Brief overview of probability theory

Hamed Shahbazi, Fall 2018, CS-534

Reasoning under Uncertainty

- Probabilities quantify uncertainty regarding the occurrence of events.

  Examples:

  - The probability of an email to be spam.
  - Observing a blip on radar screen, the distribution over the location of the corresponding target.

Definitions

- Sample Space (Ω) : Set of all possible outcomes for an experiment.

  Examples:

  - Rolling a die: Ω = {1, 2, 3, 4, 5, 6}
  - Tossing a coin: Ω = {H, T}
  - Deploy a network of smoke sensors to detect fires in a building: Ω = {(fire, smoke), (no fire, smoke), (fire, no smoke), (no fire, no smoke)}

- Event (A): Any subset of the sample space.
- Probability P(A): How likely the experiment's actual outcome belongs to A.

The Three Axioms of Probability Theory

For any probability P:

- P(A) $\geq$ 0 for any event A .
- P(Ω) = 1 (collectively exhaustive).
- P(A ∪ B) = P(A) + P(B) for any disjoint events A and B. (mutually exclusive).

Example:

	fire	*no fire*
smoke	0.002	0.003
no smoke	0.001	0.994

$$
\begin{aligned}
& P(\{(\mathit{fire}, \mathit{smoke}), (\mathit{no\ fire}, \mathit{smoke})\}) \\
&= P(\{(\mathit{fire}, \mathit{smoke})\}) + P(\{(\mathit{no\ fire}, \mathit{smoke})\}) \\
&= 0.002 + 0.003 \\
&= 0.005
\end{aligned}
$$

Axiom Consequences

Consequence of the axioms:

- P(A) = 1 – P(A^c)
- P(ϕ) = 0
- If A $\subseteq$ B then P(A) $\leq$ P(B)
- P(A∪B) $\leq$ P(A) + P(B)
- P(A∪B) $=$ P(A) + P(B) - P(A∩B)

Conditional Probability

- Conditional probability allows us to reason with partial information.
- When $P(B) > 0$, the conditional probability of A given B is defined as

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

- It is the fraction of probability mass in B that also belongs to A.

- P(A) is called the a prior probability of A and P(A|B) is called the a posteriori probability of A given B.

Conditional Probability, Example

Deploy a network of smoke sensors to detect fires in a building.

Sample Space (Ω) =

{(fire, smoke), (no fire, smoke), (fire, no smoke), (no fire, no smoke)}

	fire	*no fire*
smoke	0.002	0.003
no smoke	0.001	0.994

$$P(\{(fire, smoke)\} \mid \{(fire, smoke), (no\ fire, smoke)\})$$
$$= \frac{P(\{(fire, smoke)\} \cap \{(fire, smoke), (no\ fire, smoke)\})}{P(\{(fire, smoke), (no\ fire, smoke)\})}$$
$$= \frac{P(\{(fire, smoke)\})}{P(\{(fire, smoke), (no\ fire, smoke)\})}$$
$$= \frac{0.002}{0.005} = 0.4$$

Product and Chain Rule

- The probability that A and B both happen is the probability that A happens times the probability that B happens, given A has occurred.

$$P(A \cap B) = P(A)\, P(B|A)$$

- Chain Rule:

$$P(\cap_{i=1}^{k} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|\cap_{i=1}^{k-1} A_i)$$

Bayes Rule

- Bayes’ rule translates causal knowledge into diagnostic knowledge:

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

Example: If A is an event that a patient has a disease, and B is the event that she displays a symptom, then P(B|A) describes a causal relationship, and P(A|B) describes a diagnostic one.

Bayes Rule, Example

In a medical diagnosis problem let:

A = Having disease

B = Test result is positive (Showing the symptom)

$P(B|A)$ = Sensitivity = 0.8, $P(B|A^c)$ = 0.1 (false alarm)

$P(A) = 0.004$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B\cap A)+P(B\cap A^C)} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} =$$

$$\frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Random Variables

- It is often useful to "pick out" aspects of the experiment's outcomes.
- A random variable X is a function from the sample space Ω.

