A Data Science Lab Project Template in R Markdown

Wenjie Wang[*]

22 February 2018

Abstract

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

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

1 Introduction

This document is designed as a template for data science lab projects. However, it can also be used as a general template in R Markdown for a single document.

The benefits of setting up a template in R Markdown are 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.

Cross-reference of mathematical equations, tables, and figures used to be a challenge when using R markdown. Usually extra packages, such as **kfigr** (Koohafkan, 2015), and extra efforts were needed for automatic and satisfactory cross-referencing. Fortunately, the arrival of the package **bookdown** (Xie, 2017) provides a much easier and more consistent syntax for cross-referencing.

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 the **bookdown** with examples. However, this is not intended as a tutorial of R Markdown or the **bookdown**. Readers are encouraged to skim the PDF or HTML output, and have a closer look at the source document of this template directly.

The rest of this project template is organized as follows: In Section 2 and Section 3, we present examples of writing mathematical equations, and mathematical environments, such as theorem, lemma, and definition, etc., respectively. Some examples for reproducing figures and including existing figures are given in Section 4. The generation of tables and other R objects is discussed in Section 5. A brief demonstration of a code chunk is given in Section 6. Several example HTML widgets and Shiny applications are given in Section 7 and Section 8, respectively. At last but not least, in Section 9, we point readers to some external resources for further reading and more advanced usage of **bookdown**.

2 Math Equations

Inline math expressions are quoted by $ in the source document, which is consistent with the syntax of LaTeX. For instance, x_i^2, $\sin(x)$, and θ are inline expressions. The equations can be simply quoted

[*]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	exm
exercise	Exercise	exr

by `$$` if no cross-reference is needed, where regular LaTeX commands under the math environment can be used. For equations that need cross-referencing, LaTeX environments for mathematical equations, such as `equation` or `align`, can be used directly. For example, Equation (1) is the well-known Euler's identity.

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

3 Math Theorem Environments

A mathematical theorem can be put inside a `theorem` chunk followed by its label. For example, the Central Limit Theorem (CLT) is presented in Theorem 3.1.

Theorem 3.1. ***(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*

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

Similarly, a lemma can be put inside a `lemma` chunk. For instance, the First Borel-Cantelli Lemma is given in Lemma 3.1.

Lemma 3.1. ***(First Borel-Cantelli Lemma)*** *Let $\{A_n\}_{n\geq 1}$ be a sequence of events with*

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

Then

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

All the available theorem environments and their label prefix designed for cross-referencing are summarized in Table 1.

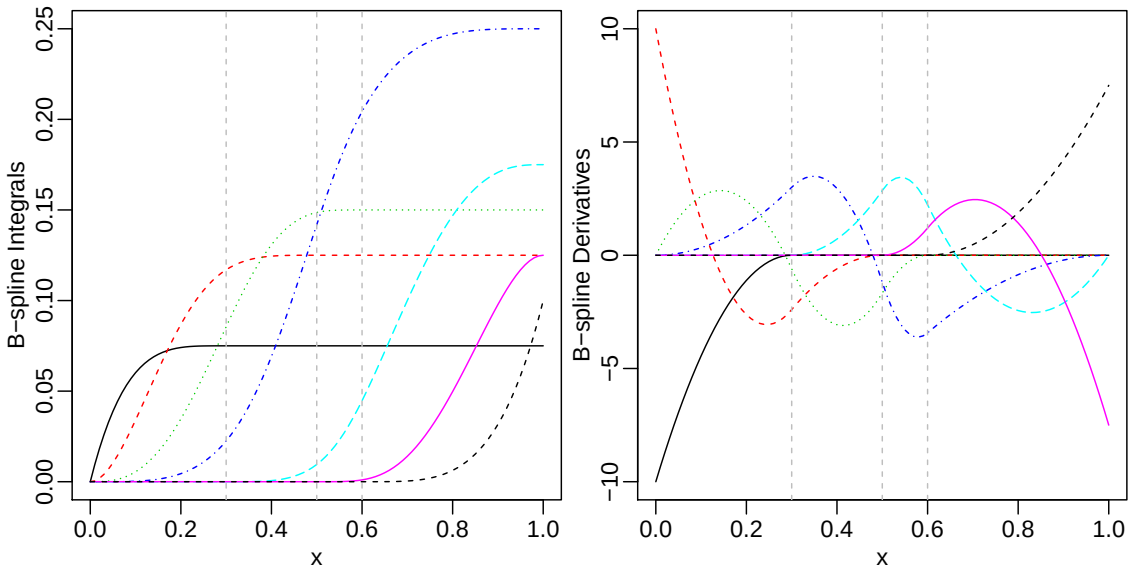

Figure 1: Integrals (left) and derivatives (right) of cubic B-splines with three internal knots.

