Iranian Journal of Fuzzy Systems

Volume 19, Number 4, (2022), pp. 165-175

Original paper

University of Sistan and Baluchestan

Copula-based Berkson measurement error models

A. R. Ziaei[1], K. Zare[2] and A. Sheikhi[3]

[1,2] *Department of Statistics, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran*
[3] *Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University, Kerman, Iran*
ar.ziaei@iau.ac.ir, karim.zare@iau.ac.ir, sheikhy.a@uk.ac.ir

Abstract

In this work, we consider the joint distribution function as well as the copula of $(X+Z, Y)$ where the random vector (X, Y, Z) is characterized by a copula $C_{X,Y,Z}$. We use this copula to analyze a Berkson measurement error model. By presenting a general form of a Berkson measurement error model with copula-dependent random variables, we investigate some of its special cases. Some theoretical results, several examples as well as a simulation study, are proposed for illustration.

Keywords: Copula, noise, perturbation of copula, measurement error.

1 Introduction and Preliminaries

Measurement accuracy of variables is the first and foremost assumption in linear models, especially regression analysis. The consistency of estimators will be dramatically affected when this assumption does not hold, and this ill-condition situation is called measurement error model or error-in-variables [10]. These models are employed when measurement of the predictor cannot be done accurately, or it is expensive or impossible to measure. To address the challenge of these measurement errors, there are two well-known types of measurement errors analysis that are categorized into *classical measurement errors* and *Berkson measurement errors.*

This paper is concerned with the Berkson-type measurement errors which has been introduced by Berkson (1950) [2], when the explanatory variable X^*, is not measured directly or precisely, and instead, another proxy, X, which is linearly related to X^*, is observed. So, only the pairs (X_i, Y_i), $i = 1, 2, \ldots, n$ are observed, and the knowledge of the explanatory variable X^*, involves error. With this regards, a linear regression equation with the Berkson error can be viewed as

$$\begin{aligned} Y &= \alpha + \beta X^* + \delta, \\ X^* &= X + Z, \end{aligned} \tag{1}$$

where X^* is the latent predictor variable and X is its surrogate variable which is measured with the error term Z. In such models X^* varies around X, i.e., the value of X^*, can be expressed as X plus a perturbation, Z. See, for example, Fuller (2009) [10] and Buonaccorsi (2010) [3]. As an example, X might represent the concentration of a toxic substance, measured at various fixed stations, while the actual exposure, X^*, of individuals varies about the accurate concentrations at the stations (see, e.g. [7]).

Also, there are many works which carry out nonlinear Berkson measurement error models, in which Wang (2004) [17], Carroll et al. (2006) [4], Chen et al. (2011) [5] and Schennach (2013) [13] are among them. Our results in this work yield linear regression as well as nonlinear. We assume that variables are associated with copulas, and types of copula determines the linearity or non-linearity regression equation.

According to the Sklar proposal, (Sklar, 1959, [16]), for any two random variables, there exists a copula function $C : [0,1]^2 \to [0,1]$, such that

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y)), \quad x, y \in \mathbb{R}, \tag{2}$$

Corresponding Author: K. Zare
Received: August 2021; Revised: March 2022; Accepted: June 2022.
https://doi.org/10.22111/IJFS.2022.7094

where $F_{X,Y}$ is the joint distribution function of X and Y and F_X, F_Y, are their marginal distribution functions respectively and denoting $C_{X,Y}(u,v)$ as the copula coupling X and Y, $C_{X,Y}$ has grounded, uniformly marginal and 2−increasing properties. See Nelsen (2006) [12] and Durante and Sempi (2016) [8] and references therein for more details. Recall that the Fréchet-Hoeffding upper bound, the Fréchet-Hoeffding lower bound and the product copula are the three basic copulas, which are denoted by $M_2(u,v) = \min(u,v)$, $W_2(u,v) = \max(0, u+v-1)$ and $\Pi_2(u,v) = uv$, respectively. These copulas can be generalized as

