Statistics for Data Science — Bonus exercises

Wintersemester 2024/25

These exercises will not be graded and do not need to be returned

1. If Ω is a finite, non-empty set, the uniform probability measure $\mathbb{P}$ on Ω is the probability measure defined by

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

for all events A. Check that $\mathbb{P}$ is indeed a probability measure.

2. An urn contains 40 balls enumerated from 1 to 40. In a lottery, 6 balls are drawn without replacement from the urn. Tickets bearing the correct sequence of numbers, up to permutation of the numbers, win a T-shirt, while the ticket with the correct ordered sequence wins a car.

   (a) Compute the probability of winning a T-shirt.

   (b) Compute the probability of winning a car.

3. John claims: "If X is a continuous random variable with PDF f_X, then for all $x \in \mathbb{R}$,

$$\mathbb{P}(X = x) = \int_x^x f_X(y)\,\mathrm{d}y = 0.$$

Therefore, for any event A,

$$\mathbb{P}(X \in A) = \mathbb{P}\Big(\bigcup_{x \in A} \{X = x\}\Big) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} 0 = 0,$$

where we used the fact that the sets $\{X = x\}$ are disjoint for different x." What do you think of John's claim? Carefully justify your answer.

4. For a real-valued random variable X, we say that $m \in \mathbb{R}$ is a median for X if $\mathbb{P}(X \geq m) \geq 1/2$ and $\mathbb{P}(X \leq m) \geq 1/2$.

   (a) If X is a continuous random variable with CDF F, show that a median is provided by $F^{-1}(1/2)$.

   (b) Compute a median for X when X is a uniform random variable. Do the same with a Gaussian random variable and an exponential random variable.

5. Let $R^2 \sim \mathrm{Exp}(1/2)$ and $\theta \sim \mathcal{U}(0, 2\pi)$. We assume that R and θ are independent. Show that $X := R\cos(\theta)$ and $Y := R\sin(\theta)$ are two i.i.d. standard normal random variables.

6. We consider the function $f\colon \mathbb{R} \to \mathbb{R}_+$ given by

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}.$$

(a) Show that f is a PDF.

(b) Let X be a random variable with PDF f. We say that X is a Cauchy random variable. Show that X is not integrable.

7. Let $X \sim \mathcal{U}(0,1)$ and define the function $f(x) = \frac{10}{3}x^{3/2}$ for $x \in [0,1]$.

   (a) Calculate $\mathbb{E}[f(X)]$ and $\mathrm{Var}(f(X))$ by hand.

   (b) Write pseudocode to demonstrate how you would approximate the expected value $\mathbb{E}[f(X)]$ using the Monte Carlo method.

   (c) Determine the sample size n required for the Monte Carlo method to ensure that the root-mean-square error for estimating $\mathbb{E}[f(X)]$ is less than 10^{-2}.

8. Assume that $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(p)$ are i.i.d. with some unknown parameter $p \in (0,1)$. Solve the maximum likelihood estimator of p and show that it converges to the true, but unknown, value of the parameter p as the sample size $n \to \infty$.

9. Let $X \sim \mathcal{N}(1,3)$ and $f(x) = 1 + 2x + x^2$.

   (a) Calculate $\mathbb{E}[f(X)]$ and $\mathrm{Var}[f(X)]$ by hand.

   (b) Implement the Monte Carlo method to approximate the expected value of f, i.e.,

$$\mathbb{E}[f(X)] \approx f_M := \frac{1}{M}\sum_{i=1}^{M} f(x_i), \qquad x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1,3).$$

   Let $M = 1, 2, 4, 8, \ldots, 256$. For each M, compute $N = 10000$ simulations (realizations) of f_M. For each M, calculate the mean and the variance of f_M over the N rounds. Visualize your results by plotting the mean and variance of f_M over the N rounds as functions of M in two separate plots.

10. Let $x_1, \ldots, x_n$ be i.i.d. copies of the random variable $x \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 10$, $\sigma = 5$, and sample size $n = 50$.

    Generate 100 *different* random samples $x_1, \ldots, x_n$ using a random number generator. Then, for each sample, compute a 95% confidence interval of the mean. For the simulation of the confidence intervals, assume that the population mean μ and population variance σ^2 used to generate the samples are *unknown*. How many confidence intervals contain the true population parameter $\mu = 10$ in your experiments?

11. Opinion polls are often conducted before parliamentary elections. In an upcoming election, two polls were conducted. In poll 1, the sample size was 1000 and 200 out of the 1000 eligible voters reported that they support the party Statistocrats. In poll 2, which was conducted later, the sample size was 1120 and 200 out of the 1120 eligible voters reported that they support Statistocrats. Based on these polls, can one conclude that support of Statistocrats has decreased?

12. (a) Give the general statistical assumptions needed for applying the Wilcoxon one sample signed rank test.
    (b) Give the null hypothesis and the two sided alternative hypothesis of the Wilcoxon one sample signed rank test.
    (c) Give the general statistical assumptions needed for applying the one sample sign test.
    (d) Give the null hypothesis and the two sided alternative hypothesis of the one sample sign test.
    (e) Compare the statistical assumptions of the Wilcoxon one sample signed rank test and the one sample sign test. Which one of these tests has milder assumptions?

