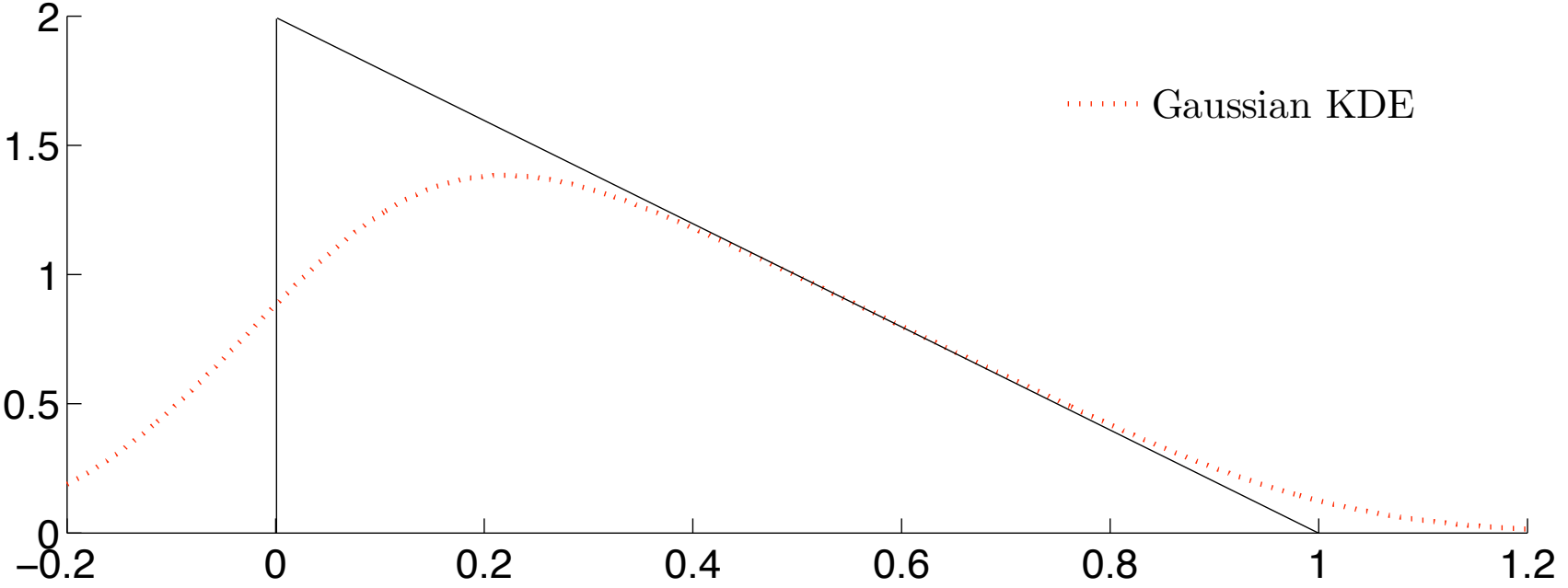


Distribution: $\frac{1}{2}\mathsf{N}\left(-2, \frac{1}{4}\right) + \frac{1}{2}\mathsf{N}\left(2, \frac{1}{4}\right)$. Samples: $N = 10^3$.

The estimate "oversmooths" the true density.

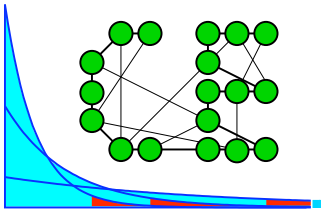


Gaussian RoT: Boundary Bias



Pdf: $f(x) = 2(1-x), x \in [0,1]$. Samples: $N = 100$.

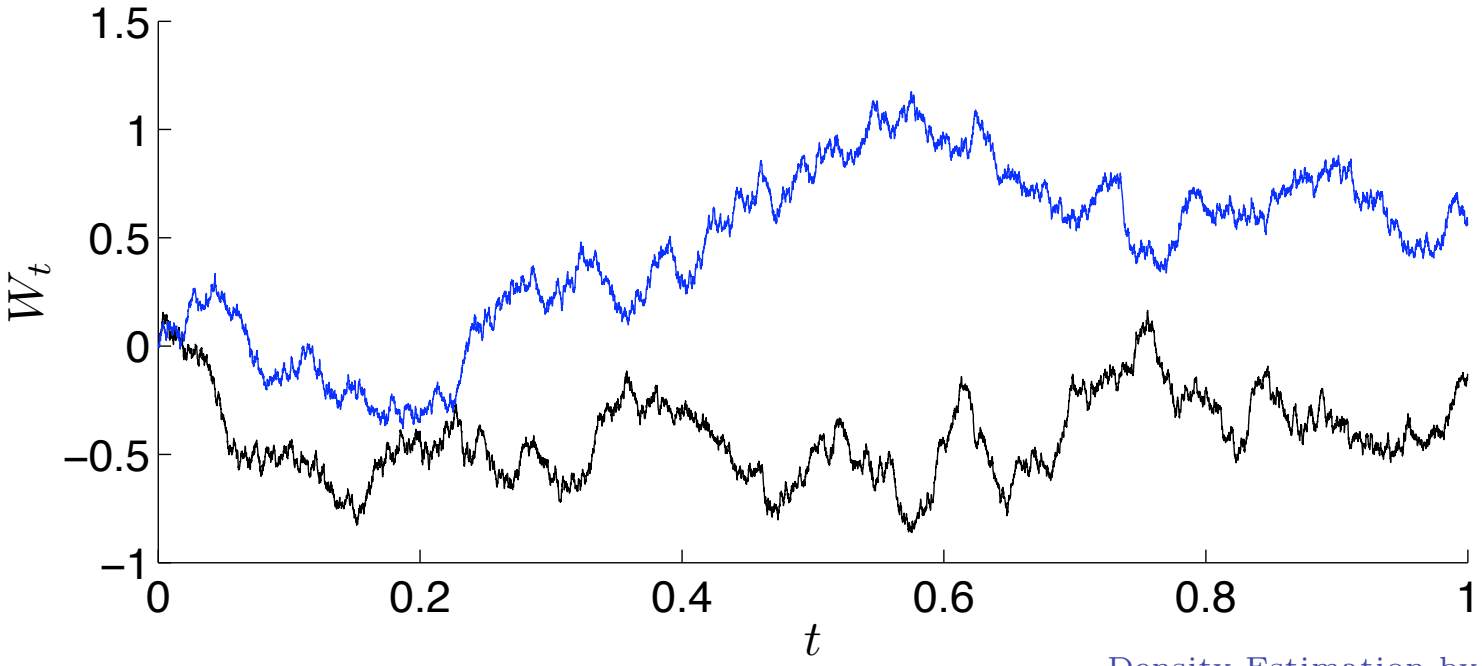
The estimate shows significant bias at the boundary 0.