$$C_2^F(u,v) = \alpha M_2(u,v) + (1-\alpha-\beta)\Pi_2(u,v) + \beta W_2(u,v),$$

where $\alpha, \beta \in [0,1]$ and $\alpha + \beta \leqslant 1$, which is a convex combination of them. Note that if $\alpha = 1$, $C_2^F(u,v)$ is the Fréchet-Hoeffding upper bound copula which depicts the complete positive dependence, if $\beta = 1$, $C_2^F(u,v)$ is the Fréchet-Hoeffding lower bound copula which depicts the complete negative dependence and if $\alpha = \beta = 0$, $C_2^F(u,v)$ is the product copula, which reveals the independence (Fréchet, 1958 [9]). A special case of this family of copulas which is known as the Fréchet-Mardia copulas, is defined as

$$C_2^{FM}(u,v) = \beta M_2(u,v) + (1-\beta)\Pi_2(u,v), \tag{3}$$

where $\beta \in [0,1]$. Another important copula family, which includes the product copula, is the Farlie-Gumbel-Morgenstern (FGM) family of copulas has the form

$$C_2^{FGM}(u,v) = uv[1+\theta(1-u)(1-v)], \tag{4}$$

where $\theta \in [-1,1]$. Evidently, (4) reduces to the product copula if $\theta = 0$.

Recently, Mesiar et al. (2019) [11] have considered that X and Y are connected by a copula $C_{X,Y}$ and have obtained the perturbed copula $C_{X+Z,Y}$ when X, Y are independent of the random variable Z. Sheikhi and Mesiar (2020) [15] studied a copula-based classical measurement error models based on perturbation of copulas. The main assumption in the Berkson measurement error model is that X and Z are dependent. By studying the perturbation of a copula $C_{X,Y}$ related to a random vector (X,Y) when the random variable X is polluted by some dependent noise Z; we apply this result to make a Berkson measurement error model (1).

The organization of this paper is as follows. Some results in perturbation of copulas are introduced in Section 2. Section 3 is devoted to the application of perturbed copulas in the Berkson measurement error model. We present some numerical analysis, including a simulation study in Section 4, and finally, some concluding remarks are given in Section 5.

2 Perturbation of Copulas

The distribution of the sum of Random variables has been extensively considered in the literature [18, 1]. Recently, Sheikhi et al. (2020) [14] have investigated the sum of copula-related random variables and have presented an extension of the Irvin-Hall distribution. Using their notation, we assume that the variables of the random vector $(X_1, X_2, \ldots, X_n)$ are connected by a copula $C_{\boldsymbol{X}}(\boldsymbol{u})$ and $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, $\boldsymbol{X}_{-n} = (X_1, X_2, \ldots, X_{n-1})^T$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^T$. Then the distribution function of $S_n = \sum_{i=1}^n X_i$ is

$$F_{S_n}(s_n) = \oint_{\boldsymbol{x}_{-n}} D_{-n}C_{\boldsymbol{X}}(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(s_n - s_{n-1}))\, \mathrm{d}F_{\boldsymbol{X}_{-n}}(\boldsymbol{x}_{-n}), \tag{5}$$

where $\boldsymbol{x}_{-n}$ is the observed vector of $\boldsymbol{X}_{-n}$ and $\oint_{\boldsymbol{x}_{-n}}$ denotes $n-1$- integrals on $x_1, x_2, ..., x_{n-1}$. If the derivative $D_{-n}C_{\boldsymbol{X}}$ exist then $D_{-n}C_{\boldsymbol{X}}(\boldsymbol{u}) = \frac{\partial^{n-1}C_{\boldsymbol{X}}(u_1,\ldots,u_n)}{\partial u_1,\ldots,u_{n-1}}$ and $D_{-n}C_{\boldsymbol{X}}(\boldsymbol{u}) = 0$ otherwise.

Also, for a special case $n = 2$, the distribution function of the sum of two random variables X, Y, which are associated to a copula $C_{X,Y}$, is given by

$$F_{X+Y}(s) = \int D_1 C_{X,Y}(F_X(x), F_Y(s-x))\mathrm{d}F_X(x) \qquad s \in \mathbb{R} \tag{6}$$

$$= \int D_2 C_{X,Y}(F_X(s-y), F_Y(y))\mathrm{d}F_Y(y), \qquad s \in \mathbb{R} \tag{7}$$

where $D_1 C_{X,Y}(u,v) = \frac{\partial C_{X,Y}(u,v)}{\partial u}$ and $D_2 C_{X,Y}(u,v) = \frac{\partial C_{X,Y}(u,v)}{\partial v}$. If these derivatives do not exist, by convention, the zero value is considered for them. It is known that when X, Y are independent random variables, then (6) and (7) reduce to the so-called the *convolution* of two distribution functions F_X and F_Y.

