Statistical and Learning Techniques in Computer Vision
Lecture 1: Random Variables

Jens Rittscher and Chuck Stewart

1 Motivation

- Imaging is a stochastic process:
  - If we take all the different sources of error into account it is hard to argue that digital images are results of deterministic processes. One can argue that the images we obtain using a digital camera, a frame grabber, a microscope, or a thermal camera are realizations of a stochastic process.
- Capturing statistics about appearance can help to make inferences:
  - Here are different distributions for pixels that come from cars, from shadows and from the background (see figure 1)

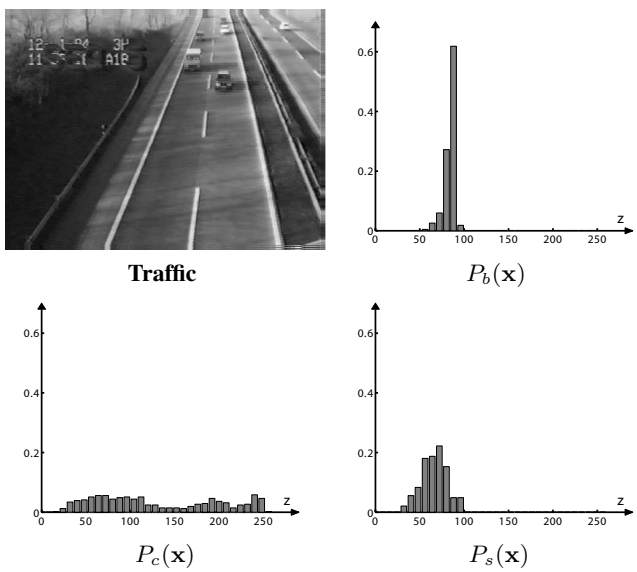

Figure 1: **Pixels as random variables.** *The image shows a sample frame of a surveillance camera installed to monitor traffic. In order to build a background model of the scene the set of pixels was devided into three different sets: background, cars, and shadow. The three graphs show the corresponding historgrams of the grey values of pixel locations from the different sets. Notice that the distribution of the background pixels $P_b(\mathbf{x})$ is sharply peaked. The distribution of greyvalues of pixels $P_c(\mathbf{x})$ that correspond to vehicles is, on the other hand, almost uniform.*

  - As another example, we can learn distributions of skin colors as an aid to face detection in images.

- Structural and appearance variations of objects are difficult to model explicitly, especially when combined with the behavior of computer vision algorithms.

2 Lecture Overview

We will go through the following material relatively quickly. Much of it is expected to be a review.

- Random variables
- Probability distributions and densities
- Mean and variance
- The Gaussian distribution
- Expectation
- Conditional probability, independence and Bayes theorem

3 A starting point: pixels as random variables

- In all of the examples above we view each pixel in the image as a random variable.
- The results of operations on pixels can also be modeled as random variables — e.g. smoothing operations, motion vector computations, or even edge detection results.
- Images are multivariate random variables.

4 Random variables

- A *random variable* is a mathematical function that assigns outcomes of a random experiment to numbers. To every outcome of this experiment we assign a number $\mathbf{x}(\xi)$ i.e.

$$\mathbf{x} : \mathcal{L} \rightarrow \mathcal{X} \tag{1}$$

  where $\mathcal{L}$ is the domain of the random variable and $\mathcal{X}$ the range. Examples include

  - Rolling a di.
  - Rolling a pair of dice. Note the difference in the domain!
  - Taking a photograph and measuring the intensity at each pixel in an image.

5 Distributions and Densities

- The *culmulative distribution function* of a random variable $\mathbf{x}$ is the function

$$F(x) = P\{\mathbf{x} \leq x\}, \tag{2}$$

  defined for very $x \in \mathcal{X}$.

- A random variable $\mathbf{x}$ is *continuous* if its distribution function, F, is continuous.

- $\mathbf{x}$ is *discrete* if F is a staircase function.

- In case F is discontinuous but not a staircase the random variable $\mathbf{x}$ is of mixed type.