Example: In drawing a card from a deck

$$\Omega = \{A\heartsuit, 2\heartsuit, \ldots, K\heartsuit, A\diamondsuit, 2\diamondsuit, \ldots, K\diamondsuit, A\clubsuit, 2\clubsuit, \ldots, K\clubsuit, A\spadesuit, 2\spadesuit, \ldots, K\spadesuit\}$$

random variable	example event
$H(\omega) = \begin{cases} \mathit{true} & \text{if } \omega \text{ is a } \heartsuit \\ \mathit{false} & \text{otherwise} \end{cases}$	$H = \mathit{true}$
$N(\omega) = \begin{cases} n & \text{if } \omega \text{ is the number } n \\ 0 & \text{otherwise} \end{cases}$	$2 < N < 6$
$F(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is a face card} \\ 0 & \text{otherwise} \end{cases}$	$F = 1$

Densities

- Let $X : \Omega \rightarrow E$ be a discrete random variable. The function $p_X : E \rightarrow R$ is the density of X if for all $x \in E$:

$$p_X(x) = P(\{\omega : X(\omega) = x\})$$

- When E is continuous, $p_X : E \rightarrow R$ is the density of X if for all $\xi \subseteq E$:

$$P(\{\omega : X(\omega) \in \xi\}) = \int_{\xi} p_X(x)\ \mathrm{d}x$$

- Note that $\int_E p_X(x)dx = 1$ for a valid density.

Densities (finite), Example

- In drawing a card:

$$\Omega = \{A\heartsuit, 2\heartsuit, \ldots, K\heartsuit, A\diamondsuit, 2\diamondsuit, \ldots, K\diamondsuit, A\clubsuit, 2\clubsuit, \ldots, K\clubsuit, A\spadesuit, 2\spadesuit, \ldots, K\spadesuit\}$$

- Let's define random variable X = n (the number of the outcome), then $\mathrm{E} = \{1, 2, \ldots 13\}$, therefore

$$p_X\ (X=2) = P(\{\omega : X(\omega) = 2\}) = 4\ /\ 52$$

Densities (infinite), Example

Let X denote the width in mm of metal pipes from an automated production line. The X has below probability density function:

$$p_X(x) = 10e^{-10(x-5.5)} \quad x \geq 5.5$$

$$p_X(x) = 0 \quad x < 5.5$$

$$P(\{w \mid 5.6 < X(w) \leq 6\}) = \int_{5.6}^{6} 10e^{-10(x-5.5)}\, dx = 0.361$$

Joint Densities

- If $X : \Omega \rightarrow \mathrm{E}$ and $Y : \Omega \rightarrow \Upsilon$ are two finite random variables, then

$p_{XY} : \mathrm{E} \times \Upsilon \rightarrow R$ is their joint density if for all $x \in \mathrm{E}$ and $y \in \Upsilon$:

$$p_{XY}(x, y) = P(\{\omega : X(\omega) = x, Y(\omega) = y\})$$

- When E or Υ are infinite, $p_{XY} : \mathrm{E} \times \Upsilon \rightarrow R$ is the joint density of X and Y if for all $\xi \subseteq \mathrm{E}$ and $\upsilon \subseteq \Upsilon$:

$$\int_{\xi} \int_{\upsilon} p_{XY}(x, y)\ \mathrm{d}y\ \mathrm{d}x = P(\{\omega : X(\omega) \in \xi, Y(\omega) \in \upsilon\})$$

Marginalization

- Marginalization refers to "summing out" the probability of a random variable *X* given the joint probability distribution of X with other variable(s).

Discrete Y

$$p_X(x) = \sum_{y \in \Upsilon} p_{XY}(x, y)$$

Continuous Y

$$p_X(x) = \int_{\Upsilon} p_{XY}(x, y) \, \mathrm{d}y$$

- Marginalization is how to ignore variables.