The following examples play important roles in the sequel.

Example 2.1. *Let X and Y be random variables uniformly distributed on $[0,1]$ linked by the Farlie-Gumbel-Morgenstern copula (4), with parameter $\theta \in [-1,1]$, then by taking the partial derivative of (4), with respect to $F_Y(y)$, we obtain*

$$D_2C_{X,Y}(F_X(t-y),F_Y(y)) = (t-y)+\theta(t-y)(1-t+y)(1-2y),$$

and by substituting $D_2C_{X,Y}(F_X(s-y),F_Y(y))$ in (7) the following two cases arise:

1. *If $0 \leqslant s \leqslant 1$:*

$$\begin{aligned} F_{X+Y}(s) &= \int_0^s D_2C_{X+Y}(F_X(s-y),F_Y(y))\mathrm{d}y \\ &= \int_0^s [(s-y)+\theta(s-y)(1-s+y)(1-2y)]\mathrm{d}y \\ &= \frac{1}{6}(3s^2+\theta(s^4-4s^3+3s^2)), \end{aligned}$$

2. *If $1 \leqslant s \leqslant 2$:*

$$\begin{aligned} F_{X+Y}(s) &= \int_{s-1}^1 D_2C_{X+Y}(F_X(s-y),F_Y(y))\mathrm{d}y + \int_0^{s-1}\mathrm{d}y \\ &= \int_0^s [(s-y)+\theta(s-y)(1-s+y)(1-2y)]\mathrm{d}y \\ &= \frac{1}{6}(12s-3s^2-6+\theta(4-4s-3s^2+4s^3-s^4)), \end{aligned}$$

which yields

$$F_{X+Y}(s) = \begin{cases} \frac{1}{6}(3s^2+\theta(s^4-4s^3+3s^2)) & 0 \leqslant s \leqslant 1 \\ \frac{1}{6}(12s-3s^2-6+\theta(4-4s-3s^2+4s^3-s^4)) & 1 \leqslant s \leqslant 2. \end{cases} \tag{8}$$

A special case of (8) occurs when $\theta = 0$ and yields the distribution of the sum of two independent uniform random variables, which is a triangular-type distribution. Also, if $s = 1$, then both two parts of (8) are the same and equal to $F_{X+Y}(s) = \frac{1}{2}$.

Now, consider the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ which is connected to the two random variables Y and Z via the $n+2$-copula $C_{\boldsymbol{X},Y,Z}(\boldsymbol{u}, v, w)$. In what follows, we discuss the joint distribution $(X_1+Z, X_2, \ldots, X_n, Y)^T$ as well as its copula. For the sake of notation or simplicity, we adopt the notation $(X_1+Z, X_2, \ldots, X_n, Y)^T = (\boldsymbol{X}+Z\boldsymbol{e}, Y)^T$ where $\boldsymbol{e} = (1,0,0,\ldots,0)^T$ is a unit basis vector.

Theorem 2.2. *If $C_{\boldsymbol{X},Y,Z}(\boldsymbol{u}, v, w)$ is the copula function of the random vector $(\boldsymbol{X}, Y, Z)$, then the joint distribution and the copula function of the vector $(\boldsymbol{X}+Z\boldsymbol{e}, Y)$ are given respectively by*

$$\int D_{n+2}C_{\boldsymbol{X},Y,Z}\big(F_{X_1}(s_1-z), F_{X_2}(s_2), ..., F_{X_n}(s_n), F_Y(t), F_Z(z)\big)\,\mathrm{d}F_Z(z), \tag{9}$$

and

$$C_{\boldsymbol{X}+Z\boldsymbol{e},Y}(\boldsymbol{u}, v) = \int D_{n+2}C_{\boldsymbol{X},Y,Z}[F_{X_1}(F^{-1}_{X_1+Z}(u_1) - F^{-1}_Z(w)), u_2, \ldots, u_n, v, w]\,\mathrm{d}w, \tag{10}$$

where $D_{n+2}C_{\boldsymbol{X},Y,Z}(\boldsymbol{u}, v, w) = \frac{\partial C_{\boldsymbol{X},Y,Z}(\boldsymbol{u},v,w)}{\partial w}$ if this derivate exists, and $D_{n+2}C_{\boldsymbol{X},Y,Z}(\boldsymbol{u}, v, w) = 0$, otherwise.

