Overview of Probability and Statistics

Michal Reinštein, Václav Hlaváč

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics
Center for Machine Perception
`http://cmp.felk.cvut.cz/~reinsmic`,
`reinstein.michal@fel.cvut.cz`

Slides: V. Hlaváč, modified: M. Reinštein, courtesy: T. Brox, V. Franc, M. Navara, M. Urban.

Outline of the talk:

- Probability vs. statistics.
- Random events.
- Probability, joint, conditional.
- Bayes theorem.
- Distribution function, density.
- Characteristics of a random variable.

Recommended reading

- A. Papoulis: Probability, Random Variables and Stochastic Processes, McGraw Hill, Edition 4, 2002.
- http://mathworld.wolfram.com/
- http://www.statsoft.com/textbook/stathome.html

Probability, statistics

- Probability: probabilistic model $\Longrightarrow$ future behavior.
  - It is a theory (tool) for purposeful decisions when the outcome of future events depends on circumstances we know only partially and the randomness plays a role.
  - An abstract model of uncertainty description and quantification of the results.
- Statistics: behavior of the system $\Longrightarrow$ probabilistic representation.
  - It is a tool for seeking a probabilistic description of real systems based on observing them and testing them.
  - It provides more: a tool for investigating the world, seeking and testing dependencies which are not apparent.
  - Two types: descriptive and inference statistics.
  - Collection, organization and analysis of data.
  - Generalization from restricted / finite samples.

Random events, concepts

An experiment with random outcome – states of nature, possibilities, experimental results, etc.

A sample space is an nonempty set Ω of all possible outcomes of the experiment.

An elementary event $\omega \in \Omega$ are elements of the sample space (outcomes of the experiment).

A space of events $\mathcal{A}$ is composed of the system of all subsets of the sample space Ω.

A random event $A \in \mathcal{A}$ is an element of the space of events.

Note: The concept of a random event was introduced in order to be able to define the probability, probability distribution, etc.

Axiomatic definition of the probability

- Ω - the sample space.
- A - the space of events.

1. $P(A) \geq 0, \quad A \in \mathcal{A}$.
2. $P(\Omega) = 1$.
3. If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$, $A \in \mathcal{A}$, $B \in \mathcal{B}$.

- If $A \subset B$ then $P(B \setminus A) = P(B) - P(A)$.
  The symbol $\setminus$ denotes the set difference.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probability

is a function P, which assigns number from the interval $\langle 0, 1\rangle$ to events and fulfils the following two conditions:

- $P(true) = 1$,
- $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$, if the events $A_n,\ n \in \mathbb{N}$, are mutually exclusive.

From these conditions, it follows:

$P(false) = 0,\qquad P(\neg A) = 1 - P(A),\qquad$ if $A \Rightarrow B$ then $P(A) \leq P(B)$.

Note: Strictly speaking, the space of events have to fulfil some additional conditions.

Joint probability, marginalization

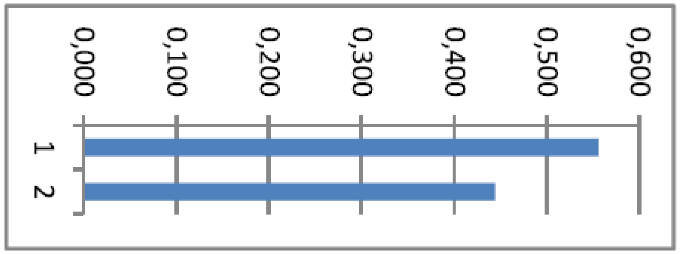
- The joint probability $P(A, B)$, also sometimes denoted $P(A \cap B)$, is the probability that events A, B co-occur.
- The joint probability is symmetric: $P(A, B) = P(B, A)$.

- Marginalization (the sum rule): $P(A) = \sum_B P(A, B)$ allows computing the probability of a single event by summing the joint probabilities over all possible events B. The probability $P(A)$ is called the marginal probability.

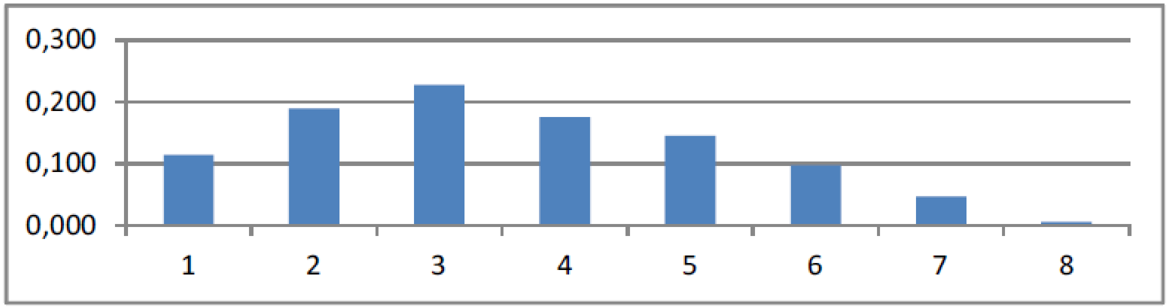
Joint probability
Contingency table, marginalization

Orienteering competition example, participants									
Age	<= 15	16-25	26-35	36-45	46-55	56-65	66-75	>= 76	Sum
Men	22	36	45	33	29	21	12	2	200
Women	19	32	37	30	23	14	5	0	160
Sum	41	68	82	63	52	35	17	2	360

Orienteering competition example, frequency									
Age	<= 15	16-25	26-35	36-45	46-55	56-65	66-75	>= 76	Sum
Men	0,061	0,100	0,125	0,092	0,081	0,058	0,033	0,006	0,556
Women	0,053	0,089	0,103	0,083	0,064	0,039	0,014	0,000	0,444
Sum	0,114	0,189	0,228	0,175	0,144	0,097	0,047	0,006	1



Marginal probability P(sex)



Marginal probability P(Age_group)

The conditional probability

- Let us have the probability representation of a system given by the joint probability $P(A, B)$.
- If an additional information is available that the event B occurred then our knowledge about the probability of the event A changes to

$$P(A|B) = \frac{P(A, B)}{P(B)},$$

which is the conditional probability of the event A under the condition B.

