A Short Introduction to the Analysis of Real Data

Qing Mai
Florida State University

February 17, 2017

Analysis of Real Data

- The analysis of real data is the driving force of the development of statistics.
  - Microarray data;
  - Sequencing data;
  - Social networks.
- Data analysis in practice is generally different from what we learn in class, where we go into technical details of theoretical studies and programming.
- We have our own way of talking...
  - First, let's assume that our data is i.i.d...
  - To discover the lineage effects, let's fit a mixed linear model!
- How should we analyze data in practice? How should we present our results?

General Guidelines

General Guidelines I

1. Set the goal of the analysis before you even look at the data.
   - What is the most important question you are trying to answer?
   - What else do you care about?
   - Don't be driven by p-values.
2. A first look at the data:
   - What are the variables? What is the response? What are the predictors?
   - What is the sample size? What is the dimension?
   - Are there missing values? Which variables are continuous? Which variables are discrete?

General Guidelines II

3. Find out as much as you can about the dataset: how was it collected? What background knowledge is available?
4. Explore the dataset with elementary techniques:
   - Summary statistics;
   - Histogram, scatter plot, box plot, etc.
   - Do you notice anything strange?
   - What assumptions are you comfortable with?
5. Analyze the data:
   - Apply an appropriate method;
   - Interpret the results. Do they make sense to you? Do you want to perform more analysis?
   - Should you do some diagnostics?

General Guidelines III

6. Summarize your results:
   - Put everything back into context;
   - Use tables and graphs whenever possible. Spend extra efforts on the labelling and caption!
   - Comment with plain English. Know your audience!

An Example

The Goal of my Analysis

- Develop a classifier that predicts whether a person has malaria based on genomics data.
- Investigate the role of some genes.
- Demonstrate that normality-based methods may not work well when data is heavy-tailed or skewed.

Analysis of a Malaria Dataset

Diagnostics of malaria is important.

- Malaria is a major contributing factor leading to increased childhood mortality;
- Malaria is a significant obstacle to sustainable economic growth in many developing countries;
- Some genes are expected to be associated with the disease status.

Analysis of a Malaria Dataset

The malaria data (Ockenhouse et al. (2006)):

- Basic task: predict whether a person is infected with malaria;
- A large set of gene expression levels were collected for this purpose.

A First Look

After a first look at this dataset, I know that:

- The response is the disease status, which is binary in this dataset.
- The variables are the gene expression levels.
- The sample size is $n = 71$, but the dimension is $p = 22,283$. Hence, we have a typical high-dimensional problem with a small sample size.
- The predictors are continuous, while the response is binary.
- There is no missing value in this dataset.

More Information...