Proof. Using the theorem of total probability we have

$$\begin{aligned} F_{\boldsymbol{X}+Z\boldsymbol{e},Y}(\boldsymbol{s}, t) &= P(X_1+Z < s_1, X_2 < s_2, \ldots, X_n < s_n, Y < t) \\ &= \int P(X_1 < s_1 - z, X_2 < s_2, \ldots, X_n < s_n, Y < t)|Z = z)\,\mathrm{d}F_Z(z) \\ &= \int D_{n+2}C_{\boldsymbol{X},Y,Z}\big(F_{X_1}(s_1-z), F_{X_2}(s_2), \ldots, F_{X_n}(s_n), F_Y(t), F_Z(z)\big)\,\mathrm{d}F_Z(z). \end{aligned}$$

Now, substituting $u_1 = F_{X_1+Z}(s_1), u_2 = F_{X_2}(s_2), \ldots, v = F_Y(t)$ and $w = F_Z(z)$, in (9), the perturbed copula of the random vector $(\boldsymbol{X}+Z\boldsymbol{e}, Y)$ will be

$$C_{\boldsymbol{X}+Z\boldsymbol{e},Y}(\boldsymbol{u}, v) = \int D_{n+2}C_{\boldsymbol{X},Y,Z}[F_{X_1}(F^{-1}_{X_1+Z}(u_1) - F^{-1}_Z(w)), u_2, \ldots, u_n, v, w]\,\mathrm{d}w,$$

which is (10). □

A special case of the above theorem will arise when we have the random vector (X, Y, Z) and its copula.[15]

Corollary 2.3. *If $C_{X,Y,Z}(u, v, w)$ is the copula function of the random vector (X, Y, Z), then the joint distribution and the copula function of the vector $(X + Z, Y)$ are given respectively by*

$$F_{X+Z,Y}(s,t) = \int D_3 C_{X,Y,Z}(F_X(s-z), F_Y(t), F_Z(z))\mathrm{d}F_Z(z), \tag{11}$$

and

$$C_{X+Z,Y}(u,v) = \int D_3 C_{X,Y,Z}[F_X(F^{-1}_{X+Z}(u) - F^{-1}_Z(w), v, w]\mathrm{d}w, \tag{12}$$

where $D_3 C_{X,Y,Z}(u, v, w) = \frac{\partial C_{X,Y,Z}(u,v,w)}{\partial w}$ if this derivate exists, and $D_3 C_{X,Y,Z}(u, v, w) = 0$, otherwise.

Proof. The proof is straightforward. □

A special case of Theorem (10) arises when the perturbed random variable Z is independent of X and Y, and is presented in the following corollary.

Corollary 2.4. *If X, Y, Z are connected by a copula $C_{X,Y,Z}$ in which, $C_{X,Y,Z} = wC_{X,Y}(u, v)$ then the joint distribution function and the copula function of $X + Z, Y$ are respectively given by*

$$F_{X+Z,Y}(t,y) = \int \int_{-\infty}^{y} D_2 C_{X,Y}(F_X(s-z), F_Y(u))\mathrm{d}u\mathrm{d}F_Z(z), \tag{13}$$

and

$$C_{X+Z,Y}(u,v) = \int \int_0^v D_2 C_{X,Y}(F_X(F^{-1}_{X+Z}(u) - z), F_Y(y))\mathrm{d}F_Y(y)\mathrm{d}F_Z(z), \tag{14}$$

where $D_2 C_{X,Y}(u, v) = \frac{\partial C_{X,Y}(u,v)}{\partial v}$ if this derivate exists, and $D_2 C_{X,Y}(u, v) = 0$ otherwise.

