EM algorithms for estimating the Bernstein copula

Xiaoling Dou[a,*], Satoshi Kuriki[a], Gwo Dong Lin[b], Donald Richards[a,c]

[a] *The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan*
[b] *Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.*
[c] *Department of Statistics, Penn State University, University Park, PA 16802, U.S.A.*

Abstract

A method that uses order statistics to construct multivariate distributions with fixed marginals and which utilizes a representation of the Bernstein copula in terms of a finite mixture distribution is proposed. Expectation-maximization (EM) algorithms to estimate the Bernstein copula are proposed, and a local convergence property is proved. Moreover, asymptotic properties of the proposed semiparametric estimators are provided. Illustrative examples are presented using three real data sets and a 3-dimensional simulated data set. These studies show that the Bernstein copula is able to represent various distributions flexibly and that the proposed EM algorithms work well for such data.

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

1. Introduction

We consider a far-reaching idea of Baker (2008), who proposed a simple and intuitive method using order statistics for constructing multivariate distributions with given marginals. Baker's idea in the case of bivariate distributions can be stated as follows: Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from cumulative distribution functions F and G, respectively, where F and G can be continuous or discrete. By sorting the two

---

*Corresponding author.
Email addresses: `xiaoling@ism.ac.jp` (Xiaoling Dou), `kuriki@ism.ac.jp` (Satoshi Kuriki), `gdlin@stat.sinica.edu.tw` (Gwo Dong Lin), `richards@stat.psu.edu` (Donald Richards)

samples, we obtain the corresponding order statistics $X_{(1)} \leq \cdots \leq X_{(m)}$ and $Y_{(1)} \leq \cdots \leq Y_{(n)}$, respectively. Furthermore, independently of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, we choose K and L, uniformly distributed random numbers from the sets $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$, respectively; then it is straightforward to show that the respective marginal distributions of $X_{(K)}$ and $Y_{(L)}$ are F and G, the same as those of X_k and Y_l. The joint distribution of $(X_{(K)}, Y_{(L)})$, the pair of the Kth and Lth smallest order statistics from the individual X- and Y-samples, respectively, is called *Baker's bivariate distribution*, and Baker's multivariate distribution can be defined in a similar way.

According to the above construction, we see that Baker's bivariate distribution is parameterized by an $m \times n$ matrix parameter $R = (r_{k,l})$, where

$$r_{k,l} = \Pr(K = k, L = l), \quad 1 \leq k \leq m, \ 1 \leq l \leq n.$$

Because the marginal distributions of K and L are both uniform, we find that R satisfies the conditions,

$$\sum_{l=1}^{n} r_{k,l} = \frac{1}{m}, \quad \sum_{k=1}^{m} r_{k,l} = \frac{1}{n}, \quad r_{k,l} \geq 0 \quad \text{for } 1 \leq k \leq m, \ 1 \leq l \leq n. \quad (1)$$

If $r_{k,l} = 1/(mn)$ for all k and l, that is, if K and L are independent then $X_{(K)}$ and $Y_{(L)}$ are also independent. Otherwise, K and L are not independent and $(X_{(K)}, Y_{(L)})$ is a correlated bivariate random variable.

Let $F_{k:m}$ and $G_{l:n}$ be the marginal distribution functions of the order statistics $X_{(k)}$ and $Y_{(l)}$, respectively. Then, Baker's distribution is a finite mixture distribution of mn-components with distribution function

$$H(x, y; R) = \Pr\big(X_{(K)} \leq x, \ Y_{(L)} \leq y\big) = \sum_{k=1}^{m} \sum_{l=1}^{n} r_{k,l} F_{k:m}(x) G_{l:n}(y). \quad (2)$$

It is well-known that the distribution functions of order statistics can be described in terms of the Bernstein polynomials; see Baker (2008). Let the Bernstein polynomial and its cumulative 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 $F_{k:m}$ and $G_{l:n}$ can be expressed as $F_{k:m}(x) = mB_{k-1,m-1}(F(x))$ and $G_{l:n}(y) = nB_{l-1,n-1}(G(y))$ (Hwang and Lin,

1984, Eq. (1)). Substituting these results into (2), we have

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

where

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

We now recall that a 2-dimensional copula is an arbitrary bivariate distribution function on $[0, 1]^2$ whose marginals are the uniform distribution on $[0, 1]$. The importance of copulas in the study of multivariate distributions stems from Sklar's theorem: Any multivariate distribution function can be represented by a copula evaluated at the corresponding marginal distribution functions (Joe, 2001; Nelsen, 2006).

In fact, the function $C(u, v; R)$ in (4) is a copula, and therefore (3) expresses Baker's distribution explicitly in terms of a copula with arguments F and G and parameter R; hence, (3) provides an explicit formulation of Sklar's theorem for Baker's distribution. In this paper, we call $C(u, v; R)$ the *Bernstein copula* (Sancetta and Satchell, 2004).

It is also well-known that copulas are useful for describing multivariate distributions, and many copulas have been proposed for that purpose. Since a copula is a distribution function, we refer to its density as a *copula density*.

Among the class of copulas, the Bernstein copula has two remarkable features. First, because of the Weierstrass approximation theorem, any 2-dimensional copula can be approximated uniformly on $[0, 1]^2$ by the Bernstein copula density

$$c(u, v; R) = \frac{\partial^2}{\partial u \partial v} C(u, v; R), \tag{5}$$

when m and n are sufficiently large (Kingsley, 1951). Therefore, any continuous bivariate density function can be approximated by the density arising from the Bernstein copula. Taking advantage of this result, Sancetta and Satchell (2004) proposed an empirical Bernstein copula density estimator and studied its consistency in mean-square-error, and Janssen et al. (2012) derived the almost sure consistency and asymptotic normality properties of the empirical Bernstein copula density estimator.

A second remarkable feature of the Bernstein copula is that it is a finite mixture distribution, as stated in (2) or (4). Following the definition of Baker's distribution, it is easy to generate random numbers from the joint

distribution of $(X_{(K)}, Y_{(L)})$. Because copulas are used not only for analyzing existing data, but also for making predictions through Monte Carlo simulation, the simplicity of data generation increases the importance of the Bernstein copula for practical applications. Moreover, the finite mixture nature of the distribution function (3) allows us to apply the expectation-maximization (EM) algorithm to estimate parameters, and we propose such estimation methods in this paper.

We remark that the EM algorithm is a widely-used method for deriving maximum likelihood estimators (MLEs) numerically. For the purposes of estimating the Bernstein copula, we prefer the EM algorithm to direct methods such as the Newton-Raphson or quasi-Newton methods, because the EM algorithm has the advantages of being easy to implement, of not requiring the computation of gradients or Hessians, and of being generally stable and not overly sensitive to starting values even when there exist multiple local maxima of the likelihood function (McLachlan and Krishnan, 2008).

The paper is organized as follows. In Section 2, EM algorithms for estimating parameters are proposed in various settings. For the first EM algorithm, we prove the local convergence of the M-step, and asymptotic properties of the proposed estimators as semiparametric estimators are provided. In Section 3, we provide illustrative examples based on Baker's distribution, and we illustrate behavior of the proposed algorithms using real-world and simulated data. Further, some additional properties of the proposed algorithms and two topics for future research are discussed in Section 4, and we provide related mathematical details in the Appendix.

2. EM algorithms based on the pseudo-likelihood function

As we have noted, Baker's distribution is a finite mixture distribution, and hence EM algorithm methodology can be applied for maximum likelihood estimation (McLachlan and Peel, 2000). Throughout this paper, we assume that the marginal distributions F and G have been estimated in advance, and we shall treat them in the subsequent analysis as known functions; this widely-used two-stage estimation procedure is referred to as the *semiparametric method* (Genest et al., 1995; Charpentier et al., 2007; Kim et al., 2007; Choroś et al., 2010).

On the basis of a random sample of size N on (X, Y), let F_N and G_N denote the marginal empirical distributions of X and Y. Throughout the paper, we take F and G to be estimated by $NF_N/(N+1)$ and $NG_N/(N+1)$,

respectively. If f and g, the corresponding density functions of F and G, exist then we estimate them with kernel estimators (see Section 3). The likelihood function with F, G, f and g replaced by their corresponding estimators is called the *pseudo-likelihood function*.

2.1. The continuous case

In this subsection, we suppose that X and Y are continuous random variables, and that F and G are absolutely continuous with densities f and g, respectively. The density functions of their kth and lth smallest order statistics, based on random samples of sizes m and n respectively, can be written as

$$
\begin{aligned}
f_{k:m}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}F_{k:m}(x) = mb_{k-1,m-1}(F(x))f(x), \\
g_{l:n}(y) &= \frac{\mathrm{d}}{\mathrm{d}y}G_{l:n}(y) = nb_{l-1,n-1}(G(y))g(y).
\end{aligned}
\tag{6}
$$

It now follows from (2) that the density of Baker's bivariate distribution can be written as

$$
h(x, y; R) = \sum_{k=1}^{m}\sum_{l=1}^{n} r_{k,l}f_{k:m}(x)g_{l:n}(y).
\tag{7}
$$

By applying (5) to (4), we obtain the copula density

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

and then it follows from Sklar's theorem that the density (7) has an alternative expression,

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

Suppose that an independent, identically distributed (i.i.d.) sample (x_i, y_i), $i = 1, \ldots, N$, is obtained from Baker's distribution (7). 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 $\Pr(K_i = k, L_i = l) = r_{k,l}$, $k \in \{1, \ldots, m\}$, $l \in \{1, \ldots, n\}$, $i = 1, \ldots, N$. We also define an $m \times n$ matrix $\tau_i = (\tau_{i,k,l})$ as a dummy variable with elements

$$
\tau_{i,k,l} =
\begin{cases}
1, & \text{if } (K_i, L_i) = (k, l), \\
0, & \text{if } (K_i, L_i) \neq (k, l)
\end{cases}
\tag{8}
$$

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

$$\prod_{i=1}^{N} \prod_{k=1}^{m} \prod_{l=1}^{n} \{r_{k,l} f_{k:m}(x_i) g_{l:n}(y_i)\}^{\tau_{i,k,l}}. \tag{9}$$

The E-step in the EM algorithm calculates the conditional expectation of $\tau_{i,k,l}$ given (x_i, y_i), $i = 1, \ldots, N$; that is,

$$\begin{aligned} \widehat{\tau}_{i,k,l} &= E\big[\tau_{i,k,l} \,|\, (x_i, y_i)_{1 \le i \le N}; R\big] \\ &= \frac{r_{k,l} f_{k:m}(x_i) g_{l:n}(y_i)}{h(x_i, y_i; R)} \\ &= \frac{r_{k,l} b_{k-1,m-1}(F(x_i)) b_{l-1,n-1}(G(y_i))}{c(F(x_i), G(y_i); R)}. \end{aligned} \tag{10}$$

The M-step maximizes the logarithm of the likelihood (9) with respect to $r_{k,l}$ by assuming $\tau_{i,k,l} = \widehat{\tau}_{i,k,l}$. The logarithm of the expectation of (9) divided by N is

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{m} \sum_{l=1}^{n} \widehat{\tau}_{i,k,l} \log(r_{k,l} f_{k:m}(x_i) g_{l:n}(y_i)) = \sum_{k=1}^{m} \sum_{l=1}^{n} \bar{\tau}_{k,l} \log r_{k,l} + \text{const.}, \tag{11}$$

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

Maximizing the function (11) is a convex problem which has a unique maximizer $R^* = (r_{k,l}^*)$ because (11) is a proper concave function in $r_{k,l}$ and the region for $R = (r_{k,l})$ defined by (1) is convex. Moreover, if $\bar{\tau}_{k,l} > 0$ for all k, l then the maximizer R^* is a (relative) interior point of the region (1); in that case, the maximizer R^* is obtained by the Lagrange multiplier method under the conditions $\sum_{l=1}^{n} r_{k,l} = 1/m$, $\sum_{k=1}^{m} r_{k,l} = 1/n$ for all k and l.

