Joint Probability Density Function

Probability and Statistics for Data Science

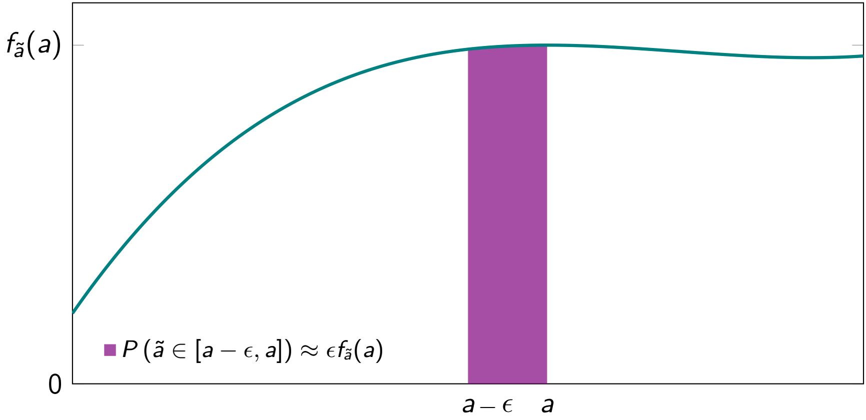
Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

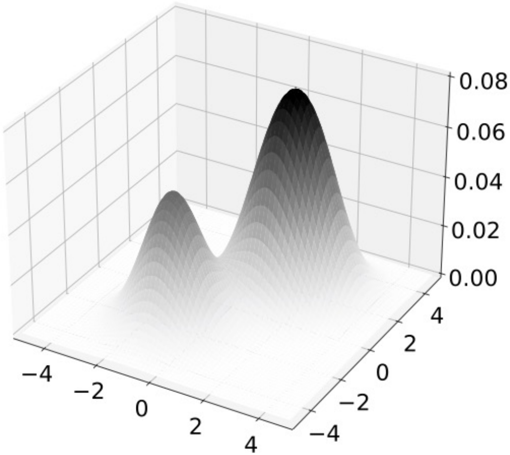
Goal

Define probability density for multiple random variables

Probability density function



Probability density $f_{\tilde{a},\tilde{b}}(a, b)$ at $\begin{bmatrix} a \\ b \end{bmatrix}$



$$\mathrm{P}\left(\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} \in [a-\epsilon, a] \times [b-\epsilon, b]\right) \approx \epsilon^2 f_{\tilde{a},\tilde{b}}(a, b)$$

Probability density as a derivative

$$
\begin{aligned}
& f_{\tilde{a},\tilde{b}}(a, b) \\
& := \lim_{\epsilon \to 0} \frac{\mathrm{P}\left(a - \epsilon < \tilde{a} \leq a, b - \epsilon < \tilde{b} \leq b\right)}{\epsilon^2} \\
& = \lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a, b) - F_{\tilde{a},\tilde{b}}(a - \epsilon, b) - F_{\tilde{a},\tilde{b}}(a, b - \epsilon) + F_{\tilde{a},\tilde{b}}(a - \epsilon, b - \epsilon)}{\epsilon^2} \\
& = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a, b) - F_{\tilde{a},\tilde{b}}(a - \epsilon, b)}{\epsilon} - \lim_{\epsilon \to 0} \frac{F_{\tilde{a},\tilde{b}}(a, b - \epsilon) - F_{\tilde{a},\tilde{b}}(a - \epsilon, b - \epsilon)}{\epsilon} \right) \\
& = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\frac{\partial F_{\tilde{a},\tilde{b}}(a, b)}{\partial a} - \frac{\partial F_{\tilde{a},\tilde{b}}(a, b - \epsilon)}{\partial a} \right) \\
& = \frac{\partial^2 F_{\tilde{a},\tilde{b}}(a, b)}{\partial a \, \partial b}
\end{aligned}
$$

Joint pdf

The joint pdf of $\tilde{a}$ and $\tilde{b}$ is

$$f_{\tilde{a},\tilde{b}}(a, b) := \frac{\partial^2 F_{\tilde{a},\tilde{b}}(a, b)}{\partial a \partial b}$$

The joint pdf of a d-dimensional vector $\tilde{x}$ is

$$f_{\tilde{x}}(x) := \frac{\partial^d F_{\tilde{x}}(x)}{\partial x[1]\, \partial x[2] \cdots \partial x[d]}$$

