# Some Concepts of Probability (Review)

Volker Tresp

Summer 2021

## Definition

- There are different way to define what a probability stands for
- Mathematically, the most rigorous definition is based on Kolmogorov axioms and probability theory is a mathematical discipline
- For beginners it is more important to obtain an intuition, and the definition I present is based on a relative frequency; in statistics, probability theory is applied to problems of the real world
- We start with an example

## Example: Students in Munich

- Let's assume that there are $\tilde{N} = 50000$ students in Munich. This set is called the *population*
- $\tilde{N}$ is the size of the population, often assumed to be infinite
- Formally, I put the all 50000 students in an urn (bag)
- I randomly select a student: this is called an *(atomic) event* or an *experiment* and defines a *random process*
- The selected student is an *outcome* of the experiment

# Sample

- A particular student will be picked with elementary probability $1/\tilde{N}$
- Performing the experiment $N$ times produces a sample (training data set) $D$ of size $N$
- An analysis of the sample can give us insight about the population (statistical inference)
- Sampling *with replacement*: I return the student to the urn after the experiment
- Sampling *without replacement*: I do not return the student to the urn after the experiment

# Random Variable

- A particular student has a height attribute *(tiny, small, medium, large, huge)*
- The height $H$ is called a *random variable* with states
  $h \in \{$*tiny, small, medium, large, huge*$\}$
- A random variable is a variable (more precisely a function of the outcome of the random experiment), whose value depends on the result of a random process
- Thus at each experiment I measure a particular $h$

# Probability

- Then the *probability* that a randomly picked student has height $H = h$ is defined as

$$P(H = h) = \lim_{N \to \infty} \frac{N_h}{N}$$

  with $0 \leq P(H = h) \leq 1$

- $N_h$ is the number of times that a selected student is observed to have height $H = h$

# Sample / Training Data

- I can estimate

$$\widehat{P}(H = h) = \frac{N_h}{N} \approx P(H = h)$$

- In statistics one is interested in how well $\widehat{P}(H = h)$ (the probability estimate derived from the sample) approximates $P(H = h)$ (the probability in the population)

- Note the importance of the definition of a population: $P(H = h)$ might be different, when I consider individuals in Munich or Germany

- Thus the population plays an important role in a statistical analysis

## Statistics and Probability

- *Probability* is a mathematical discipline developed as an abstract model and its conclusions are *deductions* based on *axioms* (Kolmogorov axioms)
- *Statistics* deals with the application of the theory to real problems and its conclusions are *inferences* or *inductions*, based on observations (Papoulis: Probability, Random variables, and Stochastic Processes)
- *Frequentist or classical statistics* and *Bayesian statistics* apply probability in slightly different ways

# Joint Probabilities

- Now assume that we also measure weight (size) $S$ with weight attributes *very light, light, normal, heavy, very heavy*. Thus $S$ is a second random variable

- Similarly

$$P(S = s) = \lim_{N \to \infty} \frac{N_s}{N}$$

- We can also count co-occurrences

$$P(H = h, S = s) = \lim_{N \to \infty} \frac{N_{h,s}}{N}$$

This is called the *joint probability distribution* of $H$ and $S$

# Marginal Probabilities

- It is obvious that we can calculate the *marginal probability* $P(H = h)$ from the joint probabilities

$$P(H = h) = \lim_{N \to \infty} \frac{\sum_s N_{h,s}}{N}$$

$$= \sum_s P(H = h, S = s)$$

- This is called marginalization
- I can calculate the marginal probability from the joint probability (without going back to the counts)

## Conditional Probabilities

- One is often interested in the *conditional probability*. Let's assume that I am interested in the probability distribution of $S$ for a given height $H = h$. Since I need a different normalization I get

$$P(S = s|H = h) = \lim_{N \to \infty} \frac{N_{h,s}}{N_h}$$

So I count the co-occurrences, but I normalize by $N_h$

# Conditional Probabilities (cont'd)

- Then,

$$P(S = s|H = h) = \frac{P(H = h, S = s)}{P(H = h)}$$

- Relationship to machine learning: $H = h$ is the *input* and $S = s$ is the *output*
- Conditioning is closely related to the definition of a population: $P(S = s|H = h)$ is the same as $P(S = s)$ in a population which is restricted to students with $H = h$