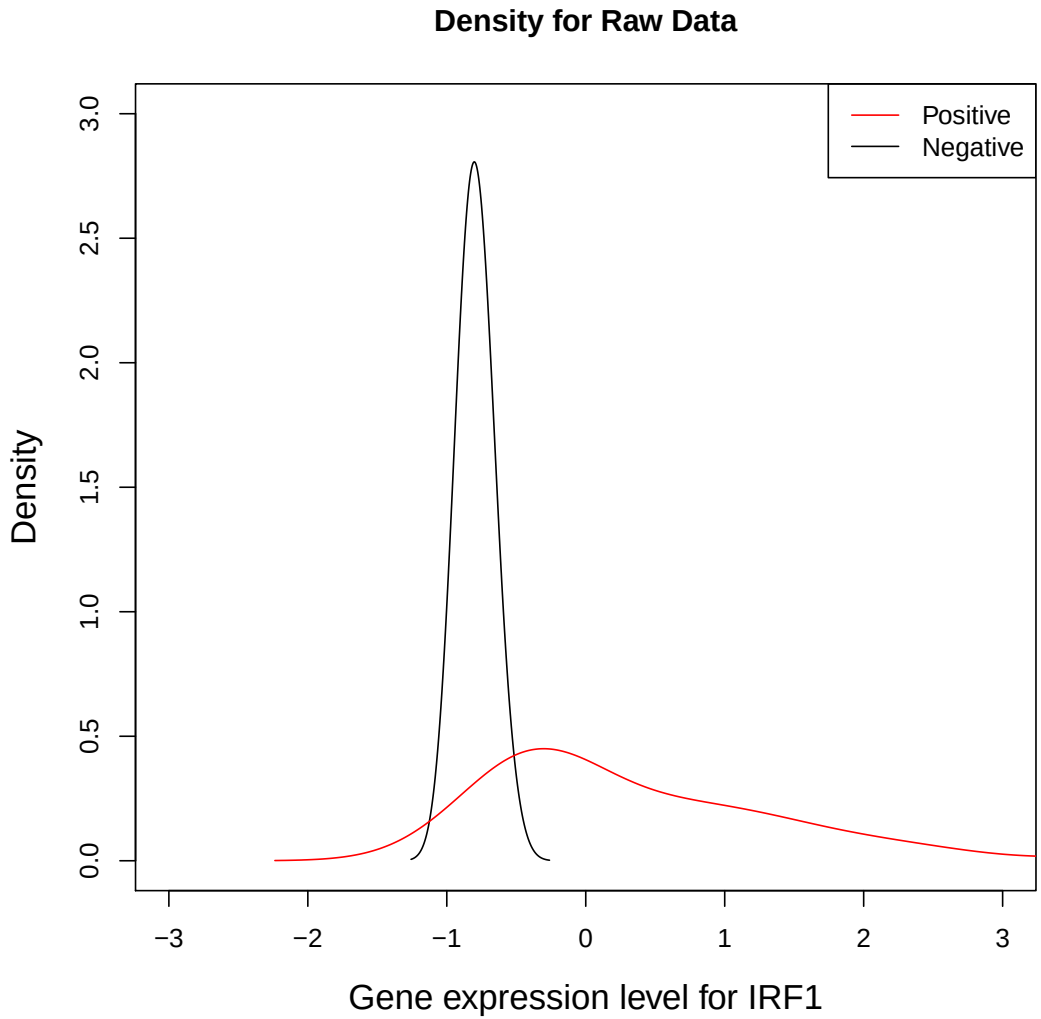
Additional information on this dataset:

- The dataset was collected on 22 patients in west Africa and 49 volunteers. Hence, this dataset is partially observational.
- All the healthy people in this dataset are volunteers.
- The gene `IRF1` included in this dataset controls the immune response in human. So we should probably expect it to be overexpressed in the patients.

Elementary Analysis

- This dataset is highly unbalanced. A large proportion of people in this dataset are infected with malaria.
- The gene expression levels have already been preprocessed, as some of them are negative.
- A plot of the density for `IRF1` shows that the gene expression levels are skewed.

Analysis of a Malaria Dataset



Elementary Analysis

- It seems that the gene expression levels for healthy people and patients are different, although I am reluctant to assume that the expression levels are normal within the patients.
- I will first apply a normality-based method (DSDA), and then show that it can be improved by a method that does not rely on normality assumption (SeSDA).
- I will look at the classification accuracy and the variable selection to compare the results.
- Because there is no generic training and testing set, I will split the dataset to form artificial training and testing set, in a balanced manner.

Analysis of a Malaria Dataset

	SeSDA	DSDA
Testing Error	1/35(0.59%)	6/35(0.99%)
Fitted Model Size	6(0.4)	18(1.5)

Table : Comparison of SeSDA and DSDA on the malaria dataset. The reported numbers are medians of 100 replicates, with standard errors obtained by bootstrap in parentheses.

Analysis of a Malaria Dataset

Remark

The gene `IRF1` is most frequently selected by SeSDA, but seldom by DSDA.

In other words, SeSDA is both more accurate and gives more sensible variable selection results.

Analysis of a Malaria Dataset

Does this make sense?

Analysis of a Malaria Dataset

Does this make sense?

- Note that SeSDA assumes that the variables can be transformed to be approximately normal.

