# Data Representations and Some Concepts of Probability (Review)

Volker Tresp

Winter 2024-2025

# Discriminant Model

- The probability that the output (random) variable (class probability) $Y$ assumes the value *cat*, given that the input is $\mathbf{x}$, is

$$P(Y = cat|\mathbf{x}) = \mathsf{sig}(f_w(\mathbf{x}))$$

- This is the basis for DNNs: classifying cats from no cats
- $\mathbf{x}$ can be a pixel image (with 1 Mio pixel values) or $M << 1$ Mio features derived from the image
- $P(Y = cat|\mathbf{x}) = 0.9$ is my confidence that $\mathbf{x}$ describes a cat
- $P(Y = cat|\mathbf{x}) = 0.9$ means: If I observe the same input $\mathbf{x}$ 10 times, then in 9 out of ten times it will show a cat

## Preamble

- "Thermodynamik ist ein komisches Fach. Das erste Mal, wenn man sich damit befasst, versteht man nichts davon. Beim zweiten Durcharbeiten denkt man, man haette nun alles verstanden, mit Ausnahme von ein oder zwei kleinen Details. Das dritte Mal, wenn man den Stoff durcharbeitet, bemerkt man, dass man fast gar nichts davon versteht, aber man hat sich inzwischen so daran gewoehnt, dass es einen nicht mehr stoert." Arnold Sommerfeld

## Example: Students in Munich

- Let's assume that there are $\tilde{N} = 50000$ students in Munich. The set of all students in Munich $\Omega$ is called the *population*
- $\tilde{N}$ is the size of the population, often assumed to be infinite
- Formally, I put 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*
- $\omega$: The selected student is an *outcome* of the experiment and defines a **row in the data matrix**; if Jack was selected, then $\omega = Jack$

## Sample

- A particular student will be picked with elementary probability $1/\tilde{N}$
- Performing the experiment $N$ times produces a sample (training data set) $\mathbf{D}$ 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; then, at any time, $P(\omega = Jack) = 1/\tilde{N}$; this is easier to analyse
- Sampling without replacement: I do not return the student to the urn after the experiment; this is how a normal data matrix is formed

# Random Variable

- On each selected student, we perform a measurement, i.e., height $H$, and the result (outcome) of the measurement is a value, e.g., *(tiny, small, medium, tall, huge)*; $H$ is called a random variable

- A random variable (e.g., *Height*) is a function (measurement) of the outcome (e.g., *Jack*) of the random experiment; its value is a function of the outcome; we write

$$\textit{Height(Jack)} = \textit{tall}$$

- Physics view: *Height* is the measurement type, *Jack* the entity on which the measurement is performed, and *tall* is the outcome

- Data matrix (table) view: *Height* is a name of a column in a data matrix, *Jack* the name of the row and *tall* the entry in row and column

## Probability

- In statistics, one estimates the probability from the sample (the training data)
- Then the *probability* that a randomly picked student has height $H = h$ is defined as

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

  with $0 \leq P(H = h) \leq 1$; $N \to \infty$ indicates a sampling with replacement
- $N_h$ is the number of times that a selected student is observed to have height $H = h$

# Sample / Training Data

- I can estimate

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

- This is the number of times that we observe the value of $h$ in column $H$ in the data matrix, divided by the number of observations $N$
- In statistics one is interested in how well $\hat{P}(H = h)$ (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
- Note that the randomness enters through the sampling process: Jack's height is not random

# Law of Large Numbers

- Law of Large Numbers (Bernoulli)

$$P\{|N_h/N - P(H = h)| < \epsilon\} \to 1 \quad \textit{as} \quad N \to \infty$$

## 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*

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

and another ways of writing this:

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

# Evidence

- Evidence

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

- This equation the basis for generative AI: $P(H = h)$ is a simple distribution, $P(S = s|H = h)$ is modelled by a DNN, $P(S = s)$ is a complex distribution
- Special deterministic case: If $s = f(h)$, i.e., $P(S = s|H = h) = \delta(s - f(h))$, i.e., $s$ follows deterministically from $h$

$$P(S = s) = \sum_{f(h)=s} P(H = h)$$

(Note that this is not the same as $E(S) = \sum_h f(h)P(H = h)$)

