Code-based, open-source software for teaching interactive data visualisation

Shan-I Lee

Department of Statistics
The University of Auckland

Supervisor: Dr Paul Murrell

A dissertation submitted in partial fulfillment of the requirements for the degree of BSc(Hons) in Statistics, The University of Auckland, 2017.

Abstract

The software required to implement interactive techniques in data visualisation has become readily available to statisticians with the development of web-based graphics and several open-source R packages. This paper proposes using the R packages, **plotly, crosstalk** and **shiny**, to teach a set of powerful techniques for interactive data visualisation. The value of applying linked brushing, identification, scaling, subset selection and tours, to reveal further insight during exploratory data analysis will be demonstrated. Outliers are quickly examined using linked brushing and identification to better understand how they are unusual. Different structures in the data are explored with interactive scaling and subset selection. Dynamic tours provide views of complex multivariate structures and linked brushing enables multiple representations of the data to be explored simultaneously and related together. The additional insights gained from applying interactive data visualisation in comparison to analysis with static plots alone, show that the set of interactive techniques and software identified is a powerful toolbox for statisticians to have.

For direct access to all interactive plots this paper is recommended to be read online at https://shanl33.github.io/Honours$_Report/$.

Contents

Chapter 1

Introduction

A thesis should always have an introduction. The purpose is to describe the general subject area, state the research problem of interest, outline the main results of the thesis, and put the results in context with the wider subject area and its applications.

The main body of the text must be divided into a logical scheme which is followed consistently throughout the work. It usually starts with an introduction chapter and ends with a conclusion chapter. See, for example, the table of contents on page 3.

There is strict 35-page limit for an applied mathematics dissertation, including the references but excluding appendices.

1.1 Some Basics About LaTex

1.1.1 Equations

The main strength of LaTex is mathematical typesetting.

There is a huge amount of information about LaTex on the internet. A helpful short manual, also included in this folder, is the file `latex_intro.pdf` This document gives a lot of sample LaTex commands. The file `latex-howto.tex` in this folder also contains examples of many latex commands.

We first show some simple examples of mathematical formulae using latex typesetting.

1. The basic functions: $\cos(x), \sin(x), \ln(x)$, (`$\cos(x),\sin(x),\ln(x)$`).
2. Greek letters: $\alpha\beta\gamma\delta\epsilon$... (`$\alpha\beta\gamma\delta\epsilon...$`).
3. Mathematical symbols: $\int \oint \sum \lim \bigcup \bigcap$ (`$\int\oint\sum\lim\bigcup\bigcap$`).
4. Fractions: $\frac{1}{2}, \frac{1}{2-x}$ (`$\frac{1}{2},\frac{1}{2-x}$`).

The following matrix

$$\left[\begin{array}{ccc} U_r & r & W_r \\ 0 & 1 & V_x \\ 0 & 0 & W_x \end{array} \right], \tag{1.1}$$

is generated using the `equarray` environment:

```
\begin{eqnarray}\label{eqn:matrix}
\left[
\begin{array}{ccc}
 U_{r}& r &W_{r}\\
   0 &1 &V_{x}\\
   0& 0 & W_{x}
\end{array}
\right],
\end{eqnarray}
```

The `\label{eqn:matrix}` command labels the equation with `{eqn:matrix}` which can be referred to somewhere else in the text by using `\ref{eqn:matrix}` or `\eqref{eqn:matrix}`.

The command `\notag` eliminates the numbering of the first equation,

$$\begin{aligned} \lambda^{(1)} &= tr[T^{(1)}P], \\ \lambda^{(2)} &= tr[T^{(2)}P - T^{(1)}ST^{(1)}P]. \end{aligned} \tag{1.2}$$

```
\begin{eqnarray} \label{eqn:lambda_trace}
\lambda^{(1)}&=&tr[T^{(1)}P],\notag  \\
\lambda^{(2)}&=&tr[T^{(2)}P - T^{(1)}ST^{(1)}P].
\end{eqnarray}
```

1.1.2 Itemized lists

Example of an itemized list:

- muscle and fat cells remove glucose from the blood,
- cells use glucose for protein synthesis.

```
\begin{itemize}
\item muscle and fat cells remove glucose from the blood,
\item cells use glucose for protein synthesis.
\end{itemize}
```

This can be done by an enumerated list:

1. muscle and fat cells remove glucose from the blood,
2. cells use glucose for protein synthesis.

```
\begin{enumerate}
\item muscle and fat cells remove glucose from the blood,
\item cells use glucose for protein synthesis.
\end{enumerate}
```

1.1.3 Inserting figures

You may save your Matlab figures as jpg files. Figures should be stored in the same folder as the latex files. For the graphicx package to work you usually need to ask latex to create a pdf file (e.g., command pdflatex or latexpdf).

An example of an inserted image is given in Figure **??**.

Figure 1.1: Mode shapes

1.1.4 Tables

Example of a table,

Gene	GeneID	Length
human latexin	1234	14.9 kbps
mouse latexin	2345	10.1 kbps
rat latexin	3456	9.6 kbps

Table 1.1: **title of table** - Overview of latexin genes.

```
\begin{table}[htdp]
\centering
\begin{tabular}{ccc}
% ccc means 3 columns, all centered; alternatives are l, r
{\bf Gene} & {\bf GeneID} & {\bf Length} \\
\hline % draws a line under the column headers
human latexin & 1234 & 14.9 kbps \\
mouse latexin & 2345 & 10.1 kbps \\
rat latexin   & 3456 & 9.6 kbps \\
\end{tabular}
\caption[title of table]{\textbf{title of table} - Overview of latexin genes.}
\label{latexin_genes} % label for cross-links with \ref{latexin_genes}
\end{table}
```

See how to add two vertical lines in the table (Simply change `{ccc}` to `{c|c|c}`)

Gene	GeneID	Length
human latexin	1234	14.9 kbps
mouse latexin	2345	10.1 kbps
rat latexin	3456	9.6 kbps

Table 1.2: **title of table** - Overview of latexin genes.

1.1.5 How to Refer to Equations, Sections, etc

1. References can be linked to equations, figures, tables or sections using the command `\ref`: Equation (**??**), Figure **??**, Table **??** and Section **??**.
   ```
   Equation~(\ref{eqn:lambda_trace}),Figure~\ref{modes},
   Table~\ref{latexin_genes2} and Section~\ref{table}.
   ```

2. Equations can be conveniently referred to using `\eqref`. See, for example, Equation (**??**).
   ```
   Equation \eqref{eqn:lambda_trace}
   ```
   Note that `\eqref` includes the round brackets by itself.

3. Citations are in a similar way but using the command `\cite`:

   [**?**], [**?**], and [**?**], or [**?**, **?**, **?**] .

   ```
   \cite{Salmond}, \cite{Stull}, and \cite{TandC},
   or  \cite{Salmond,Stull,TandC} .
   ```

   There are many different styles for writing citations – you should follow the norms for your subject are.

   A more advanced way to do citations is to use `bibtex`. This is a powerful tool and we encourage you to try it. There is plenty of information about it on the web.

Chapter 2

Methodologies and analysis

2.1 Methodologies

2.2 Analysis

Chapter 3

Discussion

3.1 Main results

3.2 Discussion

Chapter 4

Conclusions

You may add more chapters as needed in the file.

References

[1] Farge Marie, *Wavelet Transforms and Their Applications to Turbulence*, Ann. Rev. Fluid Mech. volume 24, pages 395-457, 1992.

[2] Salmond Jennifer, *Vertical Mixing of Ozone in the Very Stable Nocturnal Boundary Layer*, PhD Thesis, University of British Columbia, 2001.

[3] Stull B. Ronald, *Introduction to Boundary Layer Meteorology*, Dordrecht; Boston: Kulwer Academic Publishers, 1988.

[4] Torrence Christopher, Compo Gilbert P., *A Practical Guide to Wavelet Analysis*, Bulletin of the American Meteorological Society volume 79, pages 61-78, 1998.

Appendix A

Some extra things

This is an optional chapter for any additional material that does not fit conveniently into the body of the text (e.g., data, copies of computer programmes). Note that appendices won't necessarily be marked.