Maximum likelihood estimation of a common mean vector in the bivariate FGM copula model for meta-analysis

Graduate Institute of Statistics, National Central University, Taiwan
Jia-Han Shih
Department of Social Information, Mejiro University, Japan
Yuan-Tsung Chang
Department of Mathematical and Physical Sciences, Japan Women's University, Japan
Yoshihiko Konno
Graduate Institute of Statistics, National Central University, Taiwan
Takeshi Emura

Abstract

Bivariate meta-analysis is a method to obtain summary estimates where two outcomes are collected across different studies. However, most existing methods for bivariate meta-analysis are based on the bivariate normal model (Berkey et al. 1998; Riley 2009; Mavridis and Salanti 2011). Then it is natural to consider an alternative model which provides a different dependence pattern from the bivariate normal distribution model.

We introduce a general copula-based approach including model construction, maximum likelihood estimation, and the Fisher information matrix. In this context, we focus on the so-called Farlie-Gumbel-Morgenstern (FGM) copula which has a simple and mathematically attractive form. This form allows some special mathematical identities and tractable Fisher information matrix. These properties make the bivariate FGM copula model suitable for fixed-effects meta-analysis. This paper is based on a manuscript currently under review (Shih et al. 2018-) with additional discussions on copula models.

Keywords Asymptotic theory · Copula · Fisher information matrix · Maximum likelihood estimation · Multivariate analysis

---

Jia-Han Shih (corresponding author); E-mail: tommy355097@gmail.com
Graduate Institute of Statistics, National Central University
No. 300, Zhongda Road., Zhongli District, Taoyuan City 32001, Taiwan

1 Introduction

Bivariate meta-analysis is a widely used method in the field of educational research. For instance, a meta-analysis can be performed based on bivariate test scores which are collected across several studies (Gleser and Olkin 1994; Riley 2009). Traditionally, the bivariate meta-analysis is usually based on the bivariate normal model (Berkey et al. 1998; Riley 2009; Mavridis and Salanti 2011). We first illustrate the traditional method by giving an example based on our original data on bivariate entrance exam scores

We collect the bivariate entrance exam scores data for entering the Graduate Institute of Statistics, National Central University (NCU), Taiwan. The data consist of mathematics and statistics scores of 848 students across 5 academic years (from 2013 to 2017). We let i = 1, 2, …, 5 corresponding to 2013, 2014, …, 2017. The possible range of score is from 0 to 100 for both two subjects. Based on the individual scores, we compute the mean scores of mathematics (Y_{i1}) and statistics (Y_{i2}), and their covariance matrix ($\mathbf{C}_i$) for each i. Hence the data consist of $\{(Y_{i1}, Y_{i2}), i = 1, 2, \ldots, 5\}$. The data are the official records obtained from the Admission Division of NCU and are summarized in Figure 1.

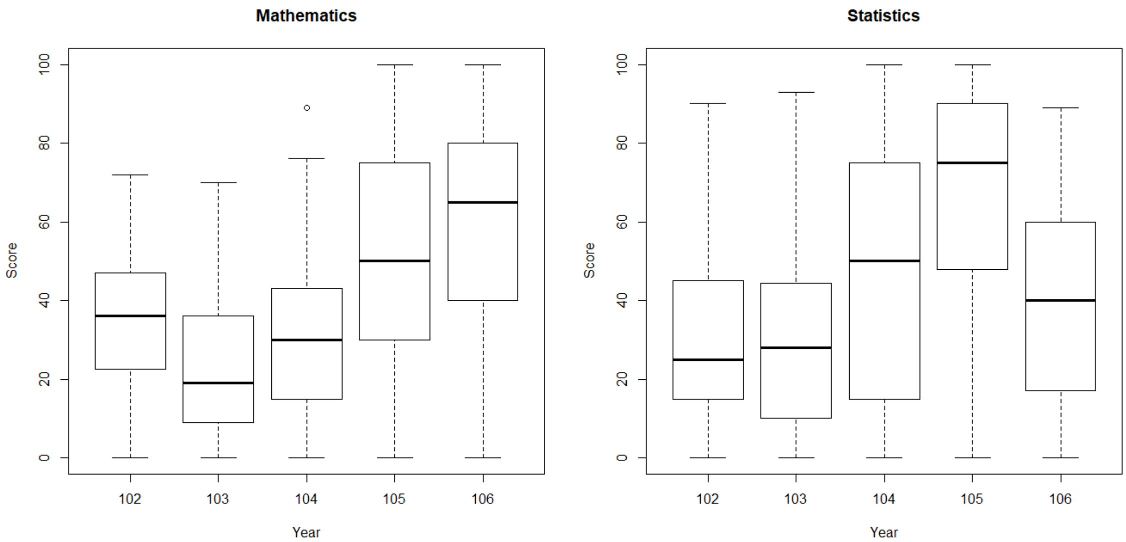

Figure 1. Boxplots of the entrance exam scores on Mathematics and Statistics across 5 academic years (from 2013 to 2017).

We fit the data to the fixed-effects bivariate normal model (Berkey et al. 1998; Mavridis and Salanti 2011). For each year i, suppose the mean scores of mathematics (Y_{i1}) and statistics (Y_{i2}) follow a bivariate normal distribution.

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim \mathrm{BVN}\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \mathbf{C}_i = \begin{bmatrix} \sigma_{i1}^2 & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \sigma_{i2}^2 \end{bmatrix} \right), \qquad i = 1, 2, \ldots, 5, \tag{1}$$

where $\rho_i \in (-1, 1)$ is the within-study correlation for year i. All the mean scores ($\mathbf{Y}_i$'s) across 5 academic years share the same common mean vector ($\boldsymbol{\mu}$). Based on the entrance exam data, the log-likelihood function is

$$\ell^{\mathrm{N}}(\boldsymbol{\mu}) = \sum_{i=1}^{5} \frac{\rho_i}{(1-\rho_i^2)} \left(\frac{Y_{i1}-\mu_1}{\sigma_{i1}} \right) \left(\frac{Y_{i2}-\mu_2}{\sigma_{i2}} \right) - \frac{1}{2} \sum_{j=1}^{2} \sum_{i=1}^{5} \frac{1}{(1-\rho_i^2)} \left(\frac{Y_{ij}-\mu_j}{\sigma_{ij}} \right)^2$$
$$-5\log(2\pi) - \sum_{j=1}^{2}\sum_{i=1}^{5} \log \sigma_{ij} - \frac{1}{2}\sum_{i=1}^{5} \log(1-\rho_i^2).$$

Then, the maximum likelihood estimator (MLE) of the common mean vector is

$$\hat{\boldsymbol{\mu}}^{\mathrm{N}} = \begin{bmatrix} \hat{\mu}_1^{\mathrm{N}} \\ \hat{\mu}_2^{\mathrm{N}} \end{bmatrix} = \underset{\boldsymbol{\mu} \in \mathrm{R}^2}{\arg\max}\, \ell^{\mathrm{N}}(\boldsymbol{\mu}) = \left(\sum_{i=1}^{5} \mathbf{C}_i^{-1} \right)^{-1} \sum_{i=1}^{5} \mathbf{C}_i^{-1} \mathbf{Y}_i .$$

Computer programs for bivariate meta-analysis are well-developed in R. Here, we apply the R command "*mvmeta*" in the R package "*mvmeta*" (Gasparrini 2018). The MLE for the common mean score of mathematics is $\hat{\mu}_1^{\mathrm{N}} = 35.83$ (95% confidence interval (CI): 34.51 – 37.16). On the other hand, the MLE for the common mean score of statistics is $\hat{\mu}_2^{\mathrm{N}} = 38.64$ (95% CI: 36.94 – 40.34). The fitted log-likelihood value is -342.65 (the left panel of Figure 2). Our analysis reveals that the mean scores of mathematics and statistics are much lower than 50. This indicates that most students perform poorly on the exam.

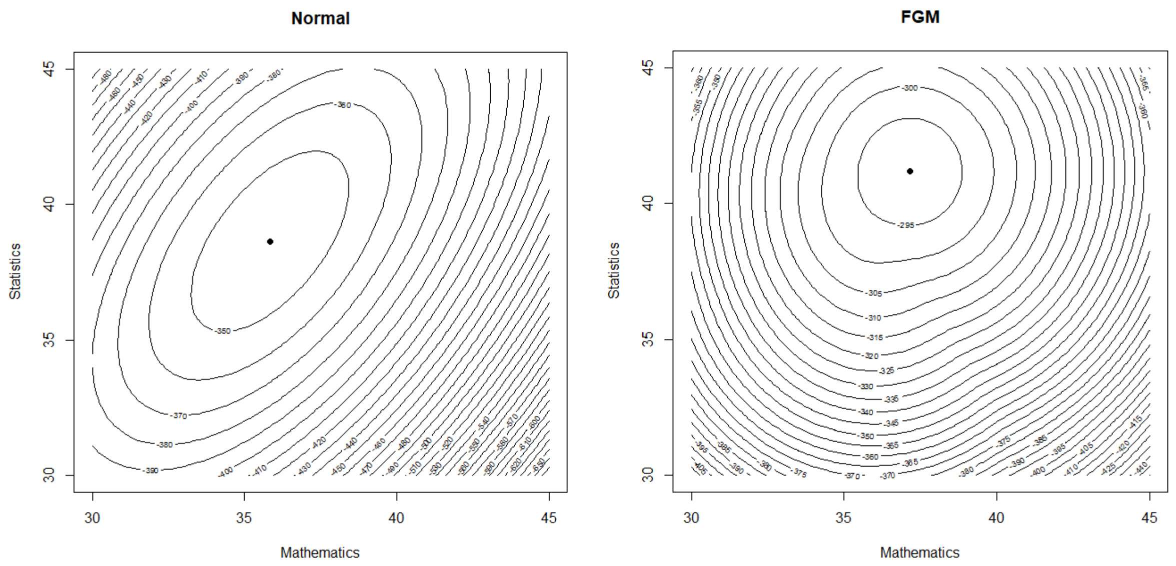

Figure 2. Contour plots of the log-likelihood function based on the entrance exam data where the points indicate the MLE of the common mean. Left panel: Bivariate normal model, Right panel: Bivariate FGM copula model.

Traditional meta-analysis based on the bivariate normal model does not allow different dependence structure between two outcomes. Then it is natural to consider an alternative model which can provide a different dependence structure from the bivariate normal model. We shall introduce a copula-based approach in the next section.

2 Copula model

A bivariate copula is a bivariate distribution function with two marginal uniform distributions on the unit interval [0, 1] (Nelsen 2006). One can model the dependence structure between two outcomes by using a copula model in which marginal models are flexibly chosen. There are many well-known copulas such as the independence (Nelsen 2006), Clayton (Clayton 1978), Frank (Genest 1987), and Gumbel (Gumbel 1960a) copulas from the Archimedean family and the normal (Gaussian) and *t*-copulas from the elliptical family.

According to Sklar's theorem (Sklar 1959), any bivariate distribution can be uniquely determined by a bivariate copula and two marginal distributions. Thus, one can represent the common mean bivariate normal model in Eq. (1) as

$$\Pr(Y_{i1} \le y_1, Y_{i2} \le y_2) = \Phi_{\rho_i}\left(\frac{y_1-\mu_1}{\sigma_{i1}}, \frac{y_2-\mu_2}{\sigma_{i2}}\right) = C^{\mathrm{N}}_{\rho_i}\left\{\Phi\left(\frac{y_1-\mu_1}{\sigma_{i1}}\right), \Phi\left(\frac{y_2-\mu_2}{\sigma_{i2}}\right)\right\},$$

where $\Phi_\rho(\cdot,\cdot)$ is the bivariate cumulative distribution function (c.d.f.) of the bivariate standard normal distribution with correlation ρ, $\Phi(\cdot)$ is the c.d.f. of $N(0,1)$, and

$$C^{\mathrm{N}}_\rho(u,v) = \Phi_\rho\{\Phi^{-1}(u), \Phi^{-1}(v)\}, \qquad 0 \le u, v \le 1$$

is the normal copula. By replacing the normal copula with any other copula, one can construct a non-bivariate normal distribution (Genest and Favre 2007).

2.1 Estimation of common mean

We introduce a general copula-based approach. Suppose that we have an arbitrary copula $C_{\boldsymbol{\theta}_i}$ indexed by possibly multiple parameters $\boldsymbol{\theta}_i$. We define the general copula model as

$$\Pr(Y_{i1} \le y_1, Y_{i2} \le y_2) = C_{\boldsymbol{\theta}_i}\left\{\Phi\left(\frac{y_1-\mu_1}{\sigma_{i1}}\right), \Phi\left(\frac{y_2-\mu_2}{\sigma_{i2}}\right)\right\}. \tag{2}$$

The corresponding density function is

$$f_{i,\boldsymbol{\mu}}(y_1, y_2) = C^{[1,1]}_{\boldsymbol{\theta}_i}\left\{\Phi\left(\frac{y_1-\mu_1}{\sigma_{i1}}\right), \Phi\left(\frac{y_2-\mu_2}{\sigma_{i2}}\right)\right\}\frac{1}{2\pi\sigma_{i1}\sigma_{i2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{2}\left(\frac{y_j-\mu_j}{\sigma_{ij}}\right)^2\right\},$$

where $C^{[1,1]}_{\boldsymbol{\theta}_i}(u,v) = \partial^2 C_{\boldsymbol{\theta}_i}(u,v)/\partial u \partial v$ is the copula density. The parameters $\boldsymbol{\theta}_i$ can be obtained by solving the moment equations

$$E_{\boldsymbol{\theta}_i}\left\{\left(\frac{Y_{i1}-\mu_1}{\sigma_{i1}}\right)\left(\frac{Y_{i2}-\mu_2}{\sigma_{i2}}\right)\right\}=\mathbf{C}_{i,12}, \qquad i=1,2,\ldots,5.$$

Based on the entrance exam data, the log-likelihood function under the model (2) is

$$\ell(\boldsymbol{\mu})=\sum_{i=1}^{5}\log\left[C_{\boldsymbol{\theta}_i}^{[1,1]}\left\{\Phi\left(\frac{Y_{i1}-\mu_1}{\sigma_{i1}}\right),\Phi\left(\frac{Y_{i2}-\mu_2}{\sigma_{i2}}\right)\right\}\right]-\frac{1}{2}\sum_{j=1}^{2}\sum_{i=1}^{5}\left(\frac{Y_{ij}-\mu_j}{\sigma_{ij}}\right)^2$$
$$-5\log(2\pi)-\sum_{j=1}^{2}\sum_{i=1}^{5}\log\sigma_{ij}.$$

Then the MLE of the common mean vector under the general copula model (2) is

$$\hat{\boldsymbol{\mu}}=\begin{bmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{bmatrix}=\underset{\boldsymbol{\mu}\in\mathrm{R}^2}{\arg\max}\,\ell(\boldsymbol{\mu}).$$

The MLE does not have closed-form expression under the general copula model. One can maximize the log-likelihood function by performing the Newton-Raphson (NR) algorithm stated below.

The Newton-Raphson algorithm

Step 1. Set the starting values as the univariate estimators

$$\mu_1^{(0)}=\left(\sum_{i=1}^{n}\frac{1}{\sigma_{i1}^2}\right)^{-1}\sum_{i=1}^{n}\frac{Y_{i1}}{\sigma_{i1}^2}, \qquad \mu_2^{(0)}=\left(\sum_{i=1}^{n}\frac{1}{\sigma_{i2}^2}\right)^{-1}\sum_{i=1}^{n}\frac{Y_{i2}}{\sigma_{i2}^2}.$$

Step 2. Repeat the Newton-Raphson iterations for $k=0,1,2,\ldots,$

$$\begin{bmatrix}\mu_1^{(k+1)}\\ \mu_2^{(k+1)}\end{bmatrix}=\begin{bmatrix}\mu_1^{(k)}\\ \mu_2^{(k)}\end{bmatrix}-\left.\begin{bmatrix}\dfrac{\partial^2\ell(\boldsymbol{\mu})}{\partial\mu_1^2} & \dfrac{\partial^2\ell(\boldsymbol{\mu})}{\partial\mu_1\partial\mu_2}\\ \dfrac{\partial^2\ell(\boldsymbol{\mu})}{\partial\mu_1\partial\mu_2} & \dfrac{\partial^2\ell(\boldsymbol{\mu})}{\partial\mu_2^2}\end{bmatrix}^{-1}\begin{bmatrix}\dfrac{\partial\ell(\boldsymbol{\mu})}{\partial\mu_1}\\ \dfrac{\partial\ell(\boldsymbol{\mu})}{\partial\mu_2}\end{bmatrix}\right|_{\mu_1=\mu_1^{(k)},\,\mu_2=\mu_2^{(k)}}.$$

- Stop if $\max\{|\mu_1^{(k+1)}-\mu_1^{(k)}|,|\mu_2^{(k+1)}-\mu_2^{(k)}|\}<10^{-5}$. The MLE is $\hat{\boldsymbol{\mu}}=(\mu_1^{(k+1)},\mu_2^{(k+1)})^{\mathrm{T}}$.

In our experience, the NR algorithm converges very quickly. The MLE can also be obtained by using some computer programs to maximize the log-likelihood function (e.g., the R command "*nlm*" or "*optim*").

Through the above frameworks, one can estimate the common mean vector under a non-bivariate normal model by specifying any copula.

2.2 Asymptotic inference

We first define the 2×2 Fisher information matrix $I_i(\boldsymbol{\mu})$ for $i=1,2,\ldots,5$ as

$$I_{i,jk}(\boldsymbol{\mu}) = E_{\boldsymbol{\mu}}\left\{ \frac{\partial \log f_{i,\boldsymbol{\mu}}(Y_{i1},Y_{i2})}{\partial \mu_j} \frac{\partial \log f_{i,\boldsymbol{\mu}}(Y_{i1},Y_{i2})}{\partial \mu_k} \right\}, \qquad j,k=1,2.$$

The Fisher information matrix is essential in a likelihood-based approach since it contains all necessary information about the asymptotic distribution of the MLE. Shih et al. (2018-) derived the expressions of the Fisher information under the FGM copula, which can be used to perform interval estimation on the common mean vector. One can also use the observed Fisher information matrix which is the negative hessian matrix of the log-likelihood function. It is automatically obtained in the convergent step of the NR algorithm.

If we consider a sample of size n in Eq. (2), we have the samples of random vectors $\mathbf{Y}_i$, $i=1,2,\ldots,5$. The samples are independent but not identically distributed (i.n.i.d.) since

$$Cov(\mathbf{Y}_i) \neq Cov(\mathbf{Y}_j), \qquad i\neq j.$$

Therefore, the usual asymptotic theory for the sum of independent and identically distributed (i.i.d.) samples cannot be applied to our setting. To verify the consistency and asymptotic normality of the MLE, we modify the regularity conditions of Bradly and Gart (1962) in a similar manner as Emura et al. (2017c). Under the modified regularity conditions, we apply the weak law of large numbers (WLLN) and the Lindeberg-Feller central limit theorem (CLT) for i.n.i.d. random variables. We also modify the proofs Lehmann and Casella (1998) who provided the detailed proofs for the consistency and asymptotic normality of the MLE under i.i.d. samples. The details can be seen from Shih et al. (2018-).

2.3 The FGM copula model

We consider the so-called Farlie-Gumbel-Morgenstern (FGM) copula which belongs to neither the Archimedean nor the elliptical family. The FGM copula is a copula derived from the FGM distribution. The FGM distribution was first proposed by Morgenstern (1956) which is also traced back to Eyraud (1936). It was later studied by Farlie (1960) and Gumbel (1960b). The FGM copula is defined as

$$C_\theta^{\mathrm{FGM}}(u,v) = uv + \theta u(1-u)v(1-v), \qquad 0\le u,v\le 1, \tag{3}$$

where $\theta\in$[-1, 1] is the dependence parameter. Eq. (3) shows that the FGM copula has a simple analytical form, therefore, it allows closed-form expressions for various dependence measures. For instance, Kendall's tau and Spearman's rho are $2\theta/9$ and $\theta/3$, respectively (Nelsen 2006). Note that Kendall's tau and Spearman's rho are dependence measures based on the concept of concordance and are free from the marginal distributions. On the other hand, the famous Pearson correlation under the uniform, normal, exponential marginal distributions are $\theta/3$, θ/π, and $\theta/4$, respectively (Schucany et al. 1978). In addition to these theoretical aspects, the FGM copula has also been used in real applications (Louzada et al. 2013; Martinez and Achcar 2014).

We obtain the common mean bivariate FGM copula model by adapting the FGM copula to the general copula model defined in Eq. (2). An interesting feature of the FGM copula is its special structure. Eq. (3) reveals that it is written as the sum of an independence copula and a function of dependence parameter θ. This property directly leads to some mathematical identities which are useful for deriving the Fisher information matrix. Under the bivariate FGM copula model, the Fisher information matrix can be decomposed into the sum of the Fisher information matrix under the independent model and the information related to the dependence parameter θ. In addition, we approximate the exact Fisher information matrix by its linear approximation which is derived by applying the Taylor expansion. Under some mild regularity conditions, the consistency and asymptotic normality of the MLE can be proved by applying the WLLN and the Lindeberg-Feller CLT for i.n.i.d. random variables. By using the asymptotic theory, one can construct a 95% CI based on the exact, approximate, or observed Fisher information matrix. All the details of these theoretical results are referred to our original paper (Shih et al. 2018-).

3 Entrance exam data revisit

We fit the entrance exam data to the common mean bivariate FGM copula model. We compute the MLE $\hat{\boldsymbol{\mu}}^{\mathrm{FGM}} = (\hat{\mu}_1^{\mathrm{FGM}}, \hat{\mu}_2^{\mathrm{FGM}})^{\mathrm{T}}$ by the NR algorithm as described in Sect. 2.1. The MLE for the common mean score of mathematics is $\hat{\mu}_1^{\mathrm{FGM}} = 37.16$ (95% CI: 35.85 – 38.48). On the other hand, the MLE for the common mean score of statistics is $\hat{\mu}_2^{\mathrm{FGM}} = 41.17$ (95% CI: 39.48 – 42.87). We only report the 95% CI based on the exact Fisher information matrix due to the similarity of all 3 constrictions. The fitted log-likelihood value is -291.80 (the right panel of Figure 2), that is greater than the log-likelihood value under the bivariate normal model (-342.65).

The estimation results on the common mean scores of mathematics are significantly different between the bivariate FGM copula and normal models. The 95% CI of an estimator does not include another estimator. The same phenomenon can be found on the common mean scores of statistics. These findings typically indicate that at least one model is not suitable for the entrance exam data.

For further investigation, we examine the individual log-likelihood values for each year. Figure 3 reveals that the poor fit of the bivariate normal model mainly caused by the poor fit for the scores in 2017. This poor fit gives a large influence since the year 2017 has the largest number of students. In this case, we suggest choosing the bivariate FGM copula model which produces a larger log-likelihood value.

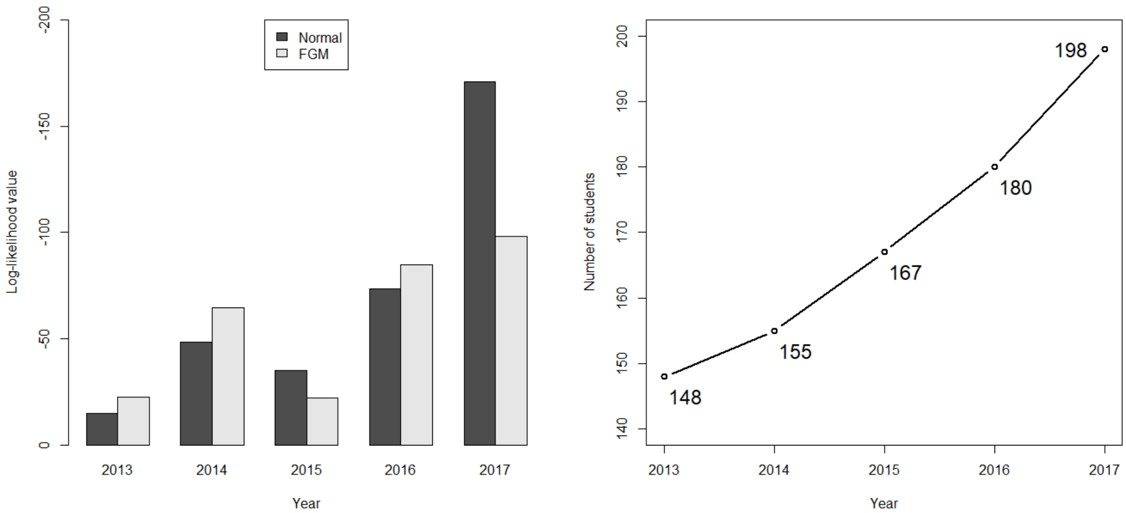

Figure 3. The individual log-likelihood values (left panel) and number of students (right panel) for each year based on the entrance exam data. Higher log-likelihood values correspond to better fit of the model.

4 Extensions

Bivariate meta-analysis has also been applied in the field of medical research, where the analysis methods usually have to deal with censoring or dependent competing risks. In such studies, the Clayton copula seems to be the most popular copula among others (Burzykowski et al. 2001; Emura et al. 2017a; Emura et al. 2017b; Emura and Chen 2018). This is due to the simple derivatives of the Clayton copula and its interpretability of the copula parameter. Still, the Fisher information matrix of the Clayton copula is not simple even under complete data (Schepsmeier and Stöber 2014). Thus, it is an interesting topic to extend the current FGM copula model and the Fisher information to incorporate censoring or dependent competing risks. Our recent paper shows that the FGM copula with the Burr III marginal models has nice tractability for a likelihood-based approach (Shih and Emura 2018), yet the expression of the Fisher information has not been given.

Acknowledgements

The authors thank the Editor, the Associate Editor, and the two anonymous reviewers from the journal *Statistics* for their helpful comments that greatly improved the original paper. Our work is supported by the government of Taiwan (MOST 105-2118-M-008-001-MY2).

References

Bradley RA, Gart JJ (1962) The asymptotic properties of ML estimators when sampling from associated populations, Biometrika 49: 205-214.

Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D (2001) Validation of surrogate end points in multiple randomized clinical trials with failure time end points, Appl Stat, 50: 405–422.

Berkey CS, Hoaglin DC, Antczak-bouckoms A, Mosteller F, Colditz GA (1998) Meta-analysis of multiple outcomes by regression with random effect, Stat Med, 17: 2537-2550.

Clayton DG (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.

Eyraud H (1936) Les principes de la mesure des correlations, Ann. Univ. Lyon III Ser. Sect. A1, 30-47.

Emura T, Nakatochi M, Murotani K, Rondeau V (2017a) A joint frailty-copula model between tumour progression and death for meta-analysis, Stat Method Med Res, 26: 2649-2666.

Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V (2017b) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model, Stat Method Med Res, doi: 10.1177/0962280216688032.

Emura T, Hu YH, Konno Y (2017c) Asymptotic inference for maximum likelihood estimators under the special exponential family with double-truncation, Stat Pap 58 (3): 877-909.

Emura T, Chen YH (2018), Analysis of survival data with dependent censoring, copula-based approaches, JSS Research Series in Statistics, Springer, Singapore.

Farlie DJG (1960) The performance of some correlation coefficients for a general bivariate distribution, Biometrika 47: 307-323.

Gumbel EJ (1960a) Distributions des valeurs extrêmes en plusiers dimensions, Publications de l'Institut de Statistique de L'Université de Paris, 9: 171-173.

Gumbel EJ (1960b) Bivariate exponential distributions, J Amer Statist Assoc, 55: 698-707.

Gleser LJ, Olkin L (1994) Stochastically dependent effect sizes. In the Handbook of Research Synthesis New York: Russel Sage Foundation.

Genest C (1987) Frank's family of bivariate distributions, Biometrika 74: 549-555.

Genest C, Favre AC (2007) Everything you always wanted to know about copula modeling but were afraid to ask, J Hydrol Eng, 12: 347-368.

Gaspaarrini A (2018) mvmeta: Multivariate and univariate meta-analysis and meta-regression, CRAN.

Lehmann EL, Casella G (1998) Theory of Point Estimation, (2nd ed.), Springer, New York.

Louzada F, Suzuki AK, Cancho VG (2013) The FGM long-term bivariate survival copula model: modeling, Bayesian estimation, and case influence diagnostics, Commun Stat Theory, 42: 673-91.

Morgenstern D (1956) Einfache Beispiele zweidimensionaler Verteilungen, Mitteilungsblatt für Mathematishe Statistik, 8: 34-235.

Mavridis D, Salanti G (2011) A practical introduction to multivariate meta-analysis, Stat Method Med Res, 22: 133-158.

Martinez EZ, Achcar JA (2014) Bayesian bivariate generalized Lindley model for survival data with a cure fraction, Comput Method Programs Biomed, 117: 145-157.

Nelsen RB (2006) An Introduction to Copulas, (2nd ed.), Springer, New York.

Riley RD (2009) Multivariate meta-analysis: the effect of ignoring within-study correlation, J R Statist Soc Ser A, 172: 789-811.

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris, 8: 229-231.

Schucany WR, Parr WC, Boyer JE (1978) Correlation structure in Farlie-Gumbel- Morgenstern distributions, Biometrika, 65: 650-653.

Schepsmeier U, Stöber J (2014). Derivatives and Fisher information of bivariate copulas. Stat Pap 55(2): 525-542

Shih JH, Emura T (2018), Likelihood-based inference for bivariate latent failure time models with competing risks under the generalized FGM copula, Comput Stat, doi: 10.1007/s00180-018-0804-0.

Shih JH, Chang YT, Konno Y, Emura T (2018-) Estimation of a common mean vector under in bivariate meta-analysis under the FGM copula, Statistics, in revision.