Conditional Density

- $p_{X|Y}(x, y) : \Xi \times \Upsilon \to \Re$ is the *conditional density of* X *given* $Y = y$ if

$$p_{X|Y}(x, y) = P(\{\omega : X(\omega) = x\} \,|\, \{\omega : Y(\omega) = y\})$$

  for all $x \in \Xi$ if Ξ is finite, or if

$$\int_\xi p_{X|Y}(x, y)\ \mathrm{d}x = P(\{\omega : X(\omega) \in \xi\} \,|\, \{\omega : Y(\omega) = y\})$$

  for all $\xi \subseteq \Xi$ if Ξ is infinite.

- Given the joint density $p_{XY}(x, y)$, we can compute $p_{X|Y}$ as follows:

$$p_{X|Y}(x, y) = \frac{p_{XY}(x, y)}{\sum_{x' \in \Xi} p_{XY}(x', y)} \qquad \text{or} \qquad p_{X|Y}(x, y) = \frac{p_{XY}(x, y)}{\int_\Xi p_{XY}(x', y)\ \mathrm{d}x'}$$

Independent Events

Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Therefore $P(A|B) = P(A)$ if $P(B) \neq 0$

Example:

Pick a random number from $\{1,2,3,\cdots,10\}$, and call it N. Suppose that all outcomes are equally likely. Let A be the event that N is less than 7, and let B be the event that N is an even number.
Are A and B independent? Yes

$$P(A)=0.6,\ P(B)=0.5,\ P(A \cap B)=0.3$$

Conditional Independent Events

- Two events A and B are **conditionally independent** given an event C with P(C) > 0 if:

$$P(A \cap B | C) = P(A|C)P(B|C)$$

Therefore

$$\begin{aligned} P(A|B, C) &= \frac{P(A \cap B|C)}{P(B|C)} \\ &= \frac{P(A|C)P(B|C)}{P(B|C)} \\ &= P(A|C). \end{aligned}$$

Conditional Independent Events

Example:

A box contains two coins: a regular coin and one fake two-headed coin (P(H)=1). I choose a coin at random and toss it twice. Define the following events.

A= First coin toss results in an H.

B= Second coin toss results in an H.

C= Coin 1 (regular) has been selected.

We have $P(A|C) = P(B|C) = \frac{1}{2}$. Also, given that Coin 1 is selected, we have $P(A \cap B|C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. To find $P(A), P(B)$, and $P(A \cap B)$, we use the law of total probability:

$$\begin{aligned} P(A) &= P(A|C)P(C) + P(A|C^c)P(C^c) \\ &= \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \\ &= \frac{3}{4}. \end{aligned}$$

Similarly, $P(B) = \frac{3}{4}$. For $P(A \cap B)$, we have

$$\begin{aligned} P(A \cap B) &= P(A \cap B|C)P(C) + P(A \cap B|C^c)P(C^c) \\ &= P(A|C)P(B|C)P(C) + P(A|C^c)P(B|C^c)P(C^c) \\ &\qquad \text{(by conditional independence of } A \text{ and } B) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot 1 \cdot \frac{1}{2} \\ &= \frac{5}{8}. \end{aligned}$$

Independent Random Variables

We say X and Y are unconditionally independent or marginally independent, denoted $X \perp Y$ if

$$X \perp Y \leftrightarrow p(X, Y) = p(X)p(Y)$$

Unconditional independence is rare, however, usually this influence is mediated via other variables rather than being direct.