13. Explain how you would visualize the following data sets.
    (a) The eye colors of students attending the Statistics for Data Science course. (Draw an example.)
    (b) The heights of male students attending the Statistics for Data Science course. (Draw an example.)
    (c) The relationship between the heights and shoe sizes of students attending the Statistics for Data Science course. (Draw an example.)
    (d) The proportions of seats by political parties in the German parliament. (Draw an example.)
    (e) The proportion of faulty products obtained from a manufacturing process. (Draw an example.)

14. Consider independent and identically distributed (i.i.d.) bivariate observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
    (a) What type of dependence can be measured using the Pearson correlation coefficient?
    (b) What type of dependence can be measured using the Spearman correlation coefficient?
    (c) Give an example of a dependence type that can neither be detected using Pearson correlation coefficient nor using Spearman correlation coefficient, but that can be detected from a scatter plot.

15. You are working as an expert in a group assisting policymakers. Policymakers believe that the best way to reduce public healthcare costs is to lower the weight of people. BMI (body mass index) under 18.5 is considered underweight, 18.5–25 is normal weight, 25–30 is overweight, and above 30 is obese. GHS (general health score) measures an individual's general health status. The value 10 indicates perfect health, and if the value is above 30, this indicates the individual has severe health problems. The BMI and GHS measurement results of seven individuals are shown below.

BMI (x)	GHS (y)
27	10
40	65
15	70
18	45
37	40
23	10
25	9

(a) Plot a scatter plot of the dataset.

(b) Calculate the Pearson correlation coefficient for the dataset.

(c) Calculate the Spearman correlation coefficient for the dataset.

(d) Describe the relationship between variables x and y and interpret the results in (a), (b), and (c).

(e) Policymakers believe that the best way to reduce public healthcare costs is to lower the weight of people. Based on this dataset alone, do you agree with the policymakers? Justify your answer.

16. Consider the multivariate linear regression model

$$y_i = b_0 + B^{\mathrm{T}} x_i + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where the elements of the 2×1 vector b_0 and 4×2 regression matrix B are unknown constants and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

(a) What does the variance inflation factor (VIF) measure?

(b) Give the definition of VIF for the explanatory variable $(x_i)_3$ and explain how it is calculated.

(c) Explain how VIF can be used in selecting explanatory variables.

17. Let $x, y, \eta \in \mathbb{R}^2$. Consider the Bayesian inverse problem

$$y = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x + \eta$$

with additive noise $\eta \sim \mathcal{N}(0, \gamma^2 I_2)$, where $I_2 \in \mathbb{R}^{2\times 2}$ is an identity matrix. Suppose that the prior distribution is given by $x \sim \mathcal{N}(0, I_2)$. What is the posterior distribution of $x|y$ if we observe $y = \begin{pmatrix} 1 & 2 \end{pmatrix}^{\mathrm{T}}$? What is the posterior covariance? What happens to the posterior distribution and posterior covariance under decreasing noise ($\gamma \downarrow 0$)?

18. Assume that $Z^0 \sim \mathcal{N}(1, 1)$ and Z^1 is given as

$$Z^1 = Z^0 + \Sigma,$$

where $\Sigma \sim \mathcal{N}(0, 1)$, and the observation model is

$$Y = Z^1 + \Xi,$$

where $\Xi \sim \mathcal{N}(0, 1)$. The random variables Z^0, Σ, and Ξ are assumed to be independent.

(a) Forecast step: What is the distribution of Z^1?

(b) Filtering step: What is the distribution of Z^1 conditioned on $Y = 1$?

(c) For the ensemble Kalman filter with perturbed observations, the analysis is obtained using the non-deterministic coupling

$$\hat{Z}^a = Z^1 - \alpha(Z^1 + \xi - y_{\text{obs}}),$$

where $\xi \sim \mathcal{N}(0,1)$ and $y_{\text{obs}} = 1$. Find the value of $\alpha \in \mathbb{R}$ such that $\hat{Z}^a$ has the same distribution as $Z^1|Y = 1$. Validate your choice of α by computing explicitly the mean and variance of $\hat{Z}^a$ and compare these with the mean and variance of $Z^1|Y = 1$.

19. Let $x_0 \sim \mathcal{N}(m_0, C)$ and consider the nonlinear dynamics

$$x_{j+1} = \Psi(x_j) + \xi_{j+1}, \quad j = 0, 1, 2, \ldots,$$

where $\xi_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, and the measurement model

$$y_{j+1} = Hx_{j+1} + \eta_{j+1}, \quad j = 0, 1, 2, \ldots,$$

where $\eta_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma)$. The random variables x, ξ, and η are assumed to be independent.

