Data Lab 2: Breast Cancer Subtyping

ACMS 40876/60876 (Spring 2026)

Due Friday, February 20, 2026 6:00 PM via GitHub

Background and Goals

In recent years, there have been many notable scientific discoveries that have been driven by unsupervised learning. One example is the discovery of breast cancer subtypes, which were identified using hierarchical clustering applied to miRNA array data. Without going into too many biological details, breast cancer is a very heterogeneous disease with many *subtypes*, or smaller groups of patients whose cancers behave similarly. Importantly, these breast cancer subtypes have different prognoses and respond differently to various treatments.

In this lab, you will conduct your own unsupervised learning analysis in order to rediscover these breast cancer subtypes. Through this process, you will gain experience with (1) applying various dimension reduction and clustering methods on real-world data and (2) developing intuition for the many different data preprocessing and modeling choices that can substantially impact your findings.

Instructions

You will be engaging with RNASeq gene expression data from breast cancer patients in The Cancer Genome Atlas.[1] In this dataset, each row corresponds to a different breast cancer patient, each column corresponds to a different gene, and each (i, j) entry in the data matrix corresponds to the expression of gene j for patient i. In total, there are $n = 1043$ patients and $p = 5000$ genes.[2] Generally speaking, you can think of high gene expression as a gene being turned "on" (i.e., high-functioning and producing lots of proteins) while low gene expression means that the gene has been turned "off" (i.e., low-functioning and producing few proteins). In what follows, you will be guided through steps to conduct an unsupervised learning analysis to rediscover breast cancer subtypes.

Note: your RNASeq gene expression dataset is different from the miRNA data used to first identify breast cancer subtypes. Consequently, do not be under the impression that you have to obtain exactly the same results as what has already been discovered about breast cancer. When you conduct your analysis, proceed as if you were a clinician/scientist and want to obtain the most *reliable* findings that are best supported by the *given* data.

As in Lab 1, please create a reproducible report with your analysis and narrated discussion. Submission details can be found at the end of this document.

[1] *Optional:* If you are curious and would like to learn more about RNASeq, here is an introductory video.

[2] The number of genes in the original dataset is actually much larger than 5000; however, we often perform a "variance-filtering" preprocessing step with this type of gene expression data. This variance-filtering removes genes with the lowest variance (e.g., those that are always turned off) and keeps the top X genes with the highest variance.

Exploratory Data Analysis

1. Conduct a brief exploratory data analysis for the provided gene expression data. At a minimum, please (a) plot the distribution of all (raw/original) gene expression values in a histogram or density plot and (b) produce one other EDA plot of your choice. Discuss your main observations or takeaways from these two EDA plots.

Dimension Reduction

2. **Dimension reduction on the raw gene expression data:**
   (a) Choose two different dimension reduction methods.[3]
   (b) To facilitate transparent and reproducible research,[4] document your choice of methods, the hyperparameters you chose, and why you made these choices.
   (c) Apply your chosen dimension reduction methods to the raw gene expression dataset, and plot the results of each dimension reduction method.[5] Briefly discuss your observations from the plots.
3. **Dimension reduction on the log-transformed gene expression data:** In genomics, it is common to transform the raw gene expression data, denoted by X, using a log-transformation by computing $\tilde{X} = log(X + 1)$. Let us refer to $\tilde{X}$ as the *log-transformed* gene expression data.
   (a) Using the same dimension reduction methods that you chose in problem (2), re-apply them on the log-transformed gene expression data $\tilde{X}$, and plot the results of each dimension reduction method.
   (b) Are the dimension reduction results similar between the different methods (applied to $\tilde{X}$)? Justify your answer using both visual and quantitative evidence.
   (c) Are the dimension reduction results applied to X similar to those applied to $\tilde{X}$? Justify your answer using both visual and quantitative evidence.
4. **Dimension reduction discussion questions:**
   (a) Based upon your observations of the dimension reduction results from problem (2) and problem (3) above, what is the issue with applying dimension reduction methods to the raw gene expression data X? How does the log transformation $log(X + 1)$ help to mitigate this issue?
   (b) Can z-score standardization (i.e., centering and scaling to have mean 0 and standard deviation 1) also mitigate this issue? Why or why not?

[3] Do NOT choose the same dimension reduction method (e.g., UMAP) with two different hyperparameters.

[4] In statistics/data science, you should always describe and detail your methodological choices such that a general reader can reproduce your analysis by simply reading the text without digging into your code.