4 Figures

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

```
x <- seq.int(0, 1, 0.01)
knots <- c(0.3, 0.5, 0.6)
ibsMat <- ibs(x, knots = knots, intercept = TRUE)
dbsMat <- dbs(x, knots = knots, intercept = TRUE)
par(mar = c(2.5, 2.5, 0.2, 0.2), mgp = c(1.5, 0.5, 0), mfrow = c(1, 2))
matplot(x, ibsMat, type = "l", ylab = "B-spline Integrals")
abline(v = knots, lty = 2, col = "gray")
matplot(x, dbsMat, type = "l", ylab = "B-spline Derivatives")
abline(v = knots, lty = 2, col = "gray")
```

It is possible that we may not wish to regenerate a plot from R code. Instead of reproducing plots on the fly, we may also include an existing figure in the document by the function `knitr::include_graghics`. Suppose we have already generated quadratic M-splines and I-splines (Ramsay, 1988) with three internal knots by **splines2** and saved the plots under directory `figs`, respectively. Then we may skip the regeneration step and include the existing plot directly as follows:

```
knitr::include_graphics(c("figs/mSpline.png", "figs/iSpline.png"))
```

In the code chunk shown above, the chunk option `out.width = '45%'` and `fig.show = 'hold'` were set so that the plots were placed side by side. 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 chunks once the code chunk is unlikely to be modified.

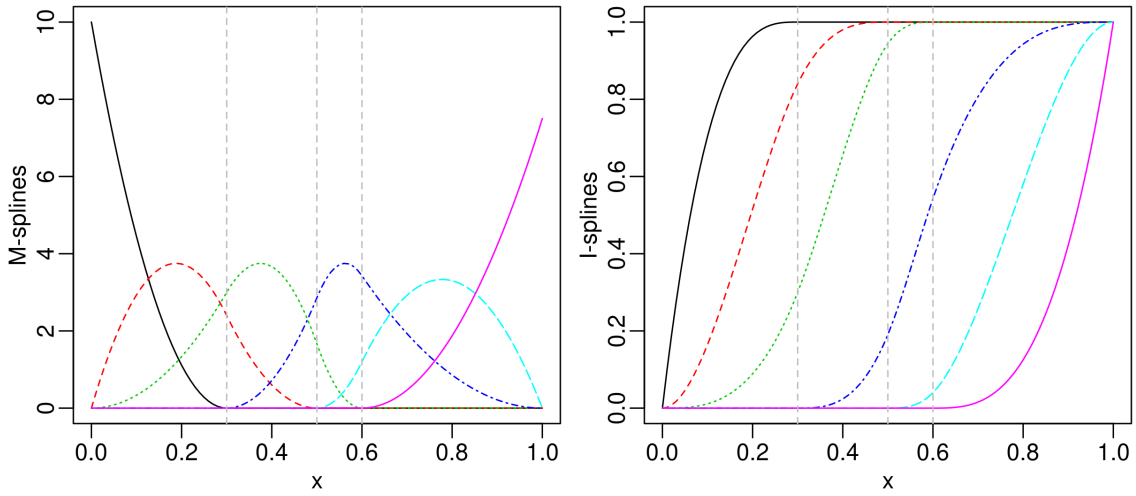

Figure 2: Quadratic M-spline (left) and I-spline (right) Bases with three internal knots.

Table 2: The first six rows of the iris dataset.

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

5 Tables and Other R objects

Tables can be similarly generated by a 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 following code chunk. Table 2 is the resulting table.

```
knitr::kable(head(iris), booktabs = TRUE,
             caption = 'The first six rows of the iris dataset.')
```

There are other R packages that can be of tremendous help in generating the Markdown source for various R objects. For example, the package **xtable** (Dahl, 2016) provides a more sophisticated support for generation of table source for LaTeX and HTML; the package **pander** (Daróczi and Tsegelskyi, 2015) provides functions for printing a variety of R objects in **pandoc**'s Markdown; the package **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.

