NON-NESTED HYPOTHESIS TESTS FOR VINE COPULAS AND STATISTICAL LEARNING TECHNIQUES IN PROCESS MONITORING

By

ZIYI CHEN

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Mathematics and Statistics

MAY 2019

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of ZIYI CHEN find it satisfactory and recommend that it be accepted.

Francis Pascual, Ph.D., Chair

Haijun Li, Ph.D.

Marc Evans, Ph.D.

ACKNOWLEDGMENTS

This dissertation would not have been possible without tremendous support from my colleagues, friends, and family. First and foremost, I am deeply indebted to express my sincere gratitude to my advisor Dr. Francis (Jave) Pascual. Jave has enlightened me with his knowledge and provided generous support in my five-year doctoral study. He has always been patient with me and walked me through this dissertation from the proposal to this final dissertation. Without him, I would not have survived the five-year doctoral study. I am also indebted to thank Dr. Marc Evans and Dr. Haijun Li, both of whom are on my Ph.D. committee, for providing invaluable suggestions and insightful comments on my research.

I would like to say thanks to my colleagues who have helped me with my career path. Dr. Nairanjana (Jan) Dasgupta has provided me with advice on professional development and opportunities to work at Axio research and Center for Interdisciplinary Statistical Eduation and Research.

I am very thankful to Dr. Mei Han, Dr. Jun Qin, Dr. Iek-Heng Chu, and Xiangfeng Rui, who have offered me great opportunities to work with them as an intern. As their intern, I have learned how to model large scale financial data. I also appreciate their suggestions about preparing for a career path.

I owe deeply to my parents, who have encouraged and supported me for all my pursuits. Even though they cannot fully grasp what I am studying in college, they have always backed me up. Their weekly checking-in phone calls have brought great warmth to me as an ordinary person besides a Ph.D. student. Without them supporting me, I doubt that I would be the way I am.

I could not imagine my life as a doctoral student without all the relationships and friendships with people I have crossed paths at WSU. These people are Yanni, Yilong, Wenlu, Jillian, Wei, Jessie and Debasmita. I appreciate Qin for her love, understanding, support, and company with me through some of the tough time.

Last but not least, I would like to thank WSU for giving me a chance to pursue my doctoral degree with sufficient fundings and support. My doctoral study is not possible without the support from WSU.

NON-NESTED HYPOTHESIS TESTS FOR VINE COPULAS AND STATISTICAL LEARNING TECHNIQUES IN PROCESS MONITORING

Abstract

by Ziyi Chen, Ph.D.
Washington State University
May 2019

Chair: Francis Pascual

In this dissertation, we first introduce three bootstrap-based non-nested hypothesis tests for regular vine-copulas. These test statistics are derived from log-likelihood ratio test statistics and Cox test statistics. This study presents the power study comparing the proposed tests with existing vine-copula non-nested hypothesis tests. Across models with varying structures of regular copulas, our hypothesis tests consistently achieve higher power. Deriving from statistical algorithms, we also propose two different control charts that can be applied to multivariate statistical process control (MSPC). The first one is based on support vector data description (SVDD). We propose a SVDD control chart using the Mahalanobis distance kernel (Mahalanobis k-chart). Mathematical illustrations and statistical comparisons are presented on the

basis of both simulations and a real example of electricity consumption. The results show that Mahalanobis k-chart can achieve lower Phase II average run length (ARL) in most shifted-process scenarios. Another control chart we propose is random oversampling gradient boosting real-time contrasts (ROGB-RTC) chart. Real-time contrasts (RTC) control charts convert the statistical process monitoring problem into a dynamic binary classification problem. But only a limited number of RTC studies have scrutinized the imbalance problem between the sample size of the reference data and that of the sliding window data. Our control chart handles the imbalance problem in terms of both classifier and monitoring statistics. Experiments show that the proposed method achieves better performance than that of the original real-time contrasts chart.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

ABSTRACT v

LIST OF TABLES ix

LIST OF FIGURES xi

1. Non-nested hypothesis test for Vine Copulas 1
   - 1.1 Introduction 1
   - 1.2 Copula and Regular Vine Copula 4
   - 1.3 Cox Test Statistic and Bootstrap Procedure 10
   - 1.4 Numerical study 17
   - 1.5 Conclusion 24

