A Data Lab Project Template in R Markdown

Wenjie Wang[*]

28 January 2017

Abstract

This is a template mainly designed for data lab projects. In this template, we review most common components in a single R Markdown document with the power of **bookdown** package and demonstrate their basic usage by examples.

Keywords: Template; R Markdown; **bookdown**; **knitr**; **Pandoc**

1 Introduction

This document is designed as a template for data lab projects. However, it can actually be used as one of general templates in R Markdown for a single document.

The motivation of setting up a template in R Markdown is due to its simple syntax and flexible output format with the help of **pandoc**. In addition, it is in favor of reproducible studies, which have been receiving increasing attention in modern research.

Instead of providing a minimal but non-informative template framework, we review most of the basic syntax of writing a single R Markdown document with the power of **bookdown** (Xie, 2017) by examples. However, it is not intended for a tutorial of R Markdown or **bookdown** package. Readers are encouraged to skim the PDF or HTML output, and have a more close look at the source document of this template directly.

The rest of this project template is organized as follows: In Section 2, we briefly discuss about the cross-referencing of R Markdown, which has a better support from package **bookdown** (Xie, 2016) than package **rmarkdown** (Allaire et al., 2016) now. In Section 3 and Section 4, we present the example of writing mathematical equations, and mathematical environment for theorem, lemma, and definition, etc., respectively. Some examples of reproducing figures and inclusion of existing ones are given in Section 5. The generation of table source and other R objects is discussed in Section 6. A brief demonstration of code chunk is given in Section 7. At last but not least, in Section 8, we point readers to some external resources for further reading and more advanced usage of **bookdown**.

2 Cross-Reference by bookdown

Cross-reference of mathematical equations, tables, and figures used to be a pain of using R markdown. Usually extra package, such as **kfigr** (Koohafkan, 2015), and certain kind of extra labor was needed for a automatic and satisfactory cross-referencing. Fortunately, package **bookdown** provides a much easier and consistent syntax for cross-referencing now.

[*] wenjie.2.wang@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

Table 1: Theorem environments in **bookdown**.

Environment	Printed Name	Label Prefix
theorem	Theorem	thm
lemma	Lemma	lem
definition	Definition	def
corollary	Corollary	cor
proposition	Proposition	prp
example	Example	ex

3 Math Equations

The inline math expressions are quoted by `$` in the source document, which is consistent with the syclit of LaTeX. For instance, x_i^2, $\sin(x)$, θ are inline expressions. The equations can be simply quoted by `$$` if no cross-reference is needed, where regular LaTeX commands under math environment can be used. For equations that need cross-referencing, LaTeX environments for mathematical equations, such as `equation`, `align`, can be used directly. For example, Equation (1) is the well-known Eular's identity.

$$e^{i\theta} = \cos(\theta) + i\sin(\theta). \tag{1}$$

4 Math Theorem Environments

The mathematical theorem can be put inside `theorem` chunk followed by its label. For example, Central Limit Theorem (CLT) is presented in Theorem 4.1.

Theorem 4.1. ***(Central Limit Theorem)*** *Let $X_1, \ldots, X_n$ be independent, identically distributed (i.i.d.) random variables with finite expectation μ, and positive, finite variance σ^2, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\frac{\bar{S}_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{L} N(0,1) \text{ as } n \to \infty.$$

All the available theorem environments for mathematical theorem, lemma, definition, etc, and their label prefix designed for cross-reference are summarized in Table 1.

The First Borel-Cantelli Lemma is given inside the `lemma` environment as shown in Lemma 4.1.

Lemma 4.1. ***(First Borel-Cantelli Lemma)*** *Let $\{A_n\}_{n\geq 1}$ be a sequence of events with*

$$\sum_n P(A_n) < \infty.$$

Then

$$P(A_n \text{ i.o.}) = P(\limsup_{n\to\infty}) = 0.$$

The Definition 4.1 is an example definition of a good template just for demonstration.