- The conditional probability is defined only for $P(B) \neq 0$.
- Product rule: $P(A, B) = P(A|B)\,P(B) = P(B|A)\,P(A)$.
- From the symmetry of the joint probability and the product rule, the Bayes theorem can be derived (to come in a more general formulation for more than two events).

Properties of the conditional probability

- $P(true|B) = 1$, $P(false|B) = 0$.
- If $A = \bigcup_{n \in \mathbb{N}} A_n$ and events $A_1, A_2, \ldots$ are mutually exclusive then $P(A|B) = \sum_{n \in \mathbb{N}} P(A_n|B)$.
- Events A, B are independent $\Leftrightarrow P(A|B) = P(A)$.
- If $B \Rightarrow A$ then $P(A|B) = 1$.
- If $B \Rightarrow \neg A$ then $P(A|B) = 0$.

Events B_i, $i \in I$, constitute a complete system of events if they are mutually exclusive and $\bigcup_{i \in I} B_i = true$.

Example: Conditional probability

Example

Consider rolling a single dice. *What is the probability that the number higher than three comes up (event A) under the conditions that the odd number came up (event B)*?

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad A = \{4, 5, 6\}, \quad B = \{1, 3, 5\}$$

$$P(A) = P(B) = \frac{1}{2}$$

$$P(A, B) = P(\{5\}) = \frac{1}{6}$$

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Example: Independent events

Events A, B are independent $\Leftrightarrow P(A, B) = P(A)\, P(B)$,
since independence means: $P(A|B) = P(A)$, $P(B|A) = P(B)$

Example

Rolling the dice once, events are: $A > 3$, event B is odd. Are A, B independent?

$$\Omega = \{1, 2, 3, 4, 5, 6\}\,, \quad A = \{4, 5, 6\}\,, \quad B = \{1, 3, 5\}$$

$$P(A) = P(B) = \frac{1}{2}$$
$$P(A, B) = P(\{5\}) = \frac{1}{6}$$
$$P(A)\, P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$P(A, B) \neq P(A)\, P(B) \Leftrightarrow$ The events are dependent.

Conditional independence

Random events A, B are conditionally independent under the condition C, if

$$P(A, B|C) = P(A|C)\,P(B|C)\,.$$

Similarly, a conditional independence of more events, random variables, etc. is defined.

Total probability theorem

Let B_i, $i \in I$, be a *complete system of events* and $\forall i \in I: P(B_i) \neq 0$.

Then for every event A holds

$$P(A) = \sum_{i \in I} P(B_i)\, P(A|B_i)\,.$$

Proof:

$$P(A) = P\left(\left(\bigcup_{i \in I} B_i\right) \cap A\right) = P\left(\bigcup_{i \in I} (B_i \cap A)\right)$$

$$= \sum_{i \in I} P(B_i \cap A) = \sum_{i \in I} P(B_i)\, P(A|B_i)\,.$$

Bayes theorem

Let B_i, $i \in I$, be a *complete system of events* and $\forall i \in I\colon P(B_i) \neq 0$.

Then for any $j \in I$ and every event A given $P(A) \neq 0$ holds

$$P(B_j|A) = \frac{P(B_j)\,P(A|B_j)}{\sum_{i \in I} P(B_i)\,P(A|B_i)}\,.$$

Proof (exploring the total probability theorem):

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j)\,P(A|B_j)}{\sum_{i \in I} P(B_i)\,P(A|B_i)}\,.$$

A special case: Bayes theorem for two events only

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)}\, P(B)}{P(A)},$$

where P(B|A) is the posterior probability and P(A|B) is the likelihood.

- This is a fundamental rule for machine learning (pattern recognition) as it allows to compute the probability of an output B given measurements A.
- The prior probability is $P(B)$ without any evidence from measurements.
- The likelihood P(A|B) evaluates the measurements given an output B. Seeking the output that maximizes the likelihood (*the most likely output*) is known as the maximum likelihood estimation (ML).
- The posterior probability P(B|A) is the probability of B after taking the measurement A into account. Its maximization leads to the maximum a-posteriori estimation (MAP).

The importance of Bayes theorem

- The probabilities $P(A|B_i)$ are estimated from experiments or from a statistical model.
- Having $P(A|B_i)$, the probabilities $P(B_i|A)$ can be computed and used to estimate, which event from B_i occurred.
- **A problem:** To determine *a posteriori* probability $P(B_i|A)$, it is needed to know the *a priori* probability $P(B_i)$.
- In a similar manner, we define the conditional probability distribution, conditional density of the continuous random variable, etc.

Example: Victim detections in USAR (1)

In *Urban Search & Rescue* (USAR) the ability of robots to reliably detect presence of a victim is crucial. How do we implement and evaluate this ability?