2. An Improved K-chart based on Support Vector Data Description 26
   - 2.1 Introduction 26
   - 2.2 Support vector methods 29
   - 2.3 K-chart 33
   - 2.4 Simulation study 38
   - 2.5 Application 49
   - 2.6 Conclusion 56

3. A Real-time Contrasts Control Chart with Random Oversampling Gradient Boosting 58
   - 3.1 Introduction 59
   - 3.2 Real-Time Contrasts 60
   - 3.3 Methods 65

3.4 Simulation Experiments 76
3.5 Conclusion 83
APPENDIX 84
BIBLIOGRAPHY 86

LIST OF TABLES

Table Page

1.1 Approximated power for bootstrap based tests, the Vuong's and the Clarke's test with sample size $n = 200$ and 500 from different copula models. 21

1.2 Approximated power for bootstrap based tests, the Vuong's and Clarke's test with sample size $n = 200$ and 500 from different tree structure copula models. 22

2.1 ARL values when β_1 shifts to $\beta_1+\delta\sigma_{\beta_1}$ in the linear profile monitoring experiment. 41

2.2 ARL performance of Gaussian k-chart and Mahalanobis k-chart with 3-dimensional t distribution data under shifts in μ_1 and μ_3. 45

2.3 ARL performance of Gaussian k-chart and Mahalanobis k-chart for 3-dimensional t-distribution data with shifts in σ_1^2 and σ_3^2. 47

2.4 ARL values when average hourly usage is increased by from $\boldsymbol{\mu}$ to $\boldsymbol{\mu} + \delta \times \boldsymbol{\sigma}$. 53

2.5 ARL values fixed hourly usage reading is a random error around a constant value. 56

3.1 ARL performance of hybrid ROGB-RTC and RF-RTC with multivariate normal distribution when $N_0 = 1000$ and $N_w = 10$. 78

3.2 ARL performance of oversampling ROGB-RTC with multivariate normal distribution when $N_0 = 1000$ and $N_w = 10$ 80

3.3 ARL performance of oversampling ROGB-RTC with a bivariate gamma distribution when $N_0 = 1000$ and $N_w = 10$ 82

0.1 Vine copula structures and paired-copula families of 3 Vine-copula models, the Kendall's τ values of R-vine and the estimated Kendall's τ values of C-vine and D-vine in Experiment 1. For paired-copula family types, N is Normal, G is Glayton, G is Gunmbel, F is Frank, G90 and G180 are 90 and 180 degrees of rotated Gumbel. 84

0.2 Vine copula structures and paired-copula families of 3 R-vine models, the Kendall's τ values of null R-vine model and the estimated Kendall's τ values of two alternative R-vine models in Experiment 2. For paired-copula family types, N is Normal, G is Glayton, G is Gunmbel. 85

LIST OF FIGURES

Figure Page

1.1 The tree structure of a 5 dimensional R-vine copula 8

1.2 The first tree structures of C-vine and D-vine Copulas 9

1.3 The first tree structure of selected 5-dimensional R-Vine (left), C-Vine (center) and D-Vine (right) in experiment 1. Their corresponding pair-copulas are labeled on the edges. 18

2.1 Support vector classifier with the maximal margin. 30

2.2 SVDD boundaries for different pairs of parameters (σ, C). 34

2.3 The ARL performance of three control charts when β_1 shifts to $\beta_1 + \delta\sigma_{\beta_1}$ in the linear profile monitoring experiment. 42

2.4 Phase II ARLs of Gaussian k-chart and Mahalanobis k-charts for t distribution data. Top figure shows the ARLs when (μ_1, μ_2, μ_3) is shifted up to $(\mu_1 + 2\sigma_1, \mu_2, \mu_3)$. Bottom figure are the ARLs of shifts from $(\mu_1 + 2\sigma_1, \mu_2, \mu 3)$ to $(\mu_1 + 2\sigma_1, \mu_2, \mu_3 + 2\sigma_3)$. 46

2.5 Phase II ARLs of Gaussian k-chart and Mahalanobis k-chart for data under t distribution. Top figure shows the ARLs when the variance σ_1^2 is shifted up to $2\sigma_1^2$. Bottom figure are the ARLs when variances are shifted from $(2\sigma_1^2, \sigma_2^2, \sigma_3^2)$ to $(2\sigma_1^2, \sigma_2^2, 2\sigma_3^2)$. 48