Proof. It is obvious. □

Example 2.5. *Let X and Y be random variables uniformly distributed on $[0, 1]$ and random variable Z have uniform distribution on $[0, \varepsilon]$, $0 < \varepsilon \leq 1$ and $C_{X,Y,Z}(u, v, w) = uC_2(v, w)$, where $C_2(v, w)$ is the Farlie-Gumbel-Morgenstern copula then*

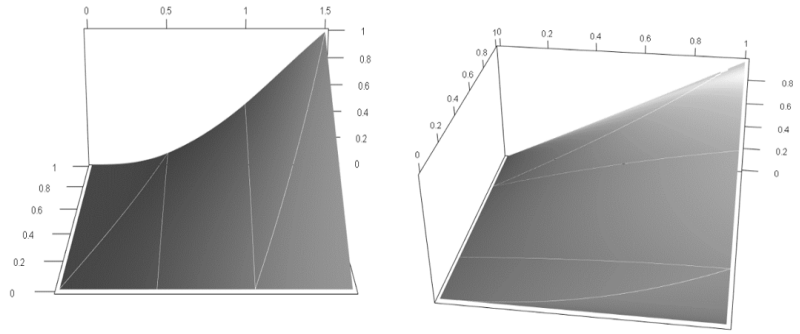
$$F_{X+Z,Y}(t,v) = \begin{cases} \frac{1}{\varepsilon^2}\left[\frac{1}{3}\varepsilon t^3 + \theta\left(\frac{1}{3}\varepsilon t^3 - \frac{1}{4}\varepsilon t^4 - \frac{1}{4}t^4 + \frac{1}{5}t^5\right)\right] & 0 < t < \varepsilon, \varepsilon v > t \\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 + \theta(\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 - \frac{1}{3}\varepsilon^2 t^3 v \\ +\frac{1}{12}\varepsilon^5 v^4 - \frac{1}{2}\varepsilon^2 t^2 v^2 + \frac{1}{4}\varepsilon^4 v^4 + \frac{1}{3}\varepsilon^2 t^3 v^2 - \frac{2}{15}\varepsilon^5 v^5)\big)] & 0 < t < \varepsilon, \varepsilon v < t \\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 + \theta(\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 - \frac{1}{3}\varepsilon^2 t^3 v \\ +\frac{1}{12}\varepsilon^5 v^4 - \frac{1}{2}\varepsilon^2 t^2 v^2 + \frac{1}{4}\varepsilon^4 v^4 + \frac{1}{3}\varepsilon^2 t^3 v^2 - \frac{2}{15}\varepsilon^5 v^5)\big)] & \varepsilon < t < 1 \\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 - \frac{1}{3}\varepsilon(t-1)^3 + \theta(\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{6}\varepsilon^4 v^3 - \\ \frac{1}{3}\varepsilon^2 t^3 v + \frac{1}{12}\varepsilon^5 v^4 - \frac{1}{2}\varepsilon^2 t^2 v^2 + \frac{1}{4}\varepsilon^4 v^4 + \frac{1}{3}\varepsilon^2 t^3 v^2 - \\ \frac{2}{15}\varepsilon^5 v^5 - \frac{1}{3}\varepsilon(t-1)^3 + \frac{1}{4}(\varepsilon+1)(t-1)^4 - \frac{1}{5}(t-1)^5)\big)] & 0 < t < 1+\varepsilon, \varepsilon v > t-1 \\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2 t^2 v - \frac{1}{2}\varepsilon^2 (t-1)^2 v + \theta(\frac{1}{2}t^2 - \\ \frac{1}{2}(t-1)^2 - \frac{1}{3}t^3 + \frac{1}{3}(t-1)^3(\varepsilon^2 v - \varepsilon^2 v^2)\big)] & 0 < t < 1+\varepsilon, \varepsilon v < t-1 \end{cases} \tag{15}$$