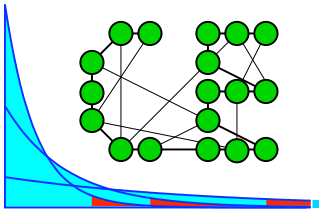


Wiener Process

The Wiener process (standard Brownian motion) is the stochastic process $\{W_t, t \geqslant 0\}$ characterized by:

1. Continuous sample paths.
2. Stationary Gaussian increments: $W_{t+s} - W_t \sim \mathsf{N}(0, s)$.
3. Independent increments.





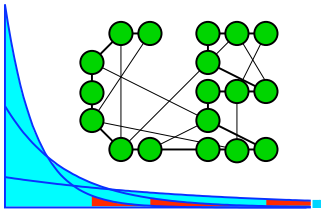
Wiener Process: Properties

The Wiener process has *many* interesting properties.
Two main properties are:

1. It is a Gaussian process: all marginal and joint distributions are Gaussian. In particular,

$$W_t \sim \mathsf{N}(0, t) \ .$$

2. It is Markov process: Conditioned on the past $\{W_s, s \leqslant t\}$, the future behaviour of the process from time t onwards is dependent only on the current state W_t.



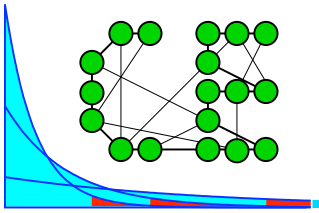
Relation with PDEs

The distribution of $(W_{s+t} \,|\, W_s = x)$ is $\mathsf{N}(x, t)$. Hence, the transition density is given by the Gaussian kernel

$$p_t(x, y) = \frac{1}{\sqrt{2\pi t}}\, \mathrm{e}^{-\frac{1}{2}\frac{(y-x)^2}{t}}, \quad t \geqslant 0, \quad x, y \in \mathbb{R}\,.$$

This satisfies the Kolmogorov backward equation

$$\frac{\partial}{\partial t} p_t(x, y) = \frac{1}{2}\frac{\partial^2}{\partial x^2} p_t(x, y) \qquad \text{(heat equation)}\,.$$



Main Idea

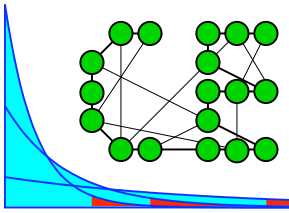
A Gaussian KDE can be viewed as the pdf of W_t, where the Wiener process $\{W_t, t \geqslant 0\}$ starts with probability $1/N$ from each of the data points $x_1, \ldots, x_N$:

$$\widehat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi t}} \mathrm{e}^{-\frac{1}{2}\frac{(x-x_i)^2}{t}} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \varphi(x, x_i; t) \ .$$

As such, $\widehat{f}(x;t)$ **satisfies the heat equation**

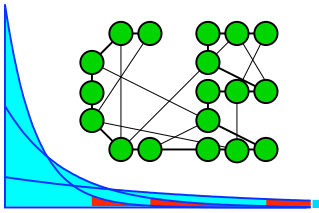
$$\frac{\partial}{\partial t}\widehat{f}(x;t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\widehat{f}(x;t), \quad t > 0, x \in \mathbb{R},$$

with $\lim_{x\to\pm\infty} \widehat{f}(x;t) = 0$ and initial condition $\widehat{f}(x;0) = \Delta(x)$, where $\Delta(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}(x)$ is the empirical density.



Advantages

- Instead of computing the Gaussian kernel density estimator $\widehat{f}(x;t)$ directly, it can be obtained by evolving the solution of the heat equation up to time t.
- This naturally extends to KDEs for finite-support distributions.
- The PDE can be solved efficiently using Fast Fourier transform (FFT) techniques.
- The bandwidth parameter has a natural interpretation (time) and automatic optimal bandwidth computation can be done efficiently.
- The idea is easily extended to more general diffusion processes.