6 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("The length of '%s' is %d." % (foo, len(foo)))
```

```
>>>  The length of 'Hello world!' is 12.
```

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

7 HTML Widgets

The **htmlwidgets** package (Vaidyanathan et al., 2016) provides a framework for easily creating R bindings to JavaScript libraries. Several R packages built based on it, such as **leaflet** (Cheng and Xie, 2016) and **DT** (Xie, 2016), enable us to embed interactive HTML widgets in the HTML output. For PDF output, a screenshot taken by the package **webshot** (Chang, 2016) will be included instead.

For example, we embed a map for the location of Department of Statistics at University of Connecticut (UConn) by **leaflet** in Figure 3.

```
urlTemplate <- "https://{s}.tile.openstreetmap.org/{z}/{x}/{y}.png"
leaflet(width = 900, height = 500) %>% addTiles(urlTemplate = urlTemplate) %>%
    setView(- 72.251113, 41.810757, zoom = 17) %>%
    addPopups(- 72.251113, 41.810757,
              '<b>Department of Statistics, UConn</b>')
```

Another example of using the package **DT** to display `mtcars` data is given here. The result is shown in Figure 4.

```
DT::datatable(mtcars)
```

8 Shiny Apps

The package **shiny** (Chang et al., 2017) is a great tool providing readers with an interactive way to explore data and results. We may easily build Shiny applications on our own, deploy, and share it online at shinyapps.io by the package **rsconnect** (Allaire, 2016). In addition to building regular applications by **Shiny**, the package **miniUI** (Cheng, 2016) provides layout function designed for Shiny applications with appropriate size on small screens.

We may embed Shiny applications in the document by `knitr::include_app`, which is mainly designed for HTML output. Similarly, a screenshot taken by **webshot** will be embedded instead for PDF output. The package **webshot** provides argument `zoom` for a possible high resolution screenshot. However, if the resolution is still not satisfactory, we may take a screenshot and include it manually by `knitr::include_graphics`.

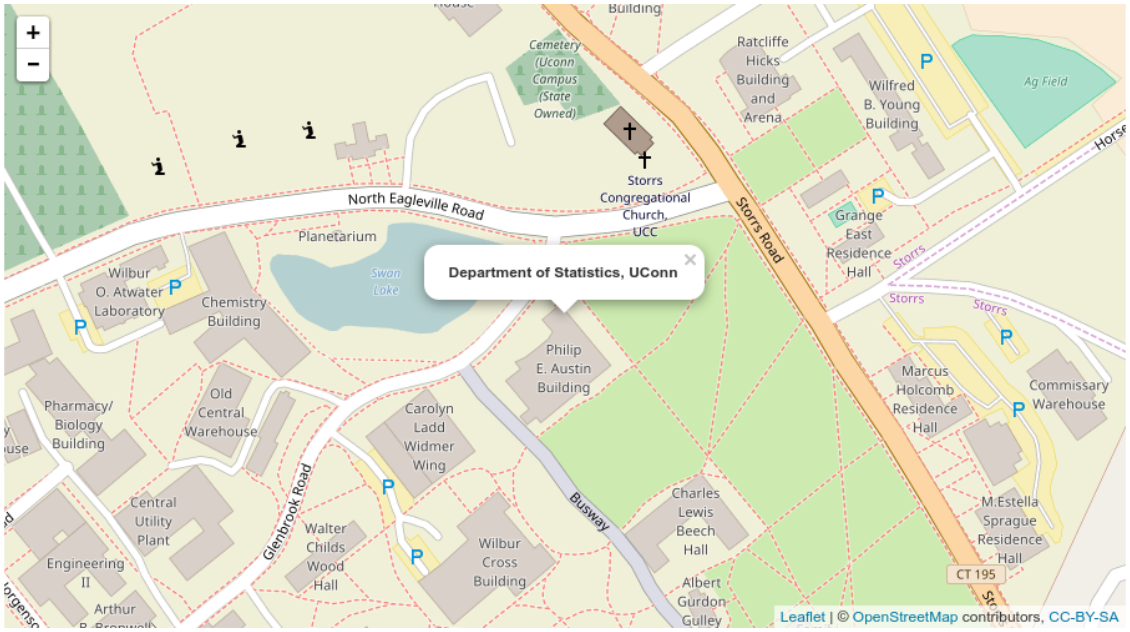

Figure 3: A map widget rendered via the **leaflet** package.

Show 10 entries Search:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4

Showing 1 to 10 of 32 entries Previous 1 2 3 4 Next

Figure 4: A table widget rendered via the **DT** package.

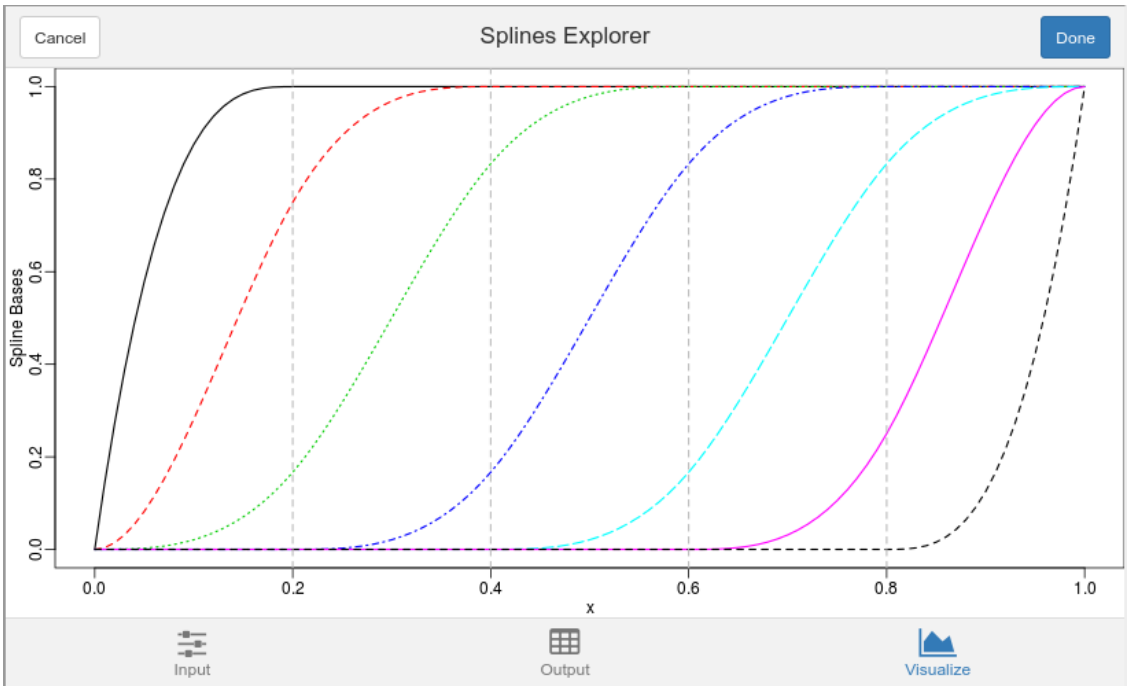

Figure 5: An example of Shiny app visualizing different spline bases available at https://wenjie-stat.shinyapps.io/minisplines2.

An example Shiny application visualizing different kind of spline bases is given in Figure 5.

