

RJafroc Documentation

Dev P. Chakraborty, PhD

2020-03-13

Contents

| | |
|--|----------|
| Preface | 5 |
| 1 Introduction | 7 |
| 1.1 References | 7 |
| 2 ROC DATA FORMAT | 9 |
| 2.1 Introduction | 9 |
| 2.2 Note to existing users | 10 |
| 2.3 The Excel data format | 10 |
| 2.4 Illustrative toy file | 10 |
| 2.5 The Truth worksheet | 10 |
| 2.6 The structure of an ROC dataset | 11 |
| 2.7 The false positive (FP) ratings | 13 |
| 2.8 The true positive (TP) ratings | 14 |
| 2.9 Correspondence between NL member of dataset and the FP work- sheet | 15 |
| 2.10 Correspondence between LL member of dataset and the TP work- sheet | 15 |
| 2.11 Correspondence using the which function | 15 |
| 2.12 References | 16 |

Preface

- This book, an extended documentation of the **RJafroc** package, is undergoing extensive edits.
- It should not be used by the casual user until I give the go ahead.
- It bypasses the file size limits of **CRAN**, currently 5 MB, which severely limits the extent of the documentation that can be included with the CRAN version of the package.
- I welcome corrections and comments by the not-so-casual-user.
- Please use the GitHub website to raise issues and comments:
 - <https://github.com/dpc10ster/RJafrocBook>

Chapter 1

Introduction

- This is the book describing the **RJafroc** package.
- The name of the book is RJafrocBook
- Modality and treatment are used interchangeably.
- Reader is a generic radiologist, or a computer aided detection algorithm, or any algorithmic “reader”
- TBA

1.1 References

Chapter 2

ROC DATA FORMAT

$$\frac{d}{dx} \left(\int_a^x f(u) du \right) = f(x)$$

$$\theta = \frac{1}{N_L N_N}$$

2.1 Introduction

- The purpose of this vignette is to explain the data format of the input Excel file and to introduce the capabilities of the function `DfReadDataFile()`. Background on observer performance methods are in my book (Chakraborty, 2017).
- I will start with Receiver Operating Characteristic (ROC) data (Metz, 1978), as this is by far the simplest paradigm.
- In the ROC paradigm the observer assigns a rating to each image. A rating is an ordered numeric label, and, in our convention, higher values represent greater certainty or **confidence level** for presence of disease. With human observers, a 5 (or 6) point rating scale is typically used, with 1 representing highest confidence for *absence* of disease and 5 (or 6) representing highest confidence for *presence* of disease. Intermediate values represent intermediate confidence levels for presence or absence of disease.
- Note that location information associated with the disease, if applicable, is not collected.
- There is no restriction to 5 or 6 ratings. With algorithmic observers, e.g., computer aided detection (CAD) algorithms, the rating could be a floating point number and have infinite precision. All that is required is that higher values correspond to greater confidence in presence of disease.

2.2 Note to existing users

- The Excel file format has recently undergone changes resulting in 4 extra `list` members in the final created `dataset` object (i.e., 12 members instead of 8).
- Code should run on the old format Excel files as the 4 extra list members are simply ignored.
- Reasons for the change will become clearer in these vignettes
- Basically they are needed for generalization to other data collection paradigms instead of crossed, for example to the split-plot data acquisition paradigm, and for better data entry error control.

2.3 The Excel data format

- The Excel file has three worksheets.
- These are named
 - `Truth`,
 - `NL` (or `FP`),
 - `LL` (or `TP`).

2.4 Illustrative toy file

- *Toy files* are artificial small datasets intended to illustrate essential features of the data format.
- The examples shown in this vignette corresponds to Excel file `inst/extdata/toyFiles/ROC/rocCr.xlsx` in the project directory.
- To view these files one needs to `clone` the source files from `GitHub`.

2.5 The Truth worksheet

- The `Truth` worksheet contains 6 columns: `CaseID`, `LesionID`, `Weight`, `ReaderID`, `ModalityID` and `Paradigm`.
- For ROC data the first five columns contain as many rows as there are cases (images) in the dataset.
- `CaseID`: unique integers, one per case, representing the cases in the dataset.
- `LesionID`: integers 0 or 1, with each 0 representing a non-diseased case and each 1 representing a diseased case.

