**Title: From Genes to Tumors: A Comparison of Feature Selection Techniques in Logistic Regression**

Feedback to: Floris, Ward and Anna.

Feedback from: Su, Valerie and Solène

**General**

The report is written in correct English and generally adheres to the word count limits. The report looks good visually (e.g., layout), the different (sub)section headings are appropriate, the references are neat, and the abstract is well-written and contains all relevant information so that someone who has not read the report can assess the contents of the report.

### **Introduction**

Explanation of the data is entirely correct, the features used in modelling are adequately discussed. The substantive research problem is adequately contextualized, and the introduction naturally leads to a clear research question.

The introduction is well-structured, adhering to the word limit of 700 words. It begins with a clear explanation of the research problem, contextualizing the phenomenon of gene dysregulation in tumor tissue. The dataset is described in detail, including the number of observations and features, effectively highlighting the challenge of high dimensionality. The different methods employed to identify differences in gene expression patterns between normal and tumor conditions are clearly explained, with a research question implicitly conveyed through the explanation of these methods. Following the discussion of the techniques and their associated limitations, adding a little summary at the end clearly stating the aim of the project (in one sentence) would improve the overall message of the introduction.

### **Methods**

The chosen analysis methods are correctly and concisely explained, as well as their rationale and relevance to this research problem. The parts about method comparison define a clear metric, and the validation approach takes into account bias-variance trade-off / overfitting / generalization to unseen data.

The methods section adheres to the word count limit of 500 words. It provides a detailed explanation of each method, including the construction of the models and their implementation using specific R packages and their corresponding functions. The comparison strategy is outlined at the end of the method, specifying that models will be evaluated using cross-validation and classification metrics, with Figure 2 illustrating the K fold cross validation approach. However, there is no mention of interpreting results for each model, such as identifying or comparing the most differentially expressed genes between the two conditions. Incorporating this aspect could provide additional insights into gene expression patterns in tumor and normal tissues. The suggestion for doing hyperparameter tuning in sparse PCA is to calculate sparsity, as it helps determine the percentage of genes per component. For example, if sumabs=0.1, it means 602 genes in one PC. For the biological interpretation part, this is a useful way to investigate the correlation of genes within each component.

The primary focus of the methods is on classification differences, but it is worth questioning whether the ranking of gene importance for classification will also be compared across techniques.

### **Results**

The results section is well-connected to the methods section (flows naturally), it focuses on the main findings while using appropriate presentation methods (i.e., text / table / figure depending on ease of understanding) with accurate descriptions.

The planned results section is well-structured and logically aligned with the preceding methods section. For the classification tasks, comparisons between methods will include the use of confusion matrices and differences in performance metrics, which will be further interpreted. Additionally, although not explicitly mentioned in the methods section, a PCA plot will be analyzed to identify important genes for comparison—an aspect directly relevant to the research question and which should be incorporated into the methods description. Investigating the key genes identified by the "normal" logistic regression model would also add value, providing a basis for further comparison between methods. Little addition that for the PCA analysis, it would be relevant to mention the number of principal components selected based on the 95% variance threshold to clearly define the PCA model used.

### **Conclusion**

The conclusion summarizes the work in an understandable way, and provides a well-defined answer to the research question that follows from the results in the report. A discussion adds context to the findings, includes theoretical interpretations beyond the factual results, and shows good understanding of issues in high-dimensional analysis and its application in substantive research.

The conclusion has not yet been finalized, but some hypotheses have been proposed. Based on these hypotheses, the research question—comparing logistic regression models with PCA-enhanced logistic regression models for analyzing gene expression data between tumor and normal tissues— is addressed.