and

$$C_{X+Z,Y}(u,v)=\begin{cases}\frac{1}{\varepsilon^2}\left[\frac{1}{3}\varepsilon A^3+\theta\left(\frac{1}{3}\varepsilon A^3-\frac{1}{4}\varepsilon A^4-\frac{1}{4}A^4+\frac{1}{5}A^5\right)\right] & 0<u<\frac{\varepsilon}{2},\varepsilon v>A\\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2A^2v-\frac{1}{6}\varepsilon^4v^3+\theta(\frac{1}{2}\varepsilon^2A^2v-\frac{1}{6}\varepsilon^4v^3-\frac{1}{3}\varepsilon^2A^3v+ & \\ \frac{1}{12}\varepsilon^5v^4-\frac{1}{2}\varepsilon^2A^2v^2+\frac{1}{4}\varepsilon^4v^4+\frac{1}{3}\varepsilon^2A^3v^2-\frac{2}{15}\varepsilon^5v^5)] & 0<u<\frac{\varepsilon}{2},\varepsilon v<A\\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2B^2v-\frac{1}{6}\varepsilon^4v^3+\theta(\frac{1}{2}\varepsilon^2B^2v-\frac{1}{6}\varepsilon^4v^3-\frac{1}{3}\varepsilon^2B^3v+ & \\ \frac{1}{12}\varepsilon^5v^4-\frac{1}{2}\varepsilon^2B^2v^2+\frac{1}{4}\varepsilon^4v^4+\frac{1}{3}\varepsilon^2B^3v^2-\frac{2}{15}\varepsilon^5v^5)] & \frac{\varepsilon}{2}<u<1-\frac{\varepsilon}{2}\\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2C^2v-\frac{1}{6}\varepsilon^4v^3-\frac{1}{3}\varepsilon(C-1)^3+ & \\ \theta(\frac{1}{2}\varepsilon^2C^2v-\frac{1}{6}\varepsilon^4v^3-\frac{1}{3}\varepsilon^2C^3v+ & \\ \frac{1}{12}\varepsilon^5v^4-\frac{1}{2}\varepsilon^2C^2v^2+\frac{1}{4}\varepsilon^4v^4+\frac{1}{3}\varepsilon^2C^3v^2- & \\ \frac{2}{15}\varepsilon^5v^5-\frac{1}{3}\varepsilon(C-1)^3+\frac{1}{4}(\varepsilon+1)(C-1)^4-\frac{1}{5}(C-1)^5)] & 1-\frac{\varepsilon}{2}<u<1,\varepsilon v>C-1\\ \frac{1}{\varepsilon^2}[\frac{1}{2}\varepsilon^2C^2v-\frac{1}{2}\varepsilon^2(C-1)^2v+\theta(\frac{1}{2}C^2-\frac{1}{2}(C-1)^2 & \\ -\frac{1}{3}C^3+\frac{1}{3}(C-1)^3)(\varepsilon^2v-\varepsilon^2v^2)] & 1-\frac{\varepsilon}{2}<u<1,\varepsilon v<C-1\end{cases}\tag{16}$$

where $A=\sqrt{2\varepsilon u}, B=u+\frac{\varepsilon}{2}$ *and* $C=\varepsilon+1-\sqrt{2\varepsilon(1-u)}$.

For detailed computations, see Appendix. Figure 1 depicts a three-dimensional visualization of the distribution of joint random variables $X+Z$ and Y as well as its copula.



1.a 1.b

Figure 1: The joint distribution and the copula function $C_{X+Z,Y}$ of Example 2.5. for $\varepsilon=0.5$ (1.a) The joint distribution function, (1.b) The copula function.

Example 2.6. *Let* X *and* Y *be random variables uniformly distributed on* $[0,1]$ *and random variable* Z *have uniform distribution on* $[0,\varepsilon]$, $0<\varepsilon\leq 1$ *and* $C_{X,Y,Z}(u,v,w)=uM_2(v,w)$, *then*

$$C_{X+Z,Y}(u,v)=\begin{cases}\frac{(2\sqrt{2\varepsilon u}-\varepsilon v)v}{2} & 0\leqslant u\leqslant\frac{\varepsilon}{2},\varepsilon v\leqslant\sqrt{2\varepsilon u}\\ u & 0\leqslant u\leqslant\frac{\varepsilon}{2},\varepsilon v\geqslant\sqrt{2\varepsilon u}\\ \frac{(2u+\varepsilon-\varepsilon v)v}{2} & \frac{\varepsilon}{2}\leqslant u\leqslant 1-\frac{\varepsilon}{2}\\ v-\frac{(\varepsilon v-g(u)+1)^2}{2\varepsilon} & 1-\frac{\varepsilon}{2}\leqslant u\leqslant 1,\varepsilon v\geqslant g(u)-1\\ v & 1-\frac{\varepsilon}{2}\leqslant u\leqslant 1,\varepsilon v\leqslant g(u)-1\end{cases}\tag{17}$$