- In the current toy dataset, the non-diseased cases are labeled 1, 2 and 3, while the diseased cases are labeled 70, 71, 72, 73 and 74. The values do not have to be consecutive integers; they need not be ordered; the only requirement is that they be **unique**.
- **Weight**: Not used for ROC data, a floating point value, typically filled in with 0 or 1.
- **ReaderID**: a **comma-separated** listing of reader labels, each represented by a **unique string**, that have interpreted the case. In the example shown below each cell has the value 0, 1, 2, 3, 4 meaning that each of the readers, represented by the strings "0", "1", "2", "3" and "4", have interpreted all cases (hence the "crossed" design). **With reader names that could be confused with integers, each cell in this column has to be text formatted as otherwise Excel will not accept it.** [Try entering 0, 1, 2, 3, 4 in a numeric formatted Excel cell.]
- The reader names could just as well have been Rdr0, Rdr1, Rdr2, Rdr3, Rdr4. The only requirement is that they be unique strings.
- Look in in the `inst/extdata/toyFiles/ROC` directory for files `rocCrStrRdrsTrts.xlsx` and `rocCrStrRdrsNonUnique.xlsx` for examples of data files using longer strings for readers. The second file generates an error because the reader names are not unique.
- **ModalityID**: a comma-separated listing of modalities (one or more modalities), each represented by a **unique string**, that are applied to each case. In the example each cell has the value "0", "1". **With treatment names that could be confused with integers, each cell has to be text formatted as otherwise Excel will not accept it.**
- The treatment names could just as well have been Trt0, Trt1. Again, the only requirement is that they be unique strings.
- **Paradigm**: this column contains two cells, **ROC** and **crossed**. It informs the software that this is an ROC dataset, and the design is crossed, meaning each reader has interpreted each case in each modality (in statistical terminology: modality and reader factors are "crossed").
- There are 5 diseased cases in the dataset (the number of 1's in the **LesionID** column of the **Truth** worksheet).
- There are 3 non-diseased cases in the dataset (the number of 0's in the **LesionID** column).
- There are 5 readers in the dataset (each cell in the **ReaderID** column contains the string 0, 1, 2, 3, 4).
- There are 2 modalities in the dataset (each cell in the **ModalityID** column contains the string 0, 1).

2.6 The structure of an ROC dataset

In the following code chunk the first statement retrieves the name of the data file, located in a hidden directory that one need not be concerned with. The

| | A | B | C | D | E | F |
|---|--------|----------|--------|-----------|------------|----------|
| | CaseID | LesionID | Weight | ReaderID | ModalityID | Paradigm |
| 1 | 1 | 0 | 0 | 0.1,2,3,4 | 0.1 | ROC |
| 2 | 2 | 0 | 0 | 0.1,2,3,4 | 0.1 | crossed |
| 3 | 3 | 0 | 0 | 0.1,2,3,4 | 0.1 | crossed |
| 4 | 70 | 1 | 1 | 0.1,2,3,4 | 0.1 | crossed |
| 5 | 71 | 1 | 1 | 0.1,2,3,4 | 0.1 | crossed |
| 6 | 72 | 1 | 1 | 0.1,2,3,4 | 0.1 | crossed |
| 7 | 73 | 1 | 1 | 0.1,2,3,4 | 0.1 | crossed |
| 8 | 74 | 1 | 1 | 0.1,2,3,4 | 0.1 | crossed |

Figure 2.1: Truth worksheet for file rocCr.xlsx

second statement reads the file using the function `DfReadDataFile()` and saves it to object `x`. The third statement shows the structure of the dataset object `x`.

```
rocCr <- system.file("extdata", "toyFiles/ROC/rocCr.xlsx",
                      package = "RJaFROC", mustWork = TRUE)
x <- DfReadDataFile(rocCr, newExcelFileFormat = TRUE)
str(x)
#> List of 12
#> $ NL           : num [1:2, 1:5, 1:8, 1] 1 3 2 3 2 2 1 2 3 2 ...
#> $ LL           : num [1:2, 1:5, 1:5, 1] 5 5 5 5 5 5 5 5 5 5 ...
#> $ lesionVector : int [1:5] 1 1 1 1 1
#> $ lesionID     : num [1:5, 1] 1 1 1 1 1
#> $ lesionWeight : num [1:5, 1] 1 1 1 1 1
#> $ dataType     : chr "ROC"
#> $ modalityID   : Named chr [1:2] "0" "1"
#> .. attr(*, "names")= chr [1:2] "0" "1"
#> $ readerID     : Named chr [1:5] "0" "1" "2" "3" ...
#> .. attr(*, "names")= chr [1:5] "0" "1" "2" "3" ...
#> $ design       : chr "CROSSED"
#> $ normalCases  : int [1:3] 1 2 3
#> $ abnormalCases: int [1:5] 70 71 72 73 74
#> $ truthTableStr: num [1:2, 1:5, 1:8, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
```