3DVAR filtering is based on the following updating scheme:

Prediction:

$$\widehat{m}_{j+1} = \Psi(m_j), \quad j = 0, 1, 2, \ldots.$$

Correction:

$$m_{j+1} = (I - KH)\widehat{m}_{j+1} + Ky_{j+1}, \quad j = 0, 1, 2, \ldots,$$

where $K = CH^{\mathrm{T}}(HCH^{\mathrm{T}} + \Gamma)^{-1}$ is the Kalman gain matrix.
(a) Show that

$$m_{j+1} = \arg\min_v \left\{ \tfrac{1}{2}(y_{j+1} - Hv)^{\mathrm{T}}\Gamma^{-1}(y_{j+1} - Hv) + \tfrac{1}{2}(v - \widehat{m}_{j+1})^{\mathrm{T}}C^{-1}(v - \widehat{m}_{j+1}) \right\}.$$

(b) If Ψ is linear, i.e., $\Psi(x) = Mx$ for some matrix M, how is the expression for m_{j+1} related to the Kalman filter?

20. The random walk Metropolis algorithm scales poorly with increasing dimension.[†] Meanwhile, the single component Gibbs sampler is computationally expensive for high-dimensional problems. A surprisingly effective alternative is the so-called *Metropolis-within-Gibbs algorithm*, which combines the powerful Gibbs sampler with the computationally inexpensive Metropolis algorithm. The algorithm to draw a sample from the d-dimensional probability density function f can be described as follows:

[†] The preconditioned Crank–Nicolson (pCN) method can be used to carry out dimension-robust sampling, but it requires careful tuning of the free parameter $\beta \in (0, 1)$.

1. Choose the initial value $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
2. Draw the next sample as follows:
   (i) Set $x = x^{(k)}$ and $j = 1$.
   (ii) Draw $t \in \mathbb{R}$ from the one-dimensional distribution
   $$f(t|y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_d) \propto f(y_1, \ldots, y_{j-1}, t, x_{j+1}, \ldots, x_d)$$
   by performing one step of the Metropolis algorithm and set $y_j = t$.
   (iii) If $j = d$, set $y = (y_1, \ldots, y_d)$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (ii).
3. Set $x^{(k+1)} = y$, increase $k \leftarrow k + 1$ and return to step 2.

Suppose that we are interested in estimating a signal $g\colon [0,1] \to \mathbb{R}$ from noisy, blurred observations modeled by

$$y_i = y(s_i) = \int_0^1 K(s_i, t) g(t)\,\mathrm{d}t + \varepsilon_i, \quad i \in \{1, \ldots, k\}, \tag{1}$$

where $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \ldots, k\}$, the blurring kernel is

$$K(s,t) = \exp\bigg(-\frac{1}{2 \cdot 0.05^2}(s-t)^2\bigg),$$

and we have i.i.d. Gaussian measurement noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 10^{-3}$. As discussed during the lectures, the integral equation (1) can be discretized using the midpoint rule with points $t_j = \frac{j}{d} - \frac{1}{2d}$, $j \in \{1, \ldots, d\}$, to obtain the linear measurement model

$$y = Ax + \varepsilon, \tag{2}$$

where $y = [y_1, \ldots, y_k]^{\mathrm{T}} \in \mathbb{R}^k$ is the measurement, $A = \big(\frac{1}{d}K(s_i, t_j)\big)_{\substack{i=1,\ldots,k\\ j=1,\ldots,d}} \in \mathbb{R}^{k\times d}$ is the system matrix, and $x = [g(t_1), \ldots, g(t_d)]^{\mathrm{T}} \in \mathbb{R}^d$ is the unknown.

Download the file `signal.mat` from the course website. The file contains the objects `y`, `A`, and `t` corresponding to the noisy, blurred signal y, the system matrix A, and the vector t, respectively. The file can be imported in Python with the command

```
data = scipy.io.loadmat('signal.mat')
```

and you can access the objects by calling `data['y']`, `data['A']`, and `data['t']`. Note that $k = d = 100$.

Suppose that we know *a priori* that the true signal x corresponds to a piecewise constant function $g\colon [0,1] \to \mathbb{R}$. A reasonable choice for the prior would then be the so-called *anisotropic total variation prior*

$$f(x) \propto \exp\bigg(-\lambda \sum_{k=1}^{d} |x_{k+1} - x_k|\bigg), \quad \lambda > 0, \tag{3}$$

where we assume periodic boundary conditions, i.e., $x_{d+1} = x_1$.

Your task is as follows:

Write down the posterior density $f(x|y)$ for the unknown parameter $x \in \mathbb{R}^d$ in (2) using the prior (3) with $\lambda = 100$. Use the Metropolis-within-Gibbs algorithm with random walk Metropolis step size $\gamma = 0.05$ to draw a sample of size $N = 10^4$ from the posterior density, and approximate the CM estimator $\hat{x}_{\text{CM}}$ of the unknown parameter x by computing the sample average. Finally, visualize the approximate CM estimator you obtained by plotting it as a function of `t`.

Hint: Your reconstruction should look a bit like the boxcar function $\mathbf{1}_{[0.3,0.7]}$, which was the function used to generate the measurement data.