- If $f(h)$ is invertible, with $h = g(s)$, $P(S = s) = P(H = g(s))$; $g(s)$ is called the **encoder** and $f(h)$ is called the **decoder** or **generator**

# Independent Random Variables

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

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

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

# Simplified Notation

- The expression $P(X = x)$ is often simplified as $P(x)$
- Thus instead of writing $P(H = 185cm)$, we write $P(185cm)$
- Joint: $P(X = x, Y = y) \equiv P(x, y)$
- Marginalization: $P(Y = y) = \sum_x P(X = x, Y = y)$ becomes

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

- Sometimes $X$ stands for the event $X = x$ with some unspecified $x$; thus one sees also $P(X)$, $P(X, Y)$, and

$$P(X) = \sum_X P(X, Y)$$

# 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

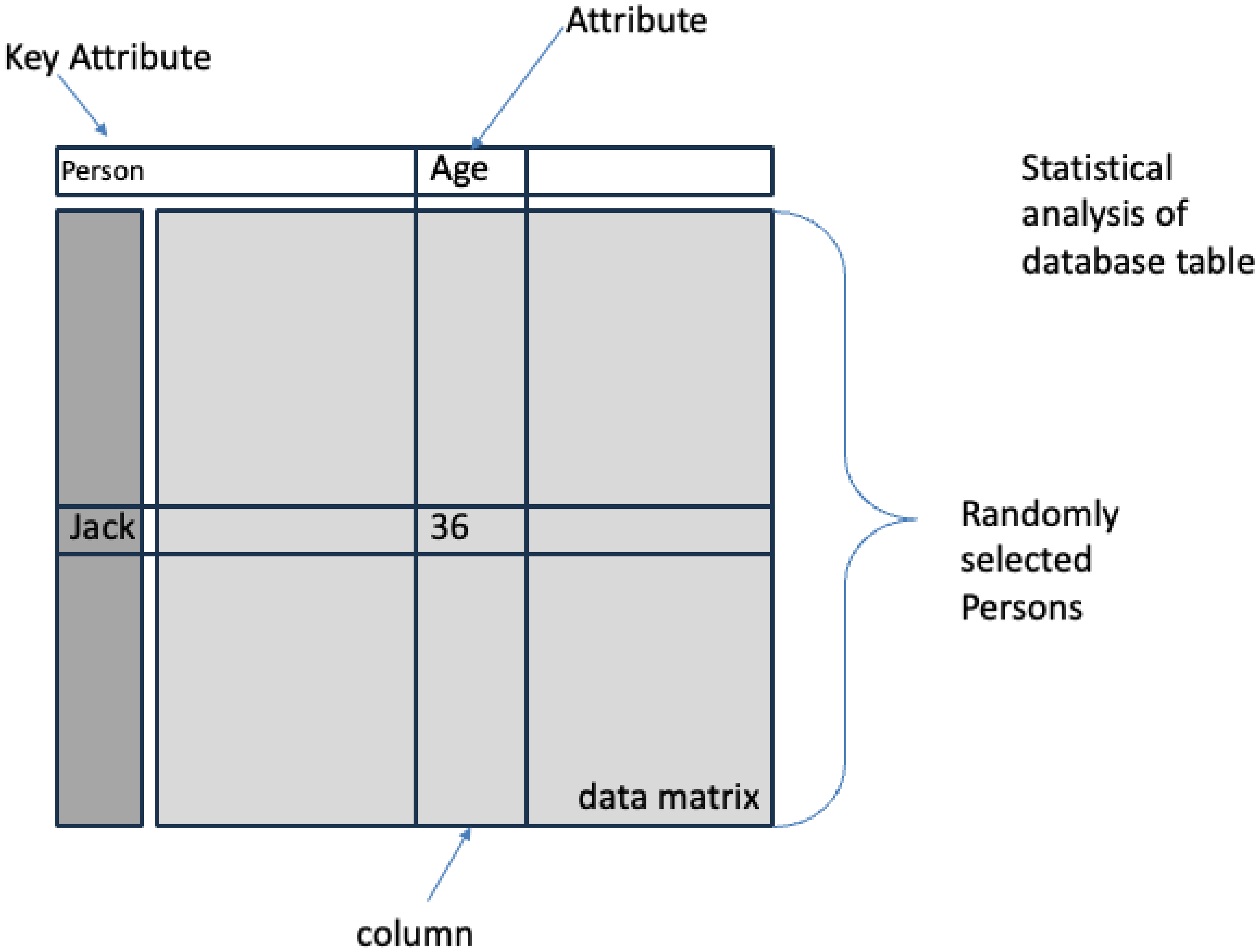
$$P(x) = \sum_y P(x,y)$$

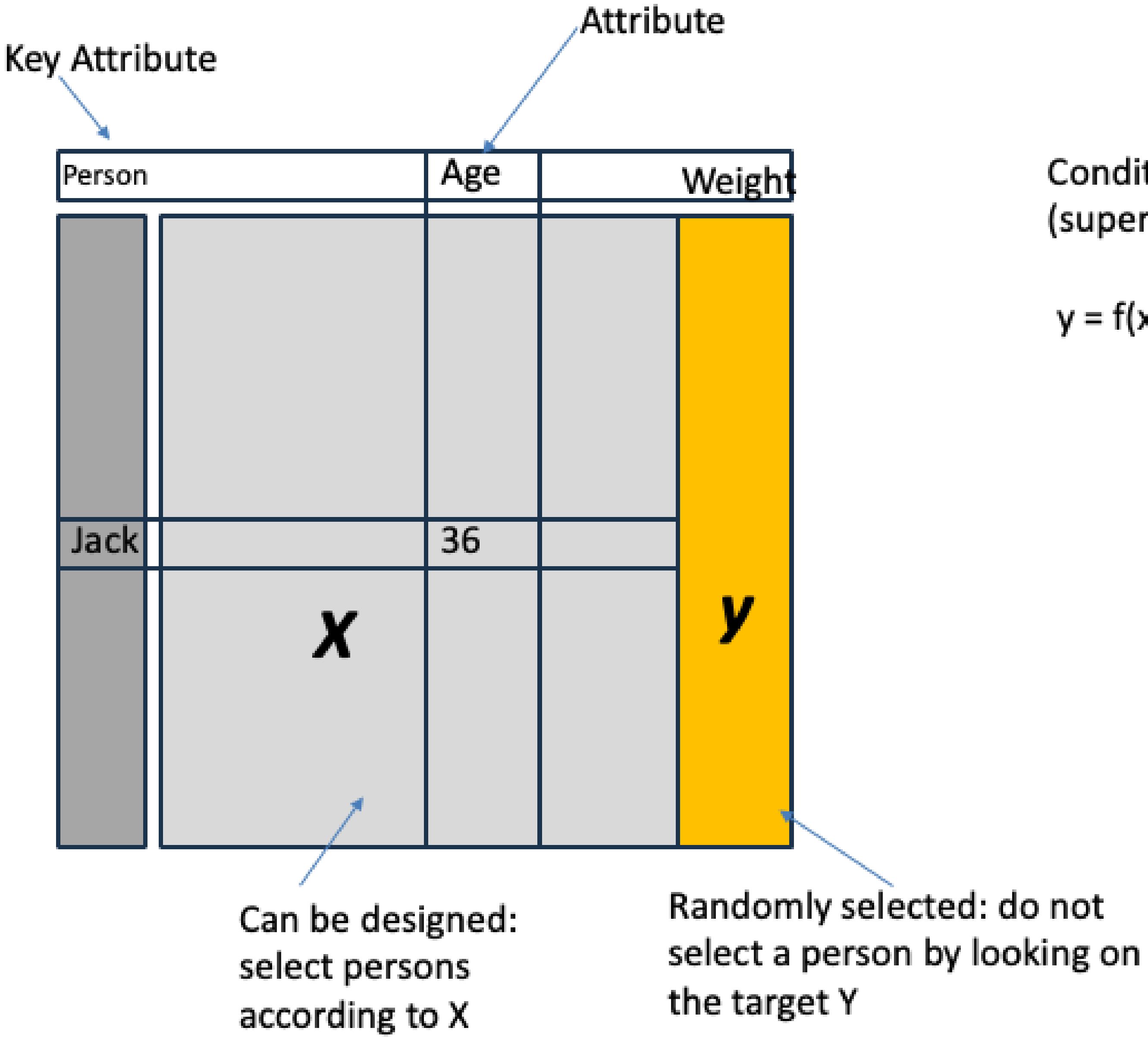
- Independent random variables

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