We introduce Lagrange multipliers μ_k and λ_l, and proceed to maximize

$$L = \sum_{k=1}^{m} \sum_{l=1}^{n} \bar{\tau}_{k,l} \log r_{k,l} - \sum_{k} \mu_k \left(\sum_{l} r_{k,l} - \frac{1}{m} \right) - \sum_{l} \lambda_l \left(\sum_{k} r_{k,l} - \frac{1}{n} \right)$$

with respect to $r_{k,l}$, μ_k and λ_l. Then, the maximizers $r_{k,l}^*$, μ_k^* and λ_l^* are obtained as the solution of

$$\frac{\partial L}{\partial r_{k,l}} = \frac{\bar{\tau}_{k,l}}{r_{k,l}} - \mu_k - \lambda_l = 0$$

subject to the restrictions in (1).

To find μ_k^* and λ_l^* satisfying

$$r_{k,l} = \frac{\bar{\bar{\tau}}_{k,l}}{\mu_k + \lambda_l} > 0 \tag{12}$$

as well as the restriction (1), we propose the following procedure:

Algorithm 2.1.

Step M0: Set $\mu_k^{(0)} = 1/2$ and $t = 0$.

Step M1: For fixed $\boldsymbol{\mu}^{(t)} = \left(\mu_1^{(t)}, \ldots, \mu_m^{(t)}\right)'$, and for $1 \leq l \leq n$, find $\lambda_l^{(t)}$ numerically as a unique solution λ_l of

$$\sum_{k=1}^{m} \frac{\bar{\bar{\tau}}_{k,l}}{\mu_k^{(t)} + \lambda_l} = \frac{1}{n} \quad \text{such that} \quad \lambda_l > -\min_k\left(\mu_k^{(t)}\right).$$

Step M2: For fixed $\boldsymbol{\lambda}^{(t)} = \left(\lambda_1^{(t)}, \ldots, \lambda_n^{(t)}\right)'$, and for $1 \leq k \leq m$, find $\widetilde{\mu}_k^{(t)}$ numerically as a unique solution $\widetilde{\mu}_k$ of

$$\sum_{l=1}^{n} \frac{\bar{\bar{\tau}}_{k,l}}{\widetilde{\mu}_k + \lambda_l^{(t)}} = \frac{1}{m} \quad \text{such that} \quad \widetilde{\mu}_k > -\min_l\left(\lambda_l^{(t)}\right).$$

Step M3: Let

$$\mu_k^{(t)} = \widetilde{\mu}_k^{(t)} - \frac{1}{m}\left(\sum_{k=1}^{m} \widetilde{\mu}_k^{(t)} - \sum_{k=1}^{m} \mu_k^{(0)}\right), \quad 1 \leq k \leq m.$$

Increase the counter t by 1, and repeat Steps M1–M3 until (12) converges.

Remark 2.1. *In Step M1, the equation $\sum_{k=1}^{m} \bar{\bar{\tau}}_{k,l}/(\mu_k^{(t)} + \lambda_l) = 1/n$ in λ_l has at most m solutions. The solution of (12) necessarily satisfies $\mu_k + \lambda_l > 0$ for all (k, l), and hence, a solution λ_l can be chosen to satisfy $\lambda_l > -\min_k\left(\mu_k^{(t)}\right)$. This solution is unique because the function $\sum_{k=1}^{m} \bar{\bar{\tau}}_{k,l}/(\mu_k^{(t)} + \lambda_l)$ is monotonically decreasing in λ_l and takes the values ∞, as $\lambda_l \downarrow -\min_k\left(\mu_k^{(t)}\right)$, and 0, as $\lambda_l \uparrow \infty$. Such λ_l can be found numerically by the bisection method (Dennis and Schnabel, 1996). Moreover, similar remarks apply to Step M2.*

Remark 2.2. *We note that if μ_k and λ_l are solutions of (12) then $\mu_k + c$ and $\lambda_l - c$ are also solutions of (12) for any constant c. Therefore, Step M3 is needed to remove this redundancy.*

The following proposition states that Algorithm 2.1 converges locally, and a proof is given in the Appendix. An empirical study suggests that this algorithm also has the global convergence property (see Section 3.1), however it remains an open problem to derive a proof of the global convergence property.

Proposition 2.1. *Suppose that $\bar{\tau}_{k,l} > 0$. Then Algorithm 2.1 has the property of locally linear convergence. That is, there exist positive constants c and d such that if $\|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^*\| \leq c$ and $\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}^*\| \leq d$, then the sequences $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ $(t = 0, 1, \ldots)$ generated from Algorithm 2.1 converge to the solutions $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^*$, respectively. Moreover, convergence is attained with the convergence rates $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\| = O(\nu^t)$ and $\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*\| = O(\nu^t)$, as $t \to \infty$, for a positive constant $\nu \in (0, 1)$, and the constants c, d and ν depend on $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$.*

To apply the EM algorithm, we will use as an initial value of $r_{k,l}$ the estimator given by Sancetta and Satchell (2004) and Janssen et al. (2012) (see Remark 2.4). Then the EM algorithm is summarized as follows.

Algorithm 2.2.
Step 0: Set $r_{k,l}$ equal to $\widetilde{r}_{k,l}$ in (13).
Step 1: Find $\widehat{\tau}_{i,k,l}$ by (10) (E-step).
Step 2: Update $r_{k,l}$ by Algorithm 2.1, Steps M0–M3 (M-step).
Repeat Steps 1 and 2 until $\widehat{\tau}_{i,k,l}$ converges.

Note that this algorithm can be extended to Baker's distributions with three or more variables.

Remark 2.3. *In (10), a common factor $f(x_i)g(y_i)$ is canceled in the numerator and the denominator. This is reasonable because the EM algorithm should be equivalent to the one based on the sample $(F(x_i), G(y_i))$, $i = 1, \ldots, N$, having uniform marginals.*

Genest et al. (1995) developed an asymptotic theory for semiparametric estimation of copulas based on the pseudo-likelihood function, and Tsukahara (2005) later extended that theory to M-estimation. Their results can be applied to our problem. Let

$$c_u(u, v; R) = mn \sum_{k=1}^{m} \sum_{l=1}^{n} r_{k,l} \frac{\mathrm{d}}{\mathrm{d}u} b_{k-1,m-1}(u) b_{l-1,n-1}(v),$$

$$c_v(u, v; R) = mn \sum_{k=1}^{m} \sum_{l=1}^{n} r_{k,l} b_{k-1,m-1}(u) \frac{\mathrm{d}}{\mathrm{d}v} b_{l-1,n-1}(v)$$

be the derivatives of $c(u, v)$ with respect to u and v. To calculate these derivatives, we can use the formula

$$\frac{\mathrm{d}}{\mathrm{d}u}b_{k,n}(u) = n\{b_{k-1,n-1}(u) - b_{k,n-1}(u)\},$$

where we set $b_{-1,n-1}(u) = b_{n,n-1}(u) \equiv 0$ to initialize the recurrence relation.

Proposition 2.2. *Suppose that the true value of the parameter $R = (r_{k,l})$ in (1) satisfies $r_{k,l} > 0$. As $N \to \infty$ the MLE, $\widehat{R}$, of R is an asymptotically normally $\sqrt{N}$-consistent estimator.*

Further, a consistent estimator of $\mathrm{NVar}(\widehat{R}) = \Sigma = (\sigma_{(k,l),(k',l')})$, an $mn \times mn$ matrix with lexicographic index (k, l), is given by $\widehat{\Sigma} = B^{+}SB^{+}$, where B and S are the sample covariance matrices of the $mn \times 1$ pseudo-observation vectors $\boldsymbol{u}_i = (u_{i,(k,l)})$ and $\boldsymbol{v}_i = (v_{i,(k,l)})$ defined by

$$u_{i,(k,l)} = \frac{mn\, b_{k-1,m-1}(F(x_i))b_{l-1,n-1}(G(y_i))}{c(F(x_i), G(y_i); \widehat{R})},$$

$$
\begin{aligned}
v_{i,(k,l)} = {}& u_{i,(k,l)} \\
& - \frac{mn}{N} \sum_{j:x_i \leq x_j} \frac{b_{k-1,m-1}(F(x_j))b_{l-1,n-1}(G(y_j))c_u(F(x_j), G(y_j); \widehat{R})}{c(F(x_j), G(y_j); \widehat{R})^2} \\
& - \frac{mn}{N} \sum_{j:y_i \leq y_j} \frac{b_{k-1,m-1}(F(x_j))b_{l-1,n-1}(G(y_j))c_v(F(x_j), G(y_j); \widehat{R})}{c(F(x_j), G(y_j); \widehat{R})^2},
\end{aligned}
$$

$i = 1, \ldots, N$, respectively, and B^{+} is the Moore-Penrose pseudo-inverse matrix of B.