- In the above code chunk flag `newExcelFileFormat` is set to `TRUE` as otherwise columns D - F in the `Truth` worksheet are ignored and the dataset is assumed to be crossed, with `dataType` automatically determined from the contents of the FP and TP worksheets.
- Flag `newExcelFileFormat = FALSE` is for compatibility with older JAFROC format Excel files, which did not have these columns in the `Truth` worksheet. Its usage is deprecated.
- The dataset object `x` is a `list` variable with 12 members.
- The `x$NL` member, with dimension `[2, 5, 8, 1]`, contains the ratings of normal cases. The extra values in the third dimension, filled with `NA`s,

are needed for compatibility with FROC datasets, as unlike ROC, false positives are possible on diseased cases.

- The `x$LL`, with dimension $[2, 5, 5, 1]$, contains the ratings of abnormal cases.
- The `x$lesionVector` member is a vector with 5 ones representing the 5 diseased cases in the dataset.
- The `x$lesionID` member is an array with 5 ones.
- The `x$lesionWeight` member is an array with 5 ones.
- The `lesionVector`, `lesionID` and `lesionWeight` members are not used for ROC datasets. They are there for compatibility with FROC datasets.
- The `dataType` member indicates that this is an ROC dataset.
- The `x$modalityID` member is a vector with two elements "0" and "1", naming the two modalities.
- The `x$readerID` member is a vector with five elements "0", "1", "2", "3" and "4", naming the five readers.
- The `x$design` member is `CROSSED`; specifies the dataset design, which is "CROSSED".
- The `x$normalCases` member lists the integer names of the normal cases, 1, 2, 3.
- The `x$abnormalCases` member lists the integer names of the abnormal cases, 70, 71, 72, 73, 74.
- The `x$truthTableStr` member quantifies the structure of the dataset, as explained in **Chapter 00 Vignette #3-#5**.

2.7 The false positive (FP) ratings

These are found in the FP or NL worksheet, see below.

| ReaderID | ModalityID | CaseID | FP_Rating |
|----------|------------|--------|-----------|
| 0 | 0 | 1 | 0 |
| 0 | 0 | 2 | 1 |
| 0 | 0 | 3 | 2 |
| 0 | 0 | 4 | 3 |
| 0 | 0 | 5 | 4 |
| 0 | 0 | 6 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 2 | 1 |
| 1 | 0 | 3 | 2 |
| 1 | 0 | 4 | 3 |
| 1 | 0 | 5 | 4 |
| 1 | 0 | 6 | 0 |
| 2 | 0 | 1 | 0 |
| 2 | 0 | 2 | 1 |
| 2 | 0 | 3 | 2 |
| 2 | 0 | 4 | 3 |
| 2 | 0 | 5 | 4 |
| 2 | 0 | 6 | 0 |
| 3 | 0 | 1 | 0 |
| 3 | 0 | 2 | 1 |
| 3 | 0 | 3 | 2 |
| 3 | 0 | 4 | 3 |
| 3 | 0 | 5 | 4 |
| 3 | 0 | 6 | 0 |
| 4 | 0 | 1 | 0 |
| 4 | 0 | 2 | 1 |
| 4 | 0 | 3 | 2 |
| 4 | 0 | 4 | 3 |
| 4 | 0 | 5 | 4 |
| 4 | 0 | 6 | 0 |
| 0 | 1 | 1 | 0 |
| 0 | 1 | 2 | 1 |
| 0 | 1 | 3 | 2 |
| 0 | 1 | 4 | 3 |
| 0 | 1 | 5 | 4 |
| 0 | 1 | 6 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 2 | 1 |
| 1 | 1 | 3 | 2 |
| 1 | 1 | 4 | 3 |
| 1 | 1 | 5 | 4 |
| 1 | 1 | 6 | 0 |
| 2 | 1 | 1 | 0 |
| 2 | 1 | 2 | 1 |
| 2 | 1 | 3 | 2 |
| 2 | 1 | 4 | 3 |
| 2 | 1 | 5 | 4 |
| 2 | 1 | 6 | 0 |
| 3 | 1 | 1 | 0 |
| 3 | 1 | 2 | 1 |
| 3 | 1 | 3 | 2 |
| 3 | 1 | 4 | 3 |
| 3 | 1 | 5 | 4 |
| 3 | 1 | 6 | 0 |
| 4 | 1 | 1 | 0 |
| 4 | 1 | 2 | 1 |
| 4 | 1 | 3 | 2 |
| 4 | 1 | 4 | 3 |
| 4 | 1 | 5 | 4 |
| 4 | 1 | 6 | 0 |