Using the joint pdf to compute probabilities

For any 2D Borel set $B \subseteq \mathbb{R}^2$

$$\mathrm{P}\left((\tilde{a}, \tilde{b}) \in B\right) = \int_{(a,b)\in B} f_{\tilde{a},\tilde{b}}(a, b)\, \mathrm{d}a\, \mathrm{d}b$$

For any d-dimensional Borel set $B \subseteq \mathbb{R}^d$

$$\mathrm{P}(\tilde{x} \in B) = \int_{x\in B} f_{\tilde{x}}(x)\, \mathrm{d}x$$

Properties

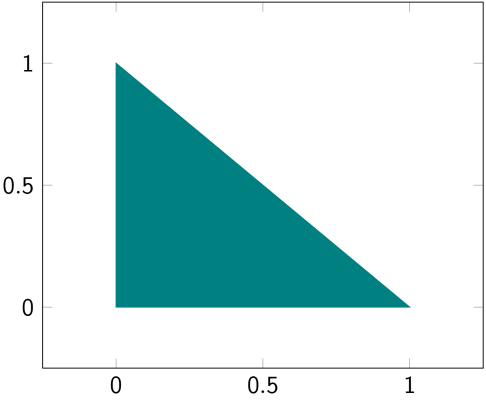
$$\int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\, \mathrm{d}a\, \mathrm{d}b = 1$$

$$\int_{x[1]=-\infty}^{\infty} \cdots \int_{x[d]=-\infty}^{\infty} f_{\tilde{x}}(x)\, \mathrm{d}x = 1$$

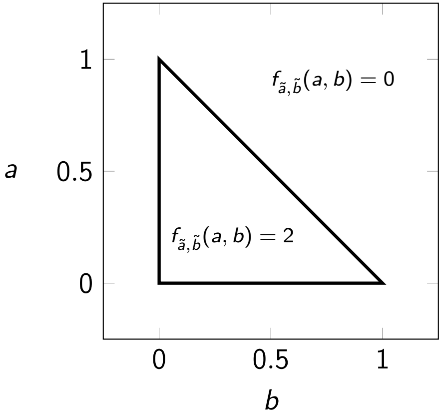
Nonnegative? Yes, because joint cdf $F_{\tilde{a},\tilde{b}}$ is non-decreasing in a and b

Any nonnegative function that integrates to 1 is a valid joint pdf

Triangle lake: Joint pdf?

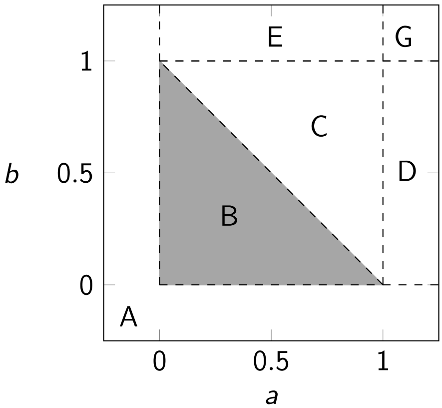


Triangle lake



$$
\begin{aligned}
\mathrm{P}(\{\tilde{a} \geq 0.6, \tilde{b} \leq 0.2\}) &= \int_{b=0}^{0.2} \int_{a=0.6}^{1-b} 2 \, \mathrm{d}a \, \mathrm{d}b \\
&= \int_{b=0}^{0.2} 2\,(0.4 - b) \, \mathrm{d}b = 0.12
\end{aligned}
$$

Joint cdf for $(a, b) \in B$



$$F_{\tilde{a},\tilde{b}}(a, b) = \int_{u=0}^{b} \int_{v=0}^{a} 2 \, \mathrm{d}v \, \mathrm{d}u$$
$$= 2ab$$

Joint cdf

$$F_{\tilde{a},\tilde{b}}(a,b) = \begin{cases} 0 & \text{if } a < 0 \text{ or } b < 0, \\ 2ab, & \text{if } a \geq 0, b \geq 0, a + b \leq 1, \\ 2a + 2b - b^2 - a^2 - 1, & \text{if } a \leq 1, b \leq 1, a + b \geq 1, \\ 2b - b^2, & \text{if } a \geq 1, 0 \leq b \leq 1, \\ 2a - a^2, & \text{if } 0 \leq a \leq 1, b \geq 1, \\ 1, & \text{if } a \geq 1, b \geq 1. \end{cases}$$

Estimating a pdf from data

We need it to be nonnegative and integrate to one

We cannot use empirical probabilities, probability of each data point is zero!

