On the Mixture Gaussian Copula to study the suitability of Diagnostic Tests

A. Nanthakumar

Department of Mathematics, SUNY-Oswego, Oswego, NY 13126
Corresponding Author: ampala.nanthakumar@oswego.edu

Received: 22 April 2013 / Revised: 19 July 2013 / Published: 15 November 2013

ABSTRACT

In this article, we develop a new copula "Mixture Gaussian Copula" to study the suitability of the diagnostic tests in the same manner as the ROC curves are used in similar studies.

Keywords: Markov model, Copula model, ROC curve

1 Introduction

For some time now, the Markov models and the Hidden Markov models have been in use in speech recognition, meteorology, biometry and many other fields. In biometry, especially in the context of epidemiological models and DNA sequencing, and micro-array analysis, the Hidden Markov model is commonly in use (see Ibe (2009) for the details and the literature). On the other hand, the copulas are used as a general way of formulating a multivariate distribution in such a way that the dependence can be infused in a reasonable manner. This is based on a simple idea that the joint distribution can be represented as a transformation of the underlying marginal distributions (see Sklar 1959). There are several copulas and each differ according to the strength of the dependence and the direction of the association. The Copula models fall either under the family of Archimedean Copulas or the non-Archimedean Copulas. The Gaussian Copulas belong to the non-Archimedean family of Copulas. Here in this paper, we consider the Gaussian Copula (to be more specific, the mixture of Gaussian Copulas). For the literature review, the interested readers are referred to Nelson (2006).

In the recent past, Krazanowski and Hand (2009), Pepe (2003), Zhou et al (2002) and others have investigated the use of Receiver Operating Characteristic Curve

(ROC curve) in the context of screening and diagnostic testing in the medical field. Pundir (2011), Gonen (2007), Shultz (1995) used this ROC curve to study the effectiveness of a single and multiple variable based medical diagnostic test. This paper discusses the use of both the Markov model and the Copula model to structure a medical diagnostic test while using the ROC curve in order to compare the effectiveness of two medical diagnostic tests. Furthermore, this is a computationally less intensive method while maintaining a fair degree of precision in diagnosis.

We divide the paper into many sections. First, we present the methodology based on the Markov chain to model the (disease, no disease) states. Next, we use the Copulas especially the mixture Gaussian Copula to model the distribution of the responses that we gather from the sample. The ROC curve based analysis of the responses and the robustness of the estimates is presented in the last section.

2 Modeling

2.1 Methodology

Here, the main objective is to study the suitability of a diagnostic test to classify a person as either healthy (H) or disease stricken (D). Suppose that at time 1, $100.(1-a)$ % of a population is healthy and the others are stricken by a disease. Due to fear that this disease may spread to the healthy population, every person in the population is advised to undergo a treatment for this disease. In order to ascertain the effectiveness of this treatment, we take two measurements one before the treatment and the other after the treatment from the same group of people. Let us say that the pre-treatment and the post-treatment measurements were taken at time 1 and 2 respectively and the transition from time 1 to time 2 occurs according to a Markov Chain with the transition probability matrix given by

$$\begin{pmatrix} b & 1-b \\ c & 1-c \end{pmatrix}$$

where, $0 \le b \le 1$ and $0 \le c \le 1$.

So, the percentage of healthy people at time 2 is $100.((b.(1-a)+c.a))$. Furthermore, assume that at time 1, the measurements from the healthy group follow a normal distribution $N\left(\mu_{11}, {\sigma_{11}}^2\right)$ and the measurements from the disease group follow a $N\left(\mu_{12}, {\sigma_{12}}^2\right)$. Similarly, at time 2, the measurements from the

healthy group follow a $N\left(\mu_{21}, {\sigma_{21}}^2\right)$ and the measurements from the disease group follow a $N\left(\mu_{22}, {\sigma_{22}}^2\right)$.