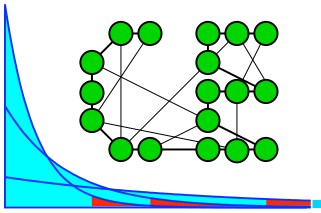


Finite-support Distributions

Suppose the pdf f is known to be 0 outside $\mathscr{X} = [0,1]$.
Within the PDE framework all we have to do is solve the heat equation over the finite domain $[0,1]$ with initial condition $\widehat{f}(x;0) = \Delta(x)$ and Neumann boundary conditions

$$\frac{\partial}{\partial x}\widehat{f}(x;t)\Big|_{x=1} = \frac{\partial}{\partial x}\widehat{f}(x;t)\Big|_{x=0} = 0\ .$$

The boundary condition ensures that $\frac{d}{dt}\int_{\mathscr{X}} \widehat{f}(x;t)\,\mathrm{d}x = 0$, from where it follows that $\int_{\mathscr{X}} \widehat{f}(x;t)\,\mathrm{d}x = \int_{\mathscr{X}} \widehat{f}(x;0)\,\mathrm{d}x = 1$ for all $t \geqslant 0$.



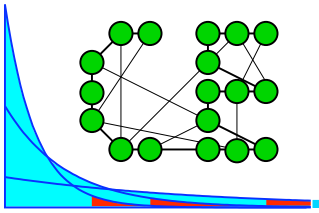
Solution

The analytical solution of this PDE in this case is:

$$\widehat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N}\theta(x,x_i;t), \quad x\in[0,1], \qquad \text{(KDE-}\theta\text{)}$$

where the kernel θ is the theta function

$$\theta(x,x_i;t) = \sum_{k=-\infty}^{\infty}\varphi(x,2k+x_i;t)+\varphi(x,2k-x_i;t), \quad x\in[0,1].$$

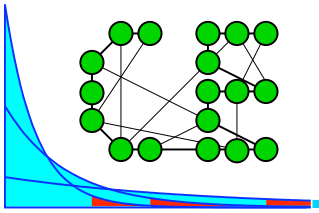


Small time/bandwidth behaviour

We have

$$\lim_{t \downarrow 0} \frac{\theta(x, x_i; t)}{\varphi(x, x_i; t)} = 1 \ .$$

Thus, for small t, the estimator (KDE-θ) behaves like the Gaussian kernel density estimator in the interior of $[0, 1]$.



Large Bandwidth Behaviour

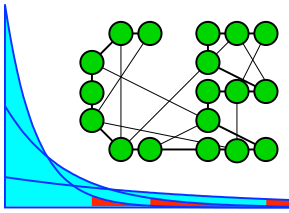
The theta function can be written as

$$\theta(x, x_i; t) = \sum_{k=-\infty}^{\infty} \mathrm{e}^{-k^2\pi^2 t/2} \cos(k\pi x) \cos(k\pi x_i)$$
$$= 1 + 2\sum_{k=1}^{\infty} \mathrm{e}^{-k^2\pi^2 t/2} \cos(k\pi x) \cos(k\pi x_i) \ . \qquad (\star)$$

We see that

$$\theta(x, x_i; t) \approx 1 + 2\,\mathrm{e}^{-\pi^2 t/2} \cos(\pi x) \cos(\pi x_i), \quad t \to \infty, \ x \in [0, 1].$$

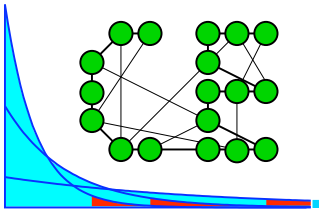
As the bandwidth becomes larger, the kernel approaches the uniform density on $[0, 1]$.



Maximum Principle

An important property of the estimator (KDE-θ) is that the number of local maxima or modes is a nonincreasing function of t. This follows from the maximum principle for parabolic PDEs.

To see this, suppose (x_0, t_0) is a local maximum, and hence $\frac{\partial^2}{\partial x^2}\widehat{f}(x_0; t_0) \leqslant 0$. Since $\widehat{f}(x; t)$ satisfies the heat equation, it follows that $\frac{\partial}{\partial t}\widehat{f}(x_0; t_0) \leqslant 0$, from which it follows that there exists an $\varepsilon > 0$ such that $\widehat{f}(x_0; t_0) \geqslant \widehat{f}(x_0; t_0 + \varepsilon)$.



Fast Evaluation of the KDE

Recall

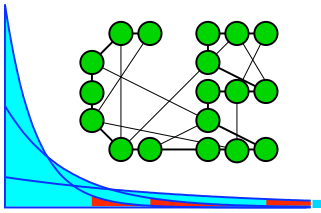
$$\widehat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N}\theta(x,x_i;t),$$

where

$$\theta(x,x_i;t) = 1 + 2\sum_{k=1}^{\infty}\mathrm{e}^{-k^2\pi^2 t/2}\cos(k\pi x)\cos(k\pi x_i)\ . \qquad (\star)$$

It follows that for large n

$$\widehat{f}(x;t) \approx \sum_{k=0}^{n-1} a_k\ \mathrm{e}^{-k^2\pi^2 t/2}\cos(k\pi x)\ ,$$

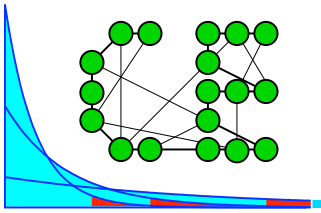


Fast Evaluation of the KDE

where the coefficients $\{a_k\}_{k=0}^{n-1}$ are given by $a_0 = 1$ and

$$\begin{aligned} a_k &= \frac{2}{N} \sum_{i=1}^{N} \cos(k\pi x_i) \\ &= 2 \int_0^1 \cos(k\pi x) \widehat{f}(x;0)\, \mathrm{d}x \\ &\approx 2 \sum_{i=0}^{n-1} \cos(k\pi \frac{i}{n})\, n\, \widehat{f}_i\, \frac{1}{n}, \end{aligned}$$

where $\widehat{f}_i$ is the number of points in $(\frac{i}{n}, \frac{i+1}{n})$, $i = 0, \ldots, n-1$.