Task 1

Assume we have a sensor S (e.g. camera) and computer vision algorithm that detects victims. We evaluated the sensor on ground truth data statistically:

- There is 20% chance of **not detecting** a victim if there is one.
- There is 10% chance of **false positive detection**.
- A priori probability of victim presence V is 60%.
- *What is the probability that there is a victim if the sensor says no victim is detected?*

Example: Victim detections in USAR (2)

Tools

What tools are we going to use?

- The Product rule: $P(A, B) = P(A|B)\,P(B) = P(B|A)\,P(A)$
- The Sum rule: $P(B) = \sum_A P(A, B) = \sum_A P(B|A)\,P(A)$
- The Bayes theorem:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_A P(B|A)\,P(A)}$$

- General inference:

$$P(V|S) = \frac{P(V, S)}{P(S)} = \frac{\sum_{A,B,C} P(S, A, B, C, V)}{\sum_{S,A,B,C} P(S, A, B, C, V)}$$

Example: Victim detections in USAR (3)

Solution

Express the sensor S measurements as conditional probability of V:

$P(S\|V)$	$S = True$	$S = False$
$V = True$	0.8	0.2
$V = False$	0.1	0.9

Express the a priori knowledge as probability:
$P(V = True) = 0.6$ and $P(V = False) = 1 - 0.6 = 0.4$

Express what we want: $P(V|S) =?$ given $S = False$ (not detecting a victim) and $V = True$ (but there is one).

Example: Victim detections in USAR (4)

- Use the tools to express what-we-want in the terms of what-we-know:

$$P(V|S) = \frac{P(V,S)}{P(S)} = \frac{P(S|V)P(V)}{\sum_V P(S,V)} = \frac{P(S|V)P(V)}{\sum_V P(S|V)\,P(V)}$$

- Substitute $S = False$ and $V = True$ and sum over V to obtain:

$$P(V|S) = \frac{P(S = False|V = True)P(V = True)}{\sum_V P(S = False|V = True)\,P(V = True)} =$$

$$= \frac{0.2 \cdot 0.6}{0.2 \cdot 0.6 + 0.9 \cdot 0.4} = 0.25$$

- **Conclusion**: if our sensors says there is no victim, we have **25%** chance of missing out someone! We need to add one more sensor ...

Example: Victim detections in USAR (5)

In *Urban Search & Rescue* (USAR) the reliability is achieved through sensor fusion: use statistics to evaluate the sensors and probability to perform the fusion.

Task 2

Assume we have a sensor S as in previous case and we add one more sensor T with the following properties:

- There is 5% chance of **not detecting** a victim if there is one.
- There is 5% chance of **false positive detection**.
- A priori probability of victim presence is the same, V is 60%.
- *What is the probability that there is a victim if both sensors confirm its presence?*

Example: Victim detections in USAR (6)

Solution

Express the sensor T measurements as conditional probability of V:

$P(T\|V)$	$T = True$	$T = False$
$V = True$	0.95	0.05
$V = False$	0.05	0.95

The a priori probability is the same:
$P(V = True) = 0.6$ and $P(V = False) = 1 - 0.6 = 0.4$

Express what we want: $P(V|S, T) =?$ given $S = True, T = True$ (both sensors see a victim) and $V = True$ (and there is one). Furthermore, we know that both sensors provide independent measurements with respect to each other.

Example: Victim detections in USAR (7)

- Naive approach using joint probability: $P(S,T,V) = P(S,T|V)P(V)$
- Conditional independence: $P(S,T|V)P(V) = P(S|V)P(T|V)P(V)$
- Applying the tools:

$$P(V|S,T) = \frac{P(V,S,T)}{P(S,T)} = \frac{P(S|V)P(T|V)P(V)}{\sum\limits_{V} P(V,S,T)} =$$

$$= \frac{P(S|V)P(T|V)P(V)}{\sum\limits_{V} P(S|V)P(T|V)\,P(V)}$$

- Substitute: $S = True, T = True$, $V = True$ and sum over V to obtain:

$$= \frac{0.8 \cdot 0.95 \cdot 0.6}{0.8 \cdot 0.95 \cdot 0.6 + 0.1 \cdot 0.05 \cdot 0.4} = 0.9956$$

- **Conclusion**: if both sensors confirm there is a victim, we have **99.56%** chance that there is a victim.

The random variable

- The random variable is an arbitrary function $X : \Omega \to \mathbb{R}$, where Ω is a sample space.
- Why is the concept of the random variable introduced? It allows to work with concepts as the distribution function, probability density, expectation (mean value), etc.
- There are two basic types of random variables:
  - Discrete – a countable number of values. *Examples: rolling a dice, the count of number of cars passing through a street in a hour.*

    The discrete probability is given as $P(X = a_i) = p(a_i)$, $i = 1, \ldots$, $\sum_i p(a_i) = 1$.
  - Continuous – values from some interval, i.e. infinite number of values. *Example: the height persons.*

    The continuous probability is given by the distribution function or the probability density function.

Distribution function of a random variable

Distribution function of the random variable X is a function $F\colon X \to [0,1]$ defined as $F(x) = P(X \leq x)$, where P is a probability.

Properties:

1. $F(x)$ is a non-decreasing function, i.e. $\forall$ pair $x_1 < x_2$ it holds $F(x_1) \leq F(x_2)$.
2. $F(X)$ is continuous from the right, i.e. it holds $\lim\limits_{h \to 0^+} F(x+h) = F(x)$.
3. - It holds for every distribution function $\lim\limits_{x \to -\infty} F(x) = 0$ a $\lim\limits_{x \to \infty} F(x) = 1$. Written more concisely: $F(-\infty) = 0$, $F(\infty) = 1$.
   - If the possible values of $F(x)$ are from the interval (a, b) then $F(a) = 0$, $F(b) = 1$.