## Simplifications for Supervised Learning

- I one is only interested in the conditional probability $P(Y|x_1, \ldots x_M)$, then $x_1, \ldots x_M$ can be "designed"
- E.g., if the input is height and the output is weight, I can select systematically people based on height, but I cannot select them based on weight
- E.g., if the input is the cause and the output is effect, I can set the cause (give medication or not) and record the outcome; but I cannot only select patients, where the medication has worked
- (Of course, for many other reasons, the selected inputs should correspond to the population I am interested in)


Key Attribute
Attribute
Person
Age
Statistical analysis of database table
Jack
36
Randomly selected Persons
data matrix
column




Conditional Probability (supervised learning)

$y = f(x)$

## 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**

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

$$\approx \frac{1}{N} \sum_{k=1}^{N} x_k = mean_x$$

(with random observations)

## Variance

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

$$var(X) = \sum_i (x_i - E(X))^2 P(X = x_i) \approx \frac{1}{N-1} \sum_i (x_i - \textit{mean}_x)^2$$

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

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

# Covariance

- **Covariance**:

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

$$\approx \frac{1}{N-1} \sum_i (x_i - \mathit{mean}_x)(y_i - \mathit{mean}_y)$$

- **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 $\mathbf{\Sigma}_{\mathbf{w}}$
- Let

$$\mathbf{y} = \mathbf{A}\mathbf{w} + \vec{\epsilon}$$

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

$$\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{A}^T + \sigma^2\mathbf{I}$$

- Special case (Gaussain distributions): $P(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \vec{\mu}_w, \mathbf{\Sigma}_{\mathbf{w}})$, $P(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{w}, \sigma^2\mathbf{I})$ then $P(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{A}\vec{\mu}_w, \mathbf{A}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{A}^T + \sigma^2\mathbf{I})$
- Special case ($\sigma^2 = 0$): $\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{A}^T$

