Expectation Maximization Algorithms for Estimating Bernstein Copula Density

Xiaoling Dou*[1], Satoshi Kuriki[1], Gwo Dong Lin[2]

[1]The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan
[2]Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.
*Corresponding author: Xiaoling Dou, e-mail: xiaoling@ism.ac.jp

Abstract

On the basis of order statistics, Baker (2008) proposed a method for constructing multivariate distributions with fixed marginals. This is another representation of the Bernstein copula. According to the construction of Baker's distribution, the Bernstein copula can be regarded as a finite mixture distribution. In this paper, we propose expectation-maximization (EM) algorithms to estimate the Bernstein copula function, and prove the local convergence property. Moreover, asymptotic properties of the proposed semiparametric estimators are provided. Illustrative examples are presented using real datasets.

Keyword: Baker's distribution, Bernstein polynomial, Density estimation, Linear convergence, Order statistic, Ordered categorical data.

1 Introduction

On the basis of order statistics, Baker (2008) proposed a simple and intuitive method for constructing multivariate distributions with fixed marginals. Let $F_{k:m}(x)$ and $G_{l:n}(y)$ be the distribution functions of the order statistics $X_{(k)}$ and $Y_{(l)}$, respectively. Baker's distribution is a finite mixture distribution of mn-components with distribution function

$$H(x,y;R) = \Pr\big(X_{(K)} \le x,\, Y_{(L)} \le y\big) = \sum_{k=1}^{m}\sum_{l=1}^{n} r_{kl} F_{k:m}(x) G_{l:n}(y) \tag{1}$$

with

$$\sum_{l=1}^{n} r_{kl} = \frac{1}{m}, \quad \sum_{k=1}^{m} r_{kl} = \frac{1}{n}, \quad r_{kl} \ge 0. \tag{2}$$

It is known that distribution functions of order statistics can be described in terms of the Bernstein polynomials. Let the Bernstein polynomial and its integral be

$$b_{k,n}(u) = \binom{n}{k} u^k (1-u)^{n-k}, \quad B_{k,n}(u) = \int_0^u b_{k,n}(t)dt, \quad u \in [0,1],$$

respectively. Then, the distribution functions $F_{k:m}(x)$ and $G_{l:n}(y)$ are expressed as $F_{k:m}(x) = mB_{k-1,m-1}(F(x))$ and $G_{l:n}(y) = nB_{l-1,n-1}(G(y))$ (see, e.g., (1) in Hwang and Lin, 1984). Substituting these into (1), we have

$$H(x,y;R) = C(F(x),G(y);R), \tag{3}$$

where

$$C(u,v;R) = mn\sum_{k=1}^{m}\sum_{l=1}^{n} r_{kl}B_{k-1,m-1}(u)B_{l-1,n-1}(v), \quad (u,v)\in[0,1]^2. \tag{4}$$

Using the property of finite mixture distribution, the expectation-maximization (EM) algorithm can be used to estimate parameters. In this paper, we propose estimation methods based on this idea.

2 EM algorithms based on pseudo-likelihood

2.1 General case

Throughout this paper, we assume that marginal distributions F and G are estimated in advance and treated as known functions in the subsequent analysis. This two-stage estimation is referred to as the semiparametric method and is widely used (e.g., Genest et al., 1995; Choroś et al, 2010). In this paper, F and G are estimated as $N/(N+1)$ times the marginal empirical distributions F_N and G_N of X and Y (where N is the sample size), and their density functions f and g, if they exist, are estimated with kernel estimators. The likelihood function with F, G, f and g replaced by their estimators is called the pseudo-likelihood function.

