Multiple Continuous Variables (Overview)

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

Plan

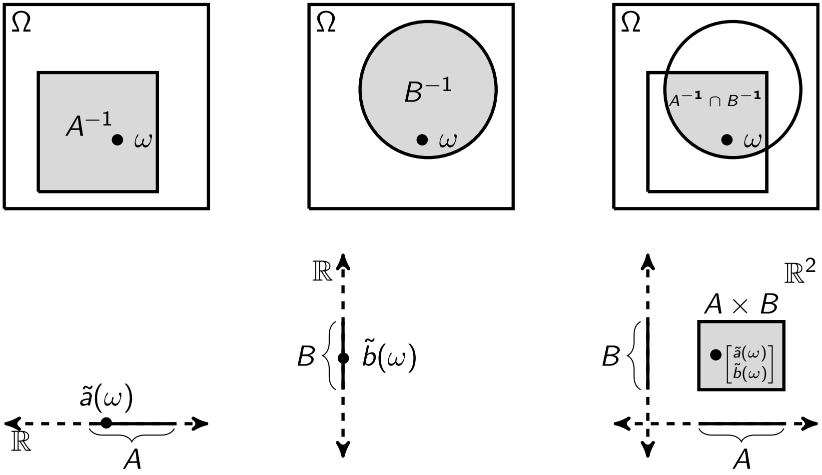
- Mathematical definition
- Joint probability density
- Marginal and conditional distributions
- Gaussian random vectors

Continuous random variables

We describe continuous random variables in terms of the probability that they belong to any interval

What about multiple continuous random variables defined on the same probability space?

Two continuous random variables



Two continuous random variables

$$
\begin{aligned}
\mathrm{P}\left(\begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix} \in A \times B\right) &:= \mathrm{P}\left(\left\{\omega \mid \tilde{a}(\omega) \in A \text{ and } \tilde{b}(\omega) \in B\right\}\right)\\
&= \mathrm{P}\left(A^{-1} \cap B^{-1}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
A^{-1} &:= \{\omega \mid \tilde{a}(\omega) \in A\},\\
B^{-1} &:= \left\{\omega \mid \tilde{b}(\omega) \in B\right\}.
\end{aligned}
$$

Higher dimensions

Let $\tilde{x} : \Omega \to \mathbb{R}^d$ be a d-dimensional vector containing d continuous random variables $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$

Defined on the same probability space $(\Omega, \mathcal{C}, \mathrm{P})$

For any d Borel sets X_1, X_2, ..., $X_d \subseteq \mathbb{R}$, the probability of the event

$$\{\omega \mid \tilde{x}(\omega) \in X_1 \times X_2 \times \cdots \times X_d\} = \cap_{i=1}^{d} \{\omega \mid \tilde{x}[i](\omega) \in X_i\}$$

is well defined

Joint cdf

The joint cdf of $\tilde{a} : \Omega \to \mathbb{R}$ and $\tilde{b} : \Omega \to \mathbb{R}$ is

$$F_{\tilde{a},\tilde{b}}(a, b) := \mathrm{P}\left(\tilde{a} \leq a, \tilde{b} \leq b\right)$$

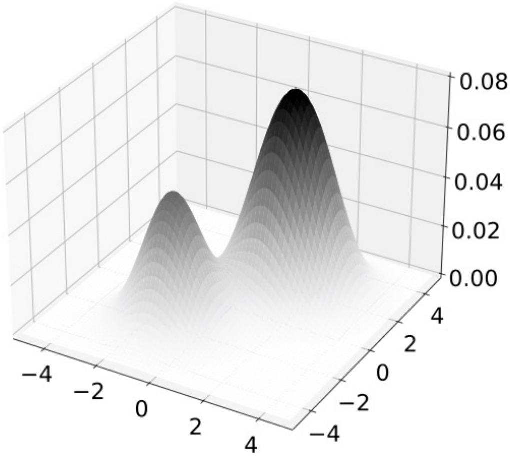
The joint cdf of a d-dimensional vector $\tilde{x} : \Omega \to \mathbb{R}^d$ is

$$F_{\tilde{x}}(x) := \mathrm{P}\left(\tilde{x}[1] \leq x[1], \tilde{x}[2] \leq x[2], \ldots, \tilde{x}[d] \leq x[d]\right)$$

Computing probabilities

$$\mathrm{P}\left(a_1 < \tilde{a} \leq a_2, b_1 < \tilde{b} \leq b_2\right)$$
$$= F_{\tilde{a},\tilde{b}}\left(a_2, b_2\right) - F_{\tilde{a},\tilde{b}}\left(a_1, b_2\right) - F_{\tilde{a},\tilde{b}}\left(a_2, b_1\right) + F_{\tilde{a},\tilde{b}}\left(a_1, b_1\right)$$

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



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

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

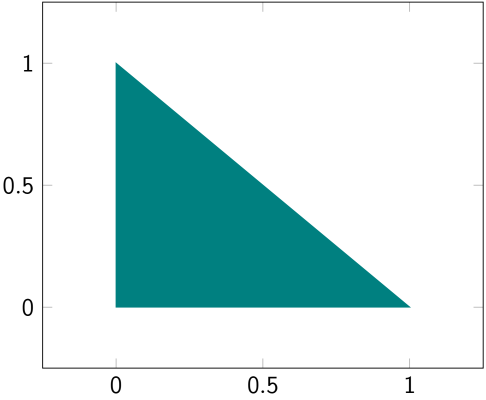
For any 2D Borel set $B \subseteq \mathbb{R}^2$

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

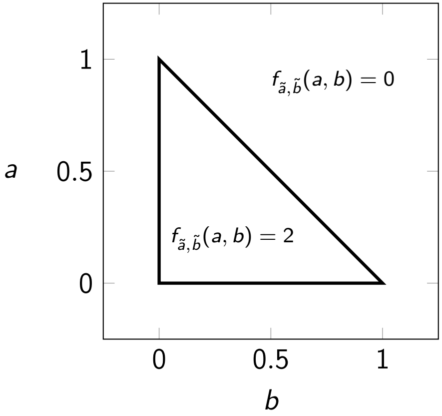
For any d-dimensional Borel set $B \subseteq \mathbb{R}^d$

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

Triangle lake: Joint pdf?



Triangle lake

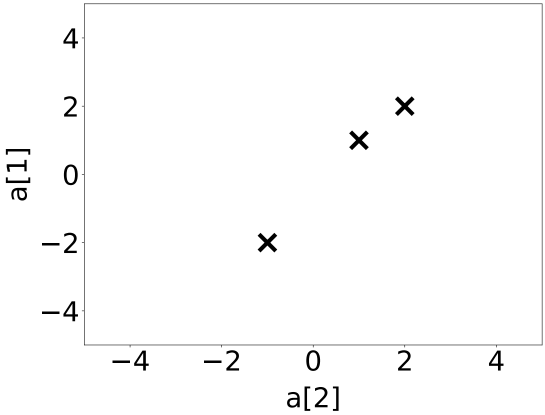


$$\mathrm{P}(\{\tilde{a} \geq 0.6, \tilde{b} \leq 0.2\}) = \int_{b=0}^{0.2} \int_{a=0.6}^{1-b} 2 \, \mathrm{d}a \, \mathrm{d}b$$
$$= \int_{b=0}^{0.2} 2\,(0.4 - b)\, \mathrm{d}b = 0.12$$