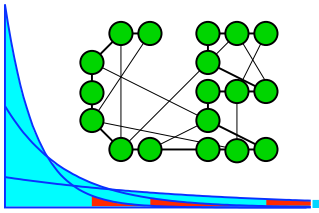


Fast Evaluation of the KDE

So, the $\{a_k, k = 0, \ldots, n-1\}$ can be calculated fast via the fast cosine transform of the $\{\widehat{f}_i, i = 0, \ldots, n-1\}$.

Moreover, $\widehat{f}(i/n), i = 0, 1, \ldots, n-1$ can be calculated fast with the inverse fast cosine transform, because

$$\widehat{f}(x;t) \approx \sum_{k=0}^{n-1} a_k \, \mathrm{e}^{-k^2\pi^2 t/2} \cos(k\pi x) \ .$$



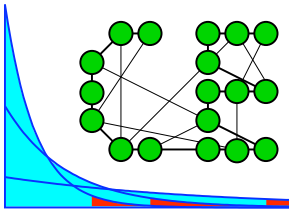
Improved Bandwidth Selection

We obtain the optimal bandwidth as a solution to a fixed-point problem

$$t = g(t) \ ,$$

where g depends on $\widehat{f}(\cdot; t)$ and can be calculated fast via the fast cosine transform.

Our approach does not assume normality of the data, unlike some other methods.

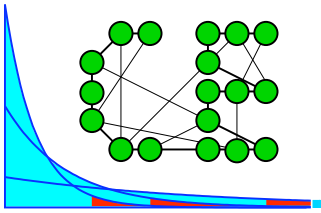


Example: Asymmetric Double Claw

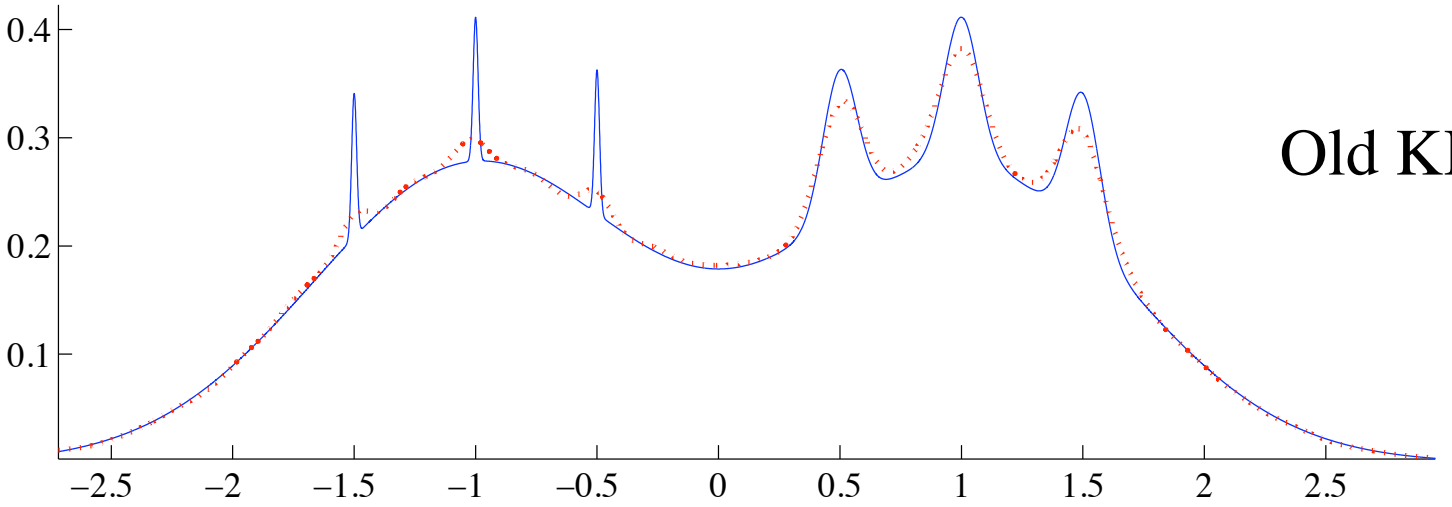
Typically, the improved plug-in method is more accurate than existing methods. For example, consider estimating the asymmetric double claw density described by:

$$\frac{46}{100}\sum_{k=0}^{1}\mathsf{N}\left(2k-1,\left(\frac{2}{3}\right)^2\right)+\frac{1}{300}\sum_{k=1}^{3}\mathsf{N}\left(-\frac{k}{2},\frac{1}{100^2}\right)$$
$$+\frac{7}{300}\sum_{k=1}^{3}\mathsf{N}\left(\frac{k}{2},\left(\frac{7}{100}\right)^2\right),$$