2.6 Electricity consumption every 15 min of client MT_158 in kW from 2011-02-01 to 2011-02-15. 50

2.7 The ARL performance of Mahalanobis k-chart and Gaussian k-chart when the mean of hourly usage is shifted from $\boldsymbol{\mu}$ to $\boldsymbol{\mu} + \delta \times \boldsymbol{\sigma}$. 54

2.8 Mahalanobis k-chart monitoring of 300 power usage samples with one standard deviation shift. 55

3.1 The framework of Real-Time Contrast. 62

3.2 An example of ROSE. 200 reference data come from a bivariate normal distribution with mean $(0, 0)$ and covariance $\boldsymbol{I}_2$. 10 window data points come from a bivariate normal distribution with mean $(0.2, 0.2)$ and covariance $\boldsymbol{I}_2$. The top plot is the scatter plot of raw data. The bottom one is the plot after using ROSE to resample 200 window data. 69

3.3 An example of a regression tree. 72

CHAPTER 1. NON-NESTED HYPOTHESIS TEST FOR VINE COPULAS

Abstract

We introduce three bootstrap-based non-nested hypothesis tests for regular vine-copulas. These test statistics are derived from log-likelihood ratio test statistics and Cox test statistics. This study presents a power study that compares the proposed tests with existing vine-copula non-nested hypothesis tests. Across models with varying structures of regular copulas, our hypothesis test procedures consistently achieve higher power.

1.1 Introduction

A copula is a multivariate distribution with a uniform marginal probability distribution. It is useful to model the dependence structure among variables since it allows to model dependencies and marginal distributions separately. Copula based models are widely used in risk management and option pricing. Model selection on copula-based models is an essential part of empirical work. However, there is a limited number of studies working on non-nested hypothesis testing on copulas, vine copulas in particular.

Suppose there are two models, say f and g, which may or may not be different from each other. If one model f cannot be derived from the other by parametric restriction or process limiting. then these two models are called "non-nested". To compare two non-nested models, one approach is to conduct a likelihood ratio test which tests the null hypothesis H_0 (model f) is significantly better than the alternative hypothesis H_1 (model g). For non-neted models, the log-likelihood ratio statistic is $\lambda = l_f - l_g$, where l_f and l_f are the log-likelihood under null and alternative models. Nevertheless, the non-nested condition violates the asymptotical Chi-square distribution when doubling log-likelihood difference. In this scenario, one can still conduct a hypothsis test based on the asymptotic distribution of test statistics. In a seminal study, Cox [15] develops an adjusted log-likelihood ratio statistic for non-nested test. In line with Cox's work, some researchers use simulation approaches to compute which still assumes an asymptotic distribution (Pesaran and Pesaran [32], [62], and Weeks [80]). Moreover, when various asymptotic ways are used to estimate the variance of Cox test statistic, they give rise to variant of Cox test statistics. In contrast to previous approaches based on Cox test statistic, Vuong [78] proposes a test based on the Kullback-Leibler information criterion $KLIC = E_0(l_f - l_g)$, where E_0 is the expected value under the true model. He proves under the regular conditions, the likelihood ratio statistic with maximum likelihood estimators of parameters is asymptotically normally distributed. Following Vuong's work, Clarke [11] proposes a distribution-free paired sign test with an asymptotic Binomial distribution under the null hypothesis.

Apart from the above approaches to find an asymptotic distribtuion, another manner is to use bootstrap (Efron[20]). Beran [3] uses a parametric bootstrap procedure to approximate the distribution of test statistics. In this case, the significance of the test is evaluated by the observed test statistic and the empirical reference distribution.

Some researchers have also combined a variety procedures to imporve the non-nested hypothesis. For example, Kapetanios and Weeks [44] combine both simulation and bootstrap procedures in their non-nested hypothesis test to adjust both the likelihood ratio and Cox test statistics. Comparing several bootstrap procedures with test statistics on asymptotic approximation, they conclude that a simple bootstrap procedure has a significantly better performance than that of complicated bootstrap procedures and asymptotic procedures.

This paper is outlined as follows. Section 1.2 introduces copula and vine copula models. Section 1.3 describes the Cox test statistic and the bootstrap procedure. In Section 1.3, we perform a simulation study with various 5-dimensional vine copulas and compared the results of different tests in terms of power. Finally, Section 1.5 summarizes the results and discusses some potential future work.