Suppose that F and G are absolutely continuous with densities f and g, respectively. The density functions of their kth and lth smallest order statistics with sample sizes m and n can be written as

$$\begin{aligned} f_{k:m}(x) &= \frac{d}{dx}F_{k:m}(x) = mb_{k-1,m-1}(F(x))f(x), \\ g_{l:n}(y) &= \frac{d}{dy}G_{l:n}(y) = nb_{l-1,n-1}(G(y))g(y). \end{aligned} \tag{5}$$

The density of Baker's bivariate distribution then can be written as

$$h(x,y;R) = \sum_{k=1}^{m}\sum_{l=1}^{n} r_{kl}f_{k:m}(x)g_{l:n}(y). \tag{6}$$

Using the copula density

$$c(u,v;R) = mn\sum_{k=1}^{m}\sum_{l=1}^{n} r_{kl}b_{k-1,m-1}(u)b_{l-1,n-1}(v),$$

density (6) has another expression

$$h(x,y;R) = c(F(x),G(y);R)f(x)g(y).$$

Suppose that an i.i.d. sample (x_i,y_i), $i=1,\ldots,N$, is obtained from Baker's distribution (6). According to the standard method for estimating a finite mixture distribution, we introduce a pair of unobserved variables (K_i,L_i) for observation i, with probability $\Pr(K_i=k, L_i=l) = r_{kl}$, $k\in\{1,\ldots,m\}$, $l\in\{1,\ldots,n\}$, $i=1,\ldots,N$. We also define an $m\times n$ matrix $\tau_i = (\tau_{i,kl})$ as a dummy variable with elements

$$\tau_{i,kl} = \begin{cases} 1 & ((K_i,L_i)=(k,l)), \\ 0 & ((K_i,L_i)\neq(k,l)). \end{cases} \tag{7}$$

Note that τ_i and (K_i,L_i) are one-to-one. The likelihood for the full dataset (x_i,y_i,τ_i), $i=1,\ldots,N$, is given by

$$\prod_{i=1}^{N}\prod_{k=1}^{m}\prod_{l=1}^{n}\left\{r_{kl}f_{k:m}(x_i)g_{l:n}(y_i)\right\}^{\tau_{i,kl}}. \tag{8}$$

The E-step in the EM algorithm calculates the conditional expectation given (x_i, y_i), $i = 1, \ldots, N$, that is,

$$\begin{aligned}\widehat{\tau}_{i,kl} &= E\big[\tau_{i,kl} \,|\, (x_i, y_i); R\big] = \frac{r_{kl} f_{k:m}(x_i) g_{l:n}(y_i)}{h(x_i, y_i; R)} \\ &= \frac{r_{kl} b_{k-1,m-1}(F(x_i)) b_{l-1,n-1}(G(y_i))}{c(F(x_i), G(y_i); R)}. \end{aligned} \tag{9}$$

The M-step maximizes the logarithm of the likelihood (8) by assuming $\tau_{i,kl} = \widehat{\tau}_{i,kl}$. The logarithm of the expectation of (8) divided by N is

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{m} \sum_{l=1}^{n} \widehat{\tau}_{i,kl} \log(r_{kl} f_{k:m}(x_i) g_{l:n}(y_i)) = \sum_{k=1}^{m} \sum_{l=1}^{n} \bar{\tau}_{kl} \log r_{kl} + \text{const.}, \tag{10}$$

where $\bar{\tau}_{kl} = \sum_{i=1}^{N} \widehat{\tau}_{i,kl}/N$.

Maximizing (10) is a convex problem and has a maximizer $R^* = (r^*_{kl})$, because (10) is a proper concave function in r_{kl} and the region of $R = (r_{kl})$ defined by (2) is convex. Moreover, if $\bar{\tau}_{kl} > 0$ for all k, l, the maximizer R^* is a (relatively) interior point of the region (2). In such a case, the maximizer R^* is obtained by the Lagrange multiplier method under the conditions $\sum_l r_{kl} = 1/m$, $\sum_k r_{kl} = 1/n$. We introduce Lagrange multipliers μ_k and λ_l, and maximize

$$L = \sum_{k=1}^{m} \sum_{l=1}^{n} \bar{\tau}_{kl} \log r_{kl} - \sum_k \mu_k \Big(\sum_l r_{kl} - \frac{1}{m} \Big) - \sum_l \lambda_l \Big(\sum_k r_{kl} - \frac{1}{n} \Big)$$