Hence, the measurement corresponding to time 1, that is S_1 follows a mixture normal distribution

$$(1-a).N\left(\mu_{11}, {\sigma_{11}}^2\right) + a.N\left(\mu_{12}, {\sigma_{12}}^2\right) \tag{1}$$

and similarly the measurements corresponding to time 2, S_2 follows a mixture normal distribution

$$(b.(1-a)+c.a).N\left(\mu_{21}, {\sigma_{21}}^2\right) + (1-b.(1-a)-c.a).N\left(\mu_{22}, {\sigma_{22}}^2\right) \tag{2}$$

Note that the measurements S_1 and S_2 are dependent and we propose a new copula "Mixture Gaussian Copula" to model the correlation structure. Generally speaking, Copulas are used for modeling the joint distributions from the marginal distributions.

2.2 Copula Modeling

In this section, we present the definition pertaining to the copulas and the formula for the Gaussian Copula.

Definition: A copula is a multivariate joint distribution defined on the k dimensional unit cube $[0,1]^k$ such that every marginal distribution is uniform on the interval $[0,1]$.

In other words, $C:[0,1]^k \to [0,1]$ is a $k-$dimensional copula if

a). $C(u) = 0$ whenever $u \in [0,1]^k$ has at least one component equal to 0.

b). $C(u) = u_i$ whenever $u \in [0,1]^k$ has all the components equal to 1 except the i^{th} one which is equal to u_i.

c). $C(u)$ is $k-$increasing.

Gaussian Copula:

$$C(u_1, u_2) = \Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right) \tag{3}$$

where $\Phi_2(x, y)$ represents the bivariate normal cumulative distribution function and $\Phi^{-1}(x)$ represents the functional inverse of the cumulative standard normal distribution function.

2.2.1 Mixture Gaussian Copula:

We define the mixture Gaussian Copula density for the four component normal mixture as follows.

$$\phi(s_1, s_2) = \frac{p_1}{2.\pi.\sigma_{11}.\sigma_{21}.\sqrt{1-\rho^2}}.e^{-\frac{1}{2.(1-\rho^2)}.\left\{\left(\frac{s_1-\mu_{11}}{\sigma_{11}}\right)^2+\left(\frac{s_2-\mu_{21}}{\sigma_{21}}\right)^2-2.\rho.\left(\frac{s_1-\mu_{11}}{\sigma_{11}}\right).\left(\frac{s_2-\mu_{21}}{\sigma_{21}}\right)\right\}}$$

$$+ \frac{p_2}{2.\pi.\sigma_{12}.\sigma_{22}.\sqrt{1-\rho^2}}.e^{\frac{-1}{2.(1-\rho^2)}.\left\{\left(\frac{s_1-\mu_{12}}{\sigma_{12}}\right)^2+\left(\frac{s_2-\mu_{22}}{\sigma_{22}}\right)^2-2.\rho.\left(\frac{s_1-\mu_{12}}{\sigma_{12}}\right).\left(\frac{s_2-\mu_{22}}{\sigma_{22}}\right)\right\}} \quad (4)$$

$$+ \frac{p_3}{2.\pi.\sigma_{11}.\sigma_{22}}.e^{\frac{-1}{2}.\left\{\left(\frac{s_1-\mu_{11}}{\sigma_{11}}\right)^2+\left(\frac{s_2-\mu_{22}}{\sigma_{22}}\right)^2\right\}} + \frac{p_4}{2.\pi.\sigma_{12}.\sigma_{21}}.e^{-\frac{1}{2}.\left\{\left(\frac{s_1-\mu_{12}}{\sigma_{12}}\right)^2+\left(\frac{s_2-\mu_{21}}{\sigma_{21}}\right)^2\right\}}$$

where, $0 \le p_i \le 1$ for $i = 1,....,4$ and $\sum_{i=1}^{4} p_i = 1.$

2.3 ROC Concept and its application

ROC Curves are used for classification in the context of assessing the performance of a diagnostic test. It is of immense help in the area of clinical diagnosis. Also, the ROC Curves can be used for comparing two or more diagnostic tests. It is a graph of the true positive rate against the false positive rate in medical diagnosis.