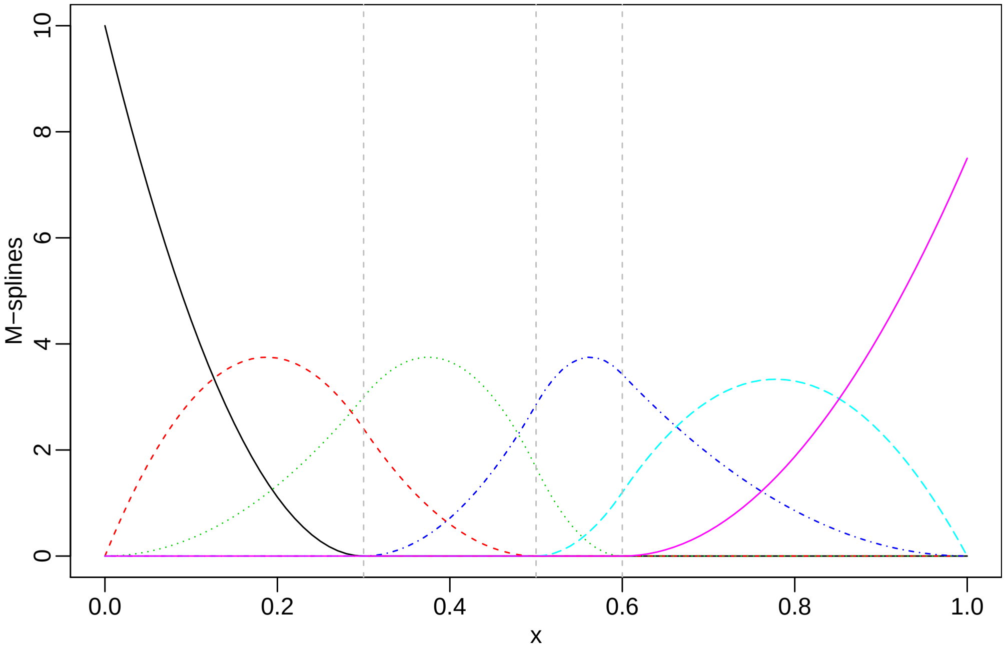

Figure 1: Quadratic M-spline Bases with three internal knots.

Definition 4.1. This is a good template if readers are able to easily follow and write their own stories based on it.

5 Figures

Figures can be generated by the code chunk within the source document. For example, the quadratic M-splines (Ramsay, 1988) with three internal knots generated by **splines2** package (Wang and Yan, 2017) are plotted by the following R code chunk. The resulting plot is shown in Figure 1.

```r
x <- seq.int(0, 1, 0.01)
knots <- c(0.3, 0.5, 0.6)
msMat <- mSpline(x, knots = knots, degree = 2, intercept = TRUE)
par(mar = c(2.5, 2.5, 0, 0), mgp = c(1.5, 0.5, 0))
matplot(x, msMat, type = "l", ylab = "M-splines")
abline(v = knots, lty = 2, col = "gray")
```

It is possible that we may not be able to regenerate the plot we want from R code. Instead of reproducing plots on the fly, we may also include an existing figure into the document by the function `knitr::include_graphics`. Suppose we have already generated a quadratic I-splines by function `splines2::iSpline` and saved the plot under directory `figs`. Then we may skip the regeneration step and include the existing plot directly as follows:

```r
knitr::include_graphics("figs/iSpline.png")
```

We may set the chunk option `echo = FALSE` so that the code chunk generating the plots are excluded from the output. Also, the chunk option `cache` can be set to be `TRUE` for time-consuming code chunk once the code chunk is unlikely to be modified.