We therefore say X and Y are conditionally independent (CI) given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y \mid Z \leftrightarrow p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Mean and Variance

- Discrete random variable X:

$$E\,[X] = \mu = \sum_{x \in E} x\, p(x)$$

$$Var\,[X] = \sigma^2 = E\,[(X - \mu)^2] = \sum_E (x - \mu)^2 p(x) dx = E[X^2] - \mu^2$$

- Continuous random variable X:

$$E\,[X] = \mu = \int_E x p(x) dx \quad \text{(if integral is finite)}$$

$$Var\,[X] = \sigma^2 = E\,[(X - \mu)^2] = \int_E (x - \mu)^2 p(x) dx = E[X^2] - \mu^2$$

Mean and Variance, Example

Binomial distribution:

- Tossing a coin n times, with p as head probability.
- Let X ~ Bin(n, p) be the number of heads. Therefore E = {0, 1, 2, … n}
- Bin(k|n, p) = $\binom{n}{k}$ p^k $(1-p)^{(n-k)}$

$$
\begin{aligned}
\mu &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= np \sum_{k=0}^{n} k \frac{(n-1)!}{(n-k)!k!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{((n-1)-(k-1))!(k-1)!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^{\ell} (1-p)^{(n-1)-\ell} && \text{with } \ell := k-1 \\
&= np \sum_{\ell=0}^{m} \binom{m}{\ell} p^{\ell} (1-p)^{m-\ell} && \text{with } m := n-1 \\
&= np(p+(1-p))^m \\
&= np
\end{aligned}
$$

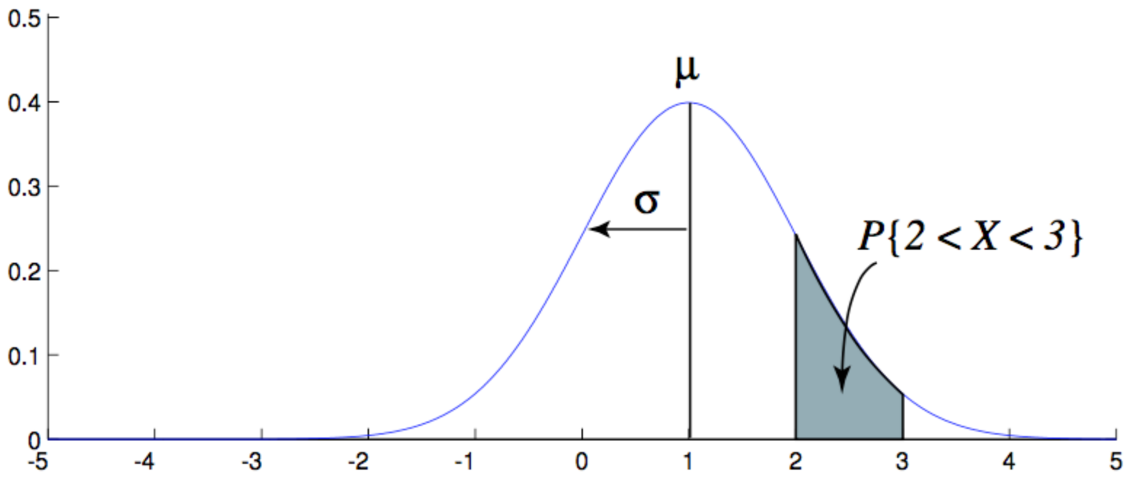
Continuous Density, Example

Gaussian distribution:

- One of the simplest densities for a real random variable.
- It can be represented by two real numbers: the mean μ and variance σ^2.

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Term $\sqrt{2\pi\sigma^2}$ is the normalization constant



Covariance and Correlation

The covariance between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as:

$$\operatorname{cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

If X is a d-dimensional random vector, its covariance matrix is defined to be the following symmetric, positive definite matrix:

$$\begin{aligned}
\operatorname{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] \\
&= \begin{pmatrix}
\operatorname{var}[X_1] & \operatorname{cov}[X_1,X_2] & \cdots & \operatorname{cov}[X_1,X_d] \\
\operatorname{cov}[X_2,X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2,X_d] \\
\vdots & \vdots & \ddots & \vdots \\
\operatorname{cov}[X_d,X_1] & \operatorname{cov}[X_d,X_2] & \cdots & \operatorname{var}[X_d]
\end{pmatrix}
\end{aligned}$$

