Parallel Computation Framework for Discrete Copula Modeling

Dhyey Mavani

Amherst College

Table of contents

Abstract

Understanding regression dependencies among discrete variables in categorical data—especially with ordinal responses—is a significant challenge in fields like finance, where the natural order of variables can unlock deeper insights into underlying distributions. While numerous model-based methods have been developed to examine these structures, there is a notable lack of flexible, model-free approaches. To address this gap, SCCRAM, a novel model-free measure based on the checkerboard copula, was introduced to identify and quantify regression dependence in multivariate categorical data involving both ordinal and nominal variables. Building upon this foundation, my thesis focuses on developing scalable implementations of SCCRAM in R and Python, utilizing parallel computing to enhance efficiency and accessibility for large-scale data analysis. Initial simulations confirm the effectiveness of these tools, providing researchers with a powerful resource for exploratory modeling and a deeper investigation into regression dependence structures within complex categorical datasets.

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

Introduction

XXX

1.1   What's in this project directory?

XXX

- `fig` XXX

> **i** Note
>
> XXX

1.2   Quarto workflow

1. XXX

   - XXX

1.3 Recommended reading before you begin editing

XXX

> ⬥ **Caution**
>
> XXX

XXX

Chapter 2

Unraveling Notion of Dependence through Copulas

2.1 Figures

Consider customizing your plot themes per-plot—as we do below to create Figure 2.1—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().

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

2.2 Tables

Your tables should be publication quality. Consider using **gt** (Iannone et al. 2024) or **kable-Extra** (Zhu 2024) to customize your tables. The **gtsummary** package (Sjoberg et al. 2021)

Source: https://www.espn.com/wnba/stats/player
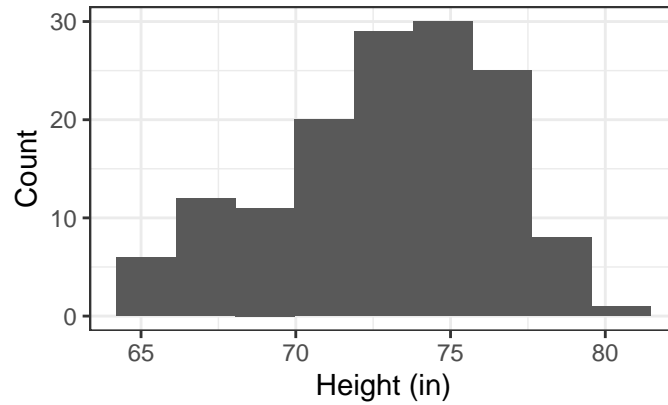
Figure 2.1: Distribution of heights of WNBA players in the 2024 season.

Table 2.1: Average WNBA player height by position.

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

may also come in handy.

Table 2.1 shows the average heights of WNBA players by position.

2.3 Chapter 2 Code

The following code was used to create Chapter 2.

2.3.1 Code within chapter

```
1   # Load packages

2   library(tidyverse)

3   library(gt)

4

5   # Set default ggplot theme for document
```

```r
6   theme_set(theme_classic())

7   # If using kableExtra tables, print blank cells instead of `NA`

8   options(knitr.kable.NA = "")

9

10  # Load data

11  load("data/temp_wnba.RData")

12  # Use Freedman-Diaconus rule to set binwidth

13  ht_bw <- 2 * IQR(wnba$height) / nrow(wnba)^(1/3)

14

15  # Create histogram of height faceted by player position

16  ggplot(wnba, aes(height)) +

17    geom_histogram(binwidth = ht_bw) +

18    labs(x = "Height (in)",

19         y = "Count",

20         caption = "Source: https://www.espn.com/wnba/stats/player") +

21    theme_bw()

22  wnba |>

23    group_by(position) |>

24    summarize(mean_ht = mean(height)) |>

25    gt() |>

26    cols_label(

27      position = "Position",

28      mean_ht = "Average height (in)"
```

```
29      ) |>

30      fmt_number(decimals = 1)

31   # ============================================================================

32   # Sample R script for thesis template

33   #

34   # Cleans temp_raw_wnba.csv dataset, which contains data pulled from

35   # https://www.espn.com/wnba/stats/player on 2024/06/19

36   #

37   # Last updated: 2024/06/19

38   # ============================================================================

39   library(tidyverse)

40

41   wnba <- read_csv("data/temp_raw_wnba.csv") |>

42      janitor::clean_names() |>

43      # Pull jersey numbers off of names and

44      # turn height text into msmt (6'4" = 6.3333)

45      mutate(jersey = str_extract(name, "[0-9]+$"),

46             name = str_remove(name, "[0-9]+$"),

47             ht_ft = parse_number(str_extract(ht, "^[0-9]")),

48             ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),

49             height = ht_ft * 12 + ht_in,

50             weight = parse_number(wt),

51             position = factor(pos,
```

```
52                                  levels = c("G", "F", "C"),

53                                  labels = c("Guard", "Forward", "Center"))) |>

54    select(-c(ht, wt, ht_ft, ht_in, pos))

55

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

2.3.2 Code sourced from external scripts

```
1   # ==============================================================================

2   # Sample R script for thesis template

3   #

4   # Cleans temp_raw_wnba.csv dataset, which contains data pulled from

5   # https://www.espn.com/wnba/stats/player on 2024/06/19

6   #

7   # Last updated: 2024/06/19

8   # ==============================================================================

9   library(tidyverse)

10

11  wnba <- read_csv("data/temp_raw_wnba.csv") |>

12    janitor::clean_names() |>

13    # Pull jersey numbers off of names and

14    # turn height text into msmt (6'4" = 6.3333)

15    mutate(jersey = str_extract(name, "[0-9]+$"),

16           name = str_remove(name, "[0-9]+$"),

17           ht_ft = parse_number(str_extract(ht, "^[0-9]")),
```

```r
18            ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),

19            height = ht_ft * 12 + ht_in,

20            weight = parse_number(wt),

21            position = factor(pos,

22                              levels = c("G", "F", "C"),

23                              labels = c("Guard", "Forward", "Center"))) |>

24      select(-c(ht, wt, ht_ft, ht_in, pos))

25

26  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.

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).

```
1  # ============================================================================
2  # Sample R script for thesis template
3  #
4  # Cleans temp_raw_wnba.csv dataset, which contains data pulled from
5  # https://www.espn.com/wnba/stats/player on 2024/06/19
6  #
7  # Last updated: 2024/06/19
```

```r
8   # ============================================================================
9   library(tidyverse)
10
11  wnba <- read_csv("data/temp_raw_wnba.csv") |>
12    janitor::clean_names() |>
13    # Pull jersey numbers off of names and
14    # turn height text into msmt (6'4" = 6.3333)
15    mutate(jersey = str_extract(name, "[0-9]+$"),
16           name = str_remove(name, "[0-9]+$"),
17           ht_ft = parse_number(str_extract(ht, "^[0-9]")),
18           ht_in = parse_number(str_extract(ht, '[0-9]+\\"$')),
19           height = ht_ft * 12 + ht_in,
20           weight = parse_number(wt),
21           position = factor(pos,
22                             levels = c("G", "F", "C"),
23                             labels = c("Guard", "Forward", "Center"))) |>
24    select(-c(ht, wt, ht_ft, ht_in, pos))
25
26  save(wnba, file = "data/temp_wnba.RData")
```

```r
1   # ============================================================================
2   # Sample R script for thesis template
3   #
4   # Doesn't do anything useful
```

```
5    #

6    # Last updated: 2024/08/24

7    # ============================================================================

8

9    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. …