The area under the ROC Curve (AUROC) is a measure of accuracy of the diagnostic test. In probabilistic terms, this is the probability for the true positive rate to exceed the false positive rate in any diagnostic test. This technique is found to be very effective in evaluating the performance of the diagnostic test.

In our case, the situation is different. However, the objective remains the same. We evaluate the performance of the test based on the probability that $S_2 < S_1$, where

S_1 and S_2 are the health related measurements at time 1 and 2 respectively. We assume that the joint distribution can be modeled as a mixture Gaussian Copula.

Furthermore, the Copulas are supposed to yield the marginal distributions when the data is collapsed. Therefore, by equating the marginal distributions, we have

$$p_1 + p_3 = 1 - a \tag{5}$$

$$p_1 + p_4 = b.(1-a) + c.a \tag{6}$$

$$p_1 + p_2 + p_3 + p_4 = 1 \tag{7}$$

As we can see, there are three equations in four variables indicating that there could be multiple solutions for the mixture Gaussian Copula modeling proportions.

Note that based on this Copula density,

$$\begin{aligned} P(S_2 < S_1) = {} & p_1.P(S_2 - S_1 < 0) + p_2.P(S_2 - S_1 < 0) + \\ & p_3.P(S_2 - S_1 < 0) + p_4.P(S_2 - S_1 < 0) \\ = {} & p_1.P\left(Z \le \frac{\mu_{11} - \mu_{21}}{\sqrt{{\sigma_{11}}^2 + {\sigma_{21}}^2 + 2.\rho.\sigma_{11}.\sigma_{21}}}\right) \\ & + p_2.P\left(Z \le \frac{\mu_{12} - \mu_{22}}{\sqrt{{\sigma_{12}}^2 + {\sigma_{22}}^2 + 2.\rho.\sigma_{12}.\sigma_{22}}}\right) \\ & + p_3.P\left(Z \le \frac{\mu_{11} - \mu_{22}}{\sqrt{{\sigma_{11}}^2 + {\sigma_{22}}^2}}\right) + p_4.P\left(Z \le \frac{\mu_{12} - \mu_{21}}{\sqrt{{\sigma_{12}}^2 + {\sigma_{21}}^2}}\right) \end{aligned} \tag{8}$$

So, we have

$$\begin{aligned} P(S_2 < S_1) = {} & p_1.\Phi\left(\frac{\mu_{11} - \mu_{21}}{\sqrt{{\sigma_{11}}^2 + {\sigma_{21}}^2 + 2.\rho.\sigma_{11}.\sigma_{21}}}\right) + \\ & p_2.\Phi\left(\frac{\mu_{12} - \mu_{22}}{\sqrt{{\sigma_{12}}^2 + {\sigma_{22}}^2 + 2.\rho.\sigma_{12}.\sigma_{22}}}\right) + \\ & p_3.\Phi\left(\frac{\mu_{11} - \mu_{22}}{\sqrt{{\sigma_{11}}^2 + {\sigma_{22}}^2}}\right) + p_4.\Phi\left(\frac{\mu_{12} - \mu_{21}}{\sqrt{{\sigma_{12}}^2 + {\sigma_{21}}^2}}\right) \end{aligned} \tag{9}$$

3 Application

3.1 A Real Application

Flu season for Children

As we know the flu seasons come and go and the children are very vulnerable beside the elderly to suffer from the flu. Some children will have to see their physicians for recovery and some others recover on their own without any treatment. Also, there are those healthy children who will not be affected at all by the flu. So, one can see a Markov-chain pattern to explain the state space of the status of the children during a flu season.

Suppose that a test is done during the flu season by taking health related measurements from the children at the outset of the flu season and again towards the end of the flu season. This is to study the effectiveness of the medical test in diagnosing flu on the children.

Let S_1 = Health related measurement at the beginning of the flu season

S_2 = Health related measurement at the end of the flu season

The estimates of $P(S_2 < S_1)$ or $P(S_2 > S_1)$ can provide a true picture of the health status if the measurement $S(S_1, S_2)$ for this study is carefully chosen.