where $g(u)=\varepsilon+1-\sqrt{2\varepsilon(1-u)}$, *and*

$$F_{X+Z,Y}(t,v)=\begin{cases}\frac{t^2}{2\varepsilon} & 0<t<\varepsilon,\varepsilon v>t\\ vt-\frac{1}{2}\varepsilon v^2 & 0<t<\varepsilon,\varepsilon v<t\\ vt-\frac{1}{2}\varepsilon v^2 & \varepsilon<t<1\\ v & 1<t<1+\varepsilon,\varepsilon v<t-1\\ \frac{1}{\varepsilon}[-\frac{1}{2}+\varepsilon tv-\frac{1}{2}\varepsilon^2v^2-\frac{1}{2}t^2+t] & 1<t<1+\varepsilon,\varepsilon v>t-1\end{cases}$$

Figure 2 presents the joint distribution function as well as the copula function of $(X+Z,Y)$ in this example.

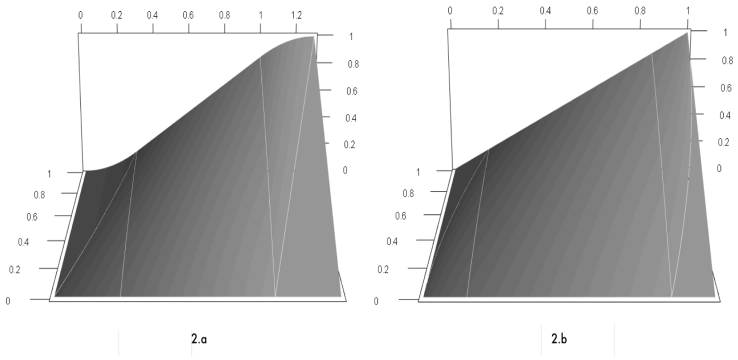


Figure 2: The joint distribution and the copula function $C_{X+Z,Y}$ of Example 2.6 for $\varepsilon = 0.3$ (2.a) The joint distribution function, (2.b) The copula function.

Example 2.7. *Let X, Y and Z are random variables uniformly distributed on $[0,1]$ while X and Z are connected by the Fréchet-Mardia copula* (3) *with parameter $\alpha \in [0,1]$ and Y, Z associated to M_2, then the joint distribution function, $F_{X+Z,Y}$ and the copula function $C_{X+Z,Y}$ are respectively given by*

$$F_{X+Z,Y}(s,y) = \begin{cases} \alpha(sy - \frac{1}{4}s^2 - \frac{1}{2}y^2) + (1-\alpha)(\frac{1}{2}s^2y - \frac{1}{6}y^3) & 0 < s < 1, \frac{s}{2} < y < s \\ \frac{1}{4}\alpha s^2 + (1-\alpha)\frac{s^3}{3} & 0 < s < 1, y > s \\ \frac{1}{2}\alpha y^2 + (1-\alpha)(\frac{1}{2}s^2y - \frac{1}{6}y^3) & 0 < s < 1, 0 < y < \frac{s}{2} \\ \frac{1}{2}\alpha y^2 + (1-\alpha)(sy - \frac{y}{2}) & 1 < s < 2, y < s-1 \\ \frac{1}{2}\alpha y^2 + (1-\alpha)(-\frac{y^3}{6} + s^2 - s - \frac{s^3}{3} + \frac{1}{2}s^2y + \frac{1}{3}) & 1 < s < 2, s-1 < y < \frac{s}{2} \\ \alpha(-\frac{s^2}{4} - \frac{y^2}{2} + sy) + & \\ (1-\alpha)(-\frac{s^3}{3} + \frac{1}{2}s^2y - \frac{y^3}{6} + s^2 - s + \frac{1}{3}) & 1 < s < 2, y > \frac{s}{2} \end{cases} \tag{18}$$