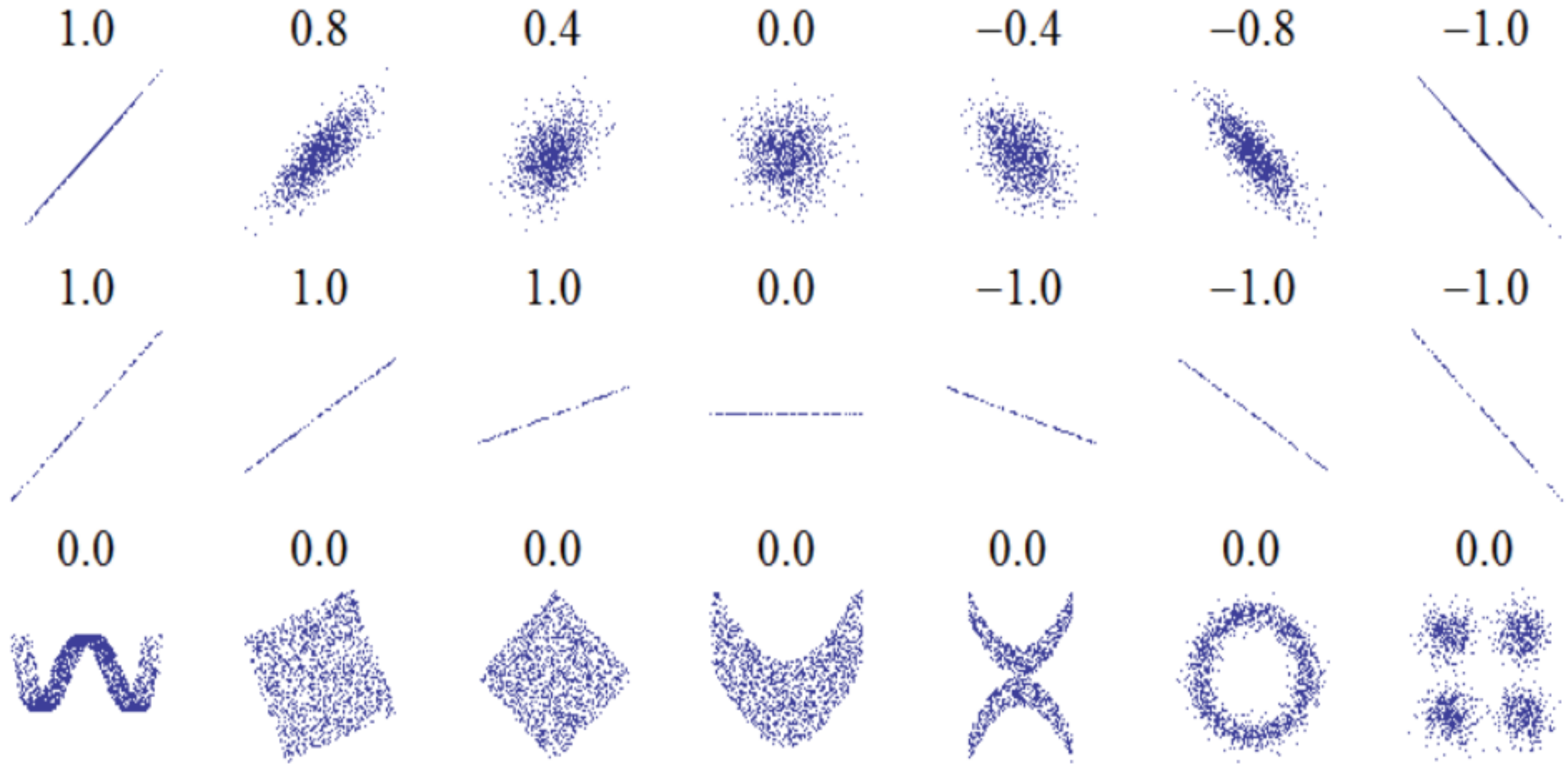
Covariance and Correlation

Covariance can be between 0 and infinity. Sometimes it is more convenient to work with a normalized measure. The correlation coefficient between X and Y is defined as:

$$\operatorname{corr}[X,Y] \triangleq \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}} \qquad -1 \le \operatorname{corr}[X,Y] \le 1$$