Kernel density estimation (KDE)

Data $X := \{x_1, x_2, \ldots, x_n\}$

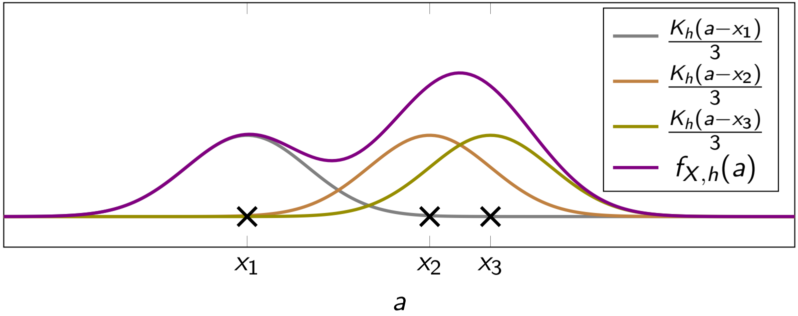
Kernel density estimate with bandwidth h is

$$f_{X,h}(a) := \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right)$$

where K is a kernel that satisfies

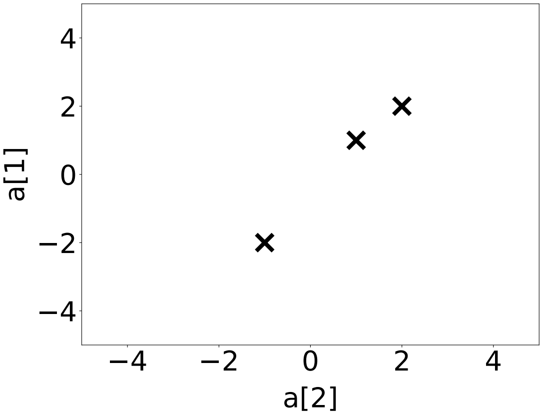
$$\begin{aligned} K(a) &\geq 0 \quad \text{for all } a \in \mathbb{R}, \\ \int_{\mathbb{R}} K(a)\, \mathrm{d}x &= 1 \end{aligned}$$