Note that the matrix B is the observed Fisher information matrix when the marginals F and G are known. As with many semiparametric estimators, $\widehat{R}$ is not efficient in the sense that its asymptotic variance is larger than that given by the Fisher information matrix when the marginals are known (unless $r_{k,l} \equiv 1/(mn)$, i.e., $c(u, v; R) = 1$); see Genest and Werker (2002) regarding the inefficiency of the Farlie-Gumbel-Morgenstern (FGM) copula estimator.

Once the estimator $\widehat{R} = (\widehat{r}_{k,l})$ has been obtained by Algorithms 2.1 and 2.2, the estimate of $h(x, y; R)$ for a fixed point (x, y) is given by

$$h(x, y; \widehat{R}) = \sum_{k=1}^{m}\sum_{l=1}^{n}\widehat{r}_{k,l}f_{k:m}(x)g_{l:n}(y)$$

9

and its asymptotic variance is evaluated as

$$\text{Var}\big(h(x, y; \widehat{R})\big) \approx \frac{1}{N} \sum_{k,k'=1}^{m} \sum_{l,l'=1}^{n} f_{k:m}(x) f_{k':m}(x) g_{l:n}(y) g_{l':n}(y) \widehat{\sigma}_{(k,l),(k',l')},$$

where $\widehat{\sigma}_{(k,l),(k',l')}$ is the $((k, l), (k', l'))$th element of an $mn \times mn$ matrix $\widehat{\Sigma} = (\widehat{\sigma}_{(k,l),(k',l')})$.

Remark 2.4. *Sancetta and Satchell (2004) and Janssen et al. (2012) proposed estimating $r_{k,l}$ by*

$$\widetilde{r}_{k,l} = \#\left\{ i \; \Big| \; \frac{k-1}{m} < \frac{N}{N+1} F_N(x_i) \leq \frac{k}{m}, \; \frac{l-1}{n} < \frac{N}{N+1} G_N(y_i) \leq \frac{l}{n} \right\} \Big/ N \tag{13}$$

as $m, n \to \infty$, where the sample size $N \to \infty$ also. For the case in which m and n are fixed, this estimator is inconsistent in our setting because

$$\lim_{N \to \infty} E\big[\widetilde{r}_{k,l}\big] = C\Big(\frac{k}{m}, \frac{l}{n}; R\Big) - C\Big(\frac{k}{m}, \frac{l-1}{n}; R\Big)$$
$$- C\Big(\frac{k-1}{m}, \frac{l}{n}; R\Big) + C\Big(\frac{k-1}{m}, \frac{l-1}{n}; R\Big)$$

is not equal to $r_{k,l}$, in general.

2.2. The discrete case

The EM algorithm in Section 2.1 is also applicable for the case in which both F and G are discrete distributions. Suppose that F and G are supported on discrete sets A and B, respectively. For simplicity, we suppose that A and B are finite sets, with cardinalities $|A|$ and $|B|$, respectively. Then the data (x_i, y_i), $i = 1, \ldots, N$, can be represented by an $|A| \times |B|$ ordered categorical table $(N_{a,b})$, where

$$N_{a,b} = \#\big\{ i \in \{1, \ldots, N\} \mid (x_i, y_i) = (a, b) \big\}, \quad a \in A, \; b \in B.$$

In this subsection, we modify the EM algorithm of Section 2.1 so that Baker's distribution (2) can be applied to the data $(N_{a,b})_{a \in A, \, b \in B}$.

The probability functions of X and Y are $f(a) = \Pr(X = a) = F(a) - F(a-)$, $a \in A$ and $g(b) = \Pr(Y = b) = G(b) - G(b-)$, $b \in B$, respectively.

The probability functions of $X_{(k)}$ and $Y_{(l)}$, the kth and lth order statistics, are $f_{k:m}$ and $g_{l:n}$, respectively, where

$$
\begin{aligned}
f_{k:m}(a) &= F_{k:m}(a) - F_{k:m}(a-) \\
&= m\{B_{k-1,m-1}(F(a)) - B_{k-1,m-1}(F(a-))\}, \\
g_{l:n}(b) &= G_{l:n}(b) - G_{l:n}(b-) \\
&= n\{B_{l-1,n-1}(G(b)) - B_{l-1,n-1}(G(b-))\}.
\end{aligned}
\tag{14}
$$

Using these results, we obtain the joint probability function of (X, Y) in the form

$$
h(a, b; R) = \Pr(X = a, Y = b) = \sum_{k=1}^{m}\sum_{l=1}^{n} r_{k,l} f_{k:m}(a) g_{l:n}(b).
$$

We introduce a dummy variable $\eta_{a,b,k,l} = \sum_{i:(x_i,y_i)=(a,b)} \tau_{i,k,l}$ with $\tau_{i,k,l}$ defined in (8). The likelihood for the full data (9) is rewritten as

$$
\prod_{a \in A}\prod_{b \in B}\prod_{k=1}^{m}\prod_{l=1}^{n}\{r_{k,l} f_{k:m}(a) g_{l:n}(b)\}^{\eta_{a,b,k,l}}.
$$

The E-step for updating $\eta_{a,b,k,l}$ becomes

$$
\widehat{\eta}_{a,b,k,l} = E\big[\eta_{a,b,k,l} \,|\, (N_{a,b})_{a \in A,\, b \in B}; R\big] = \frac{N_{a,b} r_{k,l} f_{k:m}(a) g_{l:n}(b)}{\sum_{k=1}^{m}\sum_{l=1}^{n} r_{k,l} f_{k:m}(a) g_{l:n}(b)}.
$$

By letting $\bar{\tau}_{k,l} = \sum_{a \in A}\sum_{b \in B} \widehat{\eta}_{a,b,k,l}/N$, the M-step is obtained in the same form in Section 2.1, which is Step 2 of Algorithm 2.2.

2.3. The mixed case

We can also resolve by the same approach the case in which one variable is continuous and the other is discrete. Suppose that X is continuous with density function f and Y is discrete with distribution function G. Then, the density function of $X_{(k)}$ and the probability function of $Y_{(l)}$ are given by $f_{k:m}$ in (6) and $g_{l:n}$ in (14), respectively. The joint density function becomes

$$
\begin{aligned}
h(x, b; R) &= \frac{\Pr(X \in (x, x + \mathrm{d}x), Y = b)}{\mathrm{d}x} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{n} r_{k,l} f_{k:m}(x) g_{l:n}(b) \\
&= mn \sum_{k=1}^{m}\sum_{l=1}^{n} r_{k,l} b_{k-1,m-1}(F(x)) f(x)\{B_{l-1,n-1}(G(b)) - B_{l-1,n-1}(G(b-))\}.
\end{aligned}
$$

11

The E-step is the updating rule,

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

and the M-step remains unchanged.

2.4. The case in which R is parameterized

If $R = (r_{k,l})$ satisfying (1) is parameterized by a lower-dimensional parameter θ as $r_{k,l} = r_{k,l}(\theta)$ then the estimation becomes simpler. For the case in which $m = n$, for instance, Baker (2008) discussed a subclass of bivariate distributions with a distribution function

$$
\begin{aligned}
H^{\pm}(x, y; q, n) &= (1 - q)F(x)G(y) + qH_n^{\pm}(x, y) \\
&= (1 - q)F(x)G(y) + qC_n^{\pm}(F(x), G(y)), \quad 0 \le q \le 1, \quad (15)
\end{aligned}
$$

where

$$
H_n^{+}(x, y) = \frac{1}{n} \sum_{k=1}^{n} F_{k:n}(x) G_{k:n}(y) = C_n^{+}(F(x), G(y)),
$$

$$
C_n^{+}(u, v) = n \sum_{k=1}^{n} B_{k-1,n-1}(u) B_{k-1,n-1}(v),
$$

and

$$
H_n^{-}(x, y) = \frac{1}{n} \sum_{k=1}^{n} F_{k:n}(x) G_{n-k+1:n}(y) = C_n^{-}(F(x), G(y)),
$$

$$
C_n^{-}(u, v) = n \sum_{k=1}^{n} B_{k-1,n-1}(u) B_{n-k,n-1}(v).
$$

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

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

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

where $\delta_{k,l}$ denotes Kronecker's delta. The term $r_{k,l}$ is parameterized by the scalar parameter q which adjusts the degree of independence between X and Y. Indeed, if $q = 0$ then X and Y are independent; and if $q > 0$ then X and Y are positively (respectively, negatively) correlated for the distribution H^+ (respectively, H^-). These models are expected to represent highly correlated distributions with fewer parameters than the original Baker's distribution.

Baker's distribution was originally proposed as an extension to the FGM distribution with the limitation that its correlation does not exceed $1/3$ for the case of continuous marginals (Schucany et al., 1978). Hence, the range of correlation of Baker's distribution has gathered attention and the extreme correlation cases, $H_n^{\pm}(x, y)$, are well-studied.

For the distribution $H_n^+(x, y)$, Lin and Huang (2010) investigated convergence conditions and the convergence rate, as $n \to \infty$, of the correlation converging to the maximum correlation of the Fréchet–Hoeffding upper bound. Dou et al. (2013) proved the TP$_2$ property and derived the limiting distribution of (X, U), $U = \sqrt{n}(F(X) - G(Y))$, where (X, Y) are distributed as $H_n^+(x, y)$. Huang et al. (2013) proved that the copula $C_n^+(u, v)$ in the largest correlation case with u, v fixed is a non-decreasing function of n.

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 values 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), $i = 1, \dots, N$, 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:

$$
\widehat{\tau}_i := \frac{(1-q)f(x_i)g(y_i)}{(1-q)f(x_i)g(y_i) + qh_n^{\pm}(x_i, y_i)} = \frac{1-q}{1 - q + qc_n^{\pm}(F(x_i), G(y_i))}, \tag{16}
$$

$i = 1, \ldots, N$.

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 Steps 1 and 2 until (q, n) converges.

The asymptotic variance of $\widehat{q}$ is evaluated approximately as $s/(N\beta^2)$, where β and s are the sample variances of the pseudo-observations u_i and v_i defined by

$$u_i = \left. \frac{-1 + c_n^{\pm}(F(x_i), G(y_i))}{1 - \widehat{q} + \widehat{q} c_n^{\pm}(F(x_i), G(y_i))} \right|_{n=\widehat{n}},$$

$$v_i = u_i - \frac{\widehat{q}}{N} \sum_{j:x_i \leq x_j} \left. \frac{\{-1 + c_n^{\pm}(F(x_j), G(y_j))\} \frac{\partial}{\partial u} c_n^{\pm}(F(x_j), G(y_j))}{\{1 - \widehat{q} + \widehat{q} c_n^{\pm}(F(x_j), G(y_j))\}^2} \right|_{n=\widehat{n}}$$

$$- \frac{\widehat{q}}{N} \sum_{j:y_i \leq y_j} \left. \frac{\{-1 + c_n^{\pm}(F(x_j), G(y_j))\} \frac{\partial}{\partial v} c_n^{\pm}(F(x_j), G(y_j))}{\{1 - \widehat{q} + \widehat{q} c_n^{\pm}(F(x_j), G(y_j))\}^2} \right|_{n=\widehat{n}},$$

$i = 1, \ldots, N$.