Any function fulfilling the above three properties can be understood as a distribution function.

Continuous distribution function

- The distribution function F is called (absolutely) continuous if a nonnegative function f (probability density) exists and it holds

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u \quad \text{for every } x \in X.$$

- The probability density fulfills

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1\,.$$

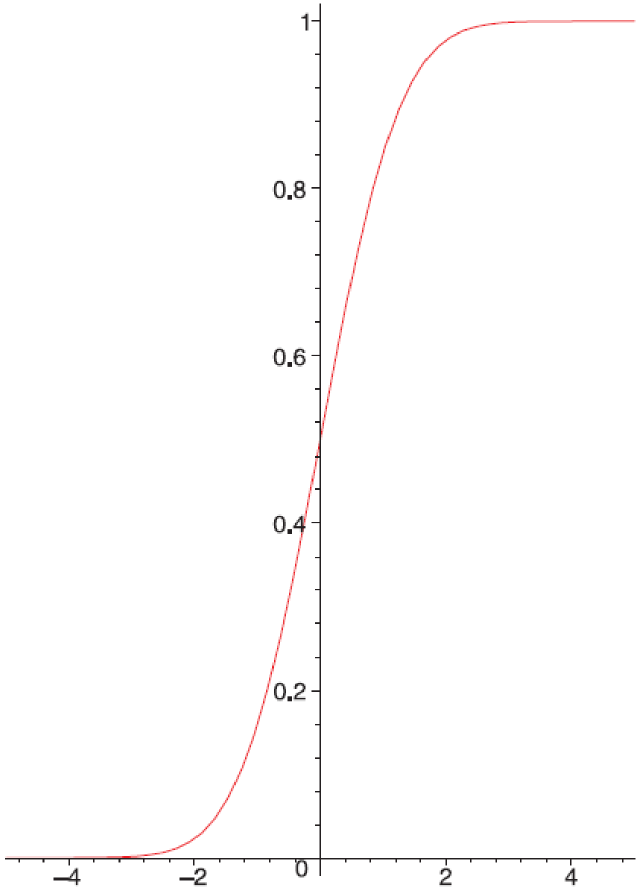
- If the derivative of $F(x)$ exists in the point x then $F'(x) = f(x)$.
- For $a, b \in \mathbb{R}$, $a < b$, it holds

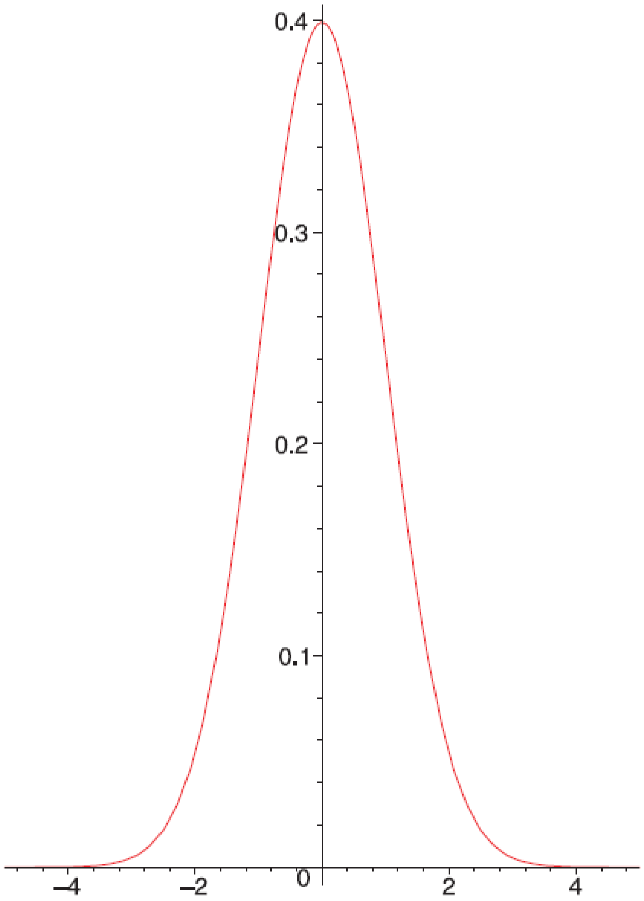
$$P(a < X < b) = \int_{a}^{b} f(x)\,\mathrm{d}x = F(b) - F(a)\,.$$

Example: Normal distribution

$$F(x)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{-x^2}{2\sigma^2}}$$





Distribution function

Probability density

The law of large numbers

The law of large numbers says that if very many independent experiments can be made then it is almost certain that the relative frequency will converge to the theoretical value of the probability density.

Jakob Bernoulli, Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis, 1713, Chapter 4.

Expectation

- (Mathematical) expectation = the average of a variable under the probability distribution.
- Continuous definition: $E(x) = \int\limits_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$.
- Discrete definition: $E(x) = \sum\limits_{x} x\, P(x)$.
- The expectation can be estimated from a N *number of samples* by $E(x) \approx \frac{1}{N} \sum\limits_{i} x_i$. The approximation becomes exact for $N \to \infty$.
- Expectation over multiple variables: $E_x(x, y) = \int\limits_{-\infty}^{\infty} (x, y)\, f(x)\, \mathrm{d}x$
- Conditional expectation: $E(x|y) = \int\limits_{-\infty}^{\infty} x\, f(x|y)\, \mathrm{d}x$.

