INSTITUT FÜR THEORETISCHE PHYSIK HEIDELBERG
PROF. DR. LUCA AMENDOLA
COMPUTATIONAL STATISTICS – SUMMER SEMESTER 2022 –

Tutors:

- Group 1: Ziyang Zheng
- Group 2: Nico Lorenz
- Group 3: Ziad Sakr
- Group 4: Adrian Schirra

Please, send (in pdf) or hand in the solution to this exercise sheet to your tutor, following their instructions, and carefully respecting the delivery date shown below. All exercise sheets will be graded. You can solve them individually or in pairs (with another student of the same tutorial group). In the latter case deliver please only one document with the solutions. Write the name of the file in the following format: X.Name1.Surname1 or X.Name1.Surname1-Name2.Surname2.pdf, with X being the number of the corresponding exercise sheet.

Exercise sheet 5

RECEIVED: May.23 - DELIVERY: May. 30

Exercise 5.1: Bivariate Gaussian (4pt)

Consider a bivariate Gaussian $p(x,y)\mathrm{d}x\mathrm{d}y$,

$$p(x,y) = \frac{1}{\sqrt{(2\pi)^2 \det C}} \exp\left(-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^t C^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right) \tag{1}$$

where C is the covariance matrix,which is in general non-diagonal, and μ_x, μ_y the means.

1. Show that for a diagonal covariance matrix the Gaussian separates, i.e. $p(x,y) = p(x)p(y)$. (2pt)

2. A conditional Gaussian distribution $p(x|y)\mathrm{d}x$ can be generated from by fixing the value y and keeping only x as the random variable. Assume you have a bivariate Gaussian measurement of the total time spent on learning about statistics with x as all the parents and y all children in Germany (among those that studied at least *some* statistics). The measurement results in $\mu_x = 1.3$ years $\sigma_x = 0.5$ years and due to the importance of statistics these days the children generation results in $\mu_y = 4.4$ years and $\sigma_y = 0.9$ years. (Assume x can be negative, so keep

the domain of integration from $-\infty$ to $+\infty$; this makes only a very small difference). If you have been already studying statistics for 4 years and the correlation coefficient is given by 0.2, the probability that your parents spent time x_1 to study statistics is $p = 10\%$. How much is x_1? How probable is it for your parents to have studied statistics for $x_2 = 1$ year ? (2pt)

Exercise 5.2: Python Exercise (6pt)

1. Find the mean, covariance matrix and correlation coefficient of the dataset "bivariate-measurements.txt" available in the the same dropbox folder where you find the recordings, sub-folder "data". You are not supposed to use the built-in function to compute the covariance matrix but to write your own function from scratch (3pt):

   (a) Calculate for both parameters x, y the the sample mean

   $$\hat{x} = \frac{1}{n}\sum_i x_i \tag{2}$$

   (b) Calculate again for both parameters the sample variance

   $$s_x^2 = \frac{1}{n-1}\sum_i (x_i - \hat{x})^2 \tag{3}$$

   (c) In the last step we can compute the covariance $\hat{V}$ and correlation coefficient r according to

   $$\hat{V}_{xy} = \frac{1}{n-1}\sum_i (x_i - \hat{x})(y_i - \hat{y}) \tag{4}$$

   $$r = \frac{\hat{V}_{xy}}{s_x s_y} \tag{5}$$

   $$\hat{V} = \begin{pmatrix} s_x^2 & r s_x s_y \\ r s_x s_y & s_y^2 \end{pmatrix} \tag{6}$$

2. Plot the data in a) two dimensional histogram by the line

   ```
   plt.hist2d(x,y,bins=50,density=True)
   ```

   and b) a two dimensional color plot. You can use scipy stats function for a multivariate Gaussian

   ```
   multi_gauss= scipy.stats.muultivariate_normal([x_bar,y_bar], [[s_x,V_xy],[V_xy,s_y]])
   ```

   in which you first set the means, followed by the covariance $\hat{V}$. To create a two dimensional colour plot you can use the following code:

```
xax,yax = np.mgrid[min(x):max(x):.01,min(y):max(y):.01]

pos= np.dstack((xax,yax))

plt.contourf(xax, yax, multigauss.pdf(pos),100)
plt.colorbar()
```

Compare a) and b) by eye: Has the estimation of the data been successful ? (3pt)

Exercise 5.3: Building a PDF Sampler (5pt)

For this exercise we want to produce random numbers distributed according to some given PDF. For that matter we would like to built a rejection sampler, in which we take two random numbers, one for the random number x itself, and one representing its probability $p(x)$. By comparing the probability of the random number to the second random number we decide whether to keep or discard the desired random number x.

First we need to take two uniform -distributed random generators. You have used such on the previous sheets. Use one generator to draw random numbers q on the interval $[0, \max(p)]$. The other one should cover your whole domain of your PDF to draw your desired data x. For our purposes use the domain $[-5, 5]$. We now want to keep or reject our drawn data x by comparing the PDF with our second random number q:

- If $q_i \leq p(x_i)$ keep x_i.
- If $q_i > p(x_i)$ discard x_i.

Show that the sampler works on a Gaussian distribution with zero mean and unit variance ($\mu = 0$ $\sigma = 1$) by creating a sample of n=10000 random numbers and show them in normalized histogram. To indicate that the function is Gaussian plot a Gaussian function with zero mean and unit variance into the histogram as well (similar to last weeks exercise 4.2). What is the sample mean and sample variance of the distribution? Does the the sample mean get closer to 0 if you increase the number of random numbers significantly (i.e. increase n by a factor of 100)?(5pt)