For the case in which both X and Y are discrete distributions, the E-step (16) in Algorithm 2.3 is replaced by

$$\widehat{\tau}_i := \frac{(1-q)f(x_i)g(y_i)}{(1-q)f(x_i)g(y_i) + (q/n) \sum_{k=1}^{n} f_{k:n}(x_i)g_{k:n}(y_i)}, \tag{17}$$

where $f(x_i) = F(x_i) - F(x_i-)$, $g(y_i) = G(y_i) - G(y_i-)$, and $f_{k:n}(x_i)$ and $g_{k:n}(y_i)$ are given in (14). If the joint distribution consists of both continuous and discrete variables, then their respective density (6) and probability function (14) should be used.

3. Four illustrative examples

In this section, we demonstrate how our algorithms perform in practical data analysis for a wide range of examples. The results show that the algorithms work well in all the illustrative examples.

3.1. Consomic mouse data

The first data set, consisting of measurements of blood concentrations of biochemical substances in mice, is available from Takada et al. (2012). We apply Algorithm 2.2 for fitting Baker's distribution (7) with continuous variables.

The data set consists of measurements of triglycerides (TG) and plasma high-density lipoprotein cholesterol (HDL) as plotted in Figure 2. The variables TG and HDL are important indicators of metabolic syndrome and are correlated with the pathogenesis of cardiovascular disease in humans. To detect the genes responsible for adiposity, TG and HDL data are taken from consomic mouse strains of 314 10-week old females. A consomic strain is an artificial inbred strain with one specified chromosome replaced by another chromosome from a different inbred strain (Takada and Shiroishi, 2012). For example, the label B6-Chr4MSM appearing in Figure 2 means that a consomic strain has all chromosomes from the mouse strain C57BL/6 (B6) except for chromosome 4, which is from the mouse strain MSM/Ms (MSM). The data are taken from 30 kinds of consomic strains including pure strains B6 and MSM, and hence are fairly heterogeneous.

Using the Gaussian kernel estimator, we first estimate the marginal density functions. The bandwidths are selected according to Silverman's "rule of thumb" (Silverman, 1986). As described in Section 2.1, we use the empirical distribution functions to approximate the (cumulative) distribution functions. The estimated marginal densities and distribution functions are shown in the left and right panels, respectively, of Figure 1. Subsequently, we estimate the Bernstein copula density (7) with the EM algorithm (Algorithm 2.2) for fixed m and n. In the estimation, we determine the matrix size of R by the Akaike information criterion (AIC). From Table 1, we find that the AIC attains its minimum value, 5210.52, when $(m, n) = (2, 3)$. Table 1 also shows that the cases in which $(m, n) = (2, 2)$ and $(m, n) = (2, 3)$ have very close AIC values; indeed, the estimated contours based on these two cases are very similar.

For the case in which $(m, n) = (2, 3)$, the initial value $\widetilde{R}$ in (13) and the MLE $\widehat{R}$ obtained as the limit of sequence starting from $\widetilde{R}$ are

$$\widetilde{R} = \begin{pmatrix} 0.232 & 0.137 & 0.118 \\ 0.099 & 0.188 & 0.226 \end{pmatrix} \quad \text{and} \quad \widehat{R} = \begin{pmatrix} 0.333 & 0.106 & 0.061 \\ 0.000 & 0.227 & 0.273 \end{pmatrix},$$

respectively. The consistent estimates of the covariance of $(\widehat{r}_{11}, \widehat{r}_{12}, \widehat{r}_{13}, \widehat{r}_{21}, \widehat{r}_{22}, \widehat{r}_{23})'$
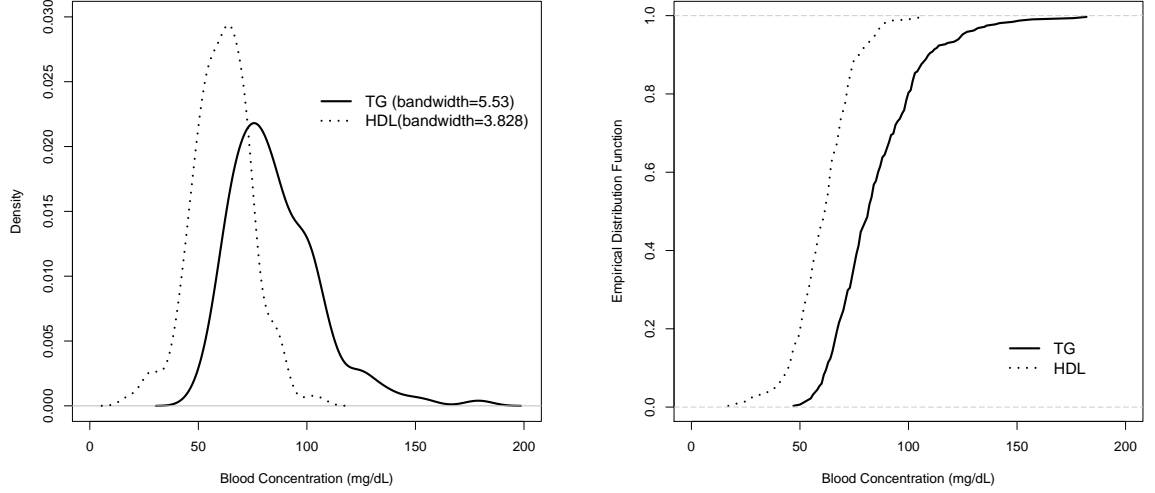
15

Figure 1: Estimated marginals of TG and HDL.
(Left: density functions. Right: cumulative distribution functions.)

calculated by Proposition 2.2 is

$$
\begin{pmatrix}
0.003 & -0.001 & -0.003 & -0.001 & 0.001 & -0.002 \\
-0.001 & 0.003 & 0.003 & -0.002 & 0.000 & 0.002 \\
-0.003 & 0.003 & 0.010 & -0.003 & -0.004 & 0.003 \\
-0.001 & -0.002 & -0.003 & 0.006 & 0.002 & -0.003 \\
0.001 & 0.000 & -0.004 & 0.002 & 0.003 & -0.002 \\
-0.002 & 0.002 & 0.003 & -0.003 & -0.002 & 0.005
\end{pmatrix}.
$$

A contour plot of the estimated joint density $h(x, y; \widehat{R})$ is shown in Figure 2.

In fitting models to the data, we checked the convergence of the algorithms when the starting points vary. Throughout the estimating procedure, two types of convergence sequences are generated; one is from Algorithm 2.1 and the other is from Algorithm 2.2. We investigate these convergences for the case in which $(m, n) = (2, 3)$ as follows.

For the first type of sequence, we varied the starting points $\boldsymbol{\mu}^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}\right)$ as $(1/2, 1/2)$ (as indicated in Algorithm 2.1) and 20 pairs of random variables distributed uniformly on $(0, 1)^2$. We find that all of these sequences converge to the same limit.

16

Table 1: AIC for female consomic mouse data.
(The minimum AIC is indicated with a box.)

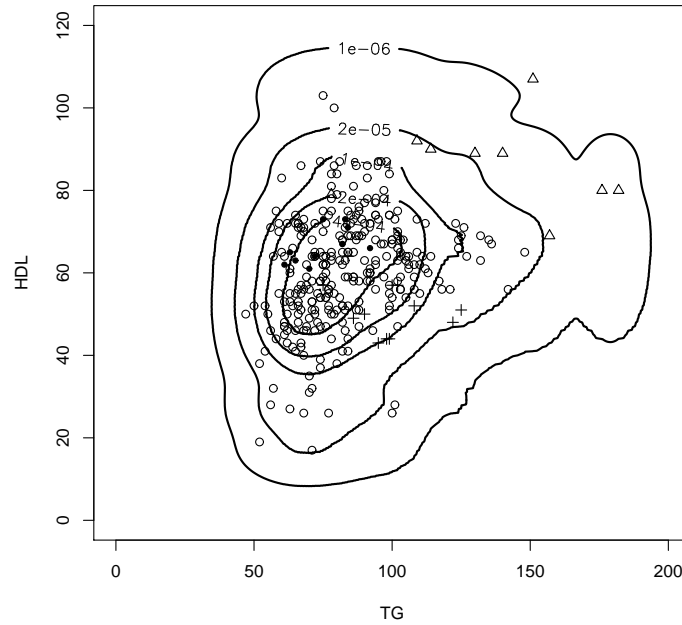| $m \setminus n$	1	2	3	4	5	6	8	10
1	5242.00	5242.00	5242.00	5242.00	5242.00	5242.00	5242.00	5242.00
2	5242.00	5210.57	5210.52	5212.15	5211.15	5210.80	5212.67	5216.23
3	5242.00	5212.55	5211.94	5214.22	5215.47	5217.64	5223.91	5230.53
4	5242.00	5214.56	5215.69	5219.16	5220.33	5224.48	5234.19	5244.29
5	5242.00	5215.37	5218.51	5223.65	5226.87	5232.10	5246.20	5259.89
6	5242.00	5216.59	5220.58	5225.99	5231.44	5238.67	5256.04	5273.77
8	5242.00	5218.77	5225.45	5233.77	5242.13	5253.45	5277.90	5302.69
10	5242.00	5221.55	5229.78	5241.92	5253.58	5268.72	5300.85	5332.22

Figure 2: TG and HDL data (female consomic mice) and estimated contour.
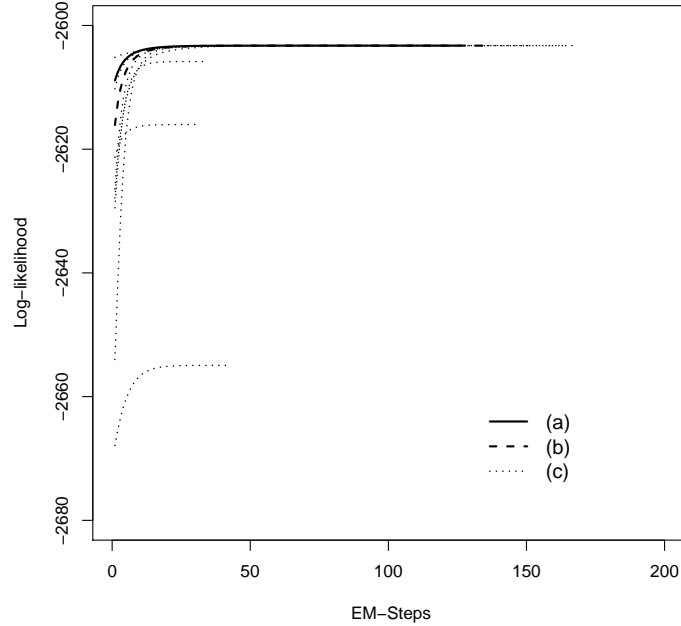(Dots: B6, Pluses: B6-Chr4MSM, Triangles: MSM, Circles: others.)