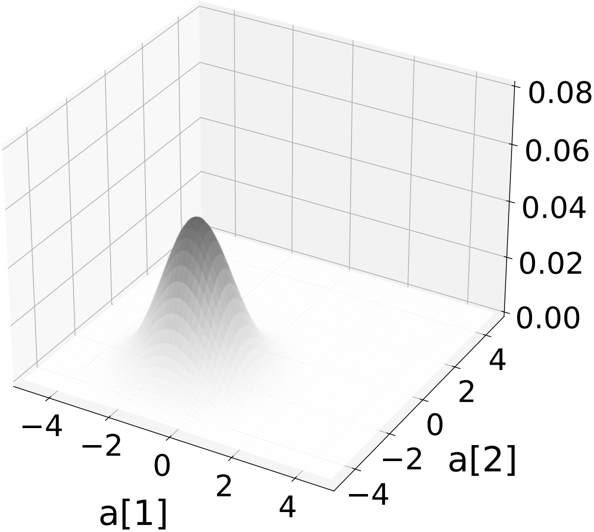
Estimating a pdf from data

We need it to be nonnegative and integrate to one

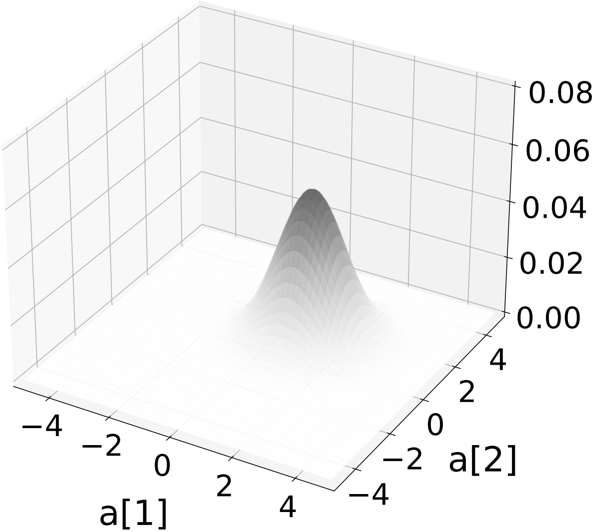
Multidimensional KDE



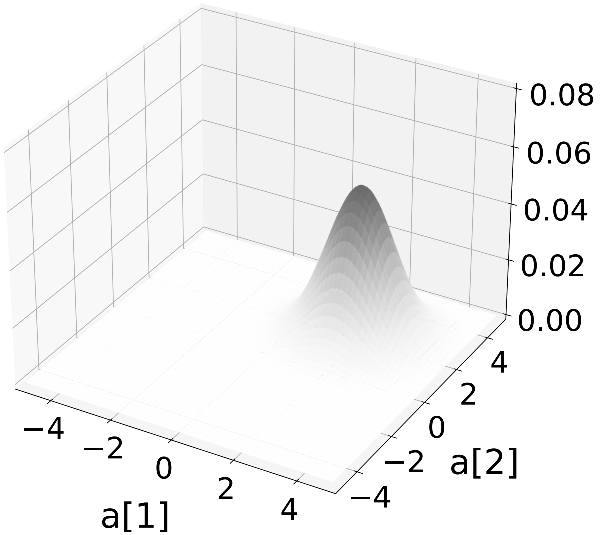
$$\frac{K(a - x_1)}{3}$$



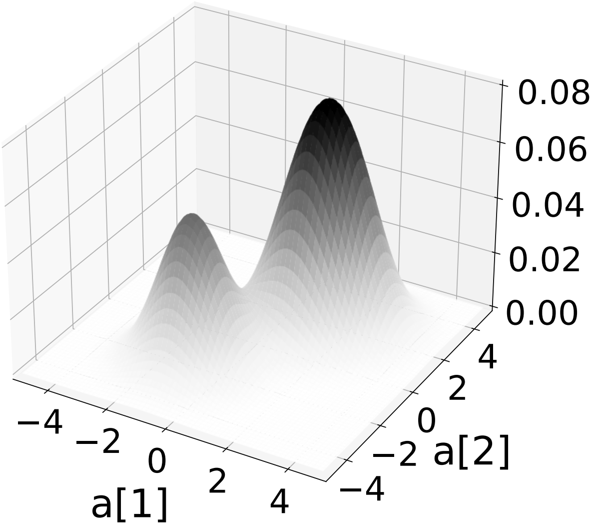
$\frac{K(a-x_2)}{3}$



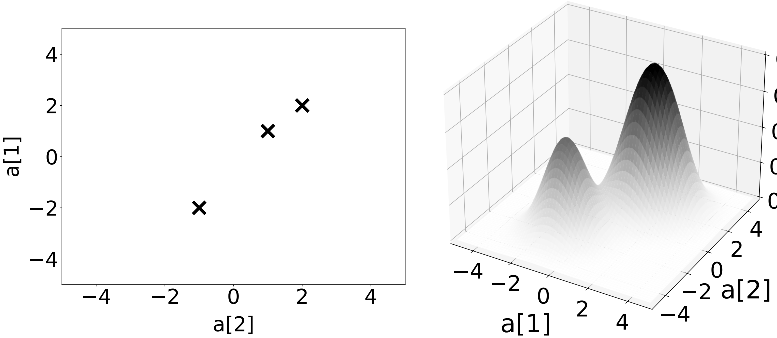
$$\frac{K(a - x_3)}{3}$$



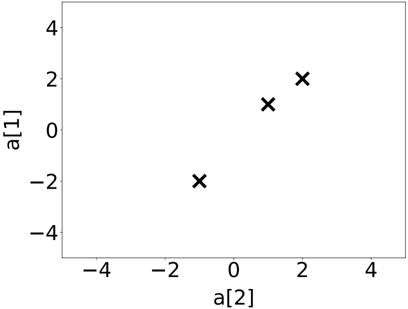
$$\frac{K(a-x_1)+K(a-x_2)+K(a-x_3)}{3}$$

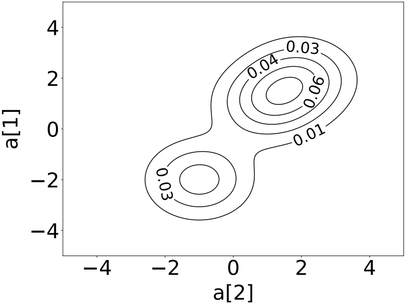


$$\frac{K(a-x_1)+K(a-x_2)+K(a-x_3)}{3}$$



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





Multidimensional KDE

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

Kernel density estimate with bandwidth h is

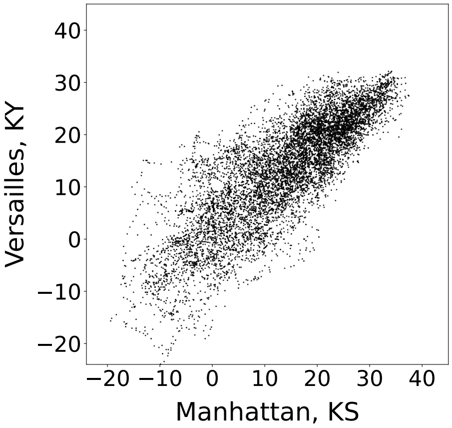
$$f_{X,h}(a) := \frac{1}{n\,h^d} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right)$$