[5] *Note:* If the computational burden of applying these dimension reduction methods to the full 5000 genes is too high for your computer, you may subset the number of genes to a smaller number (e.g., 1000 genes) for the purposes of this lab. If you choose to subset the number of genes, document how you chose which genes to keep and which genes to discard (e.g., via variance-filtering, random selection, etc.).

Clustering

5. **Clustering on the dimension-reduced data:**
   (a) Choose two different clustering methods. Again, document your choice of methods, the hyperparameters you chose, and why you made these choices.
   (b) For each clustering method chosen here and for each dimension reduction method chosen above, apply the clustering method using the dimension-reduced data from problem (3) as the input. Be sure to document the number of components you used from the dimension-reduced data, justifying your choice when possible. Plot the clustering results for $k = 2, 3, 4, 5$, where k denotes the number of clusters.
   (c) Compare and contrast the clustering results across the different choices of clustering methods, choices of dimension reduction methods, and choices of k. You must include both visual and quantitative evidence to support your discussion.
6. **Model Selection:** Use the stability-based model selection approach (with Jaccard similarity and/or adjusted Rand index) to determine what you think is the correct number (k) of breast cancer subtypes in the data. Describe the model selection procedure in enough detail such that a general reader can reproduce your analysis. Carefully justify your choice of k using both visual and quantitative evidence. [*Tip:* try your best to convince me, a skeptical reader, that your choice of k is the best choice.]
7. **Subtype Label Predictions:** Decide on your best guess of breast cancer subtype labels for two choices of k (i.e., the k you chose in problem (6) and your next best guess for k). Please save your best guess of breast cancer subtype labels to `results/best_subtype_labels.csv`. This `.csv` file should be a 1043×2 matrix of cluster labels (or integers) in the same order as the original gene expression data. Rows with the same cluster label (or integer) should correspond to samples that belong in the same cluster. An example of what this `.csv` should look like is provided here.
   - Your subtype label guesses could be the clusters from a particular dimension reduction + clustering method, a combination of many methods, or whatever creative scheme that you deem appropriate. The only requirements are that (1) you detail how these breast cancer subtype labels were computed (be sure to include sufficient detail so that a general reader can exactly reproduce your analysis), (2) you justify your analysis decisions, and (3) your code is reproducible.
   - Bonus points will be awarded to cluster label submissions that most closely align with the true breast cancer subtype labels for the true k, which will be released at the conclusion of this data lab.
8. *For ACMS 60876 students only*: How could you leverage the stability-based model selection approach to not only select the number of clusters k but also hyperparameters in the dimension reduction and/or clustering methods? Describe your proposed procedure in either words or in pseudocode. Include enough detail such that a general reader can reproduce your analysis. You do not need to actually implement this procedure.

You will be generating lots of plots throughout this lab. **Please use tools such as `patchwork` (R) or `subplots` (Python) to combine several plots into a single figure where appropriate.** This will make it easier for you to compare/contrast across plots.

Submission Details

Please push a folder named `lab2/` to your `dsip` GitHub repository **by 6:00 PM on Friday, February 20, 2026**. I will run an *automated* script that pulls from each of your GitHub repositories promptly at 6:01 PM and attempts to reproduce your report so please follow the file structure and names *exactly* with the exception that you may *add* folders as you wish.

The structure of your `lab2/` folder should follow the project structure discussed in class:

```
lab2/
├── data/
├── R/ (for R users)
├── python/ (for python users)
├── scripts/ (optional)
├── notebooks/
│   ├── lab2.qmd
│   └── lab2.html (or .pdf or other rendered output)
├── renv/ (for R users)
├── renv.lock (for R users)
├── .Rprofile (for R users)
├── environment.yml (for python users)
└── conda-lock.yml (for python users)
```

The `R/` and/or `python/` folders should contain all functions (and only functions, no scripts) necessary to reproduce your report. The `data/` folder should contain the raw data files, but **do not** push the `data/` folder to GitHub. In general, it is not good practice to store data on GitHub due to their restrictions on maximum file size (max: 100MB).

I will attempt to reproduce your report by running 'quarto render' so please be sure to include all necessary code in your repository. Keep in mind that my `data/` folder will only contain the raw data files that were initially provided to you.

A Note on Grading + Rubric

You will be graded on both the quality and reproducibility of your analysis.

A detailed rubric can be found on Canvas.

Recall the course policy regarding collaboration: Collaboration *of ideas* with the instructor and with classmates is encouraged throughout this course, with the following caveats:

- You must write up the final code and text by yourself.
- If you collaborate or use any resources other than course texts, you must explicitly acknowledge your collaborators (e.g., in writing at the end of your report) and cite the resources you used.