# Product Rule and Chain Rule

- It follows: **product rule**

$$P(S = s, H = h) = P(S = s|H = h)P(H = h)$$

$$= P(H = h|S = s)P(S = s)$$

- and **chain rule**

$$P(x_1, \ldots, x_M) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \ldots P(x_M|x_1, \ldots, x_{M-1})$$

## Bayes Formula

- If I know $P(S = s|H = h)$, does it tell me anything about $P(H = h|S = s)$? Is it the same thing?

- No, but the relationship is given by Bayes formula

# Bayes Formula (con't)

- We use the definition of a conditional probability,

$$P(H = h|S = s) = \frac{P(H = h, S = s)}{P(S = s)}$$

$$P(S = s|H = h) = \frac{P(H = h, S = s)}{P(H = h)}$$

- Thus we get *Bayes' formula*

$$P(H = h|S = s) = \frac{P(S = s|H = h)P(H = h)}{P(S = s)}$$

or

$$P(H = h|S = s) = P(S = s|H = h)\frac{P(H = h)}{P(S = s)}$$

# Independent Random Variables

- **Independence**: two random variables are independent, if,

$$P(S = s, H = h) = P(S = s)P(H = h|S = s)$$

$$= P(S = s)\ P(H = h)$$

# Summary

- Conditional probability

$$P(y|x) = \frac{P(x,y)}{P(x)} \quad \text{with} \quad P(x) > 0$$

- Product rule

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Chain rule

$$P(x_1, \ldots, x_M) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_M|x_1,\ldots,x_{M-1})$$

- Bayes' theorem

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)} \quad P(x) > 0$$

- Marginal distribution

$$P(x) = \sum_y P(x,y)$$

- Independent random variables

$$P(x, y) = P(x)P(y|x) = P(x)P(y)$$

## Marginalization and Conditioning: Basis for Probabilistic Inference

- $P(I, F, S)$ where $I = 1$ stands for influenza, $F = 1$ stands for fever, $S = 1$ stands for sneezing

- What is the probability for influenza, when the patient is sneezing, but temperature is unknown, $P(I|S)$?

- Thus I need (conditioning) $P(I = 1|S = 1) = P(I = 1, S = 1)/P(S = 1)$

- I calculate via marginalization

$$P(I = 1, S = 1) = \sum_f P(I = 1, F = f, S = 1)$$

$$P(S = 1) = \sum_i P(I = i, S = 1)$$

# Expected Values

- **Expected value**

$$E(X) = E_{P(x)}(X) = \sum_i x_i P(X = x_i)$$

# Variance

- The **Variance** of a random variable is:

$$var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

- The **Standard Deviation** is its square root:

$$stdev(X) = \sqrt{var(x)}$$

# Covariance

- **Covariance**:

$$cov(X,Y) = \sum_i \sum_j (x_i - E(X))(y_j - E(Y))P(X = x_i, Y = y_j)$$

- **Covariance matrix**:

$$\Sigma_{[XY],[XY]} = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(Y,X) & var(Y) \end{pmatrix}$$

## Covariance, Correlation, and Correlation Coefficient

- Useful identity:

$$cov(X, Y) = E(XY) - E(X)E(Y)$$

  where $E(XY)$ is the **correlation**.

- The **(Pearson) correlation coefficient** (confusing naming!) is

$$r = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

- It follows that $var(X) = E(X^2) - (E(X))^2$ and

$$var(f(X)) = E(f(X)^2) - (E(f(X)))^2$$

## More Useful Rules

- We have, independent of the correlation between $X$ and $Y$,

$$E(X+Y) = E(X) + E(Y)$$

and thus also

$$E(X^2+Y^2) = E(X^2) + E(Y^2)$$

- For the variance of the sum of random variables,

$$var(X+Y) = E[(X+Y-(E(X)+E(Y)))^2]$$

$$= E[((X-E(X))+(Y-E(Y)))^2]$$

$$= E[(X-E(X))^2] + E[(Y-E(Y))^2] + 2E[(X+E(X))(Y-E(Y)]$$

$$= var(X) + var(Y) + 2cov(X,Y)$$

- Similarly,

$$var(X-Y) = var(X) + var(Y) - 2cov(X,Y)$$

# Covariance Matrix of Linear Transformation

- Let $\mathbf{w}$ be a random vector with mean $\vec{\mu}_{\mathbf{w}}$ and covariance matrix $\Sigma_{\mathbf{w}}$
- Let

$$\mathbf{y} = \mathbf{A}\mathbf{w}$$

  where $\mathbf{A}$ is a fixed matrix.