with respect to r_{kl}, μ_k and λ_l. Then, the maximizers r^*_{kl}, μ^*_k and λ^*_l are obtained as the solution of $\partial L/\partial r_{kl} = \bar{\tau}_{kl}/r_{kl} - \mu_k - \lambda_l = 0$ and $\sum_l r_{kl} = 1/m$, $\sum_k r_{kl} = 1/n$. To find μ^*_k and λ^*_l satisfying

$$r_{kl} = \frac{\bar{\tau}_{kl}}{\mu_k + \lambda_l} \tag{11}$$

as well as the restriction (2), we propose the following procedure:

Algorithm 2.1. *Step M0. Set* $\mu_k^{(0)} = \lambda_l^{(0)} = 1/2$.

Step M1. For fixed $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$*, find* $\boldsymbol{\lambda} = (\lambda_l)$ *numerically from*

$$\sum_{k=1}^{m} \frac{\bar{\tau}_{kl}}{\mu_k + \lambda_l} = \frac{1}{n}, \quad 1 \le l \le n.$$

Step M2. For fixed $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$*, find* $\boldsymbol{\mu} = (\mu_k)$ *numerically from*

$$\sum_{l=1}^{n} \frac{\bar{\tau}_{kl}}{\mu_k + \lambda_l} = \frac{1}{m}, \quad 1 \le k \le m.$$

Step M3. Update $\boldsymbol{\mu} = (\mu_k)$ *by*

$$\mu_k := \mu_k - \frac{1}{m} \Big(\sum_{k=1}^{m} \mu_k - \sum_{k=1}^{m} \mu_k^{(0)} \Big), \quad 1 \le k \le m.$$

Repeat Steps M1–M3 until (11) converges.

The EM algorithm is summarized as follows.

Algorithm 2.2. *Step 0. Set* $r_{kl} = 1/(mn)$.

Step 1. Find $\widehat{\tau}_{i,kl}$ *by (9) (E-step).*

Step 2. Update r_{kl} *by Algorithm 2.1, Step M0–M3 (M-step).*

Repeat Steps 1 and 2 until $\widehat{\tau}_{i,kl}$ *converges.*

2.2 The case where R is parameterized

If $R = (r_{kl})$ satisfying (2) is parameterized by a lower-dimensional parameter θ as $r_{kl} = r_{kl}(\theta)$, the estimation becomes simpler. For the case of $m = n$, for instance, Baker (2008) discussed a subclass of bivariate distributions with a distribution function

$$\begin{aligned} H^{\pm}(x, y; q, n) &= (1-q)F(x)G(y) + qH_n^{\pm}(x, y) \\ &= (1-q)F(x)G(y) + qC_n^{\pm}(F(x), G(y)), \quad 0 \le q \le 1, \end{aligned} \tag{12}$$

where

$$\begin{aligned} H_n^{+}(x, y) &= \frac{1}{n}\sum_{k=1}^{n} F_{k:n}(x)G_{k:n}(y) = C_n^{+}(F(x), G(y)), \\ C_n^{+}(u, v) &= n\sum_{k=1}^{n} B_{k-1,n-1}(u)B_{k-1,n-1}(v), \end{aligned}$$

and

$$\begin{aligned} H_n^{-}(x, y) &= \frac{1}{n}\sum_{k=1}^{n} F_{k:n}(x)G_{n-k+1:n}(y) = C_n^{-}(F(x), G(y)), \\ C_n^{-}(u, v) &= n\sum_{k=1}^{n} B_{k-1,n-1}(u)B_{n-k,n-1}(v). \end{aligned}$$

The densities of $H_n^{\pm}$ and $C_n^{\pm}$ are denoted by $h_n^{\pm}$ and $c_n^{\pm}$, if they exist. $H_n^{+}(x, y)$ (or $H_n^{-}(x, y)$) is the largest positive (or negative) correlation case among Baker's distributions with $m = n$. The rank correlation of $H_n^{\pm}$ is $\pm(n-1)/(n+1)$.