3.2 Example

Here, we discuss a situation where the transition from healthy (H) to disease (D) or vice-versa takes place according to the first order Markov Chain with transition probability matrix given by

$$\begin{array}{c} \\ H \\ D \end{array} \begin{array}{c} \begin{array}{cc} H & D \end{array} \\ \begin{pmatrix} 0.95 & 0.05 \\ 0.70 & 0.30 \end{pmatrix} \end{array}$$

Note that at time 1, 80 % of the population was healthy and the remaining 20 % were disease stricken. At time 2, 90 % of the population was healthy and the remaining 10 % were disease stricken.

Test # 1

Suppose that for a diagnostic test (say Test # 1) the pre and the post treatment measurements are S_1 and S_2 respectively. Note that S_1 and S_2 were generated according to mixture normal distributions.

$$S_1 \sim 0.8\ N(4,2^2) + 0.2\ N(7,1^2)$$
$$S_2 \sim 0.9\ N(4,1.8^2) + 0.1\ N(6,1.1^2)$$

For this situation,

$$\begin{aligned} p_1 + p_3 &= 0.8 \\ p_1 + p_4 &= 0.9 \\ p_1 + p_2 + p_3 + p_4 &= 1 \end{aligned} \tag{10}$$

So, one set of solutions are

$$p_1 = 0.75,\ \ p_2 = 0.05,\ \ p_3 = 0.05,\ \ p_4 = 0.15$$

Also, note that

$$\begin{aligned} Cov(S_1, S_2) &= p_1.\,Cov(S_1, S_2) + p_2.\,Cov(S_1, S_2) + p_3.\,Cov(S_1, S_2) \\ &\quad + p_4.\,Cov(S_1, S_2) \\ &= p_1.\rho.\sigma_{11}.\sigma_{21} + p_2.\rho.\sigma_{12}.\sigma_{22} + 0 + 0 \end{aligned} \tag{11}$$

This implies,

$$\begin{aligned} 0.782 &= \hat{\rho}\,.\,\{0.75.(2).(1.8) + 0.05.(1).(1.8)\} \\ \hat{\rho} &= 0.284 \end{aligned} \tag{12}$$

This in turn implies that

$$P(S_2 < S_1) = 0.375 + 0.036 + 0.009 + 0.139 = 0.559 \tag{13}$$

Note that this is very close to the empirical estimate of 0.54 for this probability. Hence, it supports our theory that the Mixture Gaussian Copula is a very good model for checking the suitability of a certain diagnostic test to decide whether a child is healthy or not when there is a correlation structure. The diagnostic Test # 1 indicates that only 54 % of the children are healthy after the treatment when nearly 80 % should be healthy during the post-treatment period. So, diagnostic Test # 1 is not good and a different diagnostic test (say Test # 2) is needed.

Robustness Study (based on Test # 1 Results)

Here, we study the robustness of the copula based estimate of the probability for other possible choices for the mixing proportions.

p1	p2	p3	p4	Copula Probability Estimate
0.70	0.00	0.10	0.20	0.545461
0.71	0.01	0.09	0.19	0.546354
0.72	0.02	0.08	0.18	0.547256
0.73	0.03	0.07	0.17	0.548165
0.74	0.04	0.06	0.16	0.549082
0.75	0.05	0.05	0.15	0.550006
0.76	0.06	0.04	0.14	0.550937
0.77	0.07	0.03	0.13	0.551875
0.78	0.08	0.02	0.12	0.552819
0.79	0.09	0.01	0.11	0.553770
0.80	0.10	0.00	0.10	0.554726

As seen from this study, for the possible solutions that are considered here, the copula based probability estimate is robust and is about 0.55.