# Continuous Random Variables

- **Probability density**

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

- Thus

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

- The **distribution function** is

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

# Expectations for Continuous Variables

- Expected value

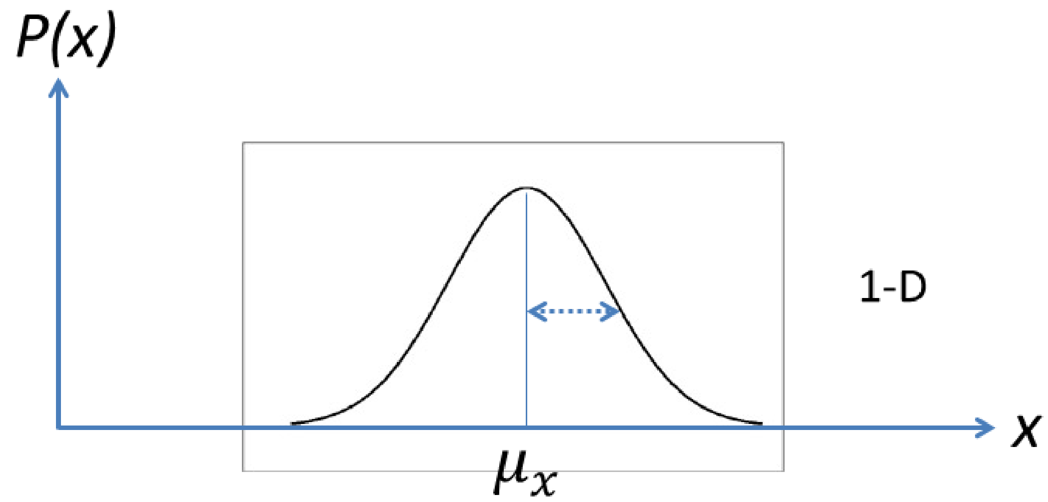
$$E(X) = E_{P(x)}(X) = \int xP(x)dx$$

- Variance

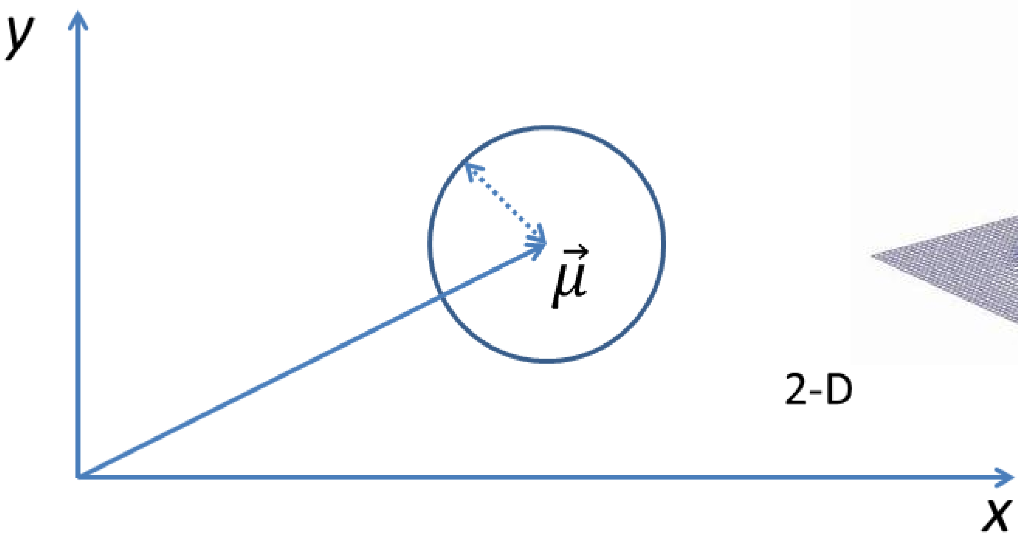
$$var(X) = \int (x - E(x))^2 P(x)dx$$

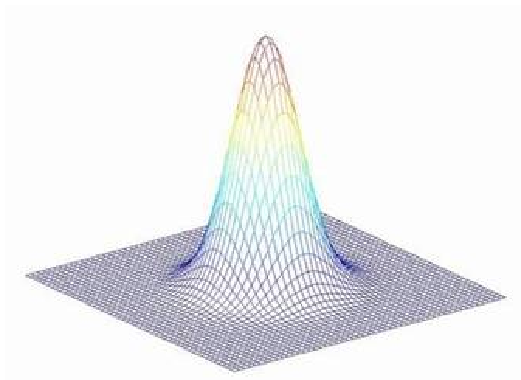
- Covariance:

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




1-D



2-D

# Joint Gaussian Distributions

- Let $\mathbf{z} = (\mathbf{x}; \mathbf{y})$, $\vec{\mu} = (\vec{\mu}_x; \vec{\mu}_y)$; thus $\mathbf{z}$ can be partitioned into $\mathbf{x}$ and $\mathbf{y}$
- With

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

we get

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

Here $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$.