Function $H^{\pm}(x, y; q, n)$ is Baker's distribution (3) with

$$r_{kl} = \begin{cases} (1-q)/n^2 + q\delta_{kl}/n & (\text{for } H^{+}), \\ (1-q)/n^2 + q\delta_{k,n-l+1}/n & (\text{for } H^{-}), \end{cases} \qquad 1 \le k, l \le n,$$

where δ_{kl} is Kronecker's delta. r_{kl} is parameterized by the scalar parameter q. The parameter q adjusts the degree of independence. When $q = 0$, X and Y are independent. When $q > 0$, X and Y are positively (or negatively) correlated for the distribution H^{+} (or H^{-}). These models are expected to represent highly correlated distributions with fewer parameters than the original Baker's distribution.

In modeling the joint distribution functions, Baker (2008) chose the parameter q by minimizing the negative log-likelihood and the Kolmogorov-Smirnov statistic for a given set of n. Here we treat n as an integer-valued parameter to be estimated, and as an alternative, we propose an EM algorithm below to estimate the parameters (q, n) simultaneously. Suppose that an i.i.d. sample (x_i, y_i), $i = 1, \ldots, N$, is obtained from the continuous distribution $H_n^{+}(x, y; q, n)$ with the density

$$\begin{aligned} h_n^{+}(x, y; q, n) =&(1-q)f(x)g(y) + qh_n^{+}(x, y) \\ =&\{1 - q + qc_n^{+}(F(x), G(y))\}f(x)g(y). \end{aligned}$$

Algorithm 2.3. *Step 0. Set* $(q, n) = (1/2, 1)$.

Step 1. E-step:

$$\widehat{\tau}_i := \frac{(1-q)f(x_i)g(y_i)}{(1-q)f(x_i)g(y_i) + qh_n^{\pm}(x_i, y_i)} = \frac{1-q}{1 - q + qc_n^{\pm}(F(x_i), G(y_i))}, \tag{13}$$

$i = 1, \ldots, N$.

Step 2. M-step:

$$q := 1 - \frac{1}{N}\sum_{i=1}^{N}\widehat{\tau}_i, \qquad n := \operatorname*{argmax}_{n\in\mathbb{N}} \sum_{i=1}^{N}(1-\widehat{\tau}_i)\log\left(c_n^{\pm}(F(X_i), G(Y_i))\right).$$

Repeat Step 1 and Step 2 until (q, n) converges.

3 Illustrative examples

3.1 Consomic mouse data

In this section, we demonstrate how our algorithms work for real data analysis. The first data are measurements of blood concentrations of biochemical substances in mice (Takada and Shiroishi, 2012). We apply Algorithm 2.2 for fitting Baker's distribution (6) with continuous variables.

The dataset taken from consomic mouse strains of 314 10-week old females consists of measurements of triglycerides (TG) and plasma high-density lipoprotein cholesterol (HDL) as plotted in Figure 1.

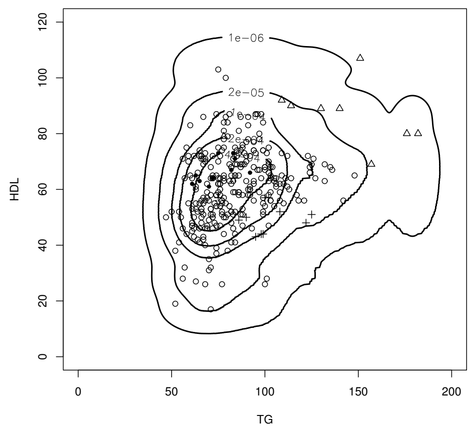

Figure 1: TG and HDL data (female consomic mice) and estimated contour.
(Dots: B6, Pluses: B6-Chr4MSM, Triangles: MSM, Circles: others.)