1.2 Copula and Regular Vine Copula

1.2.1 Definitions

Definition 1.1. *A d-dimentional copula, $C : [0,1]^d :\to [0,1]$ is a cumulative distribution function (CDF) with uniform marginals.*

We use the notation $C(\boldsymbol{u}) = C(u_1, u_2, \ldots, u_d)$ to denote a copula. The distribution condition on C has the following properties:

- $C(u_1, u_2, \ldots, u_d)$ is always non-decreasing in each component u_i.
- Since C is uniformly distributed, the i^{th} marginal distribution is obtained by setting $u_j = 1$ for all $j \neq i$,

$$C(1, \ldots, 1, u_i, 1 \ldots, 1) = u_i.$$

- For $a_i \leq b_i$, $P(U_1 \in [a_1, b_1], \ldots U_d \in [a_d, b_d])$ must be non-negative. This implies the rectangle inequality

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(u_{1,i_1}, \ldots, u_{d,i_d}) \geq 0,$$

where $u_{j,1} = a_j$ and $u_{j,2} = b_j$, for $j = 1, \ldots, d$

Any function that satisfies those 3 properties above is a copula.

1.2.2 Sklar's theorem

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a multivariate random vector with joint cdf F_X with continuous and increasing marginals. It can be easily verified that the joint distribution of $F_1(X_1), \ldots, F_d(X_d)$ is a copula. With this result, Sklar's theorem provides a theoretical foundation for obatining multivariate distributions with a copula and marginal distributions.

Theorem 1.1 (Sklar's Theorem)**.** *Consider a d-dimenstional cdf F with marginal cdf's $F_1, \ldots, F_d$. Then there exists a copula C such that*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{1.1}$$

for all $x_i \in (-\infty, \infty)$ and $i = 1, \ldots, d$.

If F_i is continuous for all $i = 1, \ldots, d$, then C is unique. Otherwise C is uniquely determined on $Ran(F_1) \times \cdots \times Ran(F_d)$, where $Ran(F_i)$ is the range of cdf F_i.

Consider a copula C and univariate cdf's $F_1, \ldots, F_d$. Then F as defined in (1.1) is a multivariate cdf with marginal cdf's $F_1, \ldots F_d$

If all marginal cdf's $F_1, \ldots, F_n$ are continuous, it can be obtained that

$$F_i(F_i^{\leftarrow}(y)) = y.$$

where $F^{\leftarrow}(x) := inf\{v : f(v) \geq x\}$. Apply above equation to (1.1), we obtain

$$C(\boldsymbol{u}) = F(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)).$$

1.2.3 Copula density and conditional distribution

The copula density can be obtained as below:

$$c(\boldsymbol{u}) := \frac{\partial^d C(\mu_1, \dots, \mu_d)}{\partial \mu_1 \dots \partial \mu_d},$$

if it exists. When the marginal cdf F_i is differentiable, we have $F_i^{\leftarrow} = F_i^{-1}$. If the copula is in the form of (1.2) and has a density, the copula density can be computed as

$$c(\boldsymbol{u}) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \dots f_d(F_d^{-1}(u_d))},$$

where f is the joint density and $f_i,\ i = 1, \dots, d$, are the marginal densities. Assuming sufficient regularity, the copula conditional cdf can be obtained by

$$P(U_2 \leq u_2 | U_1 = u_1) = \frac{\partial}{\partial u_1} C(u_1, u_2)$$

Thus, the conditional cdf of a copula can be directly derived from itself.

1.2.4 Pair-copula constructions (PCC)

Copula models are powerful tools for modeling the dependence structure. However, traditional multivariate copulas such as Archimedean copulas cannot model different dependencies between pairs of variables in high-dimension. A flexible way to construct high-dimentional copulas is to construct multivariate densities with $d(d-1)/2$ bivariate copulas (Joe [41] , Bedford and Cooke [1],[2]).