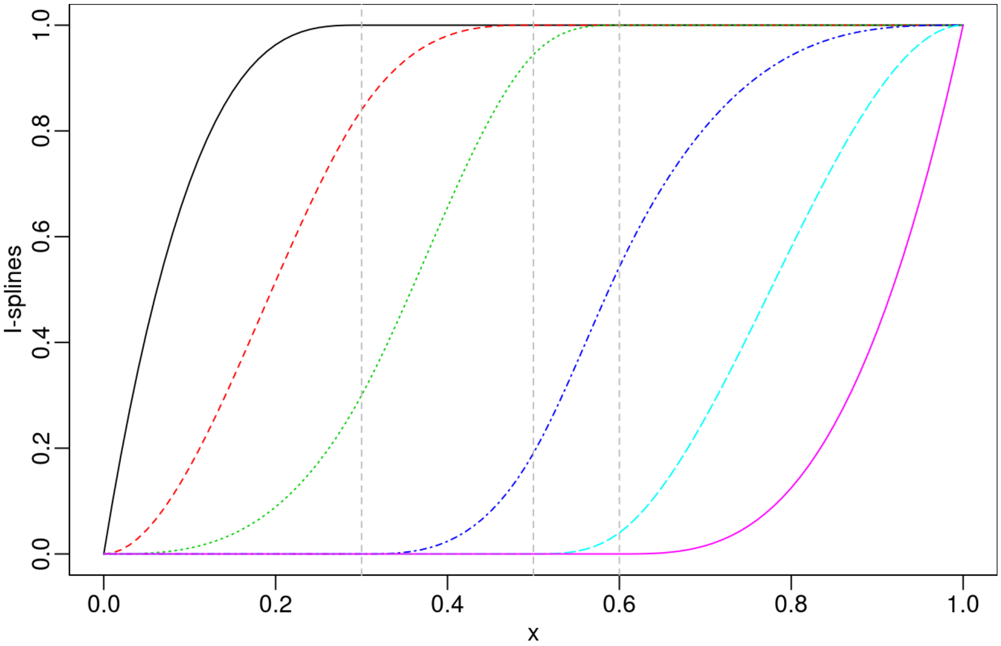

Figure 2: Quadratic I-spline Bases with three internal knots.

Table 2: First six rows of the iris data in package **datasets**.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

6 Tables and Other R objects

Tables can be similarly generated by the code chunk within the source document. Table 1 was in fact generated by function `knitr::kable`. Another simple example of table generation by `knitr::kable` is given in the follows code chunk. Table 2 is the resulting table.

```r
knitr::kable(head(iris), booktabs = TRUE,
             caption = 'First six rows of the iris data in package datasets.')
```

There are other R packages that can be of tremendous help in generating Markdown source of table and other R objects. For example, package **xtable** (Dahl, 2016) provides a more sophistic support for generation of table source for LaTeX and HTML; Package **pander** (Daróczi and Tsegelskyi, 2015) provides unctions printing a variety of R objects in **pandoc**'s Markdown; Package LaTeX **stargazer** (Hlavac, 2015) produces LaTeX code, HTML code and SCII text for well-formatted tables for results from regression models. See CRAN task view on reproducible research for a more comprehensive package list. .

7 Code Chunk

In addition to R, the code chunk can be written in a variety of other languages, such as Bash, Python, SAS, etc. by specifying the chunk option `engine`. The following code chunk is one toy example written in Python 3.

```
foo = "Hello " + "world!"
print("'" + foo + "'" + ' has %d letters.' % len(foo))
```

```
>>>  'Hello world!' has 12 letters.
```

We may set the chunk option `eval = FALSE` if we only want to present the code without evaluation.

8 Summary and Discussion

In summary, we provided the project template and reviewed the basic syntax of writing a single R Markdown document with the power and love of **bookdown** at the same time.

Xie (2017) provided a thorough introduction to **bookdown** including more advanced components, such as HTML widgets, and their usage. What's more, the manual of **Pandoc** gives all the available options that can be specified through YAML metadata section.

The template source and other associated files, such as BibTeX file, are available at the GitHub repository named datalab-templates.

Acknowledgment

We would like to thank Yihui and all the other authors, contributors for the fabulous **knitr**, **rmarkdown**, and **bookdown** packages. It would be also impossible for template to work without those fantastic open-source software: R, **pandoc**, etc.

Reference

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., and Hyndman, R. (2016), *rmarkdown: Dynamic Documents for R*, R package version 1.3.

Dahl, D. B. (2016), *xtable: Export Tables to LaTeX or HTML*, R package version 1.8-2.

Daróczi, G. and Tsegelskyi, R. (2015), *pander: An R Pandoc Writer*, R package version 0.6.0.

Hlavac, M. (2015), *stargazer: Well-Formatted Regression and Summary Statistics Tables*, Harvard University, Cambridge, USA, R package version 5.2.

Koohafkan, M. C. (2015), *kfigr: Integrated Code Chunk Anchoring and Referencing for R Markdown Documents*, R package version 1.2.

Ramsay, J. O. (1988), "Monotone Regression Splines in Action," *Statistical Science*, 425–441.

Wang, W. and Yan, J. (2017), *splines2: Regression Spline Functions and Classes Too*, R package version 0.2.4.

Xie, Y. (2016), *bookdown: Authoring Books and Technical Documents with R Markdown*, R package version 0.3.

— (2017), *bookdown: Authoring Books and Technical Documents with R Markdown*, Boca Raton, Florida: Chapman and Hall/CRC, iSBN 978-1138700109.