where K is a kernel that satisfies

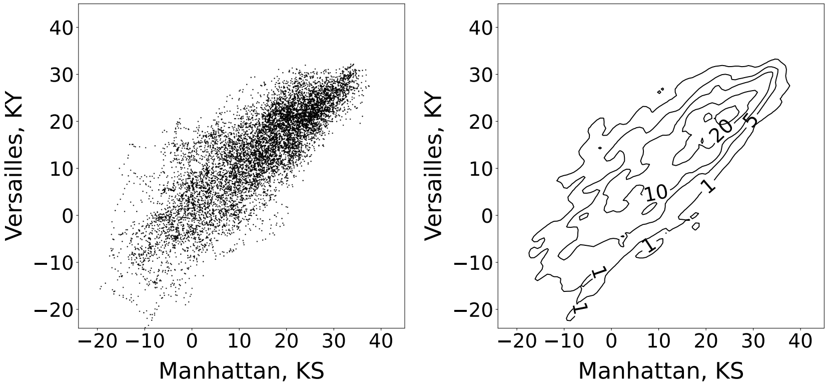
$$K(a) \geq 0 \quad \text{for all } a \in \mathbb{R}^d,$$
$$\int_{\mathbb{R}^d} K(a)\, \mathrm{d}x = 1$$

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

Temperature

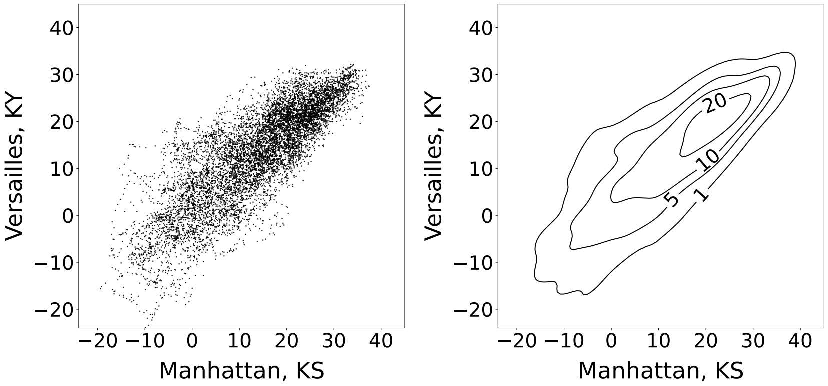


KDE ($h = 0.1$)



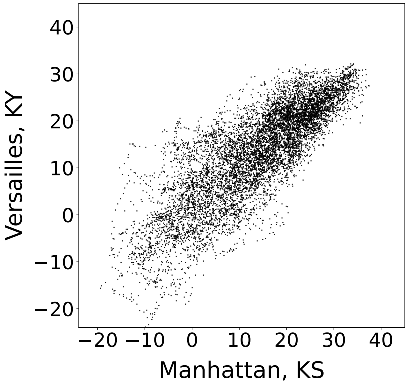
Versailles, KY
Manhattan, KS
40
30
20
10
0
−10
−20
−20
−10
0
10
20
30
40
Versailles, KY
Manhattan, KS
1
5
10
20

KDE (h = 0.25)

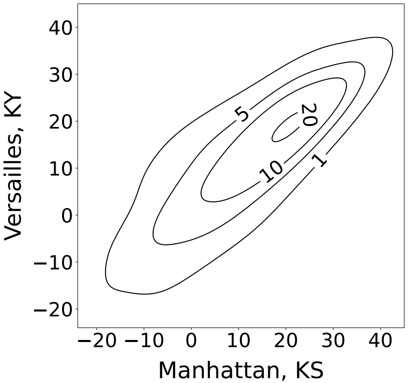


40
30
20
10
0
−10
−20
−20
−10
0
10
20
30
40
Versailles, KY
Manhattan, KS
20
10
5
1

KDE (h = 0.5)

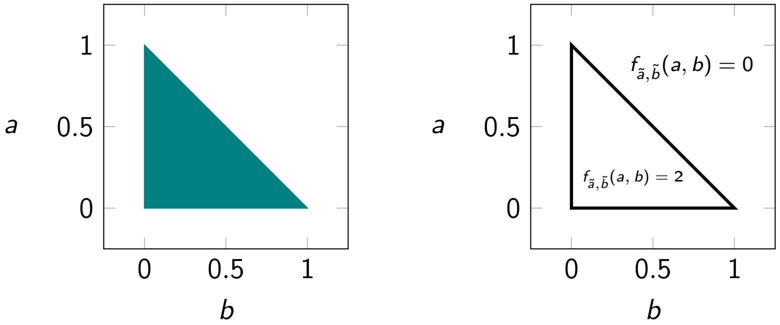


Versailles, KY
Manhattan, KS
40
30
20
10
0
−10
−20
−20
−10
0
10
20
30
40



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

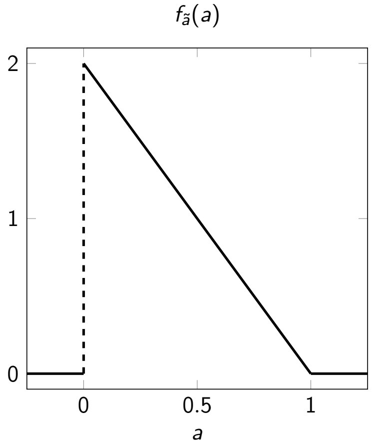
Triangle lake: Joint pdf



What if we only care about $\tilde{a}$?

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

Marginal pdf



Marginal pdf

Marginal pdf of $\tilde{a}$

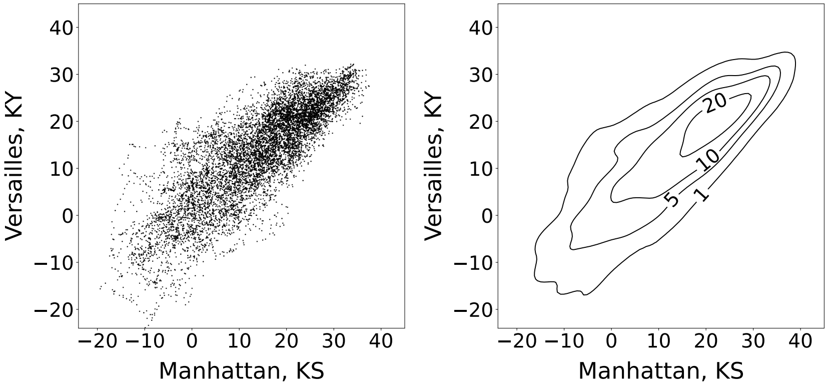
$$f_{\tilde{a}}(a) = \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\,\mathrm{d}b$$

Marginal pdf of $\tilde{x}[i]$

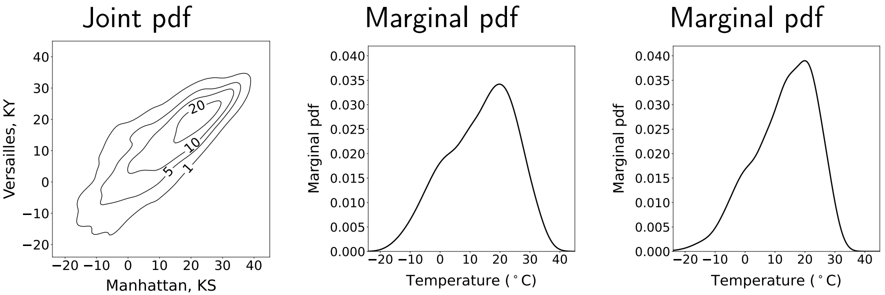
$$f_{\tilde{x}[i]}(a) =$$

$$\int_{x[1]} \cdots \int_{x[i-1]} \int_{x[i+1]} \cdots \int_{x[d]} f_{\tilde{x}}(x[1],\ldots,x[i-1],a,x[i+1],\ldots,x[d])\,\mathrm{d}x[1]\ldots\mathrm{d}x[i-1]\,\mathrm{d}x[i+1]\ldots\mathrm{d}x[d]$$

