Modelling sport results with extreme value methods

Csáfordi József András
Applied Mathematician MSc

Supervisor: Dr. Zempléni András

Contents

Previous semester

Data changes in brief

- Boston marathon
- Threshold method
- The data should be below the limit
- Each year separately
- It requires more data

Previous semester assignments

- Generalized Pareto Distribution (GPD)
- Algorithm
- Parameter table
- Diagnostic plots
- Analysis
- Estimation

In this semester, our goal is to fit a two-dimensional model to the data, from which we form a copula and thus try to estimate a best result.

Previous semester

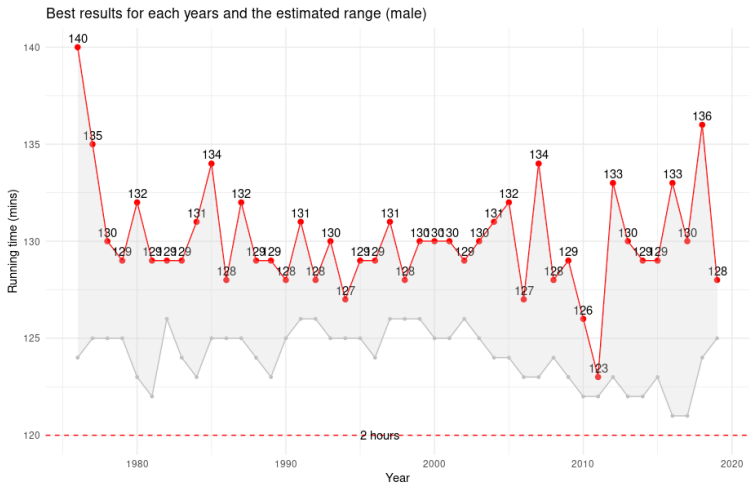


Figure: Estimated male

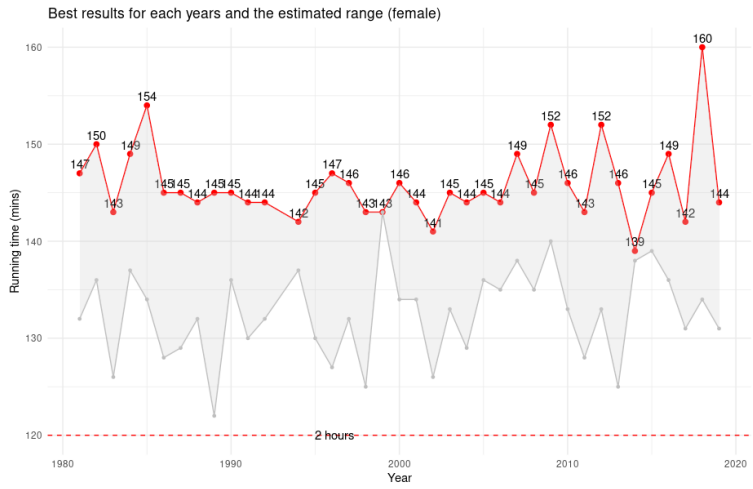
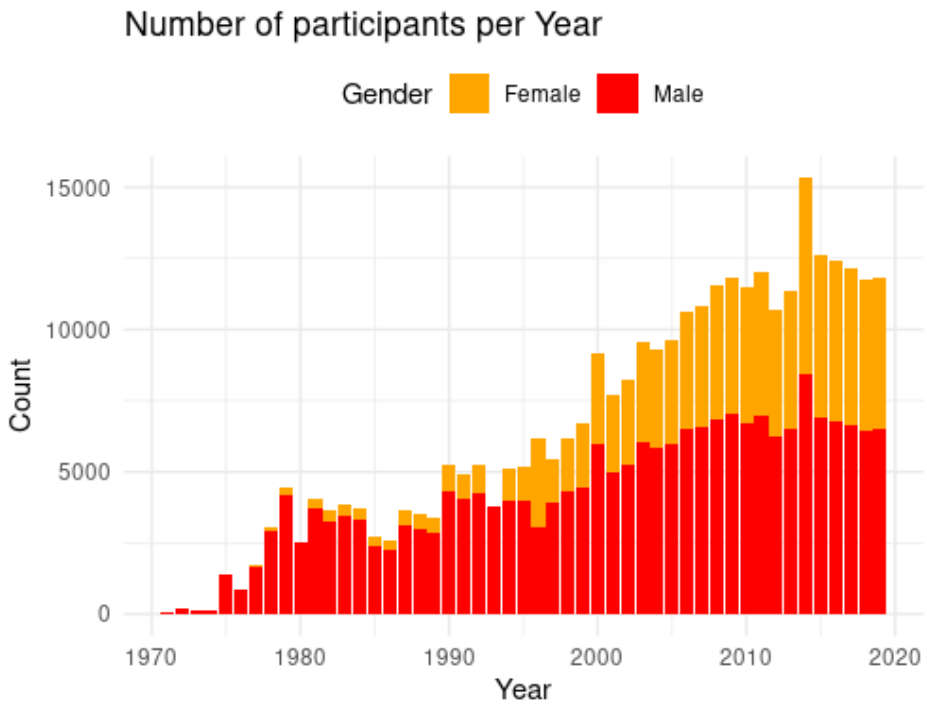