17

Figure 3: Convergence of the EM algorithm with various starting points.
(a) $R^{(0)} = \widehat{R}$; (b) $r_{k,l}^{(0)} \equiv 1/6$; (c) $\left(r_{1,1}^{(0)}, r_{1,2}^{(0)}\right) = (0.19, 0)$, $(0.15, 0.07)$, $(0.04, 0.15)$, $(0.22, 0.15)$, $(0.30, 0.19)$, $(0, 0.22)$, $(0.07, 0.26)$, $(0.11, 0.33)$.

For the second type of sequence, we have provided Figure 3 to confirm that the estimate obtained is the global maximum. This figure depicts how the likelihoods increase when the EM algorithm starts from different starting points. As the starting point $R^{(0)} = \left(r_{k,l}^{(0)}\right)$, we chose: (a) the estimator $\widetilde{R}$ in (13) of Remark 2.4 (as indicated in Algorithm 2.2); (b) $r_{k,l}^{(0)} \equiv 1/6$; and (c) 8 points randomly chosen from the R-region defined by (1) (see the legend of Figure 3). From this figure, we can see that the limit starting from the estimator $\widetilde{R}$ in (13) attains the maximum of the likelihood function.

We also conducted a maximization of the likelihood function by means of a numerical grid search for R with $(m, n) = (2, 3)$. Our calculations indicate that the maximum likelihood obtained by the EM algorithm is the global

18

Table 2: 2009 ISAT (Illinois Standards Achievement Test).
The percentage of student scores meeting or exceeding standards in
reading and mathematics, Grade 3 for $N = 2991$ schools and districts.

District name/ School name	Reading	Mathematics	County
Payson CUSD 1	78.6	88.4	Adams
Seymour Elementary School	78.6	88.4	Adams
Liberty CUSD 2	84.6	100.0	Adams
Liberty Elementary School	84.6	100.0	Adams
Central CUSD 3	63.6	87.9	Adams
Central 3-4 Middle School	63.6	87.9	Adams
CUSD 4	69.6	71.4	Adams
Greenfield Elementary School	69.6	71.4	Adams
Quincy SD 172	73.0	86.9	Adams
Adams Elementary School	67.1	79.7	Adams
Dewey Elementary School	75.4	93.4	Adams
Ellington Elementary School	90.1	94.4	Adams
$\vdots$	$\vdots$	$\vdots$	$\vdots$

maximum.

3.2. Illinois state education data

The second example is to estimate the joint density function of some 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, the percentages of students meeting or exceeding test standards are tabulated (see Table 2). The data are plotted in Figure 4 (left). Each point indicates a public school or district.

We first estimate the density functions and (cumulative) distribution functions by the kernel method and the empirical distribution function. Pearson's correlation and the sample rank correlation of the data are 0.853 and 0.851, respectively. Because of these high sample correlations, we use $H_n^+(x, y)$ in (15), the largest correlation model. The estimated density func-
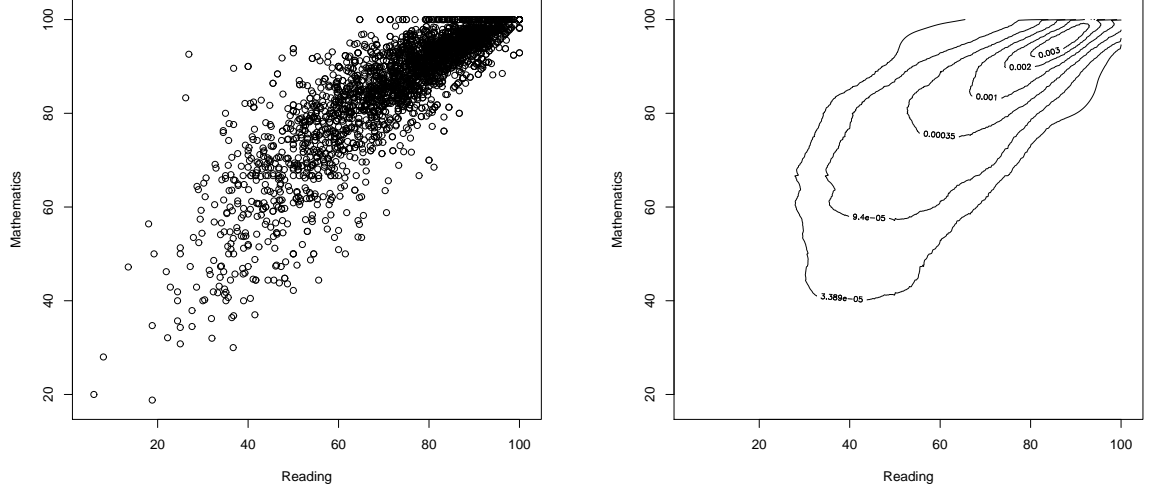
Figure 4: ISAT percent meeting or exceeding standards.
(Left: data plot. Right: estimated density contour.)

tion, $h^+(x, y; \widehat{q}, \widehat{n})$, is plotted in Figure 4 (right). Using the EM algorithm in Section 2.4, we obtain the estimates $(\widehat{q}, \widehat{n}) = (0.919, 17)$. The approximate variance of $\widehat{q}$ is 9.06×10^{-5}. The rank correlation under the estimated model is $\widehat{q}(\widehat{n} - 1)/(\widehat{n} + 1) = 0.817$.

Let $\widehat{q}(n)$ be the MLE of q under the model with n fixed as in Baker (2008). Table 3 lists the MLEs $\widehat{q}(n)$ and the corresponding log-likelihoods (profile likelihoods) for $2 \le n \le 20$, and we see that the log-likelihood is maximized at $n = 16$. Although this is different from the estimator $\widehat{n} = 17$ above, the difference between the values of the likelihood at these two estimates is small, and this shows that Algorithm 2.3 works well in this practical setting.

The analysis above assumes that the scores are continuous variables. However, as shown in Table 2, the scores are rounded off to the nearest one-tenth value. Therefore, in practice, the variables are discrete, take values $k/10$, $k = 0, 1, \ldots, 1000$, and there are many ties in this data set. Also, the number of unique values for x_i, y_i, and (x_i, y_i) are 602, 456 and 2260, respectively, among $N = 2991$ schools and districts. Applying the EM algorithm for the discrete case, i.e., Algorithm 2.3 with (16) replaced by (17), we obtain the estimates $(\widehat{q}, \widehat{n}) = (0.933, 18)$. The rank correlation under the

20

Table 3: MLE and profile log-likelihood when n is fixed. (The maximum value is indicated with a box.)

n	$\widehat{q}(n)$	Log-likelihood
2	1.000	−22580.84
5	1.000	−21812.38
10	0.980	−21477.13
12	0.970	−21439.37
14	0.949	−21422.39
15	0.939	−21418.51
16	0.929	−21416.81
17	0.919	−21416.88
18	0.909	−21418.41
19	0.899	−21421.13
20	0.889	−21424.85

discrete model is 0.835, which is slightly closer to the sample rank correlation than the one under the continuous model. In both cases, we see that the estimated rank correlations are slightly lower than the sample rank correlation of the data. However, considering the few number of parameters in the models, the differences may be acceptable.

The ISAT data set also contains the names of 102 counties to which the 2991 schools and districts belong. Since the high positive correlation may be caused by county effect, we analyze residuals obtained by simple regressions for each marginal using the name of the county as a covariate. The residuals are plotted in the left panel of Figure 5.

Similarly to the original data, the Pearson correlation and Spearman's rank correlation of the residuals are calculated as 0.835 and 0.840, respectively. Using the EM algorithm in Section 2.4, we obtain the estimates $(\widehat{q}, \widehat{n}) = (0.917, 19)$ and the joint density contour plot shown in the right panel of Figure 5. The approximate variance of $\widehat{q}$ is 0.0001. The rank correlation under the estimated model is 0.825. Baker's method of maximizing profile likelihoods gives similar results: $(\widehat{q}, \widehat{n}) = (0.919, 19)$.

It is reasonable that the correlation of residuals is still high, because the coefficients of determination of the simple regressions are not large (0.170 and

21

0.187 for Reading and Mathematics, respectively). Note that the correlation of estimated county effects for Reading and Mathematics is moderate (0.732).
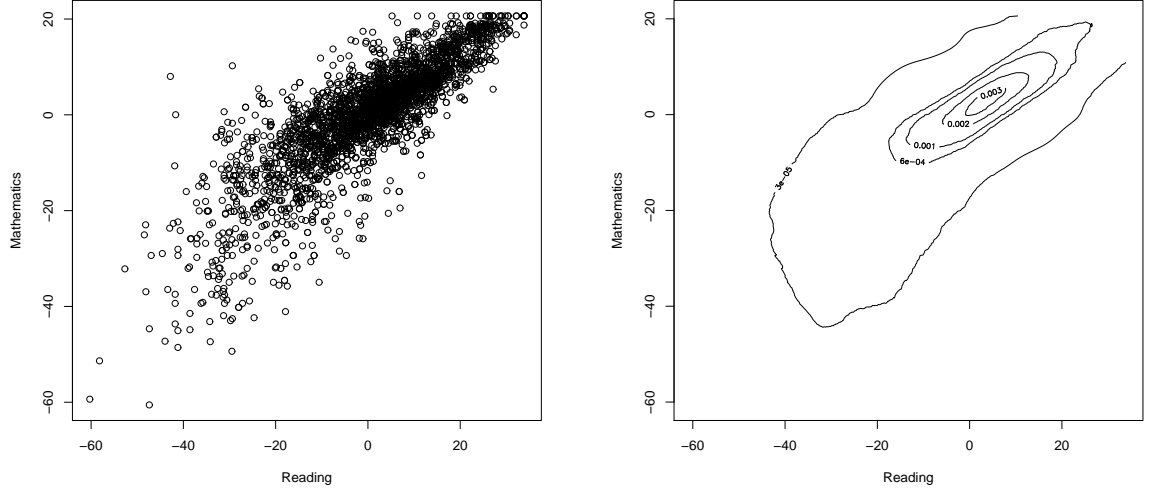

Figure 5: Residuals of ISAT percent meeting or exceeding standards. (Left: residuals by the simple regressions. Right: estimated density contour.)