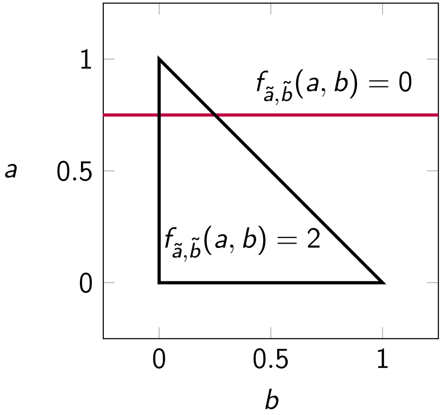
Temperature



Marginal distributions



What if we know that $\tilde{a} = 0.75$?



Conditional pdf

Conditional pdf of $\tilde{b}$ given $\tilde{a}$

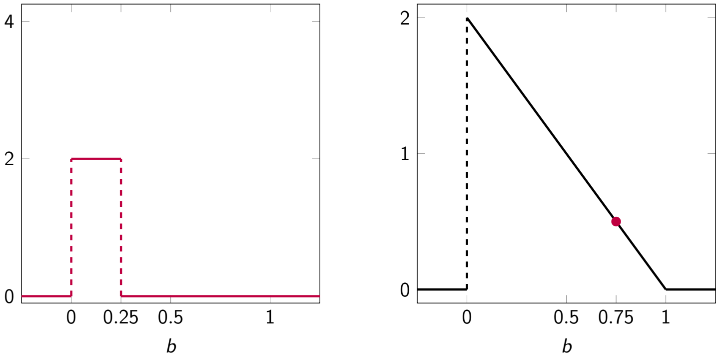
$$f_{\tilde{b}\,|\,\tilde{a}}\left(b\,|\,a\right) := \frac{f_{\tilde{a},\tilde{b}}\left(a,b\right)}{f_{\tilde{a}}\left(a\right)} \quad \text{if } f_{\tilde{a}}\left(x\right) > 0$$

Conditional pdf of $\tilde{x}[i]$ given $\tilde{x}[j] = a_j$ for $j \neq i$

$$\begin{aligned} & f_{\tilde{x}[i]\,|\,\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}\big(b\,|\,a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_d\big) \\ & = \frac{f_{\tilde{x}}\big(a_1,\ldots,a_{i-1},b,a_{i+1},\ldots,a_d\big)}{f_{\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}\big(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_d\big)} \end{aligned}$$

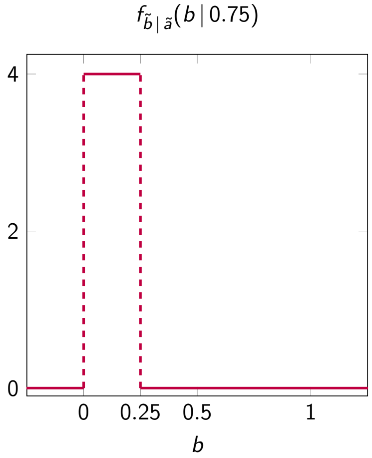
Conditional pdf

$f_{\tilde{a},\tilde{b}}(0.75, b)$ $\qquad$ $f_{\tilde{a}}(0.75)$



$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = \frac{f_{\tilde{a},\tilde{b}}(a,b)}{f_{\tilde{a}}(a)} = \frac{1}{1-a} \qquad b \in [0, 1-a]$$

Conditional pdf



Chain rule

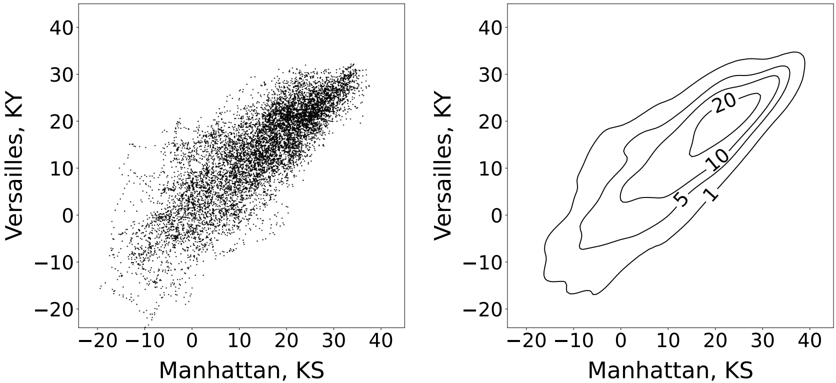
$$
\begin{aligned}
f_{\tilde{a},\tilde{b}}(a,b) &= f_{\tilde{a}}(a)\, f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) \\
&= f_{\tilde{b}}(b)\, f_{\tilde{a}\,|\,\tilde{b}}(a\,|\,b)
\end{aligned}
$$

Chain rule

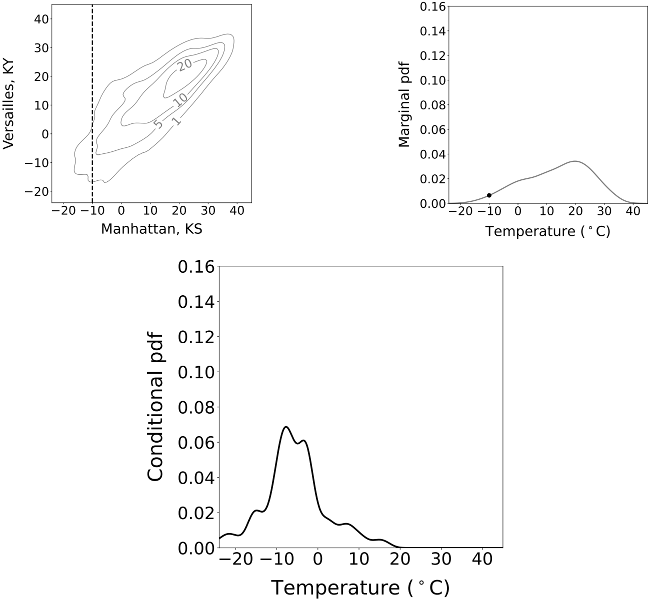
$$f_{\tilde{x}}(x) = f_{\tilde{x}[1]}(x[1]) \prod_{i=1}^{n} f_{\tilde{x}[i] \mid \tilde{x}[1],\ldots,\tilde{x}[i-1]}(x[i] \mid x[1],\ldots,x[i-1])$$

Any order works!

