Computational Statistics and Data Analysis 55 (2011) 2446–2455

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models

Jerald F. Lawless [a], Yildiz E. Yilmaz [b,*]

[a] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada N2L 3G1

[b] Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada M5T 3L9

ARTICLE INFO

Article history:
Received 29 October 2010
Received in revised form 4 February 2011
Accepted 5 February 2011
Available online 12 February 2011

Keywords:
Semiparametric maximum likelihood
Model misspecification
Pseudolikelihood
Clayton copula
Gumbel–Hougaard copula
Frank copula

ABSTRACT

We consider bivariate distributions that are specified in terms of a parametric copula function and nonparametric or semiparametric marginal distributions. The performance of two semiparametric estimation procedures based on censored data is discussed: maximum likelihood (ML) and two-stage pseudolikelihood (PML) estimation. The two-stage procedure involves less computation and it is of interest to see whether it is significantly less efficient than the full maximum likelihood approach. We also consider cases where the copula model is misspecified, in which case PML may be better. Extensive simulation studies demonstrate that in the absence of covariates, two-stage estimation is highly efficient and has significant robustness advantages for estimating marginal distributions. In some settings, involving covariates and a high degree of association between responses, ML is more efficient. For the estimation of association, PML does not offer an advantage.

1. Introduction

Copula modeling (Joe, 1997; Nelsen, 2006) has become a popular framework for the analysis of multivariate data and the literature in this area is growing rapidly. Nevertheless, as we discuss here, a number of important questions remain about the properties and performance of copula-based estimation. Our purpose in this paper is to review several such questions and to present simulation results that provide new information.

In particular, we consider the efficiency and robustness of copula-based maximum likelihood (ML) estimation relative to pseudo-maximum likelihood (PML) estimation in semiparametric models. The latter is widely used with multivariate data. It consists of first estimating the marginal distributions $F_j(y_j)$ under a "working" assumption of independence of the Y_j. Then, in a second stage, copula parameters α are estimated by maximizing a pseudolikelihood function in which the first-stage estimates of the $F_j(y_j)$ are inserted. Joe (1997, Ch. 10), Lawless (2003, Sec. 11.2.2), Chen et al. (2006), Kim et al. (2007a,b) and many others discuss this approach; Joe refers to it as the inference function for margins (IFM) method. Previous work suggests that in many parametric settings, PML is nearly as efficient as ML. However, there has been little discussion of semiparametric models in which marginal distributions are nonparametric, especially in lifetime data settings that involve censoring. Moreover, the question of the robustness of ML and PML to misspecification of the copula family has received scant attention. It is our purpose here to address these questions. Our broad conclusions are that PML is the preferable approach for marginal distribution estimation in most situations that do not involve covariates. When covariates are present,

* Corresponding address: Samuel Lunenfeld Research Institute, Prosserman Centre for Health Resesarch, Mount Sinai Hospital, 60 Murray Street, 5th floor, Room 5-225, Box #18, Toronto, Ontario, Canada M5T 3L9. Tel.: +1 416 586 4800x7538; fax: +1 416 586 8404.

E-mail addresses: jlawless@uwaterloo.ca (J.F. Lawless), yilmaz@lunenfeld.ca (Y.E. Yilmaz).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.02.008

ML can be substantially better in certain settings. For estimation of association, PML can perform worse than ML under copula model misspecification.

We now describe copula models briefly and then review work on ML and PML estimation. For simplicity, we restrict consideration to continuous bivariate distributions. Let (Y_1, Y_2) have cumulative distribution function (cdf) $F(y_1, y_2) = \Pr(Y_1 \leq y_1, Y_2 \leq y_2)$ and let $F_1(y_1) = \Pr(Y_1 \leq y_1)$ and $F_2(y_2) = \Pr(Y_2 \leq y_2)$ be the marginal cdfs. A result due to Sklar (1959) says there is for any $F(y_1, y_2)$ a unique function $C(u_1, u_2)$ such that

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)). \tag{1}$$

The function $C(u_1, u_2)$ is termed a copula, and it is a cdf on the unit square, with marginal distributions that are uniform on $(0, 1)$. The books by Joe (1997) and Nelsen (2006) examine copulas in detail and discuss many parametric families.

Copula models are widely used with lifetime or duration variables, for example, in connection with times to disease onset or death in related individuals (Hougaard et al., 1992; Hsu and Gorfine, 2006), the times to clinical events for paired organs within individuals (Huster et al., 1989) and in applications in finance, insurance and risk management (Frees and Valdez, 1998; McNeil et al., 2005; Kim et al., 2007b). In many such applications Y_1 or Y_2 may be censored in observed data (Lawless, 2003) and we consider this feature here. We also note that in lifetime applications copula models in "survival" form are often used. In this case

$$S(y_1, y_2) = \Pr(Y_1 \geq y_1, Y_2 \geq y_2) = \bar{C}(S_1(y_1), S_2(y_2)), \tag{2}$$

where $S_1(y_1) = \Pr(Y_1 \geq y_1), S_2(y_2) = \Pr(Y_2 \geq y_2)$ and $\bar{C}$ is a copula. The copulas C in (1) and $\bar{C}$ in (2) are related by $\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$.

Copula-based estimation typically uses a parametric family $C(u_1, u_2; \alpha)$ along with models for the marginal distributions. Common models in lifetime applications include the Clayton (1978) family

$$C(u_1, u_2; \alpha) = (u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-1/\alpha}, \quad \alpha > 0 \tag{3}$$

and the Gumbel–Hougaard family (Gumbel, 1960)

$$C(u_1, u_2; \alpha) = \exp\{-[(-\log u_1)^{\alpha} + (-\log u_2)^{\alpha}]^{1/\alpha}\}, \quad \alpha \geq 1. \tag{4}$$

In such families the parameters α specify the degree of association for Y_1 and Y_2, and measures such as Kendall's tau and Spearman's rho (Joe, 1997, Sec. 2.1.9) are functions of them.

In some applications $F_1 = F_2$ but for now we assume that F_1 and F_2 are unrelated. For fully parametric models in which F_1 and F_2 are specified in terms of parameters β_1 and β_2, maximum likelihood estimation of $(\beta_1, \beta_2, \alpha)$ is straightforward (e.g. Huster et al., 1989; Frees and Valdez, 1998; He and Lawless, 2003). However, in many applications there is a preference for nonparametric or (if covariates are present) semiparametric estimation of F_1 and F_2. This provides more robustness in the estimation of F_1 and F_2. In addition, Kim et al. (2007a,b) show through simulations that parametric misspecification of F_1 or F_2 can lead to significant bias in estimates of α. We focus here on nonparametric and semiparametric models.

Pseudo-maximum likelihood (PML) estimation consists of estimating F_1 and F_2 on the basis of marginal data for Y_1 and Y_2, and then the resulting estimates $\tilde{F}_1, \tilde{F}_2$ are substituted for F_1, F_2 in (1) or (2) and the likelihood function $L(\tilde{F}_1, \tilde{F}_2, \alpha)$ based on the joint observations (Y_1, Y_2) is maximized to obtain an estimate $\tilde{\alpha}$ for α. In semiparametric settings not involving covariates, this approach was introduced by Oakes (1994) and Genest et al. (1995) for uncensored data; $\tilde{F}_1$ and $\tilde{F}_2$ are then the empirical cdfs for Y_1 and Y_2. Shih and Louis (1995) proposed the same method for censored data; in this case $\tilde{F}_1$ and $\tilde{F}_2$ are Kaplan–Meier estimates. The PML approach has subsequently been used in a variety of contexts including fully parametric settings; for example, see Joe (1997, Ch. 10) and Joe (2005).

A number of questions arise naturally concerning PML estimation for semiparametric models:

(i) What is the efficiency of the marginal estimators $\tilde{F}_1, \tilde{F}_2$ relative to full semiparametric ML estimators $\hat{F}_1, \hat{F}_2$ obtained by maximizing the likelihood function $L(F_1, F_2, \alpha)$ jointly with respect to F_1, F_2 and α?
(ii) What is the efficiency of the PML estimator $\tilde{\alpha}$ relative to that of the ML estimate $\hat{\alpha}$ obtained as in (i)?
(iii) What are the effects of misspecification of the copula family $C(u_1, u_2; \alpha)$?

Full semiparametric ML estimation referred to in (i) has recently been proposed by Li et al. (2008) for Gaussian copulas and by Yilmaz and Lawless (in press) for arbitrary copulas, but they did not address efficiency relative to PML. Kim et al. (2007a) considered point (ii) as well as the effects of parametric misspecification of F_1 and F_2 on estimation of α, but only in the case of uncensored data. Their main conclusions (with regard to semiparametric estimation) are that the semiparametric PML estimator $\tilde{\alpha}$ is only a little less efficient than a fully parametric ML or PML estimator, but that the latter estimators are highly non-robust to misspecification of F_1 or F_2. Kim et al. (2007b) reached a similar conclusion when the marginal distributions are linear regression models with arbitrary error distributions. Li et al. (2008) give simulation results showing that when $F_1 = F_2 = F$ and the Gaussian copula is correct, the ML estimator $\hat{F}$ is more efficient than the marginal Kaplan–Meier PML estimator. In addition, this holds up under mild departures from the assumed Gaussian copula. On the other hand, the PML estimator $\tilde{\alpha}$ of the Pearson correlation α in the Gaussian copula model is asymptotically (semiparametric) efficient when there is no censoring and F_1 and F_2 are distinct (Klassen and Wellner, 1997; see also Genest and Werker, 2002).