3.3. Simulated trivariate data with interaction

The third example is an artificial trivariate continuous data. The data are generated by the following two steps. First, we generate data $(u_{1,i}, u_{2,i}, u_{3,i})$, $i = 1, \ldots, N$, from a trivariate Baker's distribution with the copula density

$$c(u_1, u_2, u_3) = n_1 n_2 n_3 \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \sum_{k_3=1}^{n_3} r_{k_1,k_2,k_3} \prod_{j=1}^{3} b_{k_j-1, n_j-1}(u_j). \qquad (18)$$

Here, the parameter $R = (r_{k_1,k_2,k_3})$ is defined as

$$r_{k_1,k_2,1} = \frac{1}{2n_1 n_2} \ \text{ (for all } k_1, k_2), \qquad r_{k_1,k_2,2} = \begin{cases} \frac{1}{2n_1} & (\text{if } k_1 = k_2), \\ 0 & (\text{if } k_1 \neq k_2), \end{cases}$$

with $n_1 = n_2 = 20$ and $n_3 = 2$. The sample size is chosen to be $N = 2000$. Also, we convert the uniform marginals to normal marginals by the

transformation $x_i = \Phi^{-1}(u_{1,i})$, $y_i = \Phi^{-1}(u_{2,i})$, $z_i = \Phi^{-1}(u_{3,i})$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. We then obtain random data (x_i, y_i, z_i) whose marginals have standard normal distributions.

The first row of Figure 6 depicts scatter plots for the first and second variates (X, Y) stratified with the third variable Z. The correlation between X and Y is designed to be increasing in Z, and the marginals of (X, Z) and (Y, Z) are independent. From the three panels, we can see that X and Y are almost independent when Z is small and they are highly correlated when Z is large.

We fit the Bernstein copula density (18) with an extended version of Algorithm 2.2 as well as Algorithm 2.1 for this 3-dimensional data set. The contours of the estimated density function are shown in the second row of Figure 6, and we see that the Bernstein copula represents well the characteristic of the changing correlation. For comparison, we also plot the contours estimated with the Gaussian copula in the last row of the figure even though a Gaussian copula obviously cannot adapt to the change of correlation (i.e., the 3-way interaction). Consequently, this example demonstrates the flexibility of the Bernstein copula and the usefulness of the EM algorithm for 3-dimensional data.

3.4. Trivariate uranium data

In the fourth example, we consider the uranium data set which is analyzed in Cook and Johnson (1986) and available from the R package `copula`. The data set consists of concentrations of 7 elements measured from 655 water samples collected from the Montrose Quadrangle of Western Colorado. We select three elements cobalt (Co), scandium (Sc), and cesium (Cs) as variables and fit the trivariate Bernstein copula model to the data.

We choose Cs as a stratifying variable and divide the data set into three parts according to the level of Cs. The scatter plots and the joint density contours of Co and Sc with respect to different levels of Cs are shown in Figure 7. From the figure, we can see that the correlation structure of this data set is similar to the structure of the simulation data in Section 3.3 in that the correlation of the first two variables depends on the level of the third variable. Specifically, the correlation between Co and Sc decreases when Cs increases. Because the AIC tends to result in small values of (n_1, n_2, n_3) for $\widehat{R}$ and induces a poor fit to the observed data, we investigate various choices of (n_1, n_2, n_3) in estimating $\widehat{R}$ with the algorithm for 3-dimensional data. Three examples of joint density contours with $(n_1, n_2, n_3) = (5, 5, 3)$, $(10, 10, 5)$
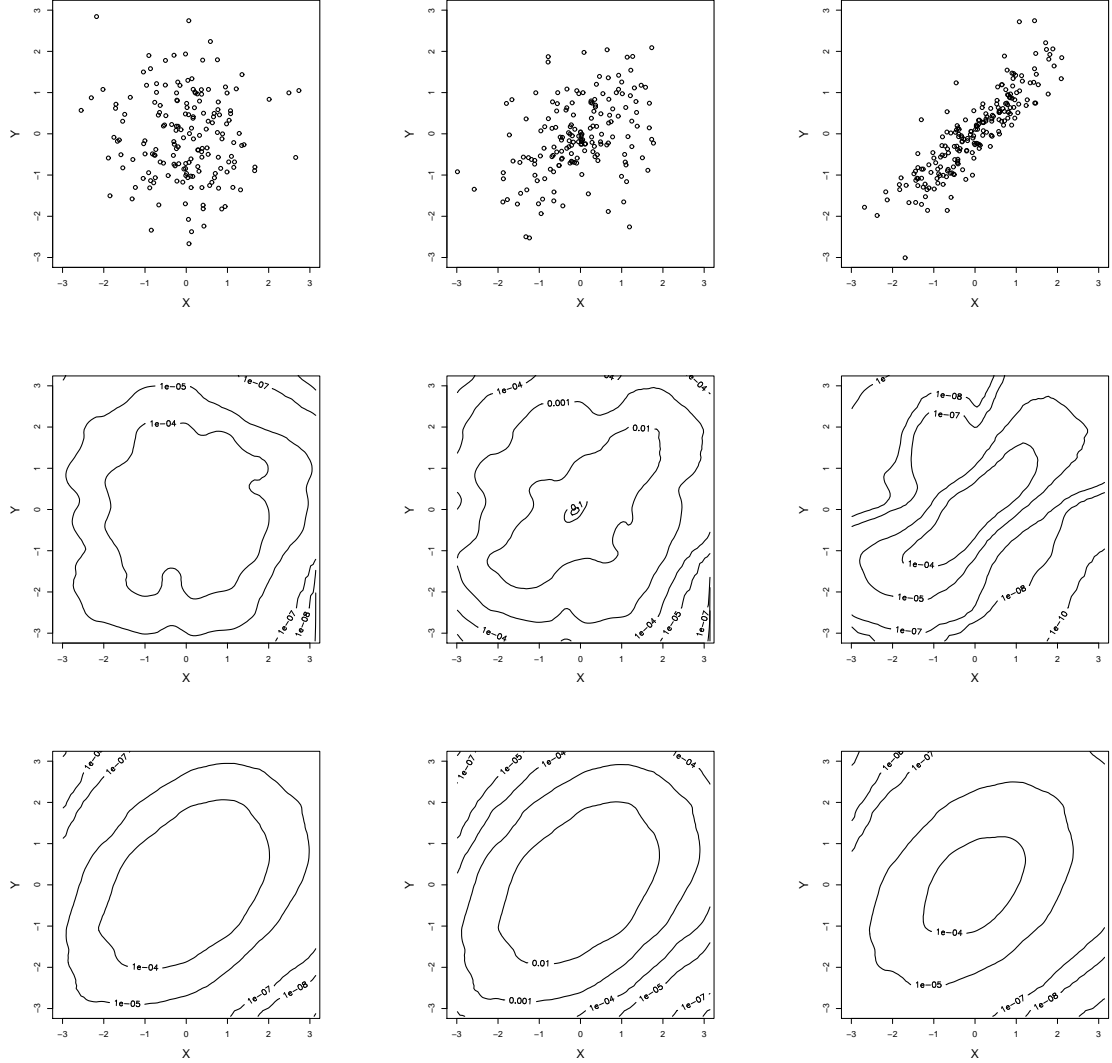
Figure 6: Simulated trivariate data and estimated contours.
Scatter plots for stratified data (first row), the estimated density with
a Bernstein copula (second row), and the estimated density with a
Gaussian copula (third row) for three cases of $(X, Y|Z)$.
Left: Z is small ($Z < 0.1$); Center: Z is moderately-sized ($0.45 < Z < 0.55$); Right: Z is large ($Z > 0.9$).

and $(15, 15, 3)$, are shown in the last three rows of Figure 7, respectively. Comparing the results, we can see that the estimated joint density contours with relatively larger sizes for $\widehat{R}$ fit the data more accurately.

4. Discussion

We have developed EM algorithms to estimate the Bernstein copula for continuous or discrete data. In this section, we provide some remarks for researchers who wish to use these algorithms for practical data analysis, and we propose further research topics.

In general, the convergence of the EM algorithm to the global maximizer (i.e., the MLE) depends on the starting point of the algorithm. In this paper, we propose to use the estimator given by Sancetta and Satchell (2004) and Janssen et al. (2012), as stated in Remark 2.4. Numerical studies in Section 3.1 provide empirical evidence that this starting point works well for calculating the MLE. However, this result has not been established theoretically, and therefore it is essential to consider a variety of starting points for the algorithm.

One of our primary objectives was to estimate the Bernstein copula. Although the estimation of marginals was not considered in this paper, it is worth noting that the estimation of marginals is of practical importance; for example, the ragged shape of the contour in Figure 2 disappears if we use a wider bandwidth in the kernel density estimation.

In Section 3.1, the AIC is used to choose the size of the matrix $R = (r_{k,l})$. However, in our method, the marginal functions F and G are also unknown and must be estimated, and for such semiparametric estimation procedures, the conventional AIC has no rationale. In the construction of the conventional AIC, the asymptotic distribution of the score function and its derivatives play crucial roles (Konishi and Kitagawa, 2008). Tsukahara (2005) developed a counterpart asymptotic theory when the marginals are estimated with empirical distribution functions; those results may be useful as building blocks for AIC methods for semiparametric copula estimation. Also, we find in the example of Section 3.4 that the AIC has a tendency to choose smaller number of parameters for this data set. Theoretical and practical methods for selecting the model may be a topic for further research.