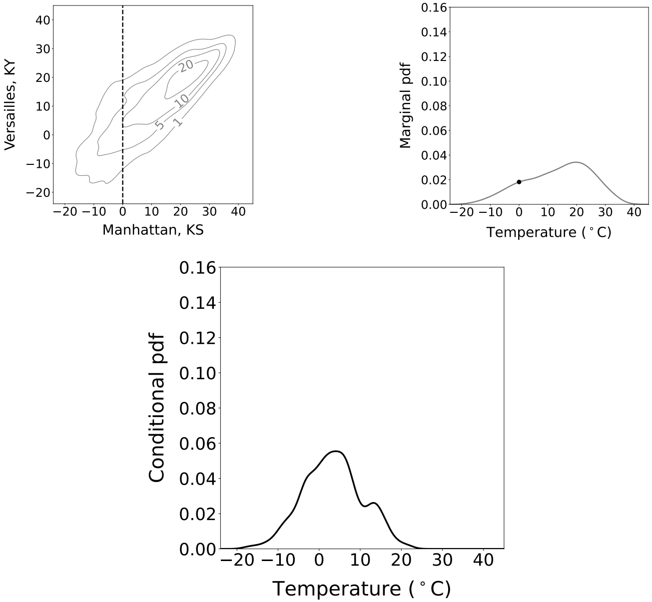
Temperature



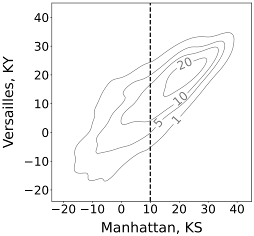
-10° in Manhattan

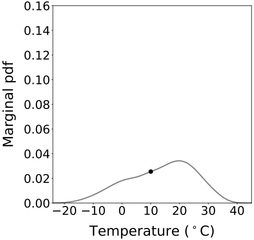


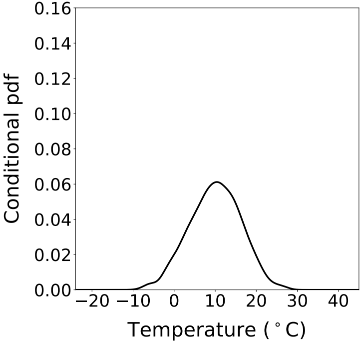
0° in Manhattan



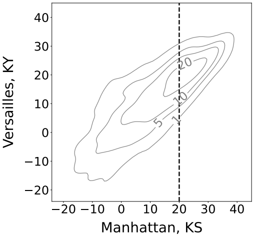
10° in Manhattan

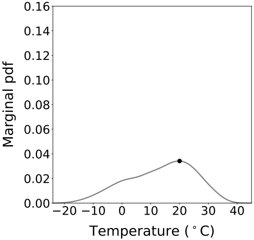


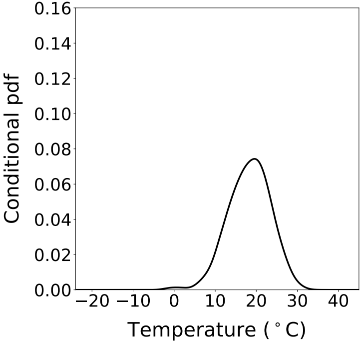




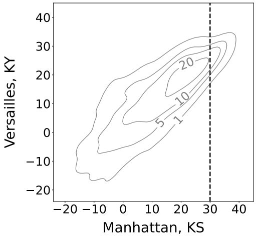
20° in Manhattan

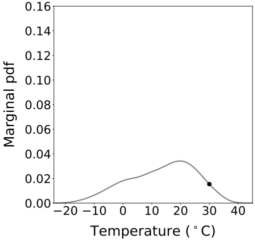


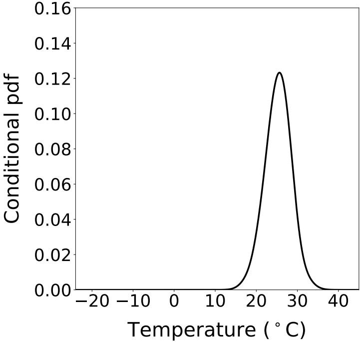




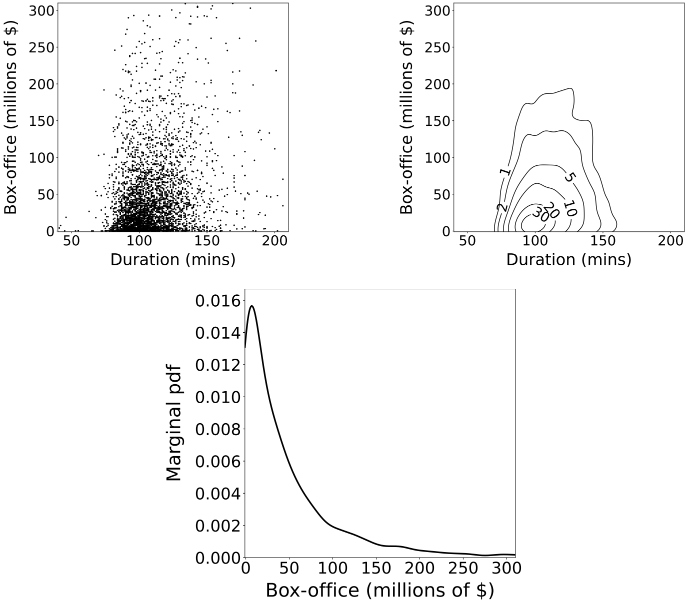
30° in Manhattan



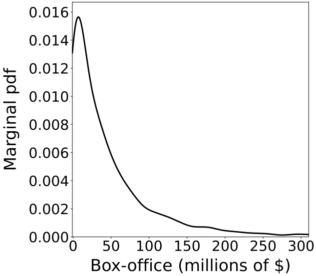


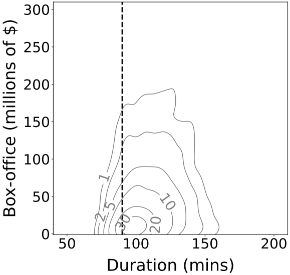


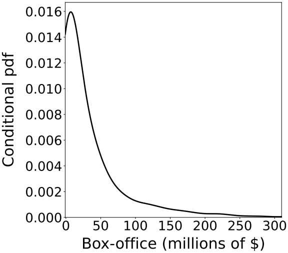
Movie length and box office



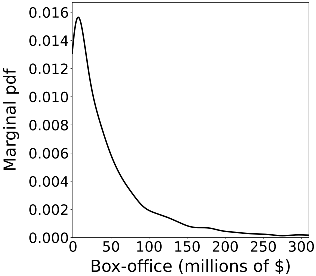
Duration = 90 mins

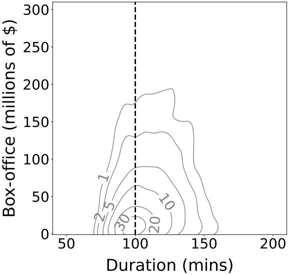


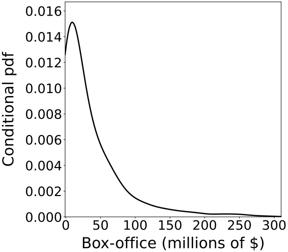




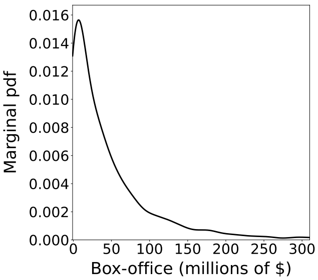
Duration = 100 mins

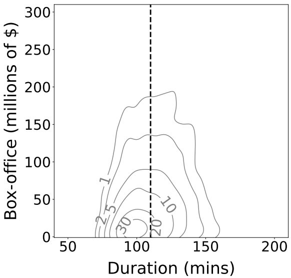


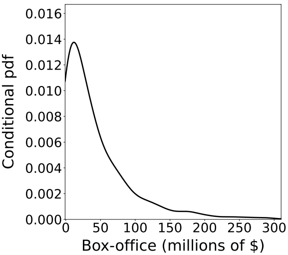




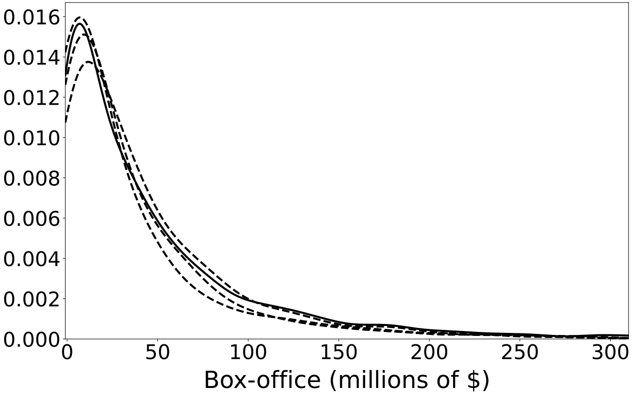
Duration = 110 mins







Marginal and conditional pdfs



Independence

The random variables $\tilde{a}$ and $\tilde{b}$ are independent if for any Borel set S and any b

$$P(\tilde{a} \in S \,|\, \tilde{b} = b) = \mathrm{P}(\tilde{a} \in S)$$

Equivalently,