Analysis of a Malaria Dataset

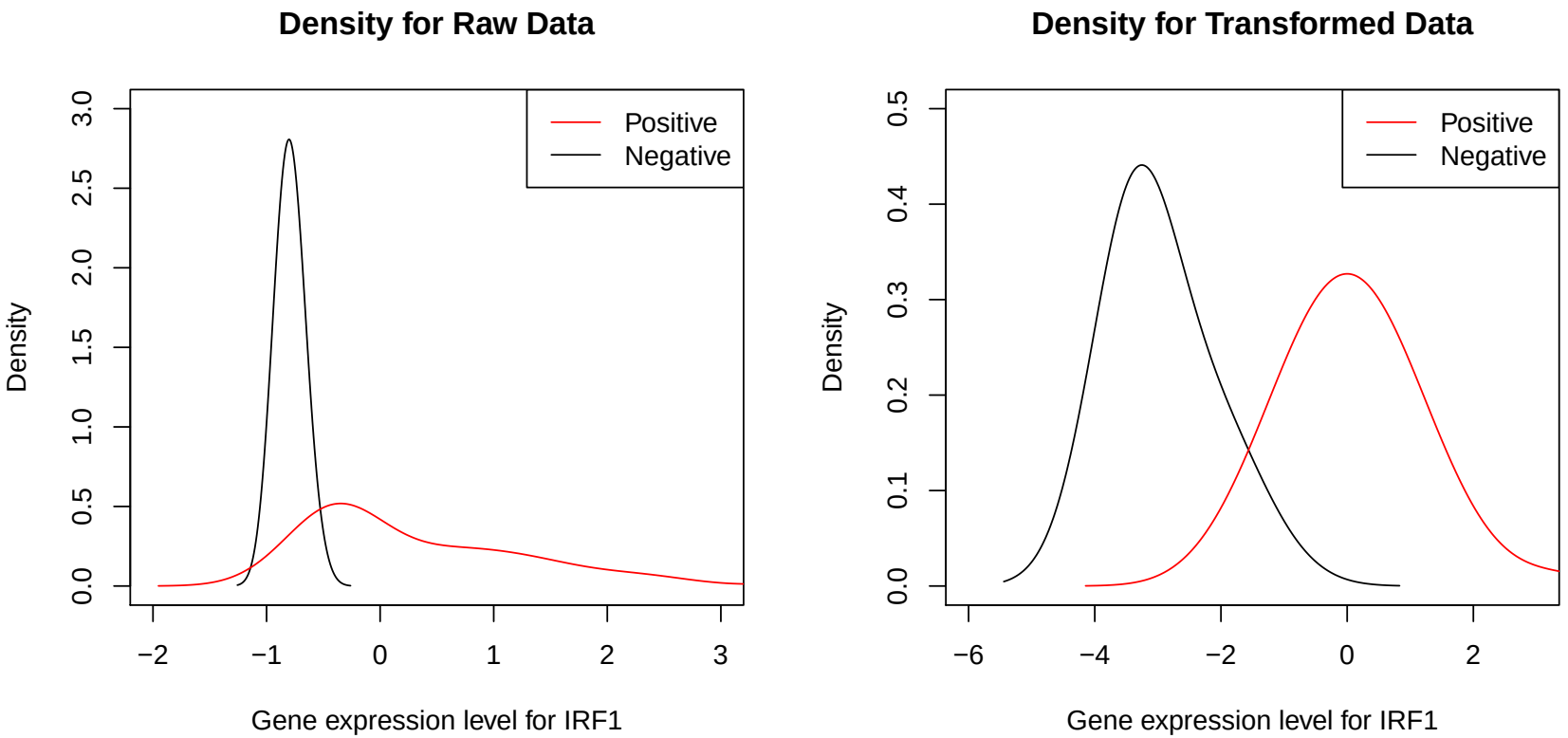


Figure : Density functions of gene `IRF1` (the 2059th gene) in the malaria data. The plot on the left displays the density function of the normalized raw data, while the one on the right is of the transformed data.

A Summary of the Results

- Malaria can be accurately diagnosed by gene expression levels.
- On a partially observational dataset, we can diagnose malaria with a median accuracy of 97.1% on independent testing sets.
- Our study confirms the important role of the gene `IRF1` in the diagnosis. The inclusion of this variable resulted in a statistically significant improvement in accuracy.
- In future statistical analysis, one should be careful with normality-based methods when variables are skewed.

Thank you!