Test # 2

Suppose that for the second diagnostic test (Test # 2) the pre treatment and the post treatment measurements are $\overline{S_1}$ and $\overline{S_2}$ respectively. Note that $\overline{S_1}$ and $\overline{S_2}$ were generated according to mixture normal distributions.

$$\overline{S}_1 \sim 0.8\ N\left(3.9, 2^2\right) + 0.2\ N\left(6.9, 1^2\right)$$

$$\overline{S}_2 \sim 0.9\ N\left(2.5, 1.8^2\right) + 0.1\ N\left(2.7, 1.1^2\right)$$

Again, note that

$$\begin{aligned} Cov\left(\overline{S}_1, \overline{S}_2\right) &= p_1.\ Cov\left(\overline{S}_1, \overline{S}_2\right) + p_2.\ Cov\left(\overline{S}_1, \overline{S}_2\right) + p_3.\ Cov\left(\overline{S}_1, \overline{S}_2\right) \\ &\quad + p_4.\ Cov\left(\overline{S}_1, \overline{S}_2\right) \\ &= p_1.\rho.\sigma_{11}.\sigma_{21} + p_2.\rho.\sigma_{12}.\sigma_{22} + 0 + 0 \end{aligned} \tag{14}$$

This implies,

$$-\ 0.629 = \hat{\rho}\ .\ \{0.75.(2).(1.8) + 0.05.(1).(1.8)\} \tag{15}$$

$$\hat{\rho} = -0.225$$

This in turn implies that

$$P\left(\overline{S_2} < \overline{S_1}\right) = 0.542 + 0.0499 + 0.0351 + 0.148 = 0.775 \qquad (16)$$

Again, note that this is very close to the empirical estimate of 0.78 for this probability. Hence, it supports our theory for the second time that the Mixture Gaussian Copula is a very good model for checking the suitability of a certain test to decide whether a child is healthy or not when there is a correlation structure. Moreover, the second test based measurement $\overline{S}$ is better than the first test based measurement S as it correctly reflects the actual situation. Actually about 80 % of the children were healthy (or were healthy due to the treatment) in the population and the copula based probability correctly points out the situation.

Robustness Study (based on Test # 2 Results)

Again, we study the robustness of the copula based estimate of the probability.

p1	p2	p3	p4	Copula Probability Estimate
0.70	0.00	0.10	0.20	0.774828
0.71	0.01	0.09	0.19	0.774811
0.72	0.02	0.08	0.18	0.774801
0.73	0.03	0.07	0.17	0.774796
0.74	0.04	0.06	0.16	0.774797
0.75	0.05	0.05	0.15	0.774803
0.76	0.06	0.04	0.14	0.774814
0.77	0.07	0.03	0.13	0.774830
0.78	0.08	0.02	0.12	0.774850
0.79	0.09	0.01	0.11	0.774874
0.80	0.10	0.00	0.10	0.774902

As seen from this study, for the solutions that are considered here, the copula based probability estimate is very robust and is about 0.775.

4 Conclusion

4.1 ROC Curve Comparison

The main feature of this article is to use the mixture Gaussian Copula to model the dependence between the observations gathered at time 1and time 2 about a disease and the ROC curves to compare the performance of the diagnostic tests. Please note that only in the context of the following ROC curves, the term “time1” represents the probability for a health related measurement collected at time1 to exceed a certain threshold. Similarly, the term “time2” represents the probability for a health related measurement collected at time 2 to exceed the same threshold. This ROC curve was drawn by having the variable “time2” on the X-axis and the variable “time1” on the Y-axis. The area under the ROC curve (AUROC) is used as a measure for comparing the diagnostic tests. Although there can be several estimates for the mixing proportions in the Mixture Gaussian Copula model, our study shows that these copula based probability estimates are fairly robust and hence there is very little difference in the area estimates under the ROC curve based on this mixture Gaussian Copula for any given diagnostic test. Moreover, the Copula based estimate seemed to agree with the empirical estimate. Here, we are comparing two different diagnostic tests, namely Test #1 and Test #2. We note that the area under the ROC curve for Test #1 is much higher than the area under the ROC curve for Test #2. So, we conclude that Test # 2 is better than Test # 1 in the context of diagnostic testing. This mixture Gaussian Copula based model is fairly precise with respect to diagnosis and is computationally less intensive.