$$\begin{aligned} F_{\tilde{a}\,|\,\tilde{b}}(a \,|\, b) &= P(\tilde{a} \leq a \,|\, \tilde{b} = b) \\ &= P(\tilde{a} \leq a) \\ &= F_{\tilde{a}}(a) \end{aligned}$$

$$f_{\tilde{a}\,|\,\tilde{b}}(a \,|\, b) = f_{\tilde{a}}(a)$$

Independence

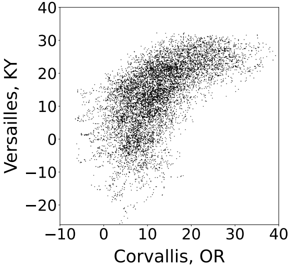
$\tilde{a}$ and $\tilde{b}$ are independent if for any a and b

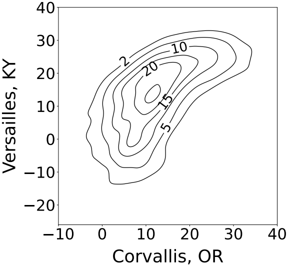
$$f_{\tilde{a},\tilde{b}}(a, b) = f_{\tilde{a}}(a) f_{\tilde{b}}(b)$$

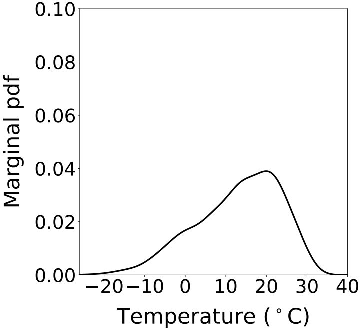
The entries of $\tilde{x}$ are independent if for all x

$$f_{\tilde{x}}(x) = \prod_{i=1}^{d} f_{\tilde{x}[i]}(x[i])$$

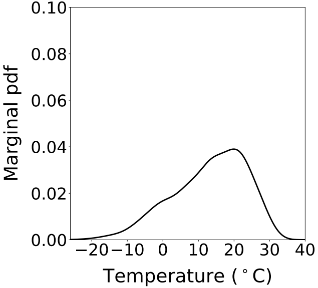
Temperature

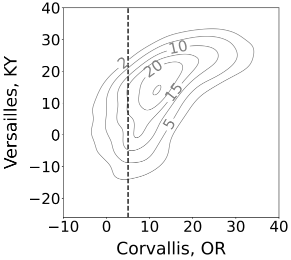


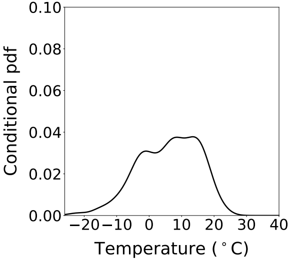




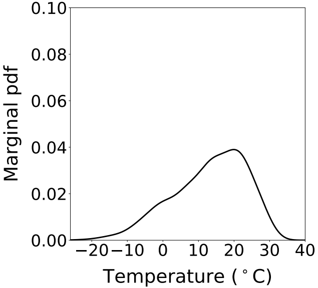
Corvallis = 5°C

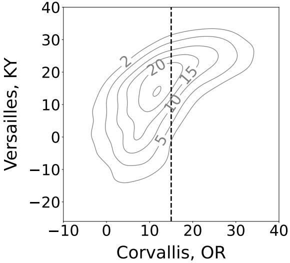


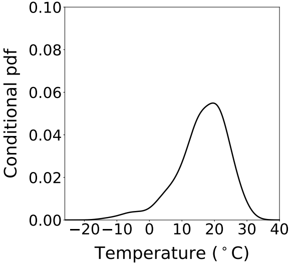




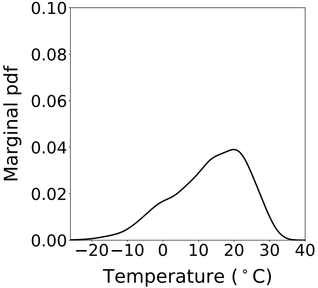
Corvallis = 15°C

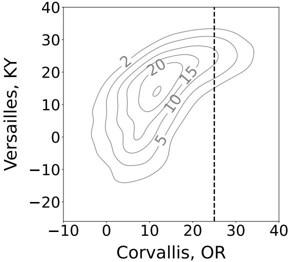


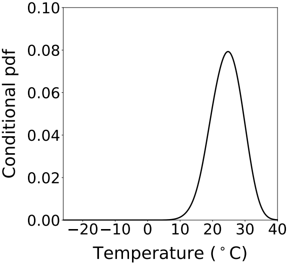




Corvallis = 25°C





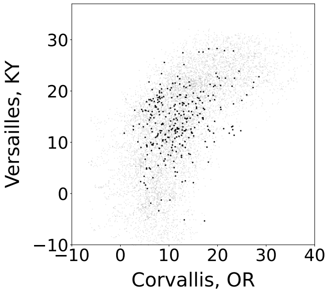


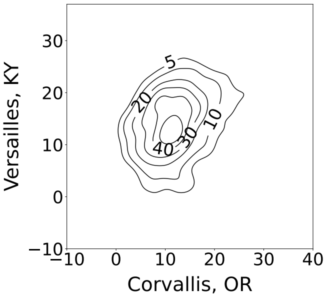
Let us condition on Manhattan

Versailles ($\tilde{v}$) and Corvallis ($\tilde{c}$) given Manhattan ($\tilde{m}$)

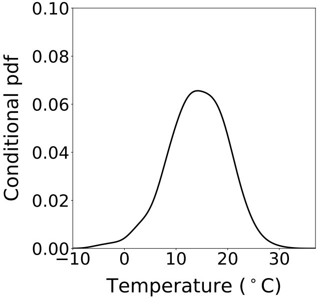
$$f_{\tilde{v},\tilde{c}\,|\,\tilde{m}}(v, c \,|\, t) = \frac{f_{\tilde{v},\tilde{c},\tilde{m}}(v, c, t)}{f_{\tilde{m}}(t)}$$