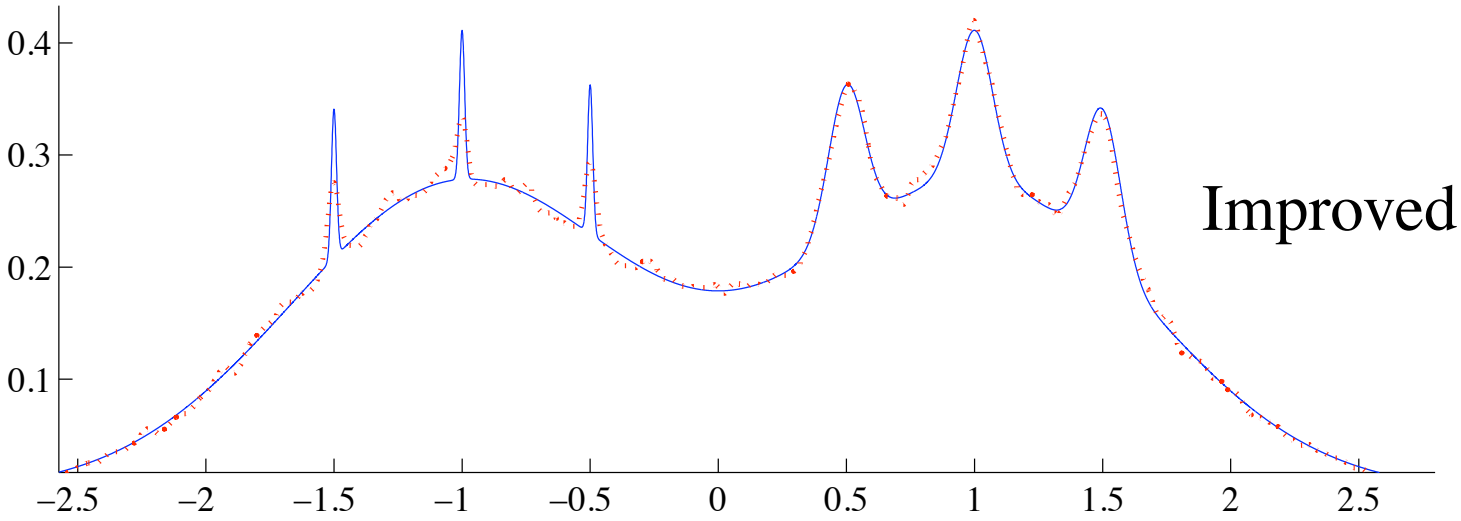
using $N = 10^5$ iid samples.



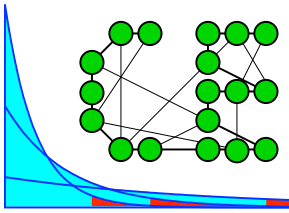
Example: Asymmetric Double Claw



Old KDE



Improved KDE



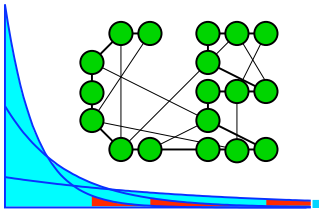
Example: Asymmetric Double Claw

To make a more rigorous comparison we use the error criterion

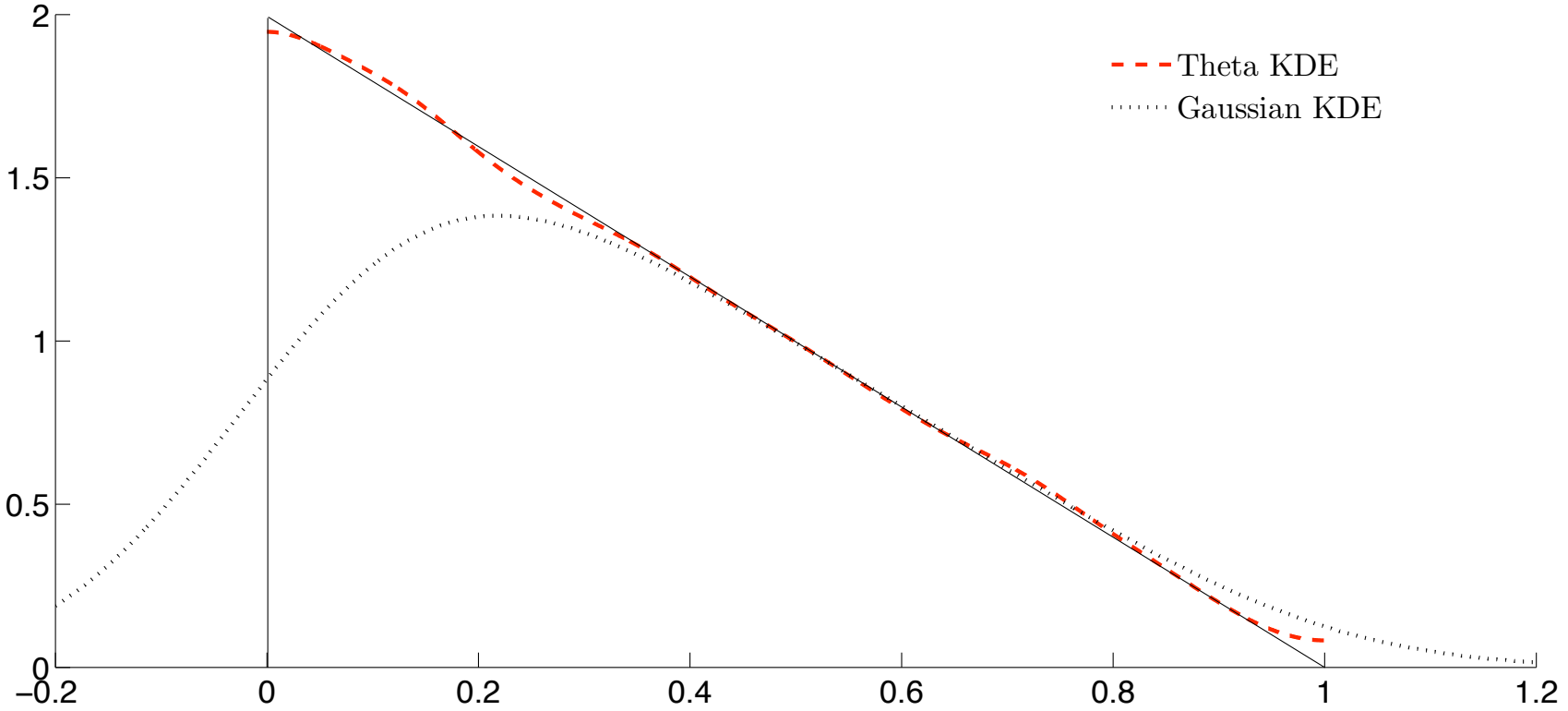
$$\text{Ratio} = \frac{\|\widehat{f}(\,\cdot\,;\widehat{t^*}\,) - f\|^2}{\|\widehat{f}(\,\cdot\,;t_{\text{LS}}) - f\|^2}\,.$$

