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

January 13, 2025

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 and data generating processes (DGPs). 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, a novel model-free measure based on the checkerboard copula, was introduced by Wei and Kim (2021) 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 and modularized implementations of discrete checkerboard copula modeling in R and Python, utilizing parallel computing to enhance efficiency and accessibility for large-scale data analysis. Initial experimentation and deployment 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 want to thank everyone who made my experience unforgettable during my thesis journey. Firstly, I want to convey my gratitude to Professor Shu-Min Liao for advising me throughout my time at Amherst College and believing in me to take on the challenge of developing a statistical software component encompassing her most recent research work. From my first research experience on campus building R-Blocks to introducing me to her research collaborator (Professor Daeyoung Kim), Prof. Liao played a pivotal role in my development. Additionally, I am indebted to Dr. Kim for his continuous encouragement and feedback while developing the software this past year.

I am also incredibly grateful to my college advisor and statistics major advisor, Professor Nicholas Horton, for always advocating for me and supporting me throughout the Amherst College experience. I also thank Professor Jun Ishii for teaching me Advanced Econometrics and Professor Amy Wagaman for teaching me Advanced Data Analysis, which helped me gain a clear and solid understanding of the foundational tools I could build on in this work.

Finally, I would like to express gratitude towards my family for their constant belief in my abilities. Special thanks to my mom, dad, sister, grandfather, and grandmother for making me capable of the opportunity to study abroad. Last but not least, I would like to thank my

friends, peers, and colleagues on campus, who took courses, worked, and played sports with me. This academic, personal, and professional growth journey would not be possible without their support.

Chapter 1

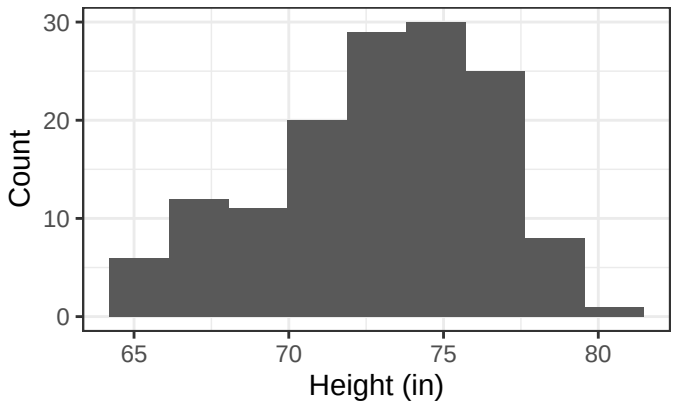
Introduction

xxx

Chapter 2

Unraveling Notion of Dependence through Copulas

Make it accessible by starting with a motivating example, then proceed towards more rigorousness gradually until challenges for discrete case are covered.



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

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

2.2 Chapter 6 Code

The following code was used to create Chapter 6.

2.2.1 Code within chapter

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

# Set default ggplot theme for document
theme_set(theme_classic())
# If using kableExtra tables, print blank cells instead of `NA`
options(knitr.kable.NA = "")

```

```r
# 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
```

```r
#
# 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))

```

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

2.2.2 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")
```

Chapter 3

Checkerboard Copula and Regression Association Measure

Start with motivation and connect with earlier chapters, then define concepts and weave in examples to solidify readers understanding.

3.1 Chapter 6 Code

The following code was used to create Chapter 6.

3.1.1 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")
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================

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

3.1.2 Code sourced from external scripts

```r
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================
```

```
8
9 print("Hello, Amherst!")
```

Chapter 4

Applications of Parallel Computing

Start with motivation, connect, and dive into details + examples of parallel computing along with use-cases.

4.1 Chapter 6 Code

The following code was used to create Chapter 6.

4.1.1 Code within chapter

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

# Set default ggplot theme for document
theme_set(theme_classic())
# If using kableExtra tables, print blank cells instead of `NA`
```

```r
options(knitr.kable.NA = "")

# Load data
load("data/temp_wnba.RData")
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================

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

4.1.2 Code sourced from external scripts

```r
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================

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

Chapter 5

Software (Package) Implementation and Testing

Start basic with motivation, and breakdown the implementation and testing phases properly while making sure that they are accessible.

5.1 Chapter 6 Code

The following code was used to create Chapter 6.

5.1.1 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")
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================

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

5.1.2 Code sourced from external scripts

```r
# ============================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ============================================================================
```

```
8
9 print("Hello, Amherst!")
```

Chapter 6

Conclusion

Like introduction, pull everything together and conclude the work!

6.1 Chapter 6 Code

The following code was used to create Chapter 6.

6.1.1 Code within chapter

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

# Set default ggplot theme for document
theme_set(theme_classic())
# If using kableExtra tables, print blank cells instead of `NA`
options(knitr.kable.NA = "")
```

```r

# Load data
load("data/temp_wnba.RData")
# ======================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ======================================================================

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

6.1.2 Code sourced from external scripts

```r
# ======================================================================
# Sample R script for thesis template
#
# Doesn't do anything useful
#
# Last updated: 2024/08/24
# ======================================================================

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

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*.

Wei, Z., and Kim, D. (2021), "On exploratory analytic method for multi-way contingency tables with an ordinal response variable and categorical explanatory variables," *Journal of Multivariate Analysis*, 186, 104793. https://doi.org/10.1016/j.jmva.2021.104793.

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. ...