Basic characteristics of a random variable

Continuous distribution	Discrete distribution
Expectation	
$E(x) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$	$E(x) = \sum_x x\, P(x)$
k-th (general) moment	
$E(x^k) = \int_{-\infty}^{\infty} x^k\, f(x)\, \mathrm{d}x$	$E(x) = \sum_x x^k\, P(x)$
k-th central moment	
$E(x^k) = \int_{-\infty}^{\infty} (x - E(x))^k\, f(x)\, \mathrm{d}x$	$E(x) = \sum_x (x - E(x))^k\, P(x)$
Dispersion	
$D(x) = \int_{-\infty}^{\infty} (x - E(x))^2\, f(x)\, \mathrm{d}x$	$E(x) = \sum_x (x - E(x))^2\, P(x)$

Covariance

The mutual covariance σ_{xy} of two random variables X, Y is

$$\sigma_{xy} = E\left((X - \mu_x)(Y - \mu_y)\right) \ .$$

The covariance matrix Σ of n variables $X_1, \ldots, X_n$ is

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & \sigma_{1n} \\ & \ddots & \\ \sigma_{n_1} & \ldots & \sigma_n^2 \end{bmatrix} \ .$$

The covariance matrix is symmetric (i.e. $\Sigma = \Sigma^\top$) and positive-semidefinite (as the covariance matrix is real valued, the positive-semidefinite means that $x^T M x \geq 0$ for all $x \in \mathbb{R}$).

Quantiles, median

- The p-quantile Q_p: $P(X < Q_p) = p$.
- The median is the p-quantile for $p = \frac{1}{2}$, i.e. $P(X < Q_p) = \frac{1}{2}$.

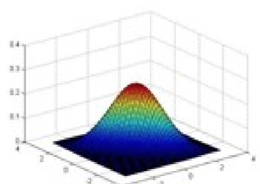
Note: Median is often used as a replacement for the mean value in robust statistics.

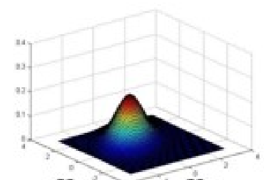
Multivariate Normal distribution (1)

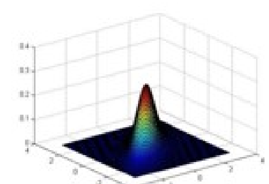
Multivariate Gaussian (Normal) distribution

Parameters μ, Σ

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$







Parameter fitting:

