Parallel Computation Framework for Discrete Copula Modeling

Dhyey Mavani

Amherst College

Submitted to the Department of Mathematics and Statistics
of Amherst College in partial fulfillment of the requirements
for the degree of Bachelor of Arts with honors.

Advisor(s):
Professor Shu-Min Liao

October 15, 2024

Table of contents

Abstract

Categorical data, particularly when involving ordinal responses, plays a crucial role in fields such as social sciences and finance, where capturing the inherent ordering of variables can lead to more meaningful inferences of the underlying distributions. A key challenge in analyzing such data lies in exploring the regression dependence structures among variables. While many model-based approaches have been developed to examine these structures, there remains a scarcity of model-free methods. Wei & Kim (2021) addressed this gap by introducing SCCRAM, a novel model-free measure based on the checkerboard copula, capable of identifying and quantifying regression dependence in multivariate categorical data with ordinal and nominal variables. This work builds on their contribution by developing scalable R and Python packages for SCCRAM, employing parallel computing to enhance accessibility and efficiency in large-scale data analysis. Initial simulations demonstrate the effectiveness of these implementations, offering researchers a robust tool for exploratory modeling and further exploration of regression dependence structures in categorical datasets for this use-case.

Acknowledgements

I would like to thank everyone who made my experience nothing short of amazing during my thesis journey. Firstly, I'd like to convey my gratitude to Professor Shu-Min Liao for advising me through this journey, and for believing in me to take on statistical software component of her most recent research work along with her collaborators. From my first research experience on campus building R-Blocks, she played a pivotal role in my journey.

I am also incredibly grateful to my college advisor, and statistics major advisor, Professor Nicholas Horton for always advocating for me and for supporting me throughout the Amherst College experience. I would also like to thank Professor Jun Ishii for teaching me Advanced Econometrics, which helped me gain a really clear understanding on the foundational tools on which I was able to build on in my work.

Finally, I would like to thank my friends and family for their constant belief in my abilities. Special thanks to my mom, dad, sister, grandfather, and grandmother for lending me and making me capable for this opportunity to study abroad. Last, but not the least, I would like to thank my friends, peers, and colleagues on campus, who took courses, worked, and played sports with me. This journey of academic, personal, and professional growth wouldn't be possible without their support.

Chapter 1

Intro to your Quarto Book thesis

This template uses a Quarto Book project to generate your thesis document. This document includes important information about how to navigate this project directory and use the Quarto Book project.

To use the Amherst College Statistics Thesis Template, do **one** of the following:

- Download the repo.

- Fork the repo.

- Using the command line or terminal, go to the directory where you want your thesis project directory to be stored and run the following to copy the contents of the repo:

  ```
  quarto use template Amherst-Statistics/thesis
  ```

1.1 What's in this project directory?

- `_extensions` folder: contains additional files for formatting the thesis. Do not delete, edit, or move this folder or its contents.
- `_quarto.yml` file. Do not rename, move, or delete this file.
  - This file controls the structure and formatting of your thesis.
  - Update this file with your thesis information, the list of qmd files to include when rendering the document, and the name of your bibliography file.
- `.qmd` files containing the contents of your thesis.
  - You should **not** delete or rename the `index.qmd` file. It must be included as named the first rendered file of your thesis. This file includes your acknowledgements, abstract, and a reproducibility statement. Be sure to edit these sections before your final thesis submission!
  - You should **not** delete the *references* or the *appendix* files (Appendix **??**, Appendix **??**) which contain syntax and information required for proper thesis formatting, but you can and should rename them as needed.
  - You should delete the *chapter* files and replace them with your own.
- `includes` folder: contains citation style file (for formatting in-text citations and the bibliography using the ASA style) and the Amherst logo included on the cover page. Do not delete, edit, or move this folder or its contents.
- `data` folder: recommended folder for storing data files
- `fig` folder: recommended folder for storing any images, graphs, etc.
- `src` folder: recommended folder for storing any code scripts.

ℹ Note

Items in the data, fig, and src folders starting with temp are used as examples within this template but can be deleted when you begin your own edits.

- After rendering your document, you will additionally see a _book folder, which will contain the rendered PDF.

- references.bib: This is a bibliography file containing bibtex-style citation information. BibTeX is a format for citations used by LaTeX, which is the language that is used in the process of rendering the document to PDF. The format uses *citation keys*, which are unique identifiers for each citation that you can customize manually or through the use of citation management software. These citation keys are what are used to make in-text cross-references throughout your thesis.