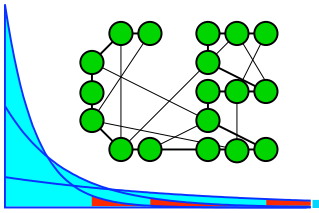
Table: Ratio of ISEs for the old and improved KDEs

N	10^4	10^5	10^6	10^7
Ratio	1.01	0.37	0.55	0.0083



Example: Boundary Bias



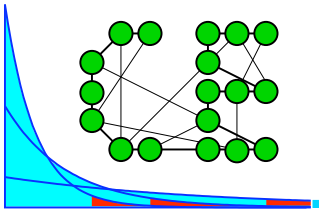


Extension: General Diffusions

Let a and p be arbitrary positive functions on $\mathscr{X} \subset \mathbb{R}$ with bounded second derivatives. Suppose that, instead of the Wiener process, we run a diffusion process $\{V_t, t > 0\}$ given by the stochastic differential equation

$$\mathrm{d}V_t = \mu(V_t)\,\mathrm{d}t + \sigma(V_t)\,\mathrm{d}W_t, \qquad \text{(SDE)}$$

where the drift coefficient $\mu(x) = \frac{a'(x)}{2p(x)}$, the diffusion coefficient $\sigma(x) = \sqrt{\frac{a(x)}{p(x)}}$, the initial state V_0 has distribution $\Delta(x)$, and $\{W_t, t > 0\}$ is a Wiener process. Obviously, if $a = 1$ and $p = 1$, we revert to the original case.



Diffusion PDE

The pdf $\widehat{f}(\cdot;t)$ of V_t satisfies the linear diffusion PDE

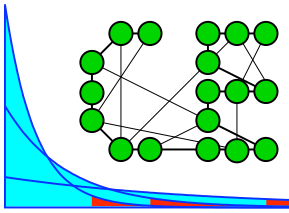
$$\frac{\partial}{\partial t}\widehat{f}(x;t) = L\widehat{f}(x;t), \quad x \in \mathscr{X},\ t > 0, \qquad \text{(LD-PDE)}$$

where L is the linear differential operator

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

with initial condition $\widehat{f}(x,0) = \Delta(x)$.

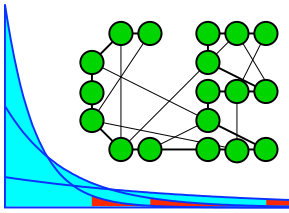
If the set $\mathscr{X}$ is bounded, we add the boundary condition $\frac{\partial}{\partial x}\left(\frac{\widehat{f}(x;t)}{p(x)}\right) = 0$ on $\partial\mathscr{X}$, which ensures that the solution $\widehat{f}(x;t)$ integrates to unity.



Diffusion PDE

- We found that the choice $a \equiv 1$ (zero drift) works best.
- The function p controls the smoothness of the KDE.
- The theta KDE provides a good candidate for p.
- When $p(x)$ is a pdf, it is the stationary pdf of the diffusion process, and

$$\lim_{t \to \infty} \widehat{f}(x;t) = p(x), \quad x \in \mathscr{X}.$$



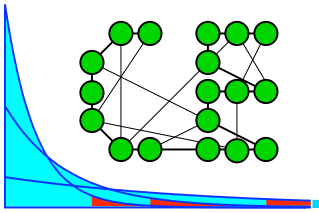
Diffusion PDE

Similar to the estimator (KDE-θ) and the Gaussian kernel density estimator, we can write the solution of (LD-PDE) as:

$$\widehat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N} \kappa(x, x_i; t),$$