- Note that

  - $F(x)$ is a monotonically non-decreasing function of x,
  - $\lim_{x\downarrow-\infty} F(x) = 0$, and
  - $\lim_{x\uparrow\infty} F(x) = 1$.

- The *empirical distribution* of a random variable $\mathbf{x}$ is constructed by performing an experiment n times and observing n values $x_1, \ldots, x_n$. Using the step function,

$$U(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0 \end{cases}, \tag{3}$$

  the empirical distribution of the random variable is

$$F_n(x) := \frac{\sum_{i=1}^{n} U(x - x_i)}{n}. \tag{4}$$

  - Stated less formally, $F_n(x)$ is the fraction of the observed values less than or equal to n.

  As $n \uparrow \infty$, $F_n \to F$.

- The derivative of $F(x)$,

$$f(x) = \frac{dF(x)}{dx}, \tag{5}$$

  is called the *density* of the random variable $\mathbf{x}$.

  - Properties of the density follow from properties of the distribution function.
  - For discrete random variables, the density is a series of impulse or probability mass functions. At every x_0 where $\lim_{x\uparrow x_0} F(x) \neq \lim_{x\downarrow x_0} F(x)$, the probability mass is just

$$p(x_0) = \lim_{x\downarrow x_0} F(x) - \lim_{x\uparrow x_0} F(x) \tag{6}$$

- Two or more random variables may be combined into a *random vector* $(\mathbf{x}_1, \ldots \mathbf{x}_k)$.
  - The cumulative distribution of $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ is
$$F(x_1, \ldots, x_k) = P(x_1 \leq \mathbf{x}_1, \ldots, x_k \leq \mathbf{x}_k), \tag{7}$$
    where on the right side of the equation the "," means "and".
  - The density is obtained by the kth partial derivative:
$$f(x_1, \ldots, x_k) = \frac{d^k F(x_1, \ldots, x_k)}{dx_1 \cdots dx_k}. \tag{8}$$

6 Mean and variance — the simplest statistics

- The *mean* of a random variable is
$$\mu(\mathbf{x}) = \int_{-\infty}^{\infty} x f(x) dx. \tag{9}$$
- The *mean* (vector) of a pair of random variables, $\mathbf{x}$ and $\mathbf{y}$, is
$$\mu \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \begin{pmatrix} x \\ y \end{pmatrix} f(x, y) dx dy \tag{10}$$
  Recall that this can be viewed as two double integrals:
$$\mu \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy \end{pmatrix} \tag{11}$$
- The mean of a discrete random variable is consequently computed as
$$\mu(\mathbf{x}) = \sum_{i=1}^{m} x_i p(x_i) \quad \text{when} \quad \mathbf{x} \in \{x_1, \ldots, x_m\}. \tag{12}$$
- The *variance* of a random variable is
$$\text{var}(\mathbf{x}) = \int_{-\infty}^{\infty} (x - \mu(\mathbf{x})) f(x) dx. \tag{13}$$
  This is sometimes written as $\sigma_{\mathbf{x}}^2$ or just σ^2.
- The *covariance matrix* of two random variables is
$$\mathbf{\Sigma}(\mathbf{x}, \mathbf{y}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \begin{pmatrix} x - \mu(\mathbf{x}) \\ y - \mu(\mathbf{y}) \end{pmatrix} \begin{pmatrix} x - \mu(\mathbf{x}) & y - \mu(\mathbf{y}) \end{pmatrix} f(x, y) dx dy \tag{14}$$
- The off-diagonal term of the resulting (symmetric) matrix,
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - \mu(\mathbf{x})][y - \mu(\mathbf{y})] f(x, y) dx dy \tag{15}$$
  is the *covariance* of $\mathbf{x}$ and $\mathbf{y}$.