  - You can rename the .bib file, just make sure to update the filename in the _quarto.yml file.
  - Consider using Zotero for managing all of the literature affiliated with your thesis. Amherst College has Zotero support and free unlimited storage through the library.
  - You can set up Zotero to automatically update your .bib file that contains all of your reference information. Video setup instructions are also available.
  - Set up an appointment with a science librarian at Amherst or attend one of the library's Zotero workshops to learn more. Your thesis advisor should also be able to help.

1.2 Quarto workflow

1. Create and edit .qmd files.

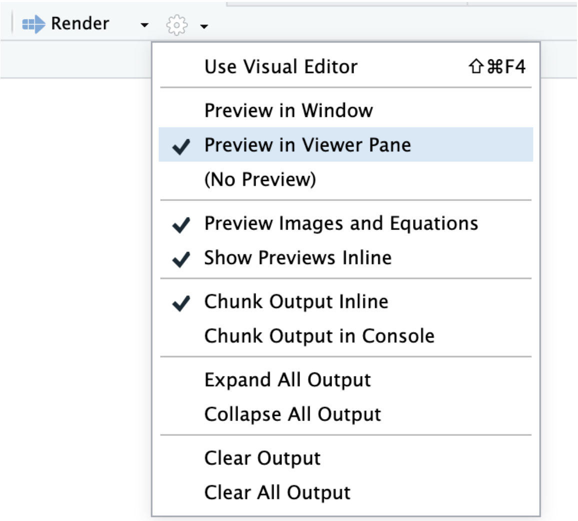
- When naming chapter and appendix files, we recommend leading the filename with the number of its rendering order.
- The `index.qmd` file must remain named `index.qmd`

2. Update the `_quarto.yml` file.

    - Add or remove `.qmd` files to the list of `chapters` or `appendices` as appropriate. You can also temporarily comment out chapters.
    - Save your own `.bib` file and make sure the filename is updated in `_quarto.yml`.

3. Render book

    - Type `cmd + shift + k` or `ctrl + shift + k` while working in any of your `.qmd` files, use the `Render` button at the top of any `.qmd` file, or use the `Render` button available under the `Build` tab.
    - If you would like the PDF to preview within the RStudio Viewer pane, click the gear icon next to the `Render` button at the top of any of your `.qmd` files and select **Preview in Viewer Pane.** It will change the default behavior until you make a different selection.



> Note
>
> Rendering any `.qmd` file will render the *entire* project (all files listed under `chapters` in the `_quarto.yml`).
>
> If you want to focus on one particular chapter without running code in or rendering other chapters, delete or comment out the other chapter files in `_quarto.yml`. The `index.qmd` file must always be included in the rendering list, though.

4. Review the changes in the PDF (located in the _book folder) to ensure the PDF output looks as desired. Here are some common issues you might want to look for:

    - No text, figures, tables, nor code extends into the margins
    - Citations and cross-references are displayed correctly and linked correctly
    - Font in figures and tables is readable
    - All figures and tables are captioned and numbered
    - Any equations that are referenced in the text should also be numbered.
    - Make appropriate use of Quarto's theorem and proof blocks when needed.
    - Though callout blocks might be useful for inserting notes to yourself or your advisor, they are likely not appropriate for your final thesis and should be excluded.
    - Figures are visually appealing (aspect ratio, appropriate use of aesthetics for contrast, etc.)
    - Headings, subheadings, lists, sublists, etc. are formatted correctly. Errors are usually due to indentation or line spacing before headings or lists.

> ℹ Note
>
> To format additional content within a list item at the same depth, the first text of the new content must align with the first text of the corresponding list item text above it. For example, the first colon `:` of this callout block is aligned with the first letter of the bullet point above.

In comparison, this text will be indented to the same depth as the text in item 4.

1.3 Recommended reading before you begin editing

Quarto is a powerful tool for authoring documents in multiple formats. Even if you are familiar with markdown, some behavior and syntax in Quarto markdown is slightly different (and often better) than R markdown.

> 🔥 Caution
>
> The code within each `.qmd` file is self-contained and rendered independently before merging the document to build the thesis, so there's no environmental memory between chapters. This means you should load required packages and datasets and set up any desired code defaults at the start of each chapter file.