An additional analysis of the data sets in Sections 3.1 and 3.2 shows that the Bernstein copula and the Gaussian copula lead to similar results for these data sets. However, for data (X, Y, Z) with 3-way interaction, such as
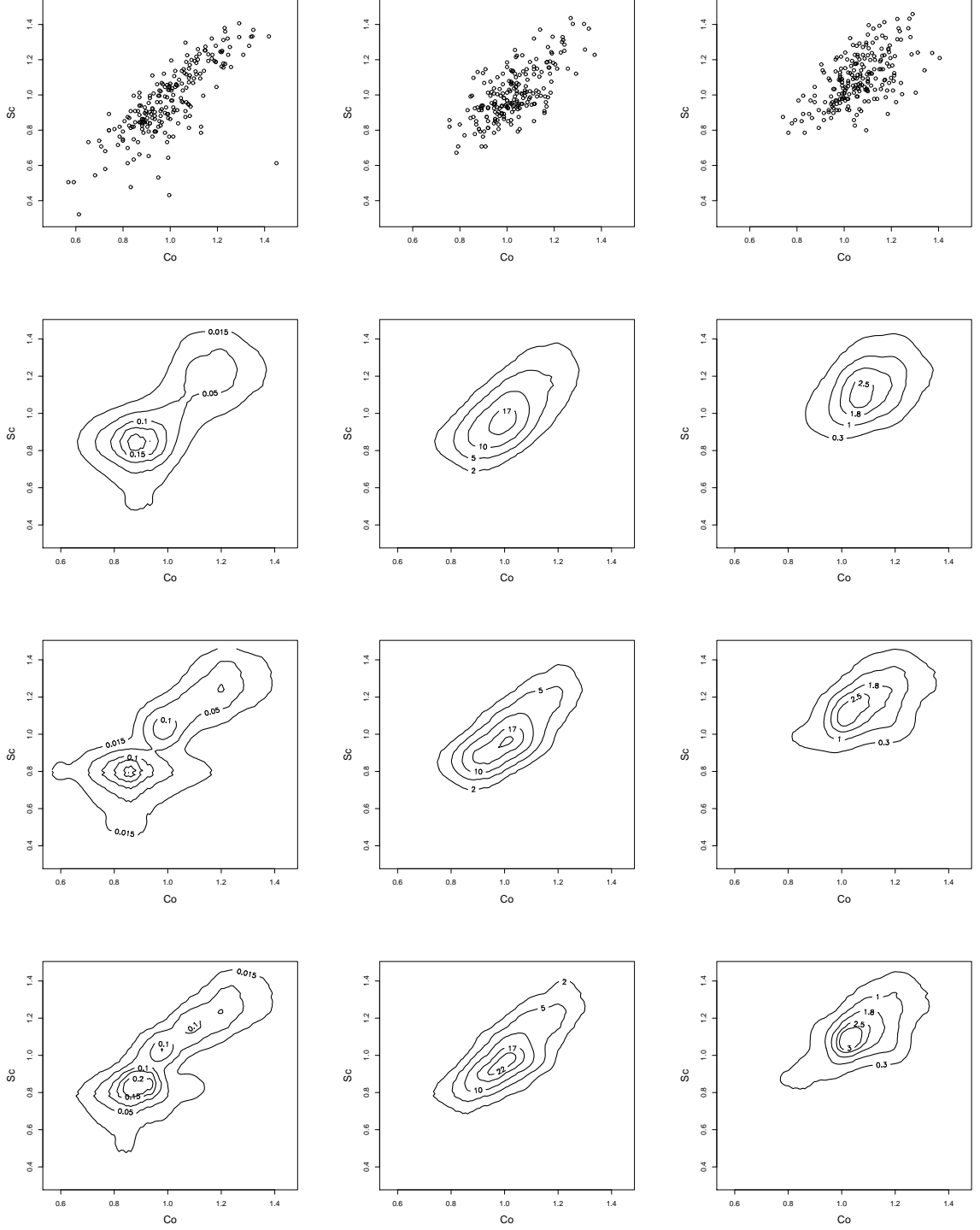
Figure 7: Scatter plots of (Co, Sc) for stratified values of Cs (first row), the estimated joint density contours with $\widehat{R}$ of sizes $5 \times 5 \times 3$ (second row), $10 \times 10 \times 5$ (third row), and $15 \times 15 \times 3$ (fourth row). Left: Cs is small (Cs $\leq$ 1.91); Center: Cs is moderately-sized (1.91 <Cs $\leq$ 2.13); Right: Cs is large (Cs > 2.13).

in Sections 3.3 and 3.4, we see that the Bernstein copula is more flexible for modeling correlation structures. This is a remarkable feature not generally possessed by other copula families such as the Gaussian, multivariate t-, and Archimedean copula families. Recently, to describe such complicated data, the vine copula method of creating a flexible copula by combining several copulas was developed (Kurowicka and Joe, 2011; Kim et al., 2013). The Bernstein copula is not only flexible in its own right, but can also be incorporated into a vine copula to model data with more complicated correlation structures. This is also a topic for further research.

Appendix A. Proof of Proposition 2.1

For vectors $\boldsymbol{\mu} = (\mu_k)_{1 \leq k \leq m}$ and $\boldsymbol{\lambda} = (\lambda_l)_{1 \leq l \leq n}$, define column vector valued functions by

$$\boldsymbol{f}(\boldsymbol{\mu}; \boldsymbol{\lambda}) = (f_k(\boldsymbol{\mu}; \boldsymbol{\lambda}))_{1 \leq k \leq m}, \quad f_k(\boldsymbol{\mu}; \boldsymbol{\lambda}) = \sum_{l=1}^{n} \frac{\bar{\tau}_{k,l}}{\mu_k + \lambda_l} - \frac{1}{m},$$

and

$$\boldsymbol{g}(\boldsymbol{\lambda}; \boldsymbol{\mu}) = (g_l(\boldsymbol{\lambda}; \boldsymbol{\mu}))_{1 \leq l \leq n}, \quad g_l(\boldsymbol{\lambda}; \boldsymbol{\mu}) = \sum_{k=1}^{m} \frac{\bar{\tau}_{k,l}}{\mu_k + \lambda_l} - \frac{1}{n}.$$

Let $\mathbb{1}_m = (\underbrace{1, \ldots, 1}_{m})'$. Each step of Algorithm 2.1 can be rewritten as follows:

Step M0. Set the initial values for $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}^0$. Let $t = 0$.

Step M1. For fixed $\boldsymbol{\mu}^{(t)}$, let $\boldsymbol{\lambda}^{(t)}$ be the solution of $\boldsymbol{g}(\boldsymbol{\lambda}; \boldsymbol{\mu}^{(t)}) = \boldsymbol{0}$.

Step M2. For fixed $\boldsymbol{\lambda}^{(t)}$, let $\widetilde{\boldsymbol{\mu}}^{(t+1)}$ be the solution of $\boldsymbol{f}(\boldsymbol{\mu}; \boldsymbol{\lambda}^{(t)}) = \boldsymbol{0}$.

Step M3. Let

$$\boldsymbol{\mu}^{(t+1)} = \widetilde{\boldsymbol{\mu}}^{(t+1)} - \frac{1}{m} \mathbb{1}_m \mathbb{1}_m' (\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^{(0)})$$

so that $\sum_k (\boldsymbol{\mu}^{(t+1)})_k = \sum_k (\boldsymbol{\mu}^{(0)})_k$.

Let $t := t + 1$ and go to Step M1, until $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ converge.

Let $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ be a solution of $\boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*) = \boldsymbol{0}$, $\boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*) = \boldsymbol{0}$. Since $(\boldsymbol{\mu}^* + r\mathbb{1}_m, \boldsymbol{\lambda}^* - r\mathbb{1}_n)$ is also a solution for arbitrary $r \in \mathbb{R}$, we assume without loss of generality that $\sum_k (\boldsymbol{\mu}^0)_k = \sum_k (\boldsymbol{\mu}^*)_k$. From Step M1, we obtain

$$
\begin{aligned}
\boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{\lambda}^{(t)}; \boldsymbol{\mu}^{(t)}) \\
&\doteq \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*) + \nabla_\lambda \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*)(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*) + \nabla_\mu \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*)(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*),
\end{aligned}
$$

and hence

$$
\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^* \doteq -(\nabla_\lambda \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*))^{-1} \nabla_\mu \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*)(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*). \tag{A.1}
$$

Here '$\doteq$' means that the difference of left-hand side and right-hand side is of the order $o\big(\max(\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\|, \|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*\|)\big)$. Similarly, from Step M2,

$$
\begin{aligned}
\boldsymbol{0} &= \boldsymbol{f}(\widetilde{\boldsymbol{\mu}}^{(t+1)}; \boldsymbol{\lambda}^{(t)}) \\
&\doteq \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*) + \nabla_\mu \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*)(\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^*) + \nabla_\lambda \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*)(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*),
\end{aligned}
$$

and hence

$$
\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^* \doteq -(\nabla_\mu \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*))^{-1} \nabla_\lambda \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*)(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*). \tag{A.2}
$$

Because $\mathbb{1}'_m \boldsymbol{\mu}^{(0)} = \sum_k (\boldsymbol{\mu}^0)_k = \sum_k (\boldsymbol{\mu}^*)_k = \mathbb{1}'_m \boldsymbol{\mu}^*$, we can rewrite Step M3 as

$$
\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^* = \widetilde{\boldsymbol{\mu}}^{(t+1)} - \frac{1}{m} \mathbb{1}_m \mathbb{1}'_m (\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^*) - \boldsymbol{\mu}^* = J(\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^*), \tag{A.3}
$$

where

$$
J = I_m - \frac{1}{m} \mathbb{1}_m \mathbb{1}'_m.
$$

Combining (A.2) and (A.3), we have

$$
\begin{aligned}
\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^* &= J(\widetilde{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}^*) \\
&\doteq -J(\nabla_\mu \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*))^{-1} \nabla_\lambda \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*)(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*). \tag{A.4}
\end{aligned}
$$

Let $C = (c_{k,l})_{m \times n}$, $G = (\mathrm{diag}(c_{k+}))_{m \times m}$, $H = (\mathrm{diag}(c_{+l}))_{n \times n}$, where

$$
c_{k,l} = \frac{\bar{\tau}_{k,l}}{((\boldsymbol{\mu}^*)_k + (\boldsymbol{\lambda}^*)_l)^2}, \quad c_{k+} = \sum_{l=1}^{n} c_{k,l}, \quad c_{+l} = \sum_{k=1}^{m} c_{k,l}.
$$