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1,X_1] & \operatorname{corr}[X_1,X_2] & \cdots & \operatorname{corr}[X_1,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d,X_1] & \operatorname{corr}[X_d,X_2] & \cdots & \operatorname{corr}[X_d,X_d] \end{pmatrix}$$

Correlation, Example

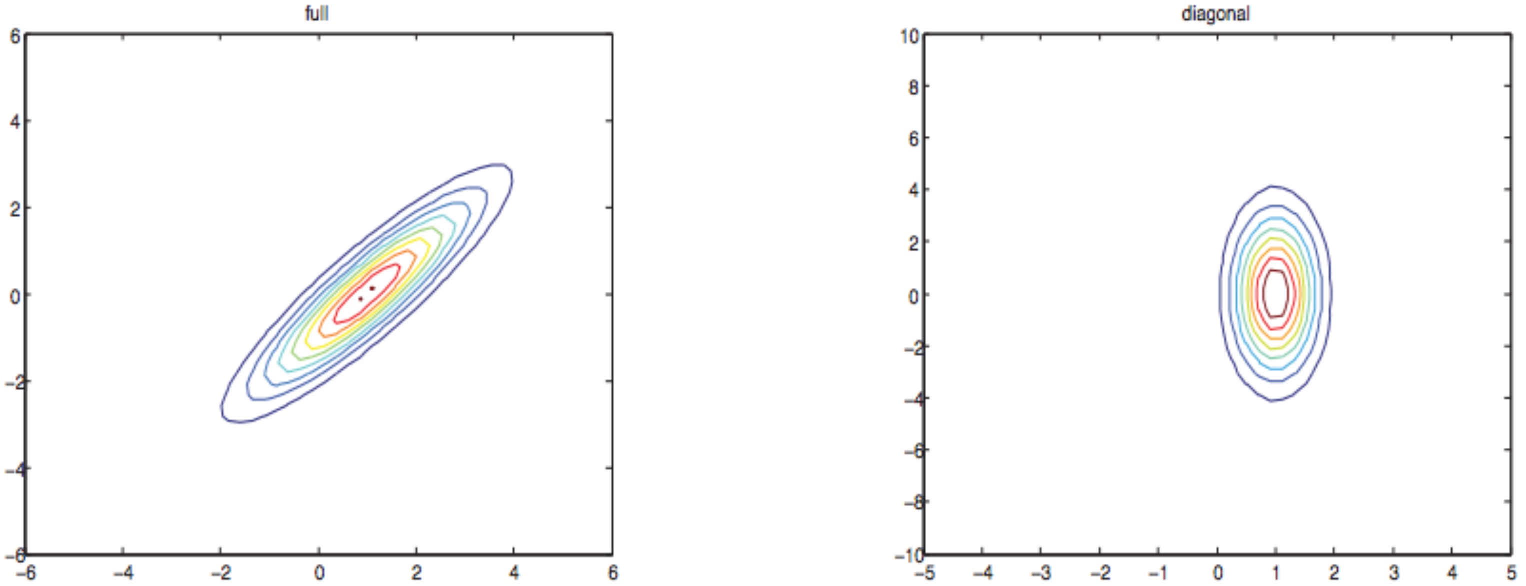


Correlation values (degree of linearity). If X and Y independent then corr(X, Y)=0

Joint Distribution, Example

- Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



(a) (b)