1. Quarto Markdown Basics, how to use inline code with the Knitr engine, and more details on including external figures.

   > ℹ Note
   >
   > Many options will work for both PDF and HTML formats, but your thesis must be submitted as a PDF. Pay attention to which elements work for both formats and which elements are only appropriate for HTML format. You need to prioritize the

> PDF rendering.

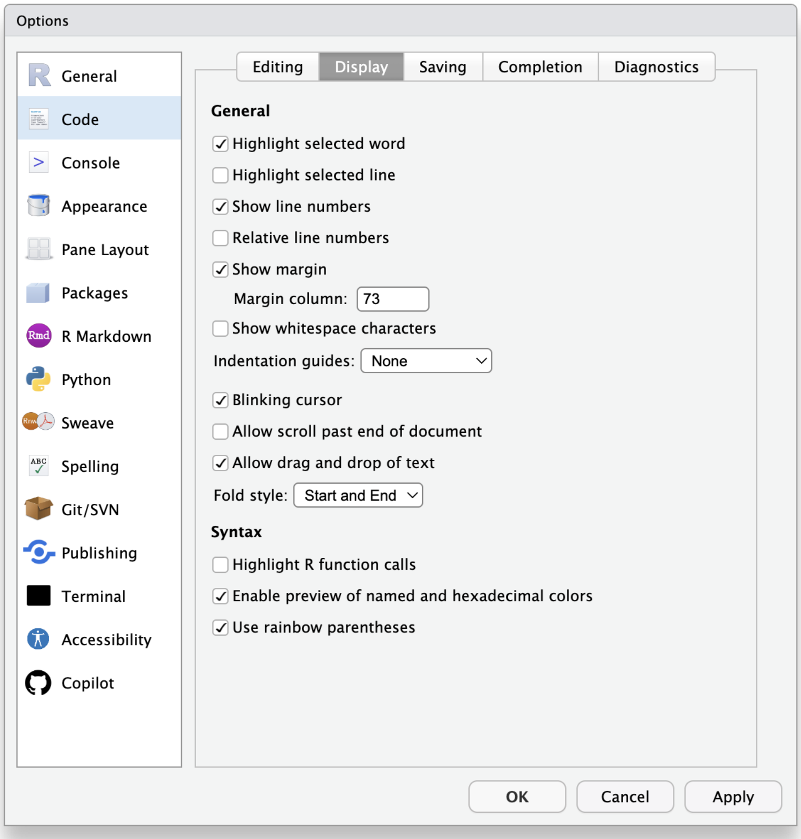
2. How to specify code chunk options in Quarto and the list of code chunk options (the Quarto Book is rendered using the KnitR engine so you can follow links on the latter page to see additional options). The document-level defaults set `echo`, `warning` and `message` to `false` (specified in `_extensions/amherst-thesis/_extension.yml`).

> ℹ Note
>
> Quarto also supports other languages such as python (see the list under the Computations section of the side menu at the linked page), but may require different YAML settings in order to execute.

Caution

To ensure all of your displayed code fits within the printed code blocks without bleeding into the margins, set up RStudio to show the margin line at 73. None of your code or comments should fall past that line within your scripts or `.qmd`. In other words, no line of code should be more than 73 characters long.

1. Go to **Tools** → **Global Options...**
2. Select the **Code** menu
3. Choose the **Display** tab
4. Check **Show margin** and set **Margin column: 73**



General

☑ Highlight selected word

☐ Highlight selected line

☑ Show line numbers

☐ Relative line numbers

☑ Show margin

Margin column: 73

☐ Show whitespace characters

Indentation guides: None

☑ Blinking cursor

☐ Allow scroll past end of document

☑ Allow drag and drop of text

Fold style: Start and End

Syntax

☐ Highlight R function calls

☑ Enable preview of named and hexadecimal colors

☑ Use rainbow parentheses

3. Basics of cross-referencing in Quarto chapters, sections, tables, figures, equations, etc.

4. How cross-referencing works in Quarto Book projects like this one.

   To reiterate their cautions here…

   > ⚠ Reserved prefixes for cross-referencing
   >
   > Unless you are creating a cross-reference, avoid using the reserved cross-reference prefixes for code cell labels (e.g. set using the `label` code cell option) and element IDs (set using a # in an attribute).
   >
   > The reserved prefixes are: `fig`, `tbl`, `lst`, `tip`, `nte`, `wrn`, `imp`, `cau`, `thm`, `lem`, `cor`, `prp`, `cnj`, `def`, `exm`, `exr`, `sol`, `rem`, `eq`, `sec`.
   >
   > You must use the reserved prefixes with their corresponding element types (start every figure label with `fig-`, start every header label with `sec-`, start every table label with `tbl-`, and so on).
   >
   > Also avoid using underscores (_) in labels and IDs as this can cause problems when rendering to PDF with LaTeX.