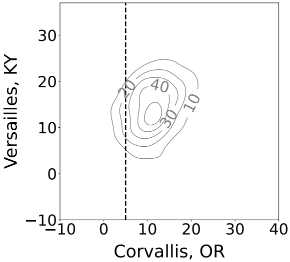
Manhattan = 15°C

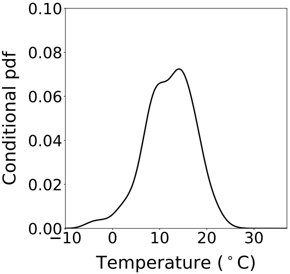




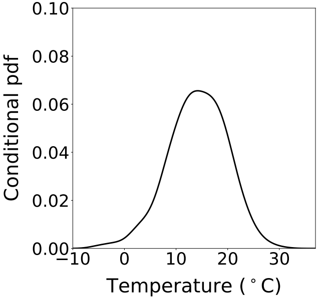
Manhattan = 15°C, Corvallis = 5°C

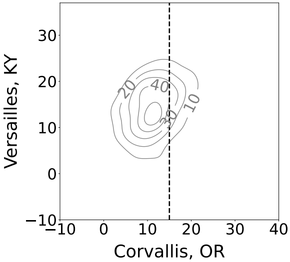


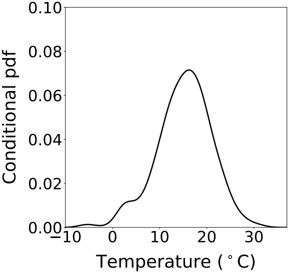




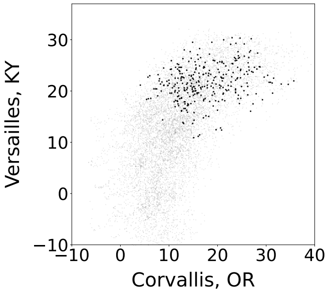
Manhattan = 15°C, Corvallis = 15°C

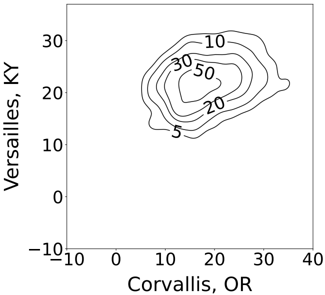




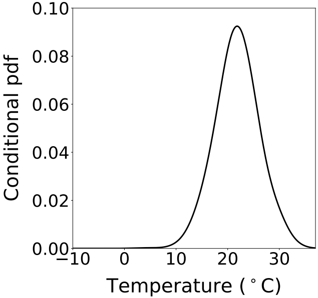


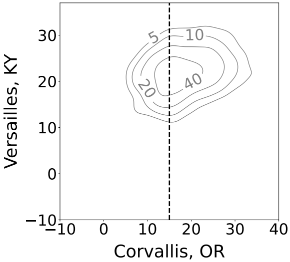
Manhattan = 25°C

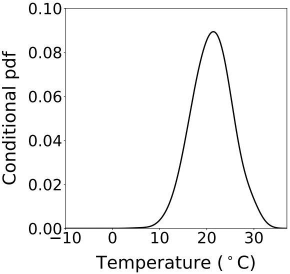




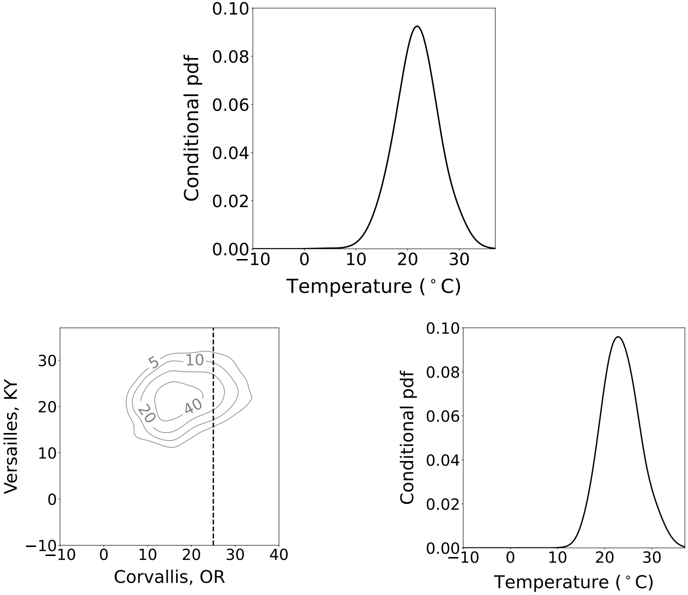
Manhattan = 25°C, Corvallis = 15°C







Manhattan = 25°C, Corvallis = 25°C



Corvallis, Manhattan, Versailles

Conditional independence

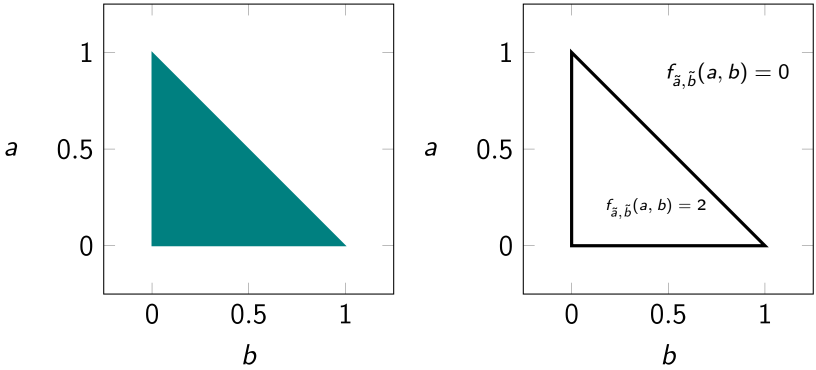
$\tilde{a}$ and $\tilde{b}$ are conditionally independent given $\tilde{c}$ if and only if

$$f_{\tilde{a},\tilde{b}\,|\,\tilde{c}}\left(a,b\,|\,c\right) = f_{\tilde{a}\,|\,\tilde{c}}\left(a\,|\,c\right) f_{\tilde{b}\,|\,\tilde{c}}\left(b\,|\,c\right) \qquad \text{for all } a,b,c$$

$\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d_1]$ are conditionally independent given $\tilde{y}$ if and only if

$$f_{\tilde{x}\,|\,\tilde{y}}\left(x\,|\,y\right) = \prod_{i=1}^{d} f_{\tilde{x}[i]\,|\,\tilde{y}}\left(x[i]\,|\,y\right), \quad \text{for all } x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}$$

