**OPTION A: DATA ANALYSIS CASE STUDY**

**Instructions**

1. Submit a compiled PDF or HTML report to Gradescope. Excluding code, the report should be at most 1200 words. Include all figures in the main text.
2. Submit the Rmarkdown or Quarto document you used to generate the report, as well as any R scripts defining functions you used. We should be able to rerun your analysis (assuming we download the data to the right directory and install the packages).

**Description**

In this assignment, you will have the chance to explore how statistical multi-omics methods can be applied to a real dataset. Your report will include three main sections: problem formulation, method application, and interpretation / reflection. You are free to choose a dataset and a statistical method that is interesting to you, but you should not simply copy an existing data analysis. We hope that, through this hands-on case study, you will come to critically understand the potential value (but also the limitations) of current multi-omics techniques.

In the problem formulation section of your report, you will describe the scientific question of interest, the data that were collected, and the way that your chosen method can bridge the two. Explain why the scientific question is interesting. Summarize the experimental design and the technologies used to generate each dataset. Identify the method you will use to analyze the data. You do not need to explain its algorithmic details, but you do need to justify its applicability to the overall scientific problem.

In the method application section of the report, walk the reader through the code used for data analysis. Describe any preprocessing steps that you took before applying the method, or if the data were already preprocessed when you downloaded it, explain what steps were taken. Do you have any evidence that the method’s assumptions are appropriately met? Did you try running the method with different hyperparameters? Save your results into a separate object / file, so that you can return to interpret them without having to rerun all the methods.

In the interpretation section, generate some tables and / or visualizations to summarize your results. Describe the main takeaways as if you were writing to a research collaborator. Were the results what you expected, or did you encounter any surprises? Finally, reflect on the overall case study. What aspects of the workflow do you think have the most room for improvement (e.g., computation time, appropriateness of modeling assumptions, complexity of preparing input data)? Explain your reasoning.

**Rubric**

Problem Context [10 points]: Fully researched and developed discussion of the context.

Implementation [10 points]: Correct and concise code implementation of the data analysis.

Interpretation [10 points]: Accurately and insightfully interprets data analysis results.

**OPTION B: BENCHMARKING EXPERIMENT**

**Instructions**

1. Submit a compiled PDF or HTML report to Gradescope. Excluding code, the report should be at most 1200 words. Include all figures in the main text.
2. Submit the Rmarkdown or Quarto document you used to generate the report, as well as any R scripts defining functions you used. We should be able to rerun your analysis (assuming we download the data to the right directory and install the packages).

**Description**

For any multi-omics problem, there are bound to be many candidate data analysis approaches. To help isolate promising directions, it is helpful design benchmarking experiments. By focusing on simple settings where the ground truth is known, these experiments can highlight methods which are sensitive and robust. This problem gives you a chance to develop such a study. You are free to choose a problem domain of your own interest. Your final report should have three main sections: a motivating discussion of the experiment’s context and metrics, details of its design and implementation, and an interpretation of its results.

In the first section, you provide context for your study. What is the “abstract” statistical problem that you would like to focus on? What is its scientific motivation, and what types of data are typically available? Describe 2 - 4 software packages that have been proposed for this analysis problem and explain the typical inputs and outputs to these packages. Finally, define metrics that can be used to compare proposed approaches.

In the second section, you will design and implement a benchmarking experiment comparing the packages that you identified above. The requirements are that you (1) compute one of your metrics of interest and (2) vary several parameters that could potentially influence performance. You may either simulate data from scratch or base the experiment on a real dataset – either way, be sure to describe your data generating procedure in detail. Next, apply your selected packages, and make sure to run at least one replicate for each setting.

In the final section, you will interpret and reflect on your results. Include at least one figure or table. Was there any noticeable “winner” among the methods you compared? Is the best method context dependent? Did you observe any settings or metrics for which the considered methods do not seem satisfactory? What would you recommend in practice, and what do you think are the most interesting opportunities for development?

**Rubric**

Problem Context [10 points]: Fully researched and developed discussion of the context.

Implementation [10 pointes]: Correct and concise code implementation of the experiment.

Interpretation [10 points]: Accurately and insightfully interprets benchmarking results.

**OPTION C: TECHNICAL REVIEW**

**Instructions**

1. Submit a compiled PDF or HTML report to Gradescope. Excluding code, the report should be at most 1200 words. Include all figures in the main text.
2. Submit the Rmarkdown or Quarto document you used to generate the report, as well as any R scripts defining functions you used. We should be able to rerun your analysis (assuming we download the data to the right directory and install the packages).

**Description**

One of the best ways to learn the details of a new method is to try explaining it using your own language and examples. In this exercise, you will explore a multi-omics method in-depth, preparing a readable introduction to your peers. Your write-up will have three main sections: problem context, method description, and code analysis.

In the problem context, summarize the scientific motivation behind the proposed method and the data sources that are typically available. What is the interpretation of these data? How are the scientific goals formulated as a statistical problem? What were some of the more basic methods that have previously been applied to the problem, and why are they unsatisfactory? You should assume the reader is someone who has taken this class, but who is not familiar with your chosen method.

In the second section, describe the mechanics of the proposed method. You may sacrifice tertiary details to emphasize the main ideas of the approach. You are encouraged to either develop a visual summary (hand-drawn is okay) of the method or prepare a toy example using highly reduced data. Rather than repeating the explanation from the original publication for the method, try to independently organize, interpret, or simplify the relevant concepts.

For the final section, first skim an existing software implementation of the method. Relate your algorithm description from the previous section with specific functions or variables in the code. Did you notice any steps of the implementation that were not discussed in the publication? What important aspect of the implementation did you find most confusing at first, and how did you eventually come to understand how it works?

**Rubric**

Problem Context [10 points]: Fully researched and developed discussion of the context.

Method Summary [10 points]: Clearly and creatively describes the method of interest.

Code Analysis [10 pointes]: Fluently between abstract statistical concepts and software implementations.