5. Citations in Quarto

6. To follow best practices, we recommend stylizing your code and text following the **tidyverse** style guide.

7. Mathematical expressions in PDF documents require LaTeX formatting within inline and display style equations. Rice University provides a convenient list of LaTeX mathematical symbols for your reference.

ℹ Note

Letters in math mode are individual variables. If you want to write text in math modes such as distribution names, words, etc, then you should use LaTeX commands to do so.

It is best practice to include any distribution names within `\operatorname{}`. For example, we might write $X \sim \operatorname{Normal}(\mu, \sigma^2)$.

When otherwise writing text within an equation, use `\text{}` or `\textit{}` for upright or italic text, respectively. For example:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Finally, to get parentheses and brackets to match the size of their contents, make use of the `\left` and `\right` commands. Compare the two following examples

1. Without `\left` and `\right`

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

2. With `\left` and `\right`

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

See the Comprehensive Quarto Guide for more advanced details. See Chapter **??** for some simple examples.

Chapter 2

Simple examples

2.1 Figures

Consider customizing your plot themes per-plot—as we do below to create Figure **??**—or changing the default `ggplot()` theme in your document within your `setup` code chunks using `ggplot2::theme_set()`.

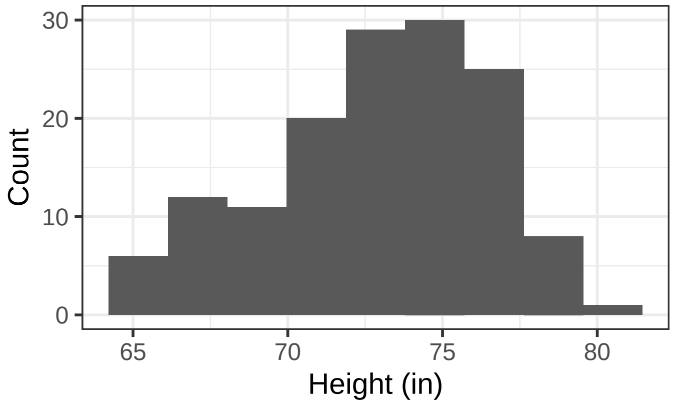
If **ggplot2** is loaded, the following code sets the default `ggplot()` theme to `theme_classic()`.

```
theme_set(theme_classic())
```

2.2 Tables

Your tables should be publication quality. Consider using **gt** (Iannone et al. 2024) or **kableExtra** (Zhu 2024) to customize your tables. The **gtsummary** package (Sjoberg et al. 2021) may also come in handy.

Table **??** shows the average heights of WNBA players by position.



Source: https://www.espn.com/wnba/stats/player

Figure 2.1: Distribution of heights of WNBA players in the 2024 season.

Table 2.1: Average WNBA player height by position.

Position	Average height (in)
Guard	70.2
Forward	74.9
Center	77.3

2.3 Chapter ?? Code

The following code was used to create Chapter **??**.

2.3.1 Code within chapter

```
# Load packages
library(tidyverse)
library(gt)

# Set default ggplot theme for document
```

```r
theme_set(theme_classic())
# If using kableExtra tables, print blank cells instead of `NA`
options(knitr.kable.NA = "")

# Load data
load("data/temp_wnba.RData")
# Use Freedman-Diaconus rule to set binwidth
ht_bw <- 2 * IQR(wnba$height) / nrow(wnba)^(1/3)

# Create histogram of height faceted by player position
ggplot(wnba, aes(height)) +
  geom_histogram(binwidth = ht_bw) +
  labs(x = "Height (in)",
       y = "Count",
       caption = "Source: https://www.espn.com/wnba/stats/player") +
  theme_bw()
wnba |>
  group_by(position) |>
  summarize(mean_ht = mean(height)) |>
  gt() |>
  cols_label(
    position = "Position",
    mean_ht = "Average height (in)"
  ) |>
  fmt_number(decimals = 1)
# =======================================================================
# Sample R script for thesis template
#
# Cleans temp_raw_wnba.csv dataset, which contains data pulled from
# https://www.espn.com/wnba/stats/player on 2024/06/19
#
# Last updated: 2024/06/19
# =======================================================================
library(tidyverse)

wnba <- read_csv("data/temp_raw_wnba.csv") |>
  janitor::clean_names() |>
  # Pull jersey numbers off of names and
  # turn height text into msmt (6'4" = 6.3333)
  mutate(jersey = str_extract(name, "[0-9]+$"),
         name = str_remove(name, "[0-9]+$"),
         ht_ft = parse_number(str_extract(ht, "^[0-9]")),
         ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),
         height = ht_ft * 12 + ht_in,
         weight = parse_number(wt),
         position = factor(pos,
```

```r
                           levels = c("G", "F", "C"),
                           labels = c("Guard", "Forward", "Center"))) |>
  select(-c(ht, wt, ht_ft, ht_in, pos))

save(wnba, file = "data/temp_wnba.RData")
```