Figure: Estimated female

Briefly about the data

1. Grows almost exponentially from 1971 to 2010
2. Big drop from 2020 (due to covid)
3. The highest number of competitors was in 2014
4. Best results:
   - 7382 seconds for men
   - 8337 seconds for women
5. 13699 seconds on average



Data histogram

About the copula

1. Tools used to model dependency between variables
2. Let F be a multivariate distribution function, and $F_1, F_2, \ldots, F_d$ be the marginal distribution functions of the individual variables. The copula, denoted by C, is a function such that: $F(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$. Where $x_1, x_2, \ldots, x_d$ are the values of the variables.
3. Briefly about the copula families I used later in my work:
   - Gaussian:
   $$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$
   - Frank:
   $$C_\theta(u, v) = -\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right),$$

Sklar's Theorem

Sklar's Theorem states that for any multivariate cumulative distribution function (CDF) H with margins $F_1, F_2, \ldots, F_n$, there exists a copula C such that for all $x_1, x_2, \ldots, x_n$ in $\mathbb{R}$:

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)).$$

If the margins F_i are continuous, then the copula C is unique. Conversely, if C is a copula and $F_1, F_2, \ldots, F_n$ are univariate CDFs, then H defined by the above equation is a multivariate CDF with margins $F_1, F_2, \ldots, F_n$.

Two-dimensional pairing

1. Pairing the male and female competitors in each year
2. The best male with the best female, the second best female with the second best male...
3. Threshold for each gender
4. Fitting Pareto distribution
5. Anderson-Darling test
6. 9200 sec for the females and 8000 sec for the males, with a p-value of 0.04 for women