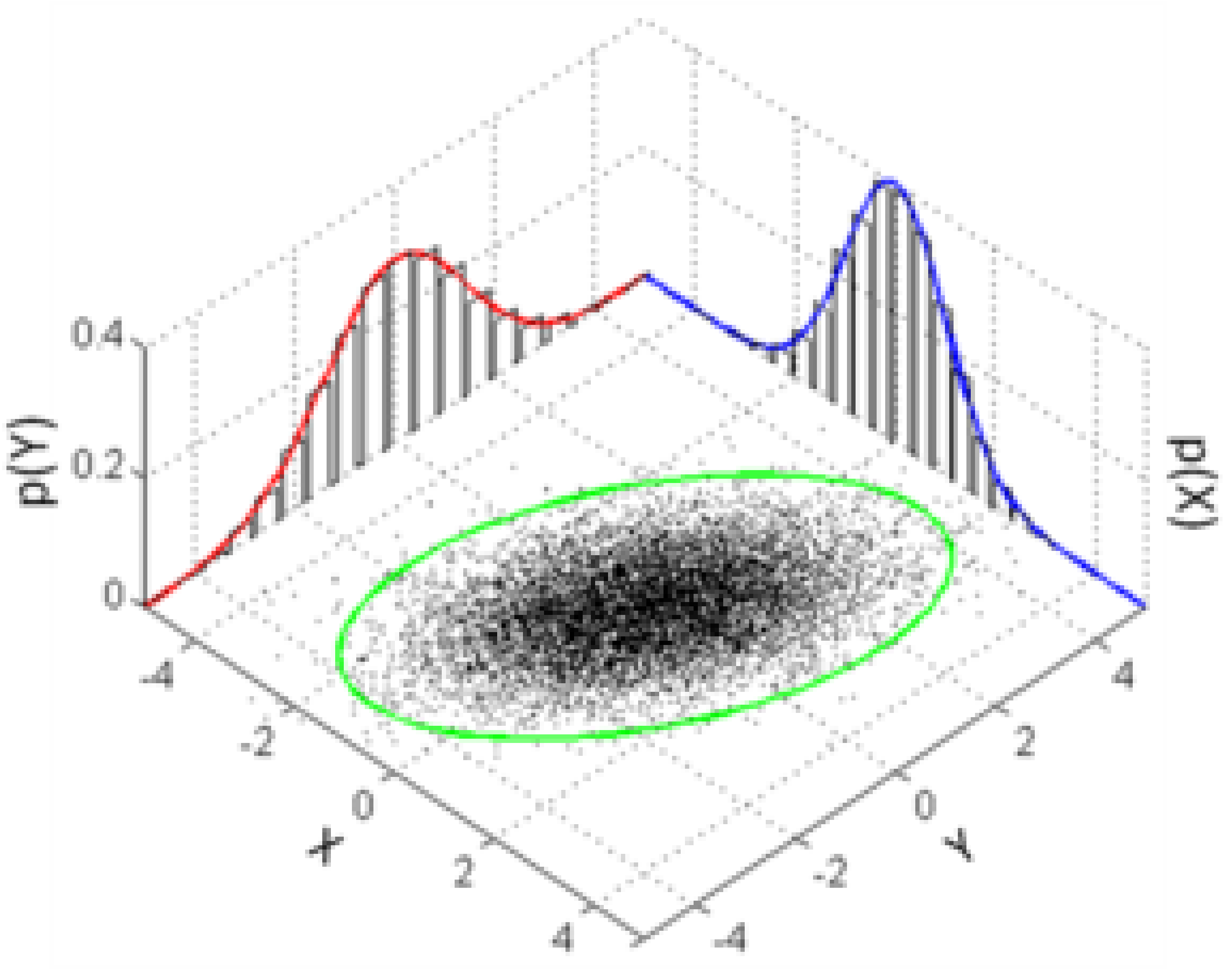
## Marginals

- For $\mathbf{x}$,

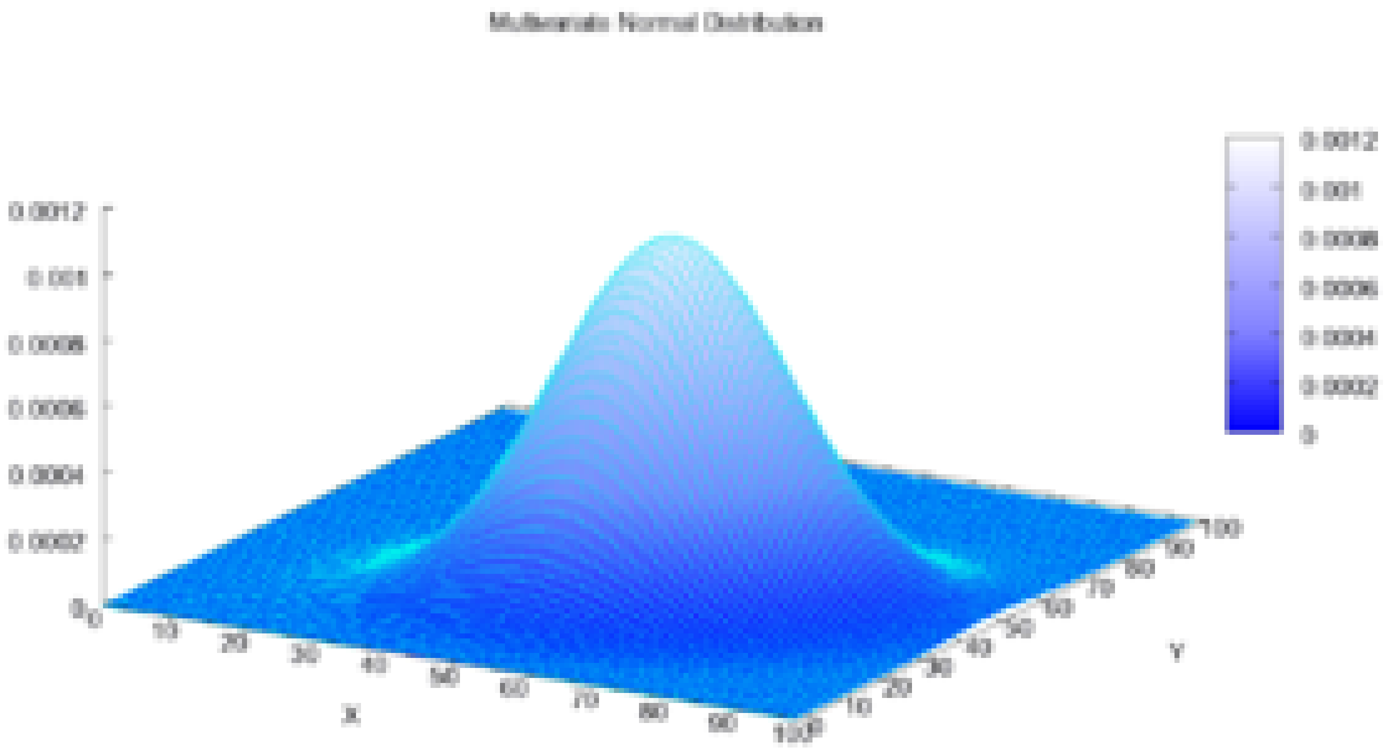
$$P(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \vec{\mu}_x, \mathbf{\Sigma}_{x,x}\right)$$

- For $\mathbf{y}$,

$$P(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}; \vec{\mu}_y, \mathbf{\Sigma}_{y,y}\right)$$



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
0
10
20
30
40
50
60
70
80
90
100
X
Y

## Conditional Densities

- For the conditionals, we get

$$P(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}; \vec{\mu}_y + \mathbf{\Sigma}_{y,x}\mathbf{\Sigma}_{x,x}^{-1}(\mathbf{x} - \vec{\mu}_x), \mathbf{\Sigma}_{y,y} - \mathbf{\Sigma}_{y,x}\mathbf{\Sigma}_{x,x}^{-1}\mathbf{\Sigma}_{x,y}\right)$$

- With $\vec{\mu}_y = 0$ and $\vec{\mu}_x = 0$, we get $E(\mathbf{y}|\mathbf{x}) = \mathbf{\Sigma}_{y,x}\mathbf{\Sigma}_{x,x}^{-1}\mathbf{x}$, which is an equation relevant for Gaussian process regression

- For noisy measurements (independent additive Gaussian noise with variance $\sigma^2$) $\mathbf{\Sigma}_{x,x} \leftarrow \mathbf{\Sigma}_{x,x} + \sigma^2\mathbf{I}$