2.3.2 Code sourced from external scripts

```r
# =======================================================================
# Sample R script for thesis template
#
# Cleans temp_raw_wnba.csv dataset, which contains data pulled from
# https://www.espn.com/wnba/stats/player on 2024/06/19
#
# Last updated: 2024/06/19
# =======================================================================
library(tidyverse)

wnba <- read_csv("data/temp_raw_wnba.csv") |>
  janitor::clean_names() |>
  # Pull jersey numbers off of names and
  # turn height text into msmt (6'4" = 6.3333)
  mutate(jersey = str_extract(name, "[0-9]+$"),
         name = str_remove(name, "[0-9]+$"),
         ht_ft = parse_number(str_extract(ht, "^[0-9]")),
         ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),
         height = ht_ft * 12 + ht_in,
         weight = parse_number(wt),
         position = factor(pos,
                           levels = c("G", "F", "C"),
                           labels = c("Guard", "Forward", "Center"))) |>
  select(-c(ht, wt, ht_ft, ht_in, pos))

save(wnba, file = "data/temp_wnba.RData")
```

References

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., Seo, J., Brevoort, K., and Roy, O. (2024), *Gt: Easily create presentation-ready display tables*.

Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., and Larmarange, J. (2021), "Reproducible summary tables with the gtsummary package," *The R Journal*, 13, 570–580. https://doi.org/10.32614/RJ-2021-053.

Ushey, K., and Wickham, H. (2024), *Renv: Project environments*.

Zhu, H. (2024), *kableExtra: Construct complex table with 'kable' and pipe syntax*.

Appendix A

Code availability

This thesis is written using Quarto with **renv** (Ushey and Wickham 2024) to create a reproducible environment. All materials (including the data sets and source files) required to reproduce this document can be found at the Github repository github.com/GITHUB-USERNAME/THESIS-REPO-NAME.

```
# =======================================================================
# Sample R script for thesis template
#
# Cleans temp_raw_wnba.csv dataset, which contains data pulled from
# https://www.espn.com/wnba/stats/player on 2024/06/19
#
# Last updated: 2024/06/19
```

```r
# =======================================================================
library(tidyverse)

wnba <- read_csv("data/temp_raw_wnba.csv") |>
  janitor::clean_names() |>
  # Pull jersey numbers off of names and
  # turn height text into msmt (6'4" = 6.3333)
  mutate(jersey = str_extract(name, "[0-9]+$"),
         name = str_remove(name, "[0-9]+$"),
         ht_ft = parse_number(str_extract(ht, "^[0-9]")),
         ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),
         height = ht_ft * 12 + ht_in,
         weight = parse_number(wt),
         position = factor(pos,
                           levels = c("G", "F", "C"),
                           labels = c("Guard", "Forward", "Center"))) |>
  select(-c(ht, wt, ht_ft, ht_in, pos))

save(wnba, file = "data/temp_wnba.RData")
```

```r
# =======================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
```

```
#
# Last updated: 2024/08/24
# =======================================================================

print("Hello, Amherst!")
```

Appendix B

Corrections

This section may be excluded if no corrections are made to your thesis after initial submission to the department and before final submission to the college.

Per the Statistics Honors Thesis Regulations:

> Corrections to theses may be made after the date on which they are due in the Department's hands. Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail, and substantial

> additions to the thesis will not be allowed. Questions about what should appear in the "Corrections" should be directed to the Chair. Electronic versions of the thesis, technical appendix, and necessary data and supplemental files must all be updated at the time of correction as well.

When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected.

1. ...
2. ...