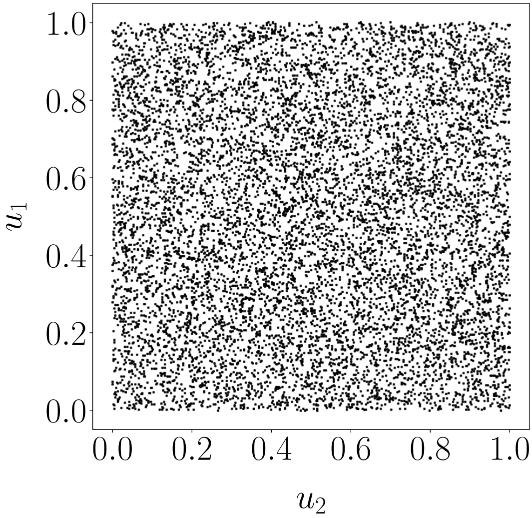
Simulating the triangle lake



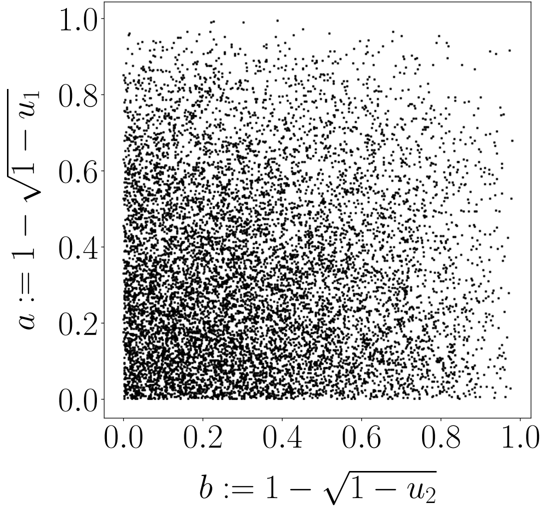
First idea

Obtain two uniform samples u_1 and u_2 and simulate $\tilde{a}$ and $\tilde{b}$ using their respective marginal cdfs

Uniform samples



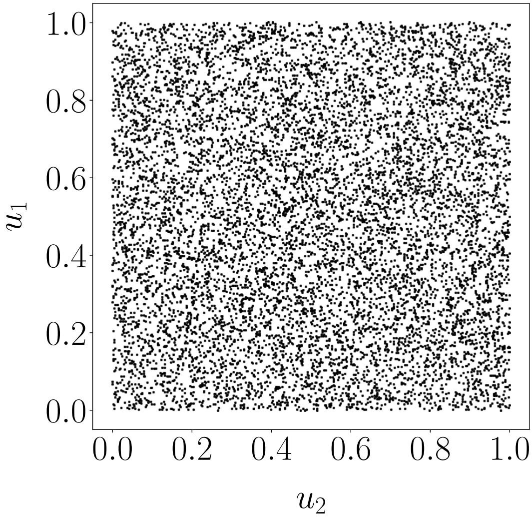
First idea



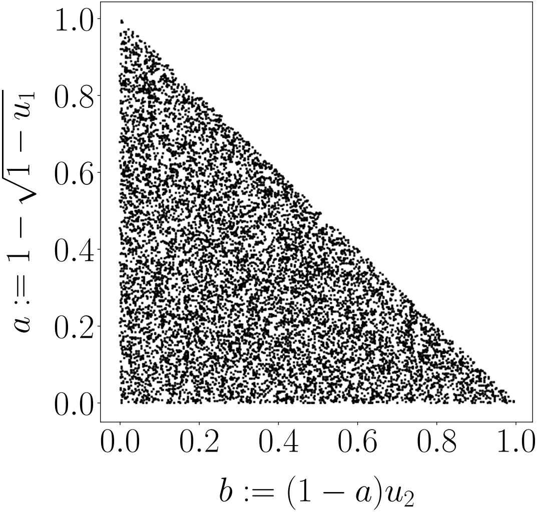
Second idea

First simulate sample a from $\tilde{a}$ and then simulate b by sampling from conditional distribution of $\tilde{b}$ given $\tilde{a} = a$

Uniform samples



Second idea



Gaussian parametric model

Goal: Use Gaussian distribution to model random vector $\tilde{x}$

Motivation: Curse of dimensionality

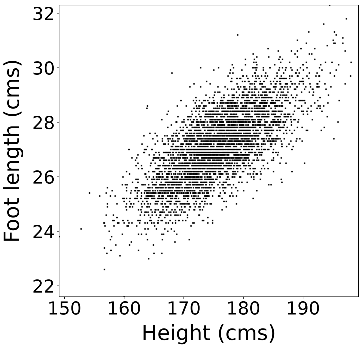
First idea

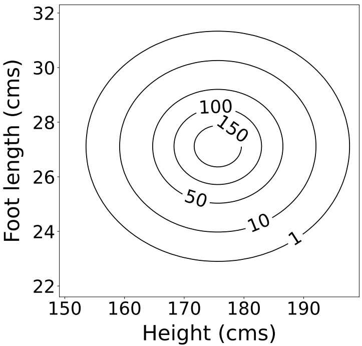
Model entries of $\tilde{x}$ as d independent Gaussian random variables with means $\mu_1, \mu_2, \ldots, \mu_d$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_d$

Joint pdf

$$
\begin{aligned}
f_{\tilde{x}}(x) &= \prod_{i=1}^{d} f_{\tilde{x}[i]}\left(\tilde{x}[i]\right) \\
&= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x[i]-\mu_i)^2}{2\sigma_i^2}\right) \\
&= \frac{1}{(2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \frac{(x[i]-\mu_i)^2}{\sigma_i^2}\right)
\end{aligned}
$$

Height and foot length





Contour surfaces

$$\left\{x \in \mathbb{R}^d \mid f_{\tilde{x}}(x) = c\right\} = \left\{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} \frac{(x[i] - \mu_i)^2}{\sigma_i^2} = c'\right\}$$

where $c' = -2 \log \left(c \, (2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sigma_i \right)$

Shape? Ellipsoid with axes along coordinate axes

How can we improve the model?

Including rotations

Additional parameters: Axes of ellipsoid $u_1, u_2, \ldots, u_d$

$$
\begin{aligned}
c' &= \sum_{i=1}^{d} \frac{{u_i}^T (x-\mu)^2}{\sigma_i^2} \\
&= (x-\mu)^T U \Lambda^{-1} U^T (x-\mu) \\
&= (x-\mu)^T \Sigma^{-1} (x-\mu)
\end{aligned}
$$

$$
U := \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \qquad \Lambda := \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}
$$

Covariance-matrix parameter $\Sigma := U \Lambda U^T$

Joint pdf

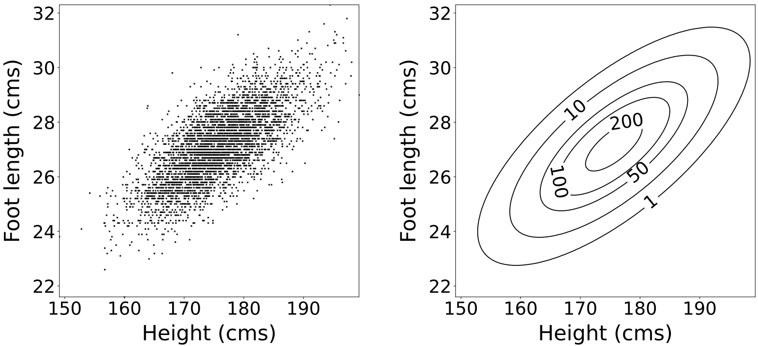
Without rotation:

$$f_{\tilde{x}}(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \frac{(x[i]-\mu_i)^2}{\sigma_i^2}\right)$$

With rotation:

$$f_{\tilde{x}}\left(x\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right)$$

With rotation



Gaussian random vector

A Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ is the mean and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix

$\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive definite (positive eigenvalues)

Properties

Marginal and conditional distributions are Gaussian

Maximum-likelihood estimation

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

$$\mu_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Sigma_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})(x_i - \mu_{\mathrm{ML}})^T$$

What have we learned

- Mathematical definition
- Joint probability density
- Marginal and conditional distributions
- Gaussian random vectors