Let $X_1, \ldots, X_d$ be a random variables with a joint distribution F and a density f. We can decompose the joint density f by

$$f(x_1, \ldots, x_d) = f(x_d|x_1, \ldots, x_{d-1})f(x_1, \ldots, x_{d-1}) \tag{1.2}$$

$$= \prod_{i=2}^{d} f(x_i|x_1, \ldots, x_{i-1}) \times f(x_1) \tag{1.3}$$

where $F(\cdot|\cdot)$ and $f(\cdot|\cdot)$ are the conditional cdf and the density function, respectively. If we apply Sklar's theorem with $d = 2$, we can derive

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2) \tag{1.4}$$

where c_{12} is a bivariate copula density. With (1.4), the conditional density of X_2 given X_1 is computed as

$$f(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \tag{1.5}$$

Using Equation (1.3), (1.4) and (1.5), the pair-copula construction in d dimensions is listed as follows:

$$f(x_1, \ldots, x_d) = \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,(i+j)|(i+1),\ldots,(i+j-1)} \cdot \prod_{k=1}^{d} f_k(x_k) \tag{1.6}$$

where $c_{i,j|i_1,\ldots,i_k} := c_{i,j|i_1,\ldots,i_k}(F(x_i|x_{i_1}, \ldots, x_{i_k}), F(x_j|x_{i_1}, \ldots, x_{i_k}))$, for $i, j, i_1, \ldots, i_k = 1, \ldots, d$ with $i < j$ and $i_1 < \ldots < i_k$.

1.2.5 Regular vine copula

Bedford and Cooke[1],[2] use trees to represent the graphic structure of multivariate copulas. Denote trees by $T_i = (V_i, E_i), i = 1, \ldots, d-1$, where V_i is the set of nodes and E_i denotes the set of edges. A vine is a sequence of trees $\nu = (T_1, \ldots, T_{d-1})$ that satisfies the following conditions:

- T_i is a tree with nodes $V_1 = 1, \ldots, d$ and edges E_1.
- For $i \geq 2$, T_i is a tree with nodes $V_i = E_{i-1}$ and edges E_i.
- Two nodes in the tree T_{i+1} are joined by an edge only if their corresponding edges in T_i share a common node.

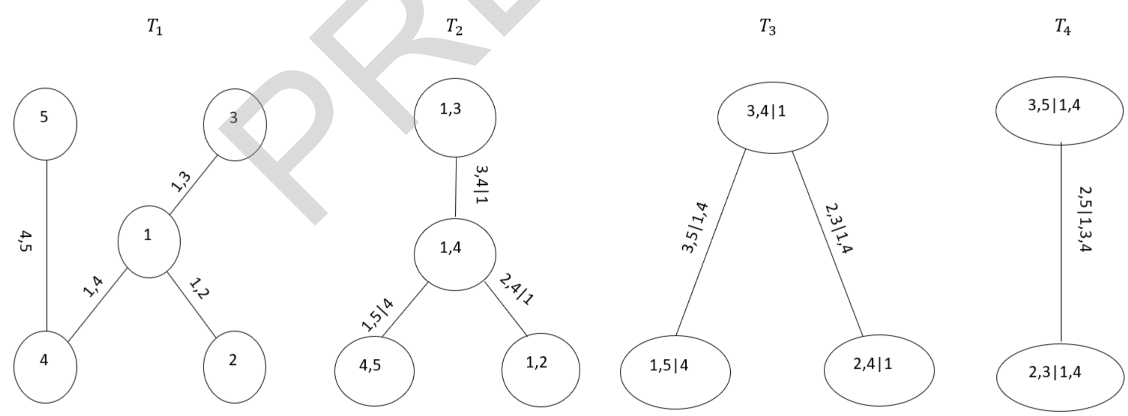

Figure 1.1: The tree structure of a 5 dimensional R-vine copula

Figure 1.1 is an example of a 5-dimensional R-vine. Tree T_1 has nodes $V_1 = 1, \ldots, 5$

and edges $E_1 = \{\{1,2\},\{1,3\},\{1,4\},\{4,5\}\}$ with unconditional pair-copulas. For trees $T_2,\ldots,T_4$, their edges $E_2,\ldots,E_4$ form conditional pair-copulas. The joint copula density of the 5-dimensional R-vine copula in Figure 1.1 can be decomposed as