Gaussian kernel, $n = 3$

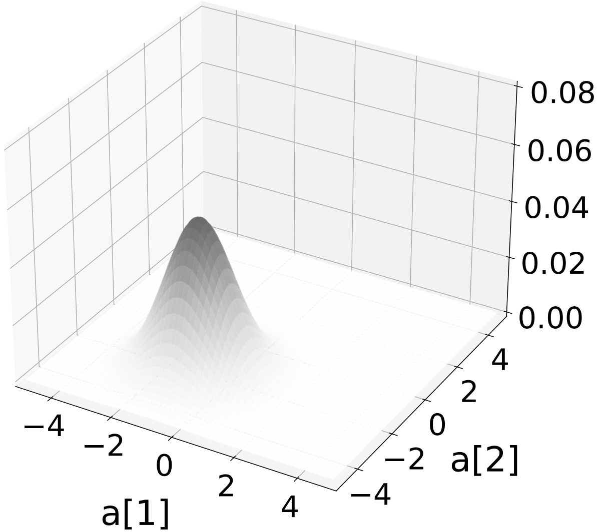


$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_3}{h}\right)$$

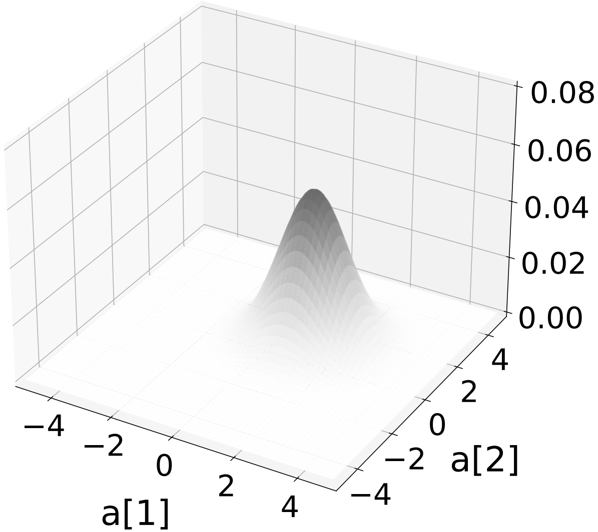
Multidimensional KDE



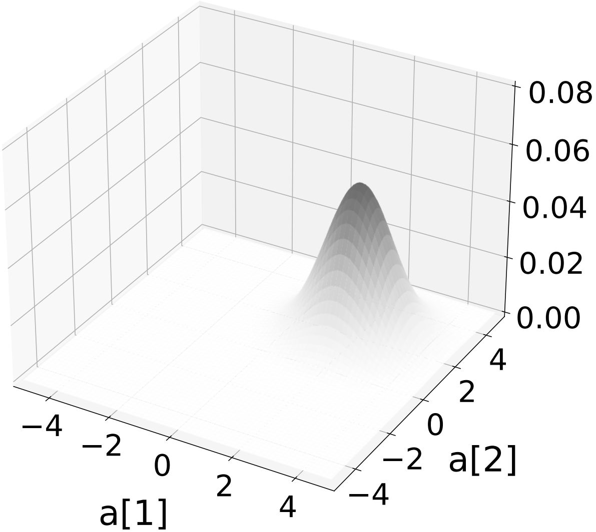
$$\frac{K(a-x_1)}{3}$$



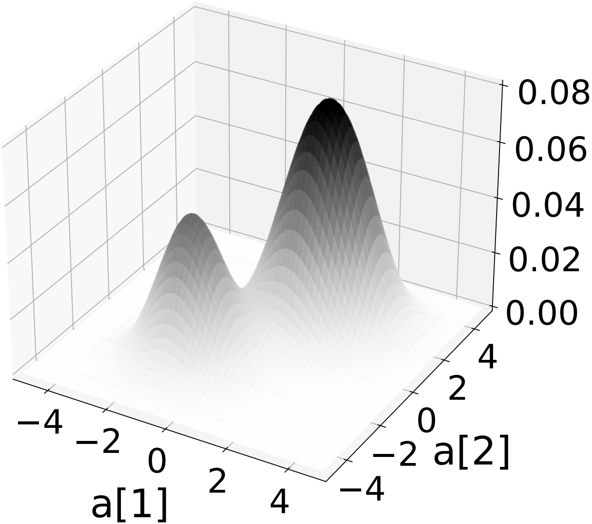
$$\frac{K(a-x_2)}{3}$$



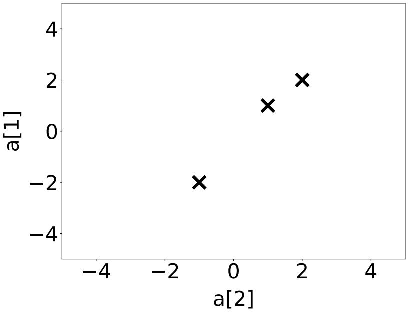
$\frac{K(a-x_3)}{3}$

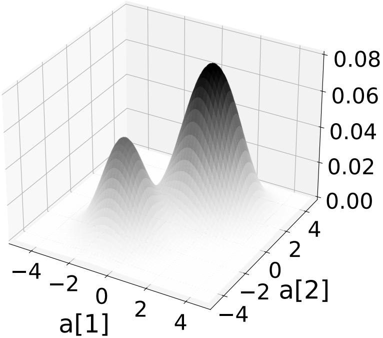


$$\frac{K(a-x_1)+K(a-x_2)+K(a-x_3)}{3}$$

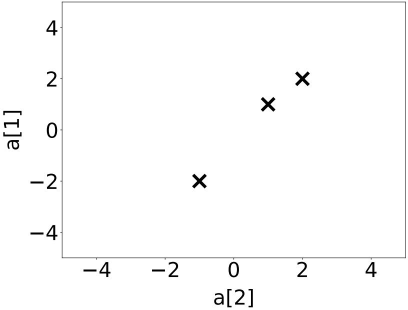


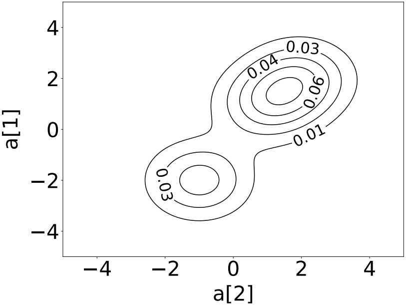
$$\frac{K(a-x_1)+K(a-x_2)+K(a-x_3)}{3}$$





$$\frac{K(a - x_1) + K(a - x_2) + K(a - x_3)}{3}$$





Multidimensional KDE

Data $X := \{x_1, x_2, \ldots, x_n\}$

Kernel density estimate with bandwidth h is

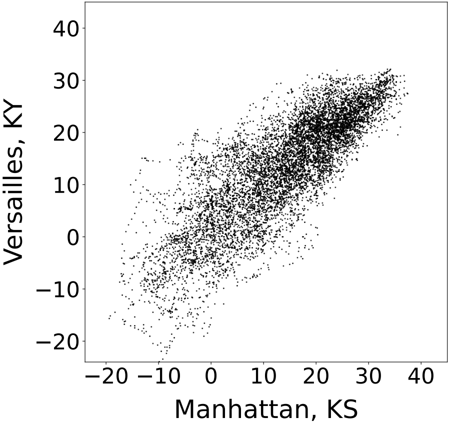
$$f_{X,h}(a) := \frac{1}{n\,h^d} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right)$$