7 Expectation

- Given a function $g(\mathbf{x})$ of a random variable $\mathbf{x}$ and the density $f(x)$ associated with $\mathbf{x}$, the *expectation* of g is

$$\mathbf{E}[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{16}$$

- The generalization to a pair or sequence of random variables is straightforward.
- We can write the mean and variance in terms of the "expectation operator":

$$\mu(\mathbf{x}) = \mathbf{E}[x] \tag{17}$$

$$\text{var}(\mathbf{x}) = \mathbf{E}[(x - \mu(\mathbf{x}))^2] \tag{18}$$

- We can think of $g(\mathbf{x})$ more generally as any function of x.
- Note that $\mathbf{y} = g(\mathbf{x})$ is a random variable if $\mathbf{x}$ is a random variable. We can use the expectation operator to compute the mean and variance of $\mathbf{y}$ based on the density of x.

8 The Normal or Gaussian Distribution

For many reasons this relatively simple distribution is the most widely used and studied.

- In case of a single variable the normal distribution is written as

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{where} \\ \mu &\quad \text{is the mean of the distribution} \\ \sigma^2 &\quad \text{is the variance, i.e.} \\ \sigma &\quad \text{the standard deviation} \end{aligned} \tag{19}$$

- The leading constant, $(\sqrt{2\pi}\sigma)^{-1}$, ensures that $\int p(x)dx = 1$, as required.
- The generalized multi-variate Gaussian is written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{20}$$

  where

  - d is the dimension of the vectors
  - $\boldsymbol{\mu}$ is the mean vector (center) of the distribution,
  - $\mathbf{\Sigma}$ is the $d \times d$, positive semi-definite, covariance matrix, and
  - $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$.

- The function
$$\triangle^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{21}$$
is called the Mahalanobis distance.
  - It is a better notion of the distance of the location $\mathbf{x}$ from the mean $\boldsymbol{\mu}$ because it takes into account the uncertainty encoded in $\boldsymbol{\Sigma}$.
  - Contours (surfaces, hypersurfaces) of constant Malanaobis distance in $\mathbb{R}^d$ are elliptical (ellipsoidal, hyperellipsoidal).
- The principal axes of the hyperellipsoids are given by the (unit) eigenvectors $\mathbf{u}_i$ of $\boldsymbol{\Sigma}$.
- In certain cases it is convenient to consider a simplified form of the Gaussian distribution in which $\boldsymbol{\Sigma}$ is a diagonal matrix.
$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2) \tag{22}$$
i.e. the components of x are statistically independent.
- The spectral decomposition of $\boldsymbol{\Sigma}$ produces a rotation matrix which can be applied to the coordinate system $\mathbf{x}$.
  - In the new coordinate system, $\boldsymbol{\Sigma}$ will be diagonal.
  - Whether or not this is a good thing to do depends on the application.
- The *central limit theorem* states that under general circumstances the mean of M random variables tends to be distributed normally, in the limit as M tends to infinity.

9 Conditional probability

- The *conditional probability* of an event $\mathcal{A}$ assuming an event $\mathcal{M}$ is
$$P(\mathcal{A}|\mathcal{M}) = \frac{P(\mathcal{A}, \mathcal{M})}{P(\mathcal{M})} \ , \tag{23}$$
where we assume that $P(\mathcal{M})$ is not 0.
- *Bayes's Theorem:* Given any events $\mathcal{A}$ and $\mathcal{B}$,
$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})} \tag{24}$$
  - The terms *a priori* and *a posteriori* are often used for the probabilities $P(\mathcal{A})$ and $P(\mathcal{A}|\mathcal{B})$.
  - We can prove Bayes's Theorem trivially from the definition of conditional probability.

- Both conditional probability and Bayes's Theorem are easily visualized using Venn diagrams.

- Two events $\mathcal{A}$ and $\mathcal{B}$ are *independent* if

$$P(\mathcal{A}, \mathcal{B}) = P(\mathcal{A})P(\mathcal{B}) \ . \tag{25}$$

  - Note how Bayes's Theorem collapses when two events are independent.
  - The concept of independence (and is complement) is fundamental, and the exercises in HW 1 will help give you a first insight.
  - Later in the semester we will often review the concept of independence and its implications.

10 Further Reading

Most of the above topics are covered in introductory textbooks on probability and statistics, as well as in background sections of recent books on pattern recognition and machine learning [Bis06] [DHS01].

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.