Exercises for Course on Modeling and System Identification (MSI)
Albert-Ludwigs-Universität Freiburg – Winter Term 2024-2025

# Exercise 2: Statistics + Parameter Estimation

**(to be returned before November 4th, 8:15)**

Prof. Dr. Moritz Diehl, Katrin Baumgärtner, Jakob Harzer, Dan Wang,
Ashwin Karichannavar, Premnath Srinivasan

In this exercise, we repeat some tools from statistics and solve a simple estimation problem using linear least squares.

## Exercise Tasks

1. ON PAPER: The covariance matrix of a vector-valued random variable $X \in \mathbb{R}^n$ with mean $\mathbb{E}\{X\} = \mu_X$ is defined by

$$\text{cov}(X) := \mathbb{E}\{(X - \mu_X)(X - \mu_X)^\top\}.$$

Prove that the covariance matrix of a vector-valued variable $Y = AX + b$ with constant $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ is given by

$$\text{cov}(Y) = A \,\text{cov}(X) A^\top.$$

(2 points)

2. ON PAPER: Let $X \in \mathbb{R}^n$ be a vector-valued random variable with mean $\mu \in \mathbb{R}^n$. Show that the covariance matrix $\text{cov}(X)$ can also be calculated by

$$\text{cov}(X) = \mathbb{E}\{XX^\top\} - \mu\mu^\top$$

(2 points)

3. ON PAPER: Suppose we are measuring a constant $x_0 \in \mathbb{R}$ perturbed by random independent noise $\epsilon$ with mean $\mu_\epsilon = 0$ and variance $\sigma_\epsilon^2 > 0$, i.e. we have

$$x = x_0 + \epsilon.$$

(a) State the mean $\mu_x$ and the variance $\sigma_x^2$ of the random variable $x$. (1 point)

(b) Let $x(n) = (x_1, \dots, x_n)$ denote a sample of $n$ observations of $x$. The sample mean is given by $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and it is an unbiased estimator of the mean $\mu_x$.
What is the variance of $\bar{x}(n)$? (1 point)

(c) Prove that the Least Squares (LS) estimate for $x_0$ is the sample mean $\bar{x}(n)$. State the minimization problem explicitly. Is it convex? (2 bonus points)

4. Consider the following experimental setup, where we measure the temperature-dependent expansion of a fluid in a long transparent pipe, such as in a traditional thermometer. We describe the length of the visible fluid with the affine model

$$m(T; \theta_1, \theta_2) = \theta_1 \cdot T + \theta_2. \tag{1}$$

where $T$ is the temperature in Celsius, and the parameters $\theta_1$ and $\theta_2$ relate to the specific expansion coefficient of the material and the length of the fluid at temperature $T = 0\,°\mathrm{C}$, respectively. Below, you find the measurements. Using the data, you will compute estimates for the parameters $\theta_1$ and $\theta_2$.

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $T(k)$ [°C] | 5 | 15 | 35 | 60 |
| $L(k)$ [cm] | 6.55 | 9.63 | 17.24 | 29.64 |

(a) CODE: Plot the measurements $T(k)$, $L(k)$ using 'x' markers. (0.5 points)

(b) ON PAPER: Using the model from above, calculate the experimental values for the parameters $\theta_1$ and $\theta_2$ by minimizing the sum of squared distances, i.e.

$$\theta_1^*, \theta_2^* = \underset{\theta_1,\,\theta_2}{\arg\min} \quad \sum_{k=1}^{4} \left(L(k) - m(T(k); \theta_1, \theta_2)\right)^2,$$

Give an analytical expression for the values of $\theta_1^*$ and $\theta_2^*$ with respect to the measurements $T(1), \ldots, T(4)$ and $L(1), \ldots, L(4)$.

*Hint: Compute the solution by setting the gradient of the objective function $f(\theta_1, \theta_2) = \sum_k (L(k) - m(T(k), \theta_1, \theta_2))^2$ with respect to the parameters ($\theta_1$, $\theta_2$) to zero, i.e. $\nabla f(\theta_1, \theta_2) = 0$. This will give you a $2 \times 2$ linear system. Check if the objective function is convex!*

CODE: Calculate the values of $\theta_1^*$ and $\theta_2^*$ using the data. Plot the fit $m(T; \theta_1^*, \theta_2^*) = \theta_1^* T + \theta_2^*$ over the range $[0, 100]$ in the same figure as before. (2 points)

(c) CODE: Now, use a third order polynomial and fit it to the data using `np.polyfit`. Again minimize the sum of squared distances to find optimal values for the coefficients of your model equation. Plot the fit in the same figure as before. (0.5 point)

(d) CODE: You take another measurement: At $T = 70\,°\mathrm{C}$ you measure a length of $L = 32.89\,\mathrm{cm}$. You can use this additional datapoint to validate your fit. Add the measurement to the existing plot.

ON PAPER: Which fit looks more reasonable to you?

*Hint: The phenomenon of fitting a model to a data set which then does not pass validation is called 'overfitting'.* (1 point)

*This sheet gives in total 10 points and 2 bonus points.*