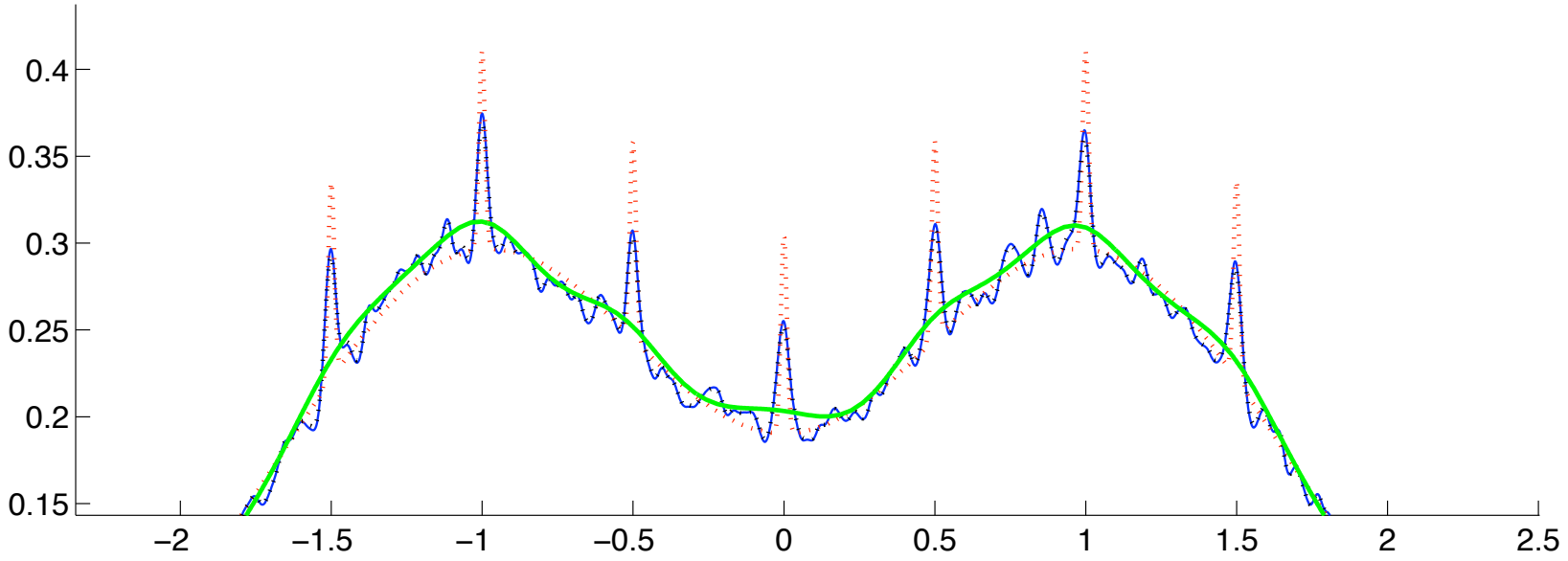
where for each fixed $y \in \mathscr{X}$ the diffusion kernel κ is the fundamental solution to the Kolmogorov forward equation:

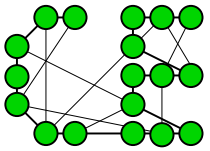
$$\begin{aligned} \frac{\partial}{\partial t}\kappa(x,y;t) &= L\kappa(x,y;t), \quad x \in \mathscr{X},\ t > 0 \\ \kappa(x,y;0) &= \delta_y(x), \quad x \in \mathscr{X}. \end{aligned}$$



Example: Symmetric Double Claw

$N = 10^5$ and ISE ratio = 0.33, the blue line shows the diffusion estimate with a pilot estimate $p\,(a = 1)$. The green line shows the Abramson estimator.





Extension: Higher Dimensions

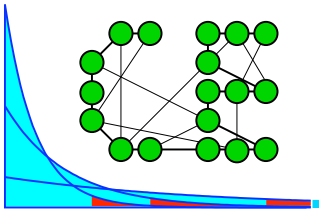
The two-dimensional version of the heat equation is:

$$\frac{\partial \widehat{f}}{\partial t}(\mathbf{x};t) = \frac{1}{2}\left(\frac{\partial^2 \widehat{f}}{\partial x_1^2}(\mathbf{x};t) + \frac{\partial^2 \widehat{f}}{\partial x_2^2}(\mathbf{x};t)\right), \ \forall t > 0, \ \mathbf{x} \in \mathscr{X}$$

$$\widehat{f}(\mathbf{x};0) = \Delta(\mathbf{x})$$

$$\mathbf{n} \cdot \nabla \widehat{f}(\mathbf{x};t) = 0, \quad \forall t > 0,$$

where $\mathbf{x} = (x_1, x_2)$ belongs to the set $\mathscr{X} \subseteq \mathbb{R}^2$, the initial condition $\Delta(\mathbf{x})$ is the empirical density of the data, and in the Neumann boundary condition $\mathbf{n}$ denotes the unit outward normal to the boundary $\partial\mathscr{X}$ at $\mathbf{x}$.