- Then $\mathbf{y}$ is a random vector with mean $\vec{\mu}_y = \mathbf{A}\vec{\mu}_w$ and covariance

$$\Sigma_{\mathbf{y}} = \mathbf{A}\Sigma_{\mathbf{w}}\mathbf{A}^T$$

## Continuous Random Variables

- **Probability density**

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \le X \le x + \Delta x)}{\Delta x}$$

- Thus

$$P(a < x < b) = \int_a^b f(x)dx$$

- The **distribution function** is

$$F(x) = \int_{-\infty}^{x} f(x)dx = P(X \le x)$$

# Expectations for Continuous Variables

- Expected value

$$E(X) = E_{P(x)}(X) = \int xP(x)dx$$

- Variance

$$var(X) = \int (x - E(x))^2 P(x)dx$$

- Covariance:

$$cov(X, Y) = \int \int (x - E(X))(y - E(Y))P(x, y)dxdy$$

# Normal (Gaussian) Distribution

- E.g., height $x$ and weight $y$ are real numbers!
- For $x$,

$$P(x) = \mathcal{N}(x; \mu_x, \Sigma_{x,x}) = \frac{1}{\sqrt{2\pi\Sigma_{x,x}}} \exp\left(-\frac{1}{2}(x-\mu_x)^T \Sigma_{x,x}^{-1}(x-\mu_x)\right)$$

$\mu_x = E(X)$, $\Sigma_{x,x} = var(X)$

- For $y$,

$$P(y) = \mathcal{N}(y; \mu_y, \Sigma_{y,y}) = \frac{1}{\sqrt{2\pi\Sigma_{y,y}}} \exp\left(-\frac{1}{2}(y-\mu_y)^T \Sigma_{y,y}^{-1}(y-\mu_y)\right)$$

$\mu_y = E(Y)$, $\Sigma_{y,y} = var(Y)$

## Joint Distribution for Independent Gaussians

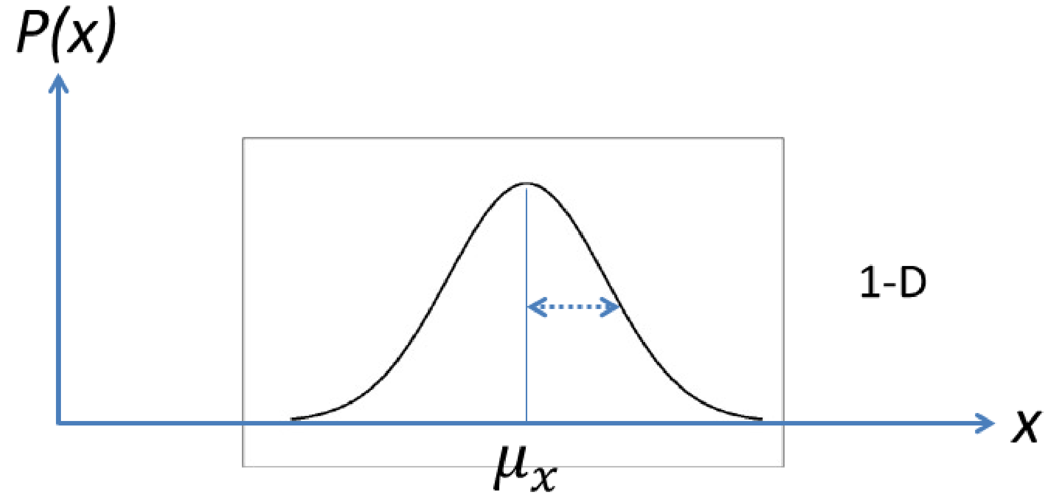
- Assume $X$ and $Y$ are independent; let $\mathbf{z} = (x; y)$, $\vec{\mu} = (\mu_x; \mu_y)$,