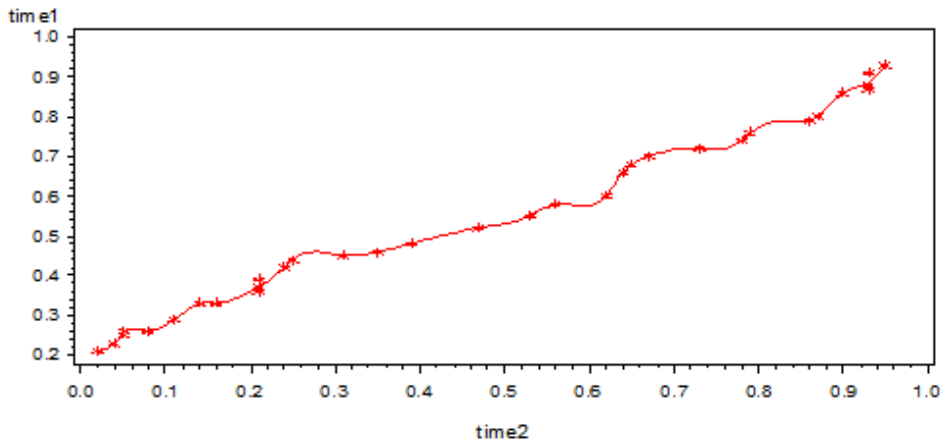

Figure 1: ROC Curve based on first set of measurements

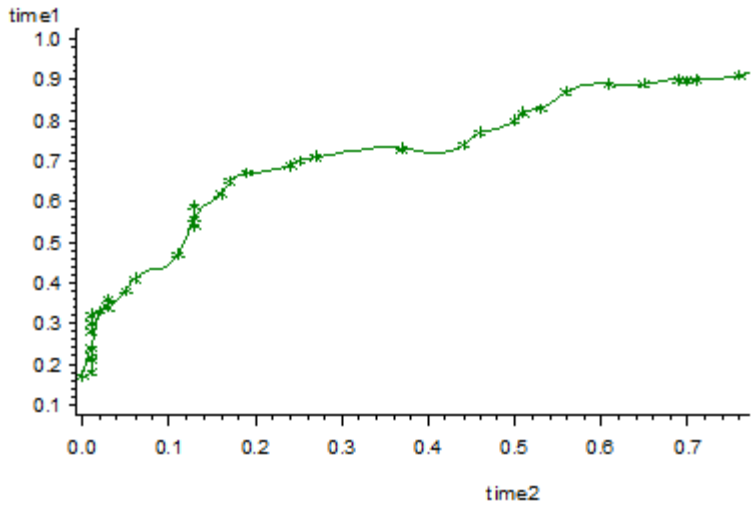

Figure 2: ROC Curve based on second set of measurements

References

1. Clayton, D.G (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151

2. Gonen, M (2007). Receiver Operating Characteristic (ROC) Curves. SUGI 31, 210, 1-18

3. Krzanowski, W.J, and Hand, D.J (2009). ROC curves for continuous data, Monographs on Statistics and Applied Probability. CRC Press, Taylor and Francis Group, New York.

4. Nelsen, R.B (1999). An Introduction to Copulas. Springer, New York.

5. Pepe, M.S. (2003). The Statistical Evaluation of medical tests for classification and prediction. Oxford Statistical Science Series, Oxford University Press.

6. Pundir, S (2011). Receiver Operating Characteristic curve for Bi-Weibull distribution and its properties. Proceedings of the Institute of Applied Statistics, Sri Lanka.

7. Shultz, E.K (1995). Multivariate Receiver Operating Characteristic Curve Analysis: Prostate Cancer Screening as an example. Clinical Chemistry 41/8(B), 1248-1255

8. Sklar, A (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de L' Institut de Statistique de L' Universite de Paris,* 8, 229-231

9. Zhou, X.H., Obuchowski, N.A and McClish, D.K (2002). Statistical methods in diagnostic medicine. Wiley, New York.