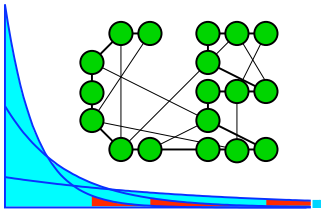


Example

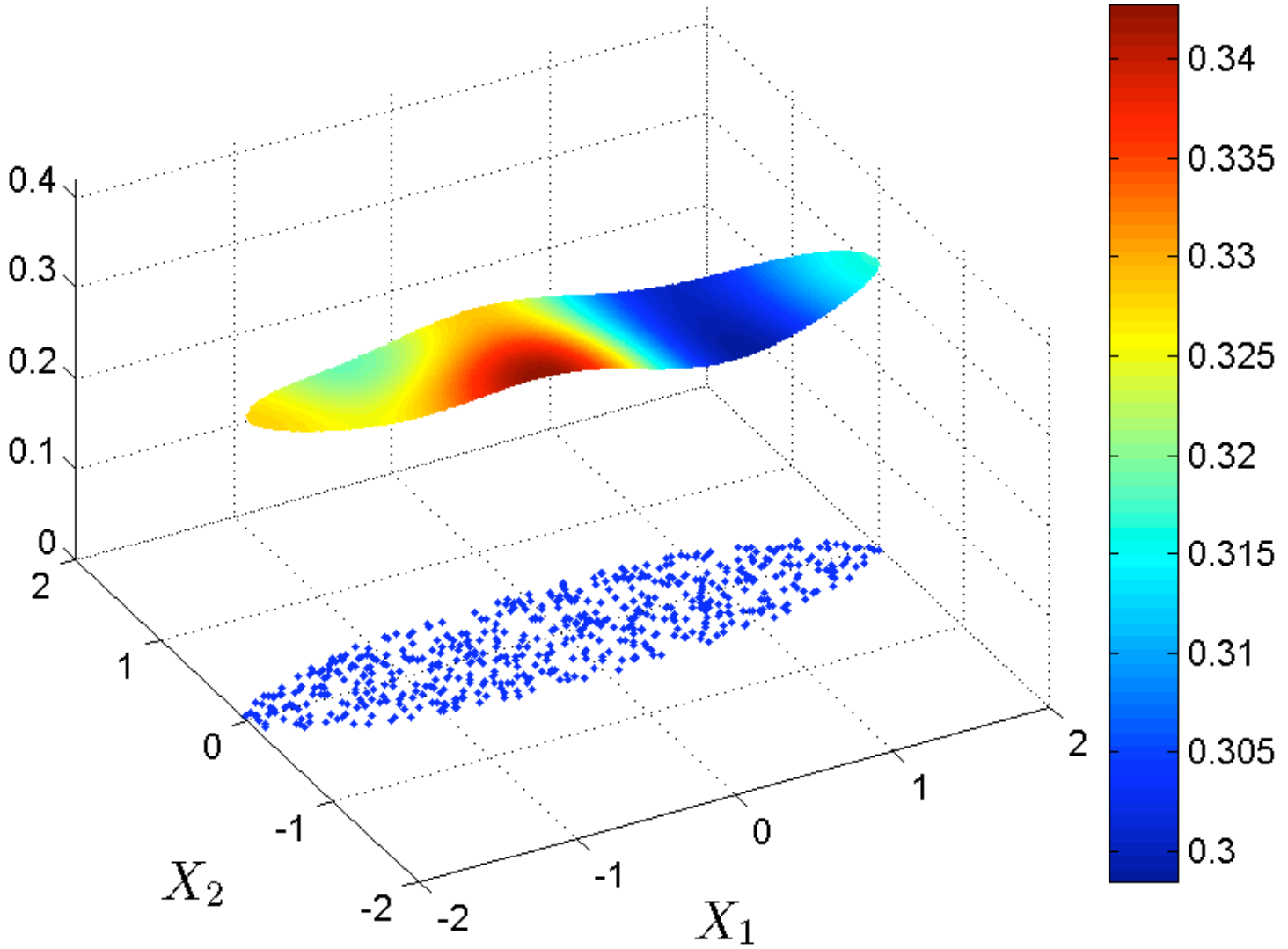
Consider the density estimation of 600 uniformly distributed points on the domain

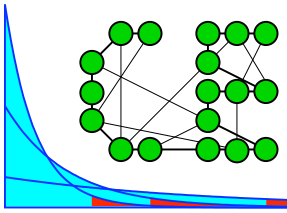
$$\mathscr{X} = \{\mathbf{x} : x_1^2 + (4x_2)^2 \leqslant 4\} \ .$$

We assume that the domain of the data $\mathscr{X}$ is known prior to the estimation.
Existing methods could not handle such two-dimensional (boundary) density estimation problems either because the geometry of the set $\mathscr{X}$ is too complex, or because the resulting estimator is not a bona-fide pdf.



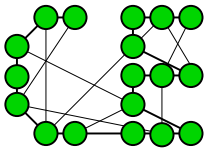
Example


0.4
0.3
0.2
0.1
0
2
1
0
-1
-2
X_2
-2
-1
0
1
2
X_1
0.34
0.335
0.33
0.325
0.32
0.315
0.31
0.305
0.3

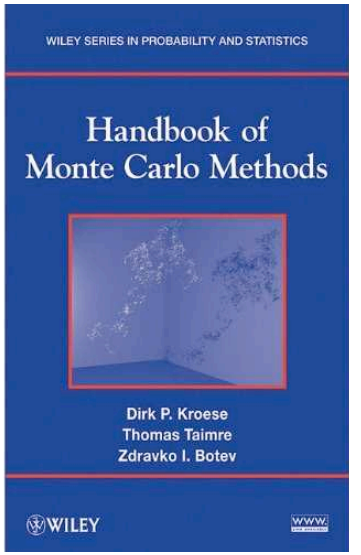


Conclusions

- The (standard) Gaussian kernel density estimate can be viewed as a solution to the Heat Equation on $\mathbb{R}$.
- Solving the heat equation on $[0, 1]$ gives a kernel density estimate in terms of theta functions.
- This formulation allows for fast computation of the bandwidth and the density estimate itself.
- We found a more accurate bandwidth selection rule.
- By considering the Kolmogorov PDE of a general diffusion process one can obtain even more accurate kernel density estimators.
- The same idea works in higher dimensions.



References

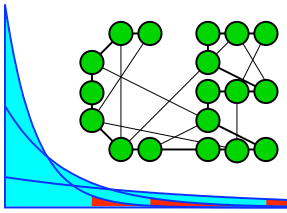


• *Handbook of Monte Carlo Methods*, D.P. Kroese, T. Taimre, Z. I. Botev, John Wiley & Sons, 2011.

• Z. I. Botev, J. F. Grotowski, D. P. Kroese. Kernel density estimation via diffusion (2010). *Ann. Stat.*, **38** (5) 2916–2957.

• B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, 1986.

• M. P. Wand, M.C. Jones. *Kernel Smoothing*, 1995.



Thank You!