$$\Sigma = \begin{pmatrix} \Sigma_{x,x} & 0 \\ 0 & \Sigma_{y,y} \end{pmatrix}$$

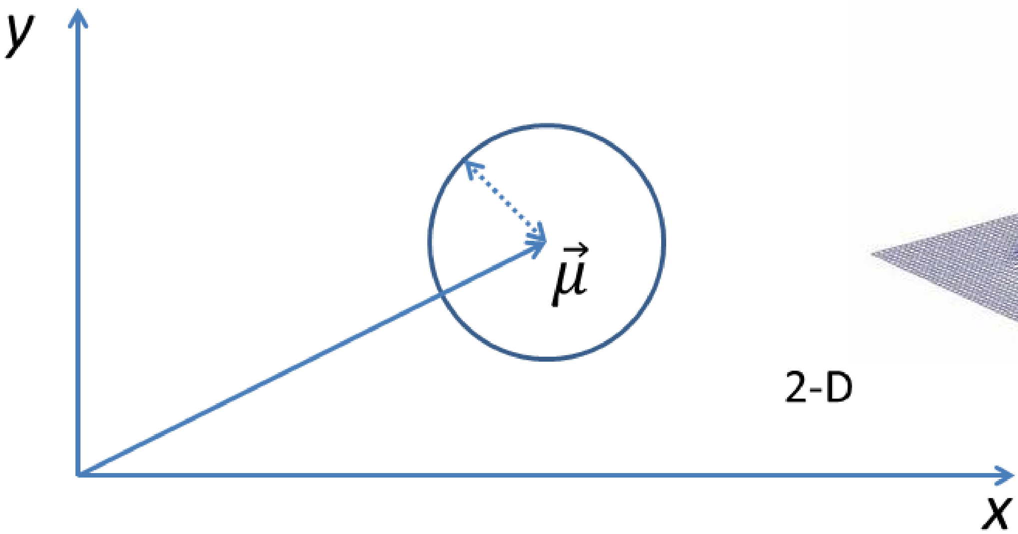
- Then,

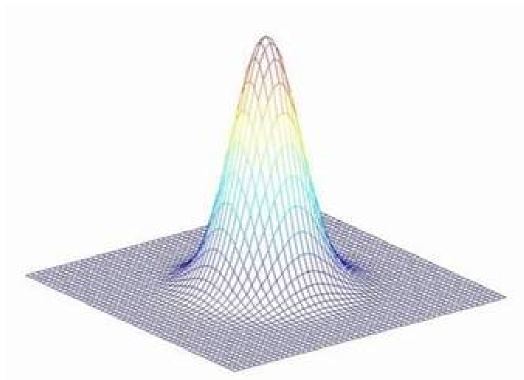
$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{M/2}\sqrt{|\Sigma|}} \exp\left( -\frac{1}{2} (\mathbf{z} - \vec{\mu})^T \Sigma^{-1} (\mathbf{z} - \vec{\mu}) \right)$$

$M$ is the dimensionality, here $M = 2$; $|\Sigma|$ is the determinant, here $|\Sigma| = \Sigma_{x,x}\Sigma_{y,y}$