Given training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

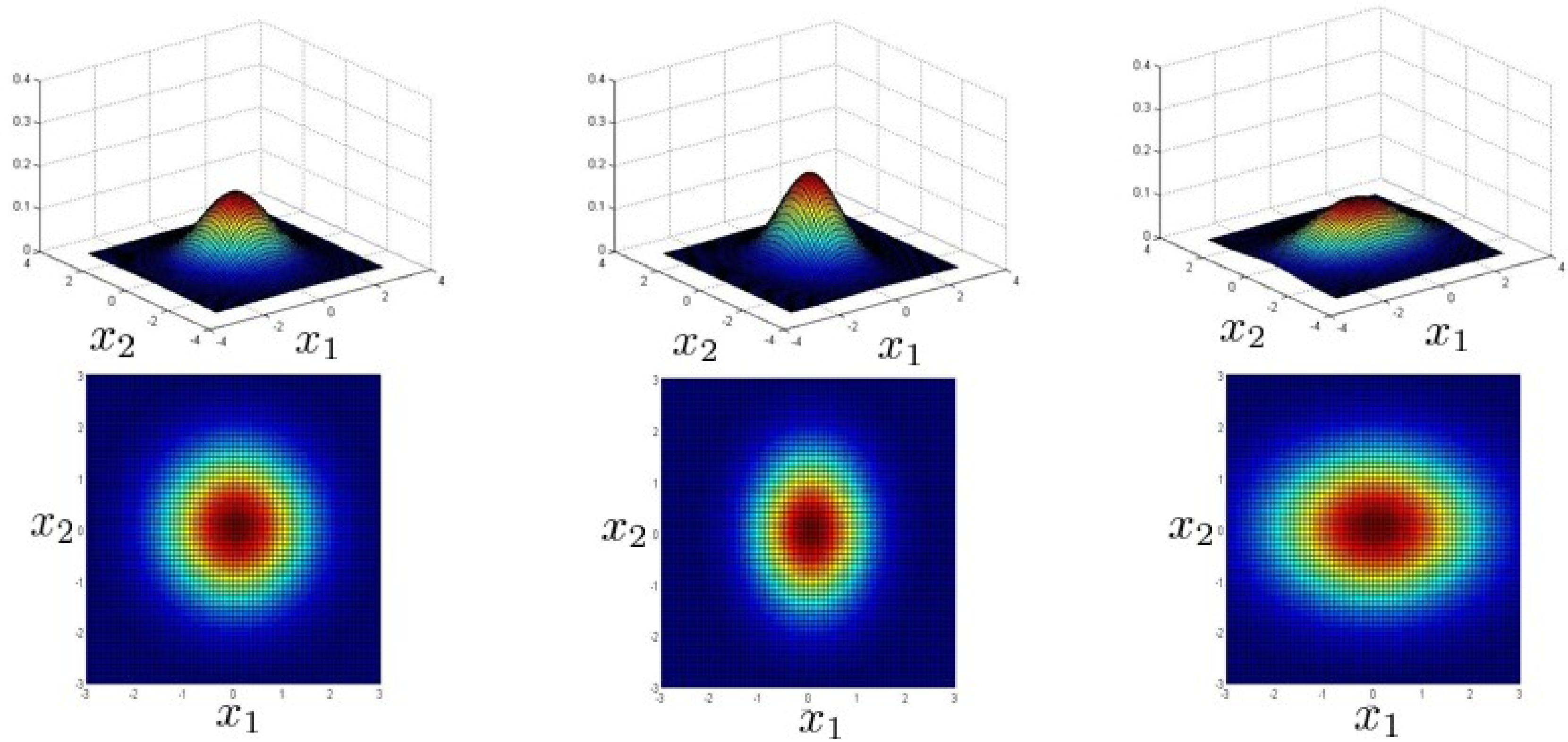
Andrew Ng

Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15

Multivariate Normal distribution (2)

Multivariate Gaussian (Normal) examples

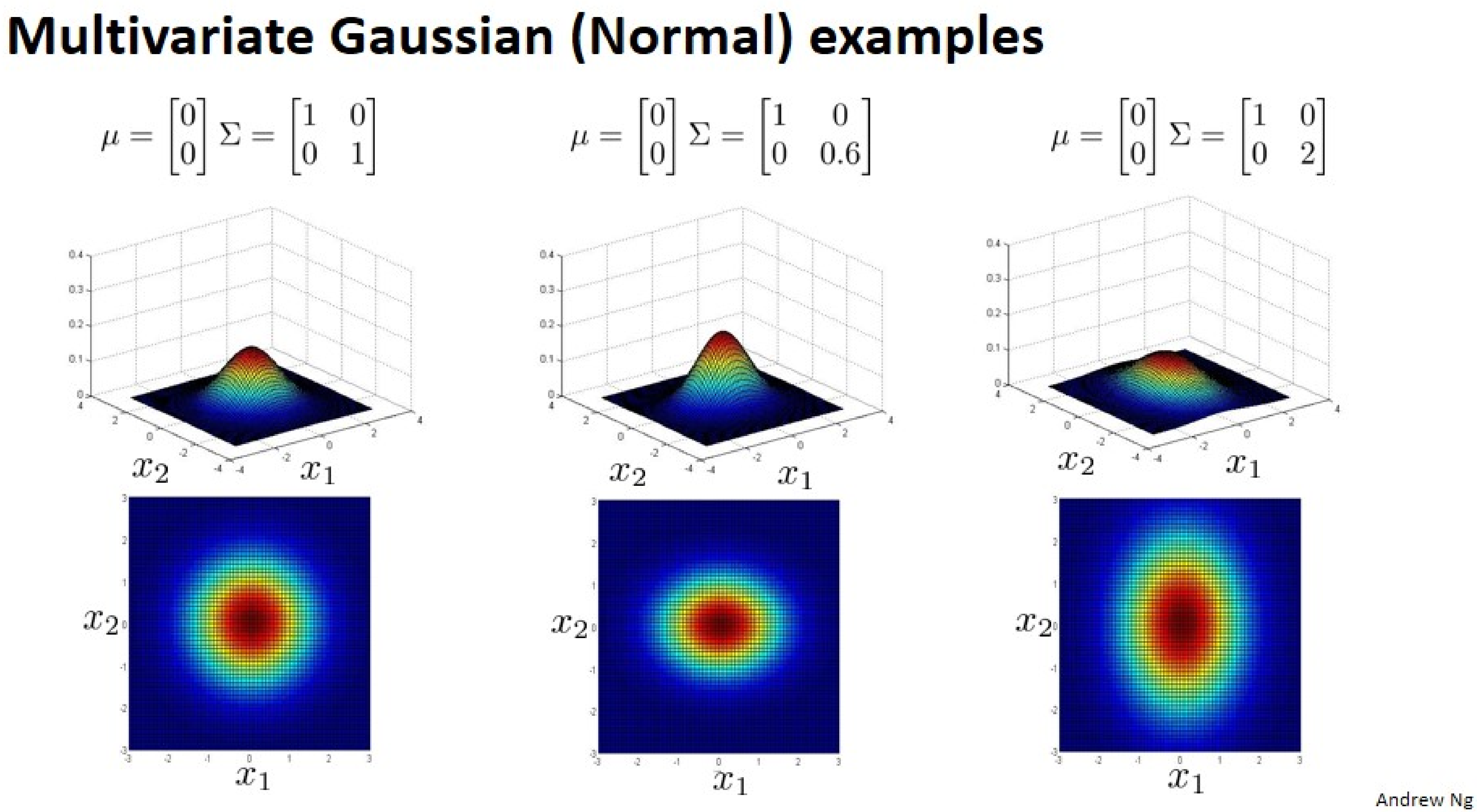
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$