$$
\begin{aligned}
&c_{12345}(u_1,\ldots,u_5)\\
&\quad = c_{1,2}(u_1,u_2)\cdot c_{1,3}(u_1,u_3)\cdot c_{1,4}(u_1,u_4)\cdot c_{4,5}(u_4,u_5)\\
&\qquad \cdot c_{1,5:4}\left(C_{1|4}(u_1|u_4), C_{5|4}(u_5|u_4)\right)\cdot c_{2,4;1}(C_{2|1}(u_2|u_1), C_{4|1}(u_4|u_1))\\
&\qquad \cdot c_{3,4;1}\left(C_{3|1}(u_3|u_1), C_{4|1}(u_4|u_1)\right)\\
&\qquad \cdot c_{2,3;1,4}\left(C_{2|1,4}(C_{2|1}(u_2|u_1), C_{4|1}(u_4|u_1)), C_{3|1,4}(C_{3|1}(u_3|u_1), C_{4|1}(u_4|u_1))\right)\\
&\qquad \cdot c_{3,5;1,4}\left(C_{3|1,4}(C_{3|1}(u_2|u_1), C_{4|1}(u_4|u_1)), C_{5|1,4}(C_{5|1}(u_5|u_1), C_{4|1}(u_4|u_1))\right)\\
&\qquad \cdot c_{2,5;1,3,4}\left(C_{2|1,3,4}(C_{2|1,4}(C_{2|1}(u_2|u_1), C_{4|1}(u_4|u_1)), C_{3|1,4}(C_{3|1}(u_3|u_1), C_{4|1}(u_4|u_1))),\right.\\
&\qquad \left. C_{5|1,3,4}(C_{3|1,4}(C_{3|1}(u_3,u_1), C_{4|1}(u_4|u_1)), C_{5|1,4}(C_{1|4}(u_1|u_4), C_{5|4}(u_5|u_4)))\right).
\end{aligned}
$$

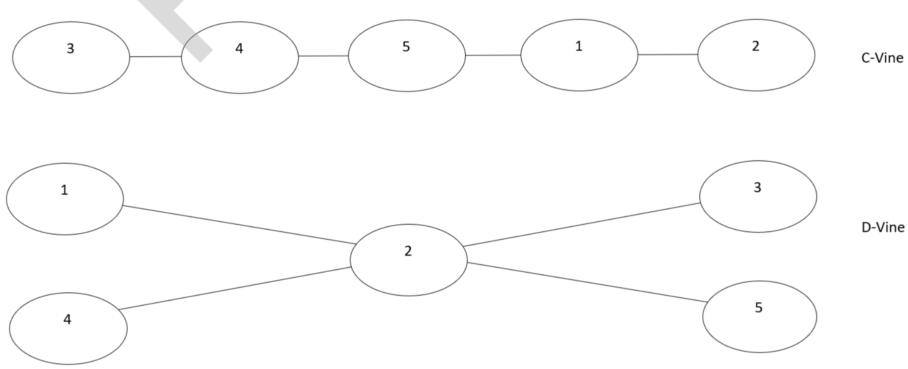

Figure 1.2: The first tree structures of C-vine and D-vine Copulas

Figure 1.2 is an example of two special R-vine tree structures, namely, C-vine and D-vine. C-vine has one root node with degree $d-1$, that connects to another $d-1$ nodes. All other nodes only have degree 1. In D-vine, two nodes have degree 1 and others have degree 2.

1.3 Cox Test Statistic and Bootstrap Procedure

1.3.1 *Cox Test statistic*

Consider two models, $F_\theta = f(\boldsymbol{y}_i|\boldsymbol{x}_i,\theta)$ and $G_\lambda = g(\boldsymbol{y}_i|\boldsymbol{x}_i,\lambda)$, where θ and λ are two vectors of unknown parameters belonging to two respective parameter spaces Θ and Λ. We write the null and alternative hypotheses of a test as

$$\boldsymbol{H}_0 : \quad F = F_\theta$$

$$\boldsymbol{H}_a : \quad F = G_\lambda$$

Here F is the true population distribution. To simplify notations, we use $f_i(\theta)$ and $g_i(\lambda)$ to represent $f(\boldsymbol{y}_i|\boldsymbol{x}_i,\theta)$ and $g(\boldsymbol{y}_i|\boldsymbol{x}_i,\lambda)$. The Cox test statistic is obtained by

$$T_f = l_f(\widehat{\theta}) - l_g(\widehat{\lambda}) - C_{fg}(\widehat{\theta},\tilde{\lambda})$$

where $l_f(\widehat{\theta})$ and $l_g(\widehat{\lambda})$ are the average log-likelihood functions

$$l_f(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log f_i(\widehat{\theta}), \quad l_g(\widehat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\log g_i(\widehat{\lambda}),$$