1-D



2-D

## Modelling Dependencies

- We get,

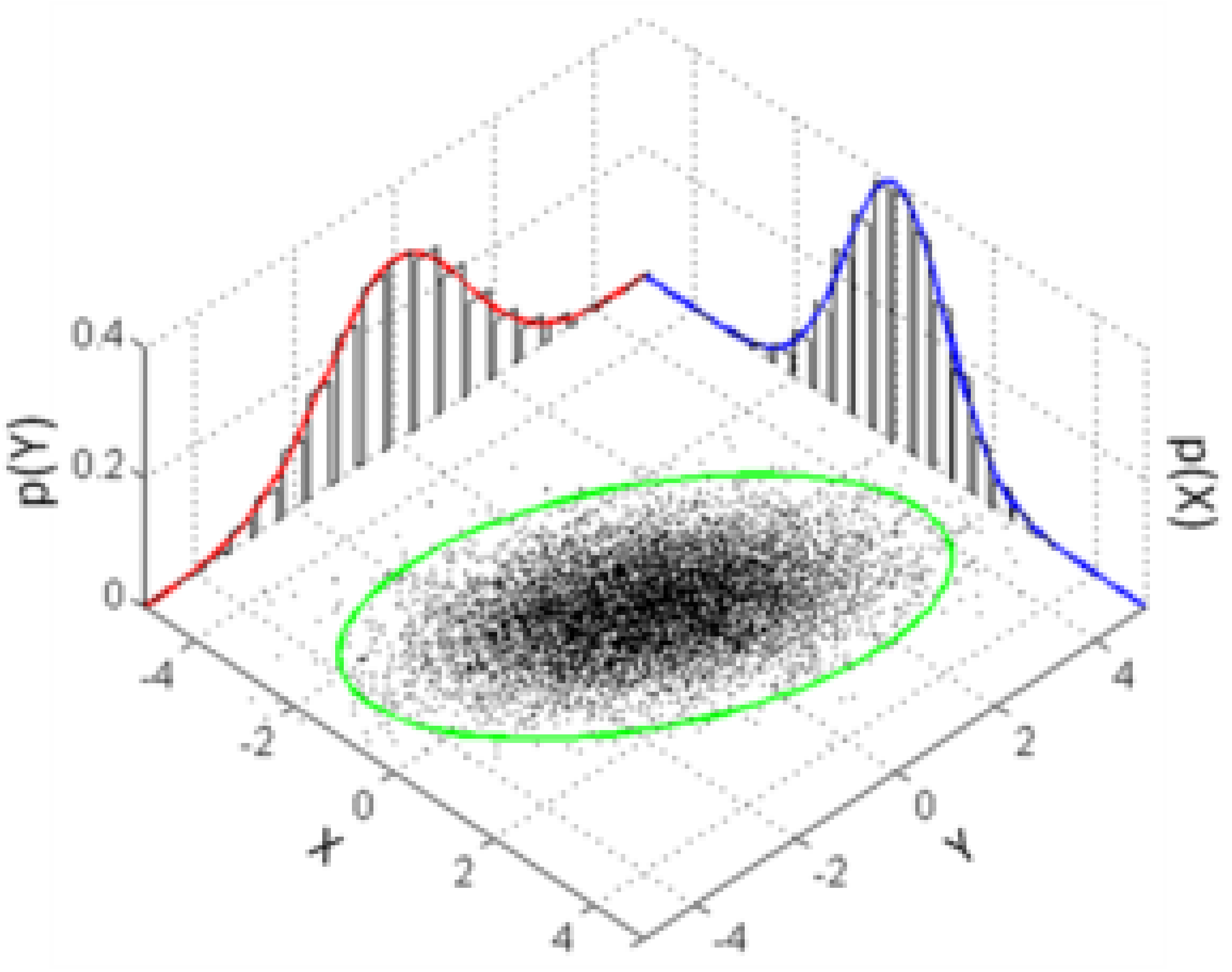
$$\Sigma = \begin{pmatrix} \Sigma_{x,x} & \Sigma_{x,y} \\ \Sigma_{y,x} & \Sigma_{y,y} \end{pmatrix}$$