Figure 2.2: FP worksheet for file rocCr.xlsx

- It consists of 4 columns, each of length 30 (= # of modalities times number of readers times number of non-diseased cases).
- **ReaderID**: the reader labels: 0, 1, 2, 3 and 4. Each reader label occurs 6 times (= # of modalities times number of non-diseased cases).

- **ModalityID**: the modality or treatment labels: 0 and 1. Each label occurs 15 times (= # of readers times number of non-diseased cases).
- **CaseID**: the case labels for non-diseased cases: 1, 2 and 3. Each label occurs 10 times (= # of modalities times # of readers).
- The label of a diseased case cannot occur in the FP worksheet. If it does the software generates an error.
- **FP_Rating**: the floating point ratings of non-diseased cases. Each row of this worksheet contains a rating corresponding to the values of **ReaderID**, **ModalityID** and **CaseID** for that row.

2.8 The true positive (TP) ratings

These are found in the TP or LL worksheet, see below.

| ReaderID | ModalityID | CaseID | LesionID | TP Rating |
|----------|------------|--------|----------|-----------|
| 0 | 0 | 70 | 1 | 5 |
| 0 | 0 | 71 | 1 | 5 |
| 0 | 0 | 72 | 1 | 5 |
| 0 | 0 | 73 | 1 | 5 |
| 0 | 0 | 74 | 1 | 5 |
| 0 | 1 | 70 | 1 | 5 |
| 0 | 1 | 71 | 1 | 5 |
| 0 | 1 | 72 | 1 | 5 |
| 0 | 1 | 73 | 1 | 5 |
| 0 | 1 | 74 | 1 | 5 |
| 1 | 0 | 70 | 1 | 5 |
| 1 | 0 | 71 | 1 | 5 |
| 1 | 0 | 72 | 1 | 5 |
| 1 | 0 | 73 | 1 | 5 |
| 1 | 0 | 74 | 1 | 5 |
| 1 | 1 | 70 | 1 | 5 |
| 1 | 1 | 71 | 1 | 5 |
| 1 | 1 | 72 | 1 | 5 |
| 1 | 1 | 73 | 1 | 5 |
| 1 | 1 | 74 | 1 | 5 |
| 2 | 0 | 70 | 1 | 5 |
| 2 | 0 | 71 | 1 | 5 |
| 2 | 0 | 72 | 1 | 5 |
| 2 | 0 | 73 | 1 | 5 |
| 2 | 0 | 74 | 1 | 5 |
| 2 | 1 | 70 | 1 | 5 |
| 2 | 1 | 71 | 1 | 5 |
| 2 | 1 | 72 | 1 | 5 |
| 2 | 1 | 73 | 1 | 5 |
| 2 | 1 | 74 | 1 | 5 |
| 3 | 0 | 70 | 1 | 5 |
| 3 | 0 | 71 | 1 | 5 |
| 3 | 0 | 72 | 1 | 5 |
| 3 | 0 | 73 | 1 | 5 |
| 3 | 0 | 74 | 1 | 5 |
| 3 | 1 | 70 | 1 | 5 |
| 3 | 1 | 71 | 1 | 5 |
| 3 | 1 | 72 | 1 | 5 |
| 3 | 1 | 73 | 1 | 5 |
| 3 | 1 | 74 | 1 | 5 |
| 4 | 0 | 70 | 1 | 5 |
| 4 | 0 | 71 | 1 | 5 |
| 4 | 0 | 72 | 1 | 5 |
| 4 | 0 | 73 | 1 | 5 |
| 4 | 0 | 74 | 1 | 5 |
| 4 | 1 | 70 | 1 | 5 |
| 4 | 1 | 71 | 1 | 5 |
| 4 | 1 | 72 | 1 | 5 |
| 4 | 1 | 73 | 1 | 5 |
| 4 | 1 | 74 | 1 | 5 |