where K is a kernel that satisfies

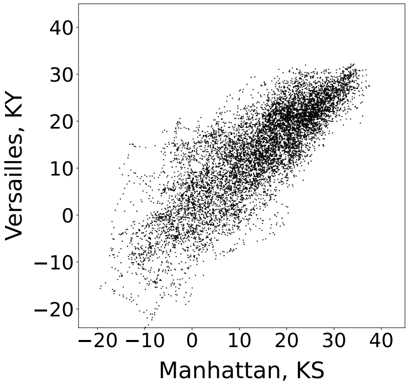
$$\begin{aligned} K(a) &\geq 0 \quad \text{for all } a \in \mathbb{R}^d, \\ \int_{\mathbb{R}^d} K(a)\,\mathrm{d}x &= 1 \end{aligned}$$

Estimate is composed of copies of the kernel centered at each data point

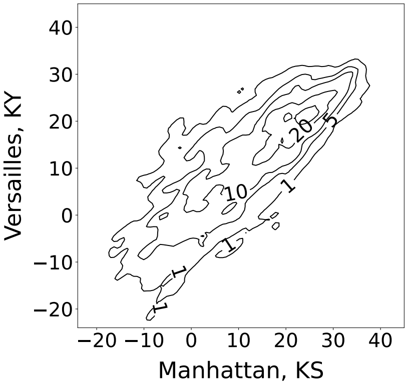
Temperature



KDE (h = 0.1)

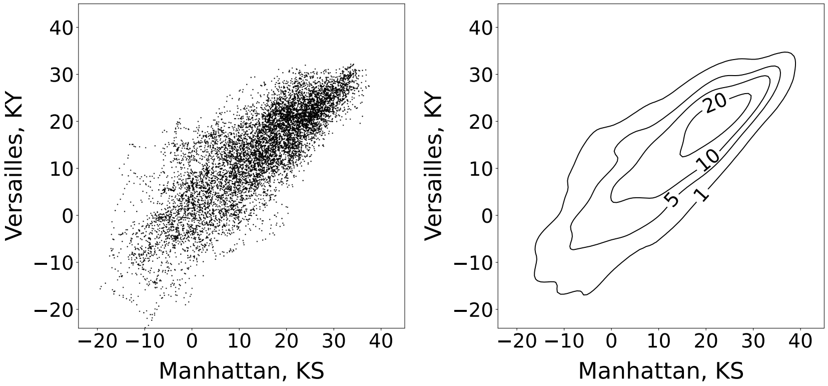


40
30
20
10
0
−10
−20
Versailles, KY
−20
−10
0
10
20
30
40
Manhattan, KS

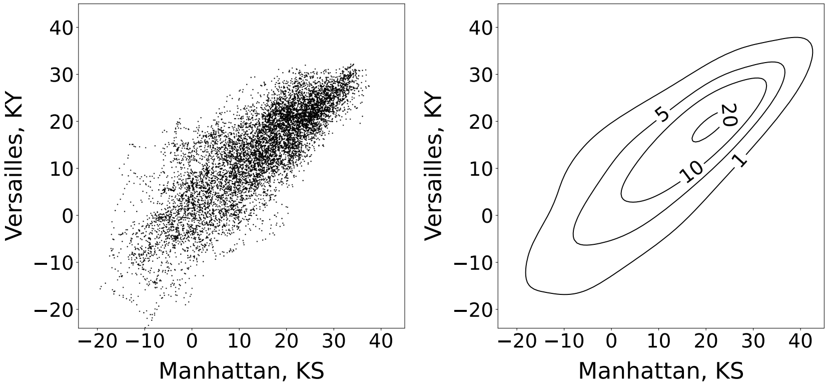


40
30
20
10
0
−10
−20
Versailles, KY
−20
−10
0
10
20
30
40
Manhattan, KS
20
5
10
1
1
1
1

KDE (h = 0.25)


Versailles, KY
Manhattan, KS
40
30
20
10
0
−10
−20
−20 −10 0 10 20 30 40
20
10
5
1

KDE (h = 0.5)



40
30
20
10
0
−10
−20
Versailles, KY
−20
−10
0
10
20
30
40
Manhattan, KS
5
20
10
1

What have we learned?

Definition and properties of joint pdf

How to estimate it from data