  and, as before,

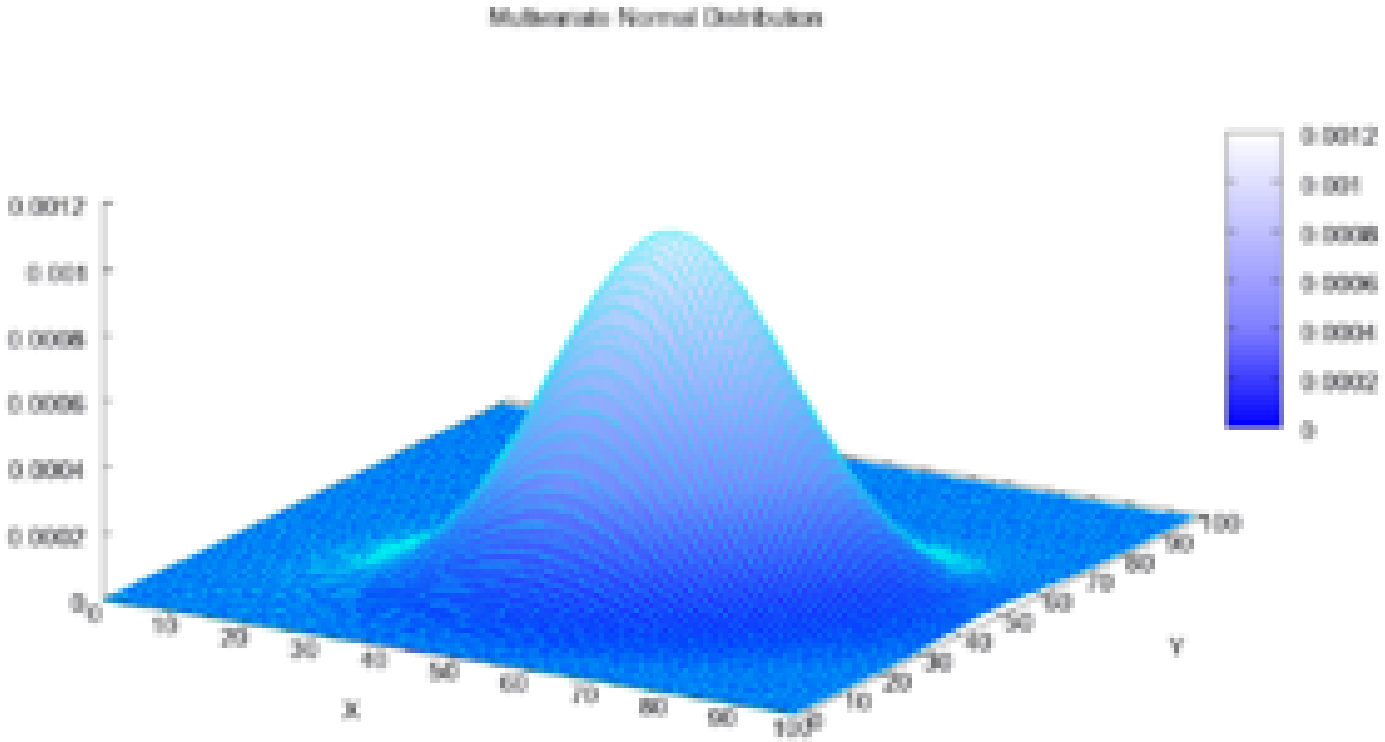
$$P(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}; \vec{\mu}, \Sigma\right) = \frac{1}{(2\pi)^{M/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\left(\mathbf{z} - \vec{\mu}\right)^T \Sigma^{-1}\left(\mathbf{z} - \vec{\mu}\right)\right)$$

  Here, $\Sigma_{x,y} = \Sigma_{y,x} = cov(X, Y)$; here $|\Sigma| = \Sigma_{x,x}\Sigma_{y,y} - \Sigma_{x,y}\Sigma_{y,x}$

- This generalizes to $M > 2$



0.4
0.2
0
p(Y)
p(X)
-4
-2
0
2
4
X
Y



Multivariate Normal Distribution
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
10
20
30
40
50
60
70
80
90
100
X
Y

## Marginal and Conditional Densities

- We already know the marginals
- For the conditionals, we get

$$P(x|y) = \mathcal{N}\left(x; \mu_x + \Sigma_{x,y}\Sigma_{y,y}^{-1}(y - \mu_y), \Sigma_{x,x} - \Sigma_{x,y}\Sigma_{y,y}^{-1}\Sigma_{y,x}\right)$$

- and

$$P(y|x) = \mathcal{N}\left(y; \mu_y + \Sigma_{y,x}\Sigma_{x,x}^{-1}(x - \mu_x), \Sigma_{y,y} - \Sigma_{y,x}\Sigma_{x,x}^{-1}\Sigma_{x,y}\right)$$