```
knitr::include_app("https://wenjie-stat.shinyapps.io/minisplines2/", "500px")
```

9 Summary and Discussion

In summary, we provided this project template and reviewed most common components and their syntax of writing a single R Markdown document with the power and love of **bookdown** and many other fantastic packages.

Xie (2017) provided a thorough introduction to **bookdown** including more advanced customization and other output formats. Additionally, the manual of **Pandoc** gives all the available options that can be specified through the YAML metadata section.

The template source and other associated files, such as BibTeX and CSS file, are available at our GitHub repository *dslab-templates*: https://github.com/statds/dslab-templates.

Acknowledgment

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

Reference

Allaire, J. (2016), *rsconnect: Deployment Interface for R Markdown Documents and Shiny Applications*, R package version 0.7.

Chang, W. (2016), *webshot: Take Screenshots of Web Pages*, R package version 0.4.0.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017), *shiny: Web Application Framework for R*, R package version 1.0.0.

Cheng, J. (2016), *miniUI: Shiny UI Widgets for Small Screens*, R package version 0.1.1.

Cheng, J. and Xie, Y. (2016), *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, R package version 1.0.1.

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.

Vaidyanathan, R., Xie, Y., Allaire, J., Cheng, J., and Russell, K. (2016), *htmlwidgets: HTML Widgets for R*, R package version 0.8.

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

Xie, Y. (2016), *DT: A Wrapper of the JavaScript Library 'DataTables'*, R package version 0.2.

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