Figure 2.3: TP worksheet for file rocCr.xlsx

- It consists of 5 columns, each of length 50 (= # of modalities times number of readers times number of diseased cases).
- **ReaderID**: the reader labels: 0, 1, 2, 3 and 4. Each reader label occurs 10 times (= # of modalities times number of diseased cases).
- **ModalityID**: the modality or treatment labels: 0 and 1. Each label occurs 25 times (= # of readers times number of diseased cases).
- **LesionID**: For an ROC dataset this column contains fifty 1's (each diseased case has one lesion).
- **CaseID**: the case labels for non-diseased cases: 70, 71, 72, 73 and 74. Each label occurs 10 times (= # of modalities times # of readers). The label of a non-diseased case cannot occur in the TP worksheet.
- **TP_Rating**: the floating point ratings of diseased cases. Each row of this worksheet contains a rating corresponding to the values of **ReaderID**, **ModalityID**, **LesionID** and **CaseID** for that row.

2.9 Correspondence between NL member of dataset and the FP worksheet

- The list member `x$NL` is an array with `dim = c(2,5,8,1)`.
 - The first dimension (2) comes from the number of modalities.
 - The second dimension (5) comes from the number of readers.
 - The third dimension (8) comes from the **total** number of cases.
 - The fourth dimension is always 1 for an ROC dataset.
- The value of `x$NL[1,5,2,1]`, i.e., 5, corresponds to row 15 of the FP table, i.e., to `ModalityID = 0`, `ReaderID = 4` and `CaseID = 2`.
- The value of `x$NL[2,3,2,1]`, i.e., 4, corresponds to row 24 of the FP table, i.e., to `ModalityID 1`, `ReaderID 2` and `CaseID 2`.
- All values for case index > 3 are `-Inf`. For example the value of `x$NL[2,3,4,1]` is `-Inf`. This is because there are only 3 non-diseased cases. The extra length is needed for compatibility with FROC datasets.

2.10 Correspondence between LL member of dataset and the TP worksheet

- The list member `x$LL` is an array with `dim = c(2,5,5,1)`.
 - The first dimension (2) comes from the number of modalities.
 - The second dimension (5) comes from the number of readers.
 - The third dimension (5) comes from the number of diseased cases.
 - The fourth dimension is always 1 for an ROC dataset.
- The value of `x$LL[1,1,5,1]`, i.e., 4, corresponds to row 6 of the TP table, i.e., to `ModalityID = 0`, `ReaderID = 0` and `CaseID = 74`.
- The value of `x$LL[1,2,2,1]`, i.e., 3, corresponds to row 8 of the TP table, i.e., to `ModalityID = 0`, `ReaderID = 1` and `CaseID = 71`.
- There are no `-Inf` values in `x$LL`: `any(x$LL == -Inf) = FALSE`.

2.11 Correspondence using the which function

- Converting from **names** to **subscripts** (indicating position in an array) can be confusing.
- The following example uses the `which` function to help out.
- The first line says that the `abnormalCase` named 70 corresponds to subscript 1 in the LL array case dimension.
- The second line prints the NL rating for `modalityID = 0`, `readerID = 1` and `normalCases = 1`.

- The third line prints the LL rating for `modalityID = 0`, `readerID = 1` and `abnormalCases = 70`.
- The last line shows what happens if one enters an invalid value for name; the result is a `numeric(0)`.
- Note that in each of these examples, the last dimension is 1 because we are dealing with an ROC dataset.
- The reader is encouraged to examine the correspondence between the NL and LL ratings and the Excel file using this method.

```
which(x$abnormalCases == 70)
#> [1] 1
x$NL[which(x$modalityID == "0"),which(x$readerID == "1"),which(x$normalCases == 1),1]
#> [1] 2
x$LL[which(x$modalityID == "0"),which(x$readerID == "1"),which(x$abnormalCases == 70),1]
#> [1] 5
x$LL[which(x$modalityID == "a"),which(x$readerID == "1"),which(x$abnormalCases == 70),1]
#> numeric(0)
```

2.12 References

Bibliography

Chakraborty, D. P. (2017). *Observer Performance Methods for Diagnostic Imaging - Foundations, Modeling, and Applications with R-Based Examples*. CRC Press, Boca Raton, FL.

Metz, C. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.