and

$$C_{X+Z,Y}(u,v) = \begin{cases} \alpha v + (1-\alpha)(vs_1 - \frac{v^2}{2}) & , 0 \leqslant u \leqslant \frac{1}{2}, 0 \leqslant v \leqslant \frac{s_1}{2} \\ \frac{s_1\alpha - v^2(1-\alpha)}{2} + s_1v(1-\alpha) & , 0 \leqslant u \leqslant \frac{1}{2}, \frac{s_1}{2} \leqslant v \leqslant s_1 \\ \frac{s_1\alpha + s_1^2(1-\alpha)}{2} & 0 \leqslant u \leqslant \frac{1}{2}, s_1 \leqslant v \leqslant 1 \\ v & \frac{1}{2} \leqslant u \leqslant 1, 0 \leqslant v \leqslant s_2 - 1 \\ \frac{(\alpha-1)(s_2-v)^2 + 2s_2 - 2\alpha(s_2-v) - 1 + \alpha}{2} & \frac{1}{2} \leqslant u \leqslant 1, s_2 - 1 \leqslant v \leqslant \frac{s_2}{2} \\ \frac{(\alpha-1)(s_2-v)^2 + 2s_2 - \alpha s_2 - 1 + \alpha}{2} & \frac{1}{2} \leqslant u \leqslant 1, \frac{s_2}{2} \leqslant v \leqslant 1, \end{cases} \tag{19}$$

where $s_1 = \frac{\sqrt{\alpha^2 + 8u(1-\alpha)} - 1}{2(1-\alpha)}$ and $s_2 = \frac{3\alpha - \sqrt{8u(\alpha-1) + \alpha^2 + 8(1-\alpha)}}{2(\alpha-1)}$. For more details we refer to Sheikhi et al.(2020) [14].

3 Copula-based measurement error models

Consider the random variables X and Y, that are linked by a copula $C_{X,Y}$, and the random variable Z, that is independent of X and Y. In this section, we use the results of the previous section to develop a copula-based Berkson measurement error model.

Consider the linear regression model with the Berkson measurement error (1). The following corollary enables us to get the desired regression function.

Corollary 3.1. *Under the assumptions of corollary 2.4, by defining $T = X + Z$, the regression equation of model* (1) *can be obtained as*

$$r_{Y|T}(t) = \int y f_{Y|T}(y|t)\mathrm{d}y = \int y D_{12}C_{X+Z,Y}(F_{X+Z}(t), F_Y(y))\mathrm{d}F_Y(y), \tag{20}$$

where $D_{12}C_{X,Y}(u,v) = \frac{\partial^2 C_{X,Y}(u,v)}{\partial u \partial v}$ if this derivate exists, and $D_{12}C_{X,Y} = 0$ otherwise.

Proof. From (1) we obtain $Y = \alpha + \beta(X + Z) + \delta$ and we have

$$\begin{aligned} r_{Y|T}(t) = \mathrm{E}_{Y|T}(y|t) &= \int y \frac{\partial}{\partial y} P(Y \leqslant y | T = t) \mathrm{d}y \\ &= \int y \frac{\partial}{\partial y} D_1 C_{T,Y}(F_T(t), F_Y(y)) \mathrm{d}y = \int y D_{12} C_{T,Y}(F_T(t), F_Y(y)) \mathrm{d}F_Y(y), \end{aligned}$$

which proves the claim. □

The following examples present a measurement error model with the assumptions of examples 2.5, 2.6, and 2.7.

Example 3.2. *With the assumptions of example 2, the measurement error model* (1) *is estimated as*

$$r_{Y|T}(t) = \begin{cases} \frac{t^2}{2\varepsilon} + \theta(\frac{t^2}{2\varepsilon} - \frac{t^3}{\varepsilon} - \frac{2t^3}{3\varepsilon^2} + \frac{2t^4}{3\varepsilon^2}) & 0 < t < \varepsilon \\ \frac{t}{2} + \theta(-\frac{t}{6} + \frac{t^2}{6}) & \varepsilon < t < 1 \\ \frac{\varepsilon}{1+\varepsilon-t}[\frac{t}{2} + (1-t)\frac{k^2}{2} + \theta(-\frac{t}{6} + \frac{t^2}{6} + \\ (\frac{t^2}{2} - \frac{3t}{2} + 1)k^2 + (2t - \frac{2t^2}{3} - \frac{4}{3})k^3)] & 1 < t < 1 + \varepsilon, \end{cases} \tag{21}$$

where $k = \frac{t-1}{\varepsilon}$. *Figure 3 depicts the simulation data as well as the fitted curves for* $\varepsilon = 0.8$ *and* $\theta = 0.3$ *(left figure) and for* $\varepsilon = 0.9$ *and* $\theta = 0.4$ *(right figure) of