Simple calculations yield

$$
-\nabla_\mu \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*) = G, \qquad -\nabla_\lambda \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*) = H,
$$

$$-\nabla_{\mu} \boldsymbol{g}(\boldsymbol{\lambda}^*; \boldsymbol{\mu}^*) = C', \qquad -\nabla_{\lambda} \boldsymbol{f}(\boldsymbol{\mu}^*; \boldsymbol{\lambda}^*) = C.$$

Therefore, (A.1) and (A.4) are rewritten as

$$\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^* \doteq -H^{-1}C'(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*), \tag{A.5}$$

$$\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^* \doteq -JG^{-1}C(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*). \tag{A.6}$$

Combining (A.5) and (A.6), we have

$$\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^* \doteq JG^{-1}CH^{-1}C'(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*), \tag{A.7}$$

$$\boldsymbol{\lambda}^{(t+1)} - \boldsymbol{\lambda}^* \doteq H^{-1}C'JG^{-1}C(\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^*). \tag{A.8}$$

To ascertain the asymptotic behavior of $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ as $t \to \infty$, we need to find the eigenvalues of the matrices $JG^{-1}CH^{-1}C'$ and $H^{-1}C'JG^{-1}C$, respectively (Hageman and Young, 1981). Let $D = G^{-1/2}CH^{-1/2}$; then we first show that the matrix D has the largest singular value $\sigma_1(D) = 1$. This is because for column vectors $\boldsymbol{u} = (u_k)_{1 \le k \le m}$ and $\boldsymbol{v} = (v_l)_{1 \le l \le n}$,

$$\boldsymbol{u}'C\boldsymbol{v} = \sum_{k,l} u_k v_l c_{k,l} \le \sqrt{\sum_{k,l} u_k^2 c_{k,l} \sum_{k,l} v_l^2 c_{k,l}} = \sqrt{\sum_k u_k^2 c_{k+} \sum_l v_l^2 c_{+l}}$$

and hence

$$\sigma_1(D) = \max_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} \boldsymbol{u}'D\boldsymbol{v} = \max_{\sum u_k^2 = \sum v_l^2 = 1} \sum_{k,l} u_k v_l \frac{c_{k,l}}{\sqrt{c_{k+}c_{+l}}}$$

$$= \max_{\sum u_k^2 c_{k+} = \sum v_l^2 c_{+l} = 1} \sum_{k,l} u_k v_l c_{k,l}$$

$$\le \max_{\sum u_k^2 c_{k+} = \sum v_l^2 c_{+l} = 1} \sqrt{\sum_k u_k^2 c_{k+} \sum_l v_l^2 c_{+l}} = 1.$$

This upper bound is attained when $\sigma_1(D) = \boldsymbol{u}_1'D\boldsymbol{v}_1$ with

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{c_{++}}} G^{1/2} \mathbb{1}_m, \qquad \boldsymbol{v}_1 = \frac{1}{\sqrt{c_{++}}} H^{1/2} \mathbb{1}_n, \qquad c_{++} = \sum_{k,l} c_{k,l},$$

because $\boldsymbol{u}_1'\boldsymbol{u}_1 = \boldsymbol{v}_1'\boldsymbol{v}_1 = 1$ and

$$\boldsymbol{u}_1'D\boldsymbol{v}_1 = \frac{1}{c_{++}} (\mathbb{1}_m'G^{1/2})G^{-1/2}CH^{-1/2}(H^{1/2}\mathbb{1}_n) = \frac{1}{c_{++}} \mathbb{1}_m'C\mathbb{1}_n = 1.$$

29

Therefore, DD' has the largest eigenvalue 1, and one of the corresponding eigenvectors is $\boldsymbol{u}_1$.

From the assumption that $\bar{\tau}_{k,l} > 0$ for all k, l, we find that DD' is a positive matrix, i.e., all elements are positive. By the Perron-Frobenius theorem (Horn and Johnson, 2013), the multiplicity of the largest eigenvalue 1 of DD' is 1, and so we obtain

$$DD' = \frac{1}{c_{++}} G^{1/2} \mathbb{1}_m \mathbb{1}'_m G^{1/2} + \sum_{k=2}^{m} \nu_k \boldsymbol{u}_k \boldsymbol{u}'_k,$$

where $1 > \nu_2 \geq \cdots \geq \nu_m \geq 0$ and $0 = \boldsymbol{u}'_k \boldsymbol{u}_1 = \boldsymbol{u}'_k G^{1/2} \mathbb{1}_m / \sqrt{c_{++}}$.

Hence, the matrix $JG^{-1}CH^{-1}C'$ appearing in (A.7) is rewritten as

$$
\begin{aligned}
JG^{-1}CH^{-1}C' &= JG^{-1/2}DD'G^{1/2} \\
&= \frac{1}{c_{++}} JG^{-1/2}G^{1/2} \mathbb{1}_m \mathbb{1}'_m G^{1/2}G^{1/2} + JA = JA, \qquad \text{(A.9)}
\end{aligned}
$$

where

$$A = G^{-1/2} \left(\sum_{k=2}^{m} \nu_k \boldsymbol{u}_k \boldsymbol{u}'_k \right) G^{1/2}.$$

This matrix JA has the same nonzero eigenvalues as those of

$$AJ = G^{-1/2} \left(\sum_{k=2}^{m} \nu_k \boldsymbol{u}_k \boldsymbol{u}'_k \right) G^{1/2} \left(I_m - \frac{1}{m} \mathbb{1}_m \mathbb{1}'_m \right) = A,$$

which has the eigenvalues $\nu_2, \ldots, \nu_m$ and 0. Here we used $0 = \boldsymbol{u}'_k G^{1/2} \mathbb{1}_m$. More precisely, it holds that

$$JAB = BN, \quad B = \left(\mathbb{1}_m, JG^{-1/2}(\boldsymbol{u}_2, \ldots, \boldsymbol{u}_m) \right),$$

where $N = \text{diag}(0, \nu_2, \ldots, \nu_m)$. The matrix B is nonsingular because

$$G^{1/2} B \begin{pmatrix} 1 & \frac{1}{m} \mathbb{1}'_m G^{-1/2}(\boldsymbol{u}_2, \ldots, \boldsymbol{u}_m) \\ \mathbf{0} & I_{m-1} \end{pmatrix} = \left(G^{1/2} \mathbb{1}_m, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m \right)$$

is nonsingular. Therefore,

$$JA = BNB^{-1}. \qquad \text{(A.10)}$$

Combining (A.10) with (A.7) and (A.9), we have $B^{-1}(\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^*) \doteq NB^{-1}(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*)$, and hence for arbitrary $\varepsilon > 0$

$$\|B^{-1}(\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^*)\| \leq (\nu_2 + \varepsilon)\,\|B^{-1}(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*)\| \qquad (A.11)$$

when t is sufficiently large.

By proceeding in a similar manner, we find that the matrix $H^{-1}C'JG^{-1}C$ in (A.8) can be diagonalized with the same eigenvalues $1 > \nu_2 \geq \cdots \geq \nu_{\min(m,n)} \geq 0$ and 0 (if $m < n$). Hence, an inequality of the same type as (A.11) holds for the sequence $\boldsymbol{\lambda}^{(t)}$. These inequalities imply the linear convergence of $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ with the rate $\nu_2 + \varepsilon$.

References

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

Charpentier, A., Fermanian, J.-D. and Scaillet, O. (2007). The estimation of copulas: Theory and practice, in *Copulas: From Theory to Application in Finance* (J. Rank ed.), Risk Books, London, 2007, pp. 35–60.

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.

Cook, R. D. and Johnson, M. E. (1986). Generalized Burr-Pareto-Logistic distribution with applications to a uranium exploration data set, *Technometrics*, **28** (2), 123–131.

Dennis, Jr., J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equation*, SIAM, Philadelphia.

Dou, X., Kuriki, S. and Lin, G. D. (2013). Dependence structures and asymptotic properties of Baker's distributions with fixed marginals, *Journal of Statistical Planning and Inference*, **143** (8), 1343–1354.

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

Genest, C. and Werker, B. J. M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models, in *Distributions with Given Marginals and Statistical Modelling* (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena eds.), Kluwer, Dordrecht, pp. 103–112.

Hageman, L. A. and Young, D. M. (1981). *Applied Iterative Methods*, Academic Press, New York.

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*, 2nd edn., Cambridge University Press, Cambridge.

Huang, J. S., Dou, X., Kuriki, S. and Lin, G. D. (2013). Dependence structure of bivariate order statistics with applications to Bayramoglu's distributions, *Journal of Multivariate Analysis*, **114**, 201–208.

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`

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2012). Large sample behavior of the Bernstein copula estimator, *Journal of Statistical Planning and Inference*, **142** (5), 1189–1197.

Joe, H. (2001). *Multivariate Models and Multivariate Dependence Concepts*, Chapman & Hall/CRC, New York.

Kim, G., Silvapulle, M. J. and Silvapulle, P. (2007). Comparison of semi-parametric and parametric methods for estimating copulas, *Computational Statistics and Data Analysis*, **51** (6), 2836–2850.

Kim, D., Kim, J.-M., Liao, S.-M. and Jung, Y.-S. (2013). Mixture of D-vine copulas for modeling dependence, *Computational Statistics and Data Analysis*, **64** (1), 1–19.

Kingsley, E. H. (1951). Bernstein polynomials for functions of two variables of class $C^{(k)}$, *Proceedings of the American Mathematical Society*, **2** (1), 64–71.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, Springer, New York.

Kurowicka, D. and Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook*, World Scientific, Singapore.

Lin, G. D. and Huang, J. S. (2010). A note on the maximum correlation for Baker's bivariate distributions with fixed marginals, *Journal of Multivariate Analysis*, **101** (9), 2227–2233.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, 2nd edn., Wiley, New York.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd edn., Springer, New York.

Sancetta, A. and Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions, *Econometric Theory*, **20** (3), 535–562.

Schucany, W. R., Parr, W. C. and Boyer, J. E. (1978). Correlation structure in Farlie-Gumbel-Morgenstern distributions, *Biometrika*, **65** (3), 650–653.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Takada, T., Mita, A., Maeno, A., Sakai, T., Shitara, H., Kikkawa, Y., Moriwaki, K., Yonekawa, H. and Shiroishi, T. (2012). Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits, *Genome Research*, **18** (3), 500–508. The data set is available at `http://molossinus.lab.nig.ac.jp/phenotype/index.html`

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

Tsukahara, H. (2005). Semiparametric estimation in copula models, *The Canadian Journal of Statistics*, **33** (3), 357–375. Erratum: ibid (2011), **39** (4), 734–735.