Andrew Ng

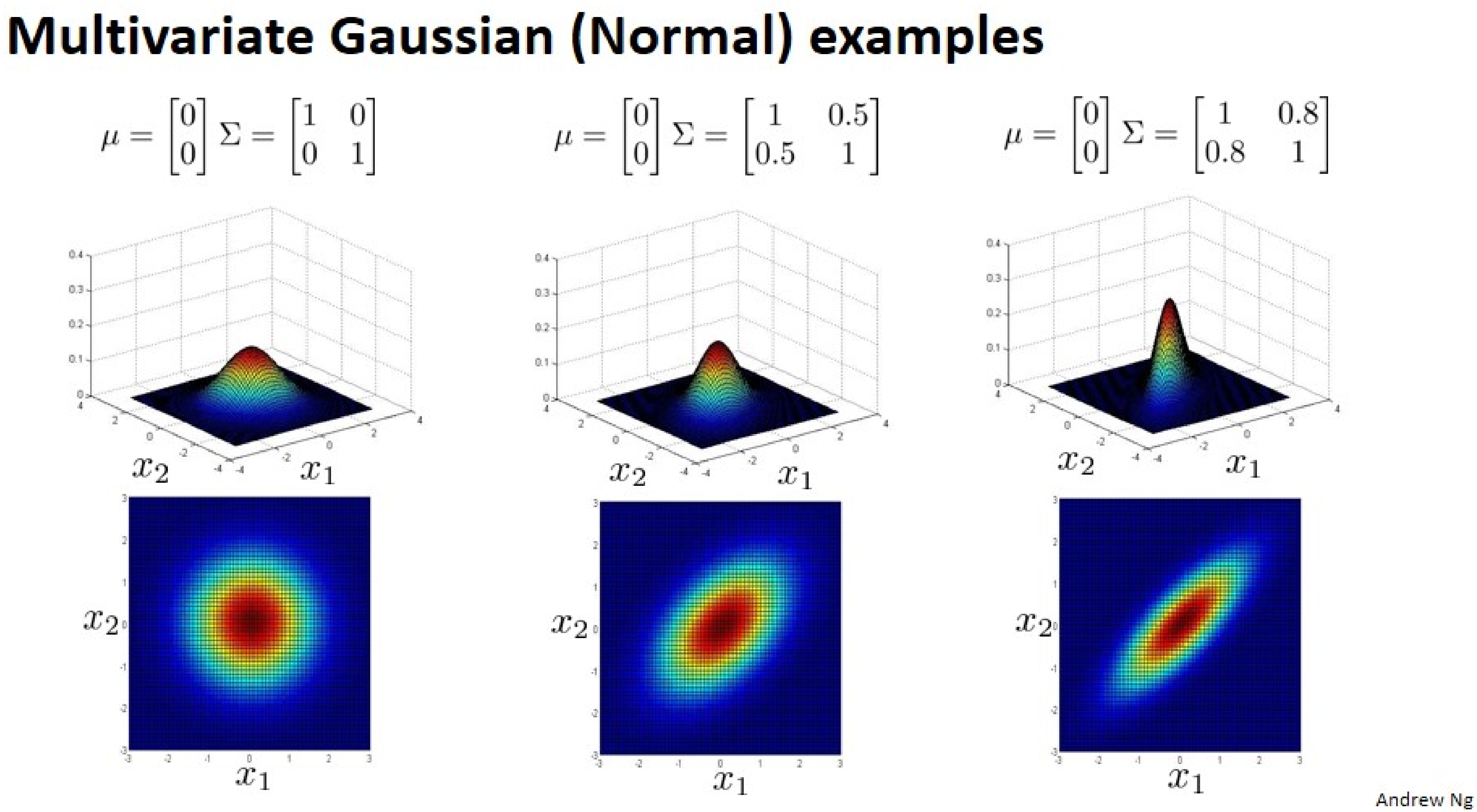
Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15

Multivariate Normal distribution (3)



Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15

Multivariate Normal distribution (4)

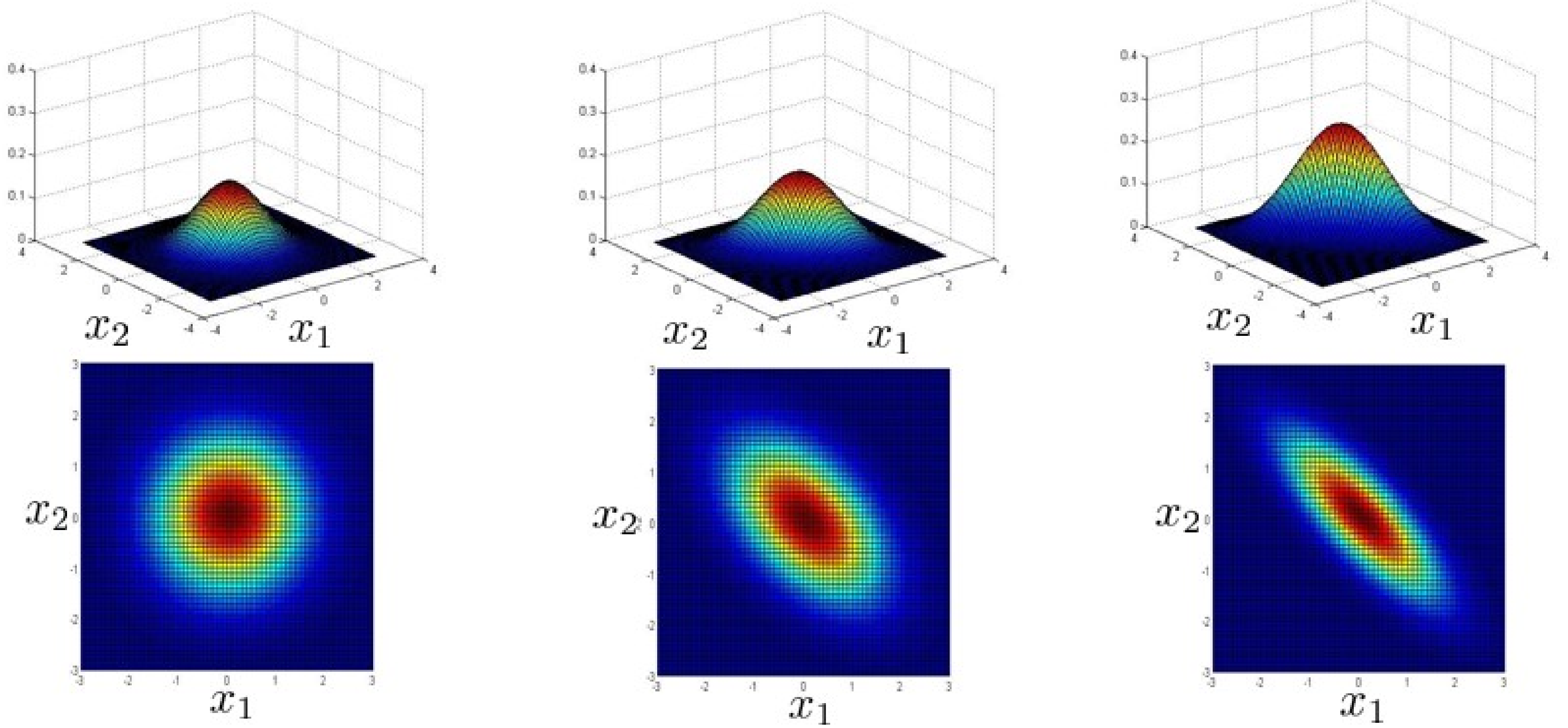


Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15

Multivariate Normal distribution (5)

Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$



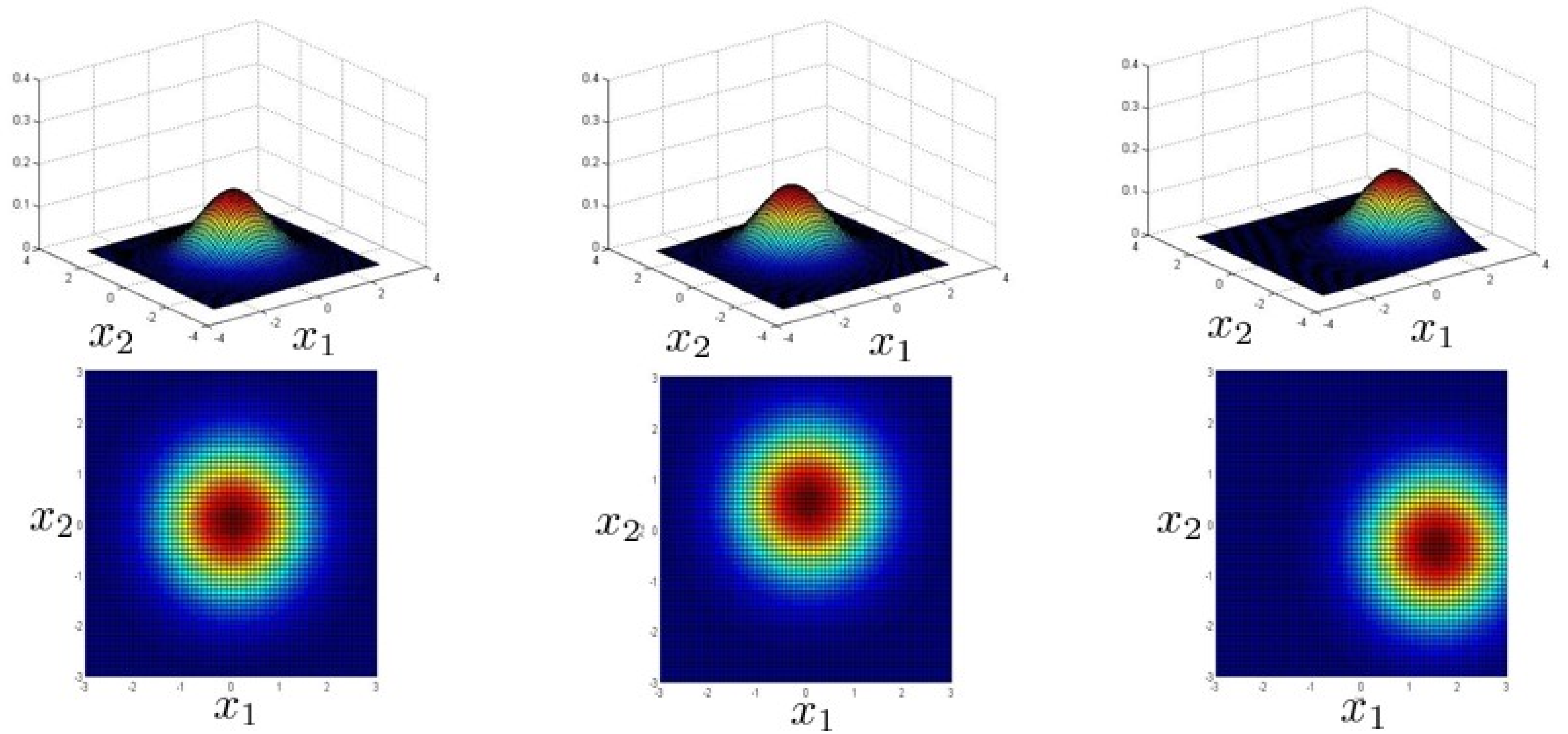
Andrew Ng

Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15

Multivariate Normal distribution (6)

Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



Andrew Ng

Source: Andrew Ng, Stanford University, Machine Learning course 2012, Lecture 15