Two-dimensional pairing and it's normalized form plot

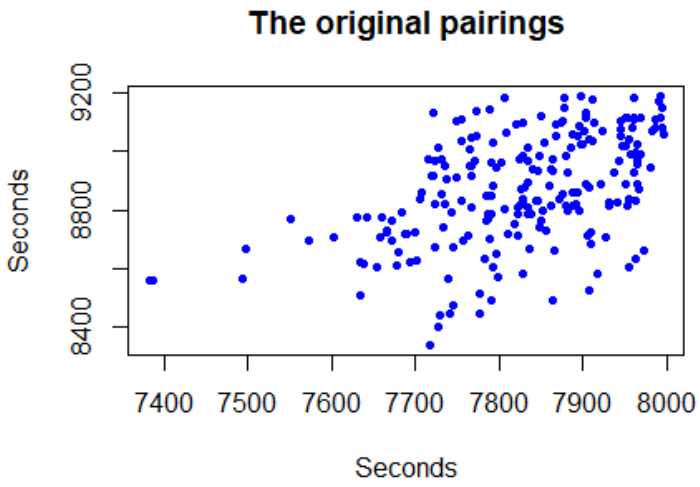


Figure: Pairing

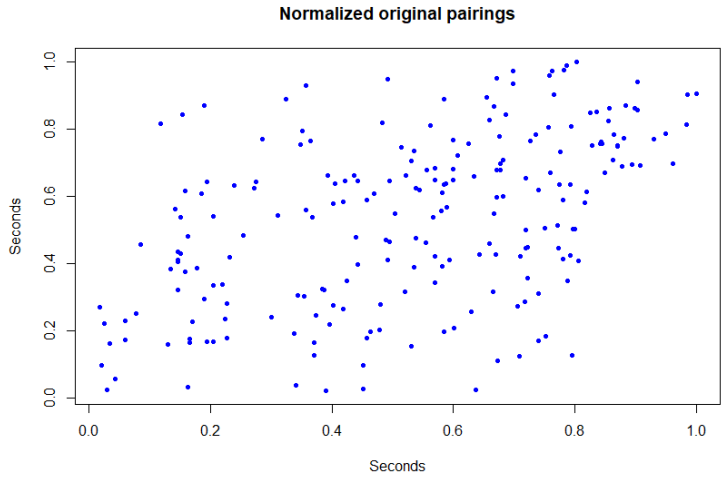


Figure: Normalized form

Copula model

- 281 final points
- Normalization
- BiCopSelect function for the copula model with the corresponding parameter and the Kendall-tau value. BiCopSelect: selects an appropriate bivariate copula family for given bivariate copula data using one of a range of methods. The corresponding parameter estimates are obtained by maximum likelihood estimation.
- Brief test using gofCopula: Goodness-of-fit tests for copulas based on the empirical process comparing the empirical copula with a parametric estimate of the copula derived under the null hypothesis.
- Gaussian and Frank copula
- Generate a copula pattern with the given parameters

Copula model - Gaussian and Frank copula

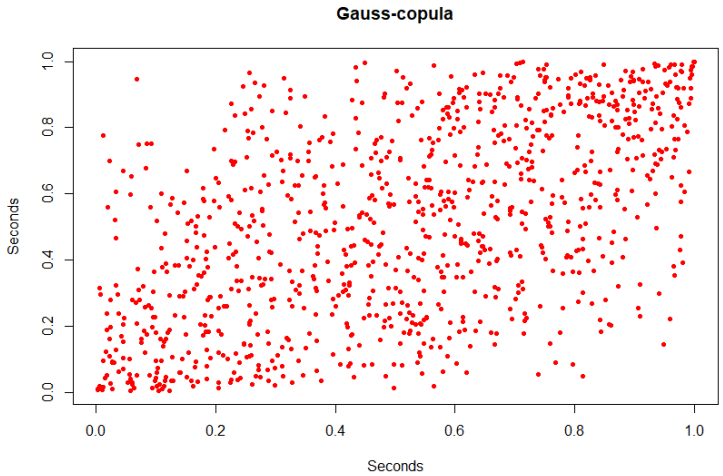


Figure: Gaussian copula

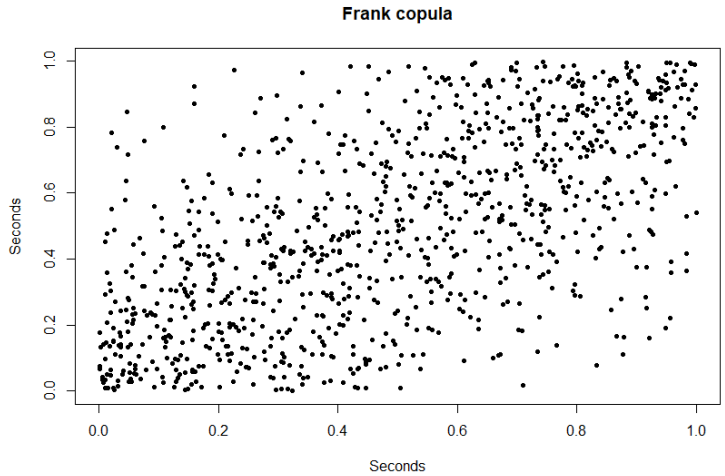
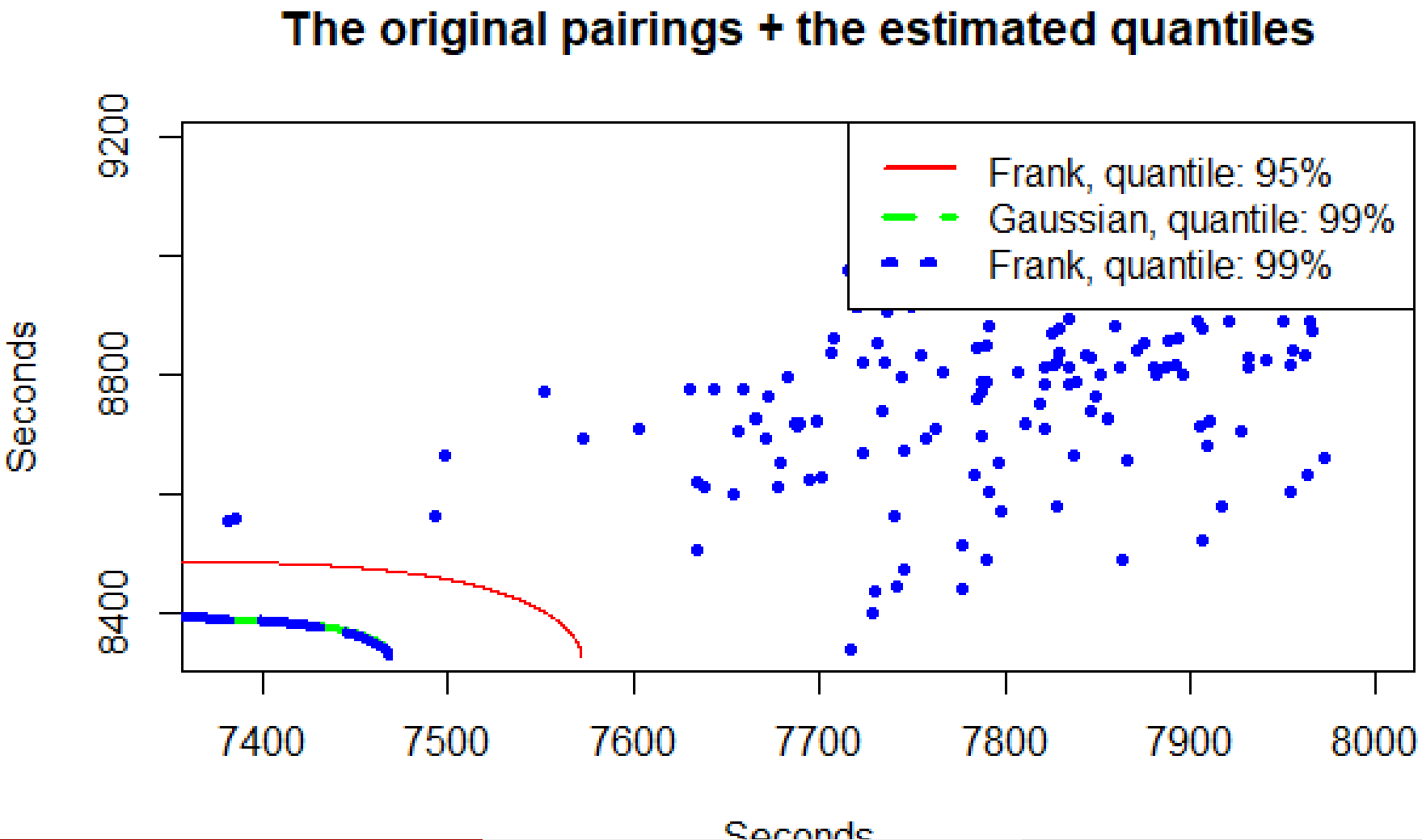


Figure: Frank copula

Copula model - back to the original scale

- Estimate based on a 100-year level
- No expected improvement in results in the near future
- Probability integral transform in each quantile for estimated line
- Transform back to the original scaling
- There is a slight improvement, but it is not considered significant compared to the 100-year level

The original pairings and the estimated quantiles



Conclusion

- Our project endeavors to predict future outcomes of the Boston Marathon
- Data challenges
- We applied the Pareto distribution however, the elusive sub-two-hour marathon time remains a distant milestone
- We constructed a two-dimensional model
- Copula analysis provided insights into gender performance dependence
- There are still tasks left: quantifying the uncertainty of the estimate, examining the dependence on the thresholds

Thank you for your attention!