Using the Gaussian kernel estimator, we first estimate the marginal density functions. As described in Section 2.1, we use the empirical distribution functions to approximate the (cumulative) distribution functions. The Bernstein copula density (6) is estimated by the EM algorithm (Algorithm 2.2). In the algorithm, the matrix size of R is determined as $(m, n) = (2, 3)$ by the Akaike information criterion (AIC). The corresponding estimate of R is obtained as

$$\widehat{R} = \begin{pmatrix} 0.333 & 0.106 & 0.061 \\ 0.000 & 0.227 & 0.273 \end{pmatrix}.$$

With this $\widehat{R}$, we estimate the density (6) as $h(x, y; \widehat{R})$. A contour plot of the estimated joint density is shown in Figure 1.

3.2 Illinois state education data

The second example is to estimate the joint density function of the Illinois Standards Achievement Test (ISAT) scores which are available from the website of the Illinois State Board of

Education. We use the ISAT performance results for reading and mathematics in grade 3 of $N = 2991$ public schools and districts in 2009. For each school or district, percentages of students meeting or exceeding test standards are observed. The data are plotted in Figure 2 (left). Each point indicates a public school or district.

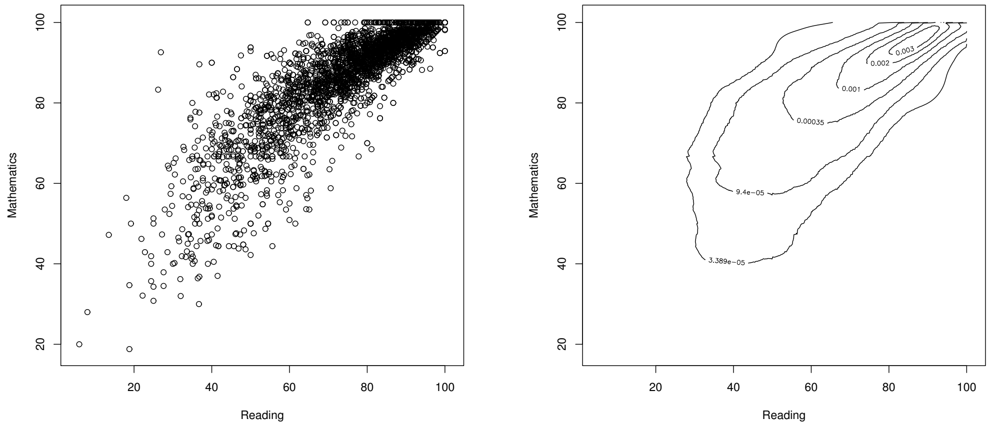

Figure 2: ISAT percent meeting or exceeding standards.
(Left: data plot. Right: estimated density contour.)

Because of the high correlation, we use the largest correlation model $H_n^+(x, y)$ in (12). The estimated density $h^+(x, y; \widehat{q} = 0.919, \widehat{n} = 17)$ is plotted in Figure 2 (right).

References

Baker, R. (2008). An order-statistics-based method for constructing multivariate distributions with fixed marginals, *Journal of Multivariate Analysis*, **99** (10), 2312–2327.

Choroś, B., Ibragimo, R. and Permiakov, E. (2010). Copula Estimation, in *Copula Theory and Its Applications* (P. Jaworski et al. eds.), Lecture Notes in Statistics 198, Springer, Heidelberg, pp. 77–90.

Genest, C., Ghoudi, K., and Rivest, L. O. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, **82** (3), 543–552.

Hwang, J. S. and Lin, G. D. (1984). Characterizations of distributions by linear combinations of moments of order statistics, *Bulletin of the Institute of Mathematics, Academia Sinica*, **12**, 179–202.

Illinois State Board of Education. *2008-09 ISAT/PSAE/ACT Performance Results*, `http://www.isbe.state.il.us/assessment/report_card.htm`

Takada, T. and Shiroishi, T. (2012). Complex quantitative traits cracked by the mouse inter-subspecific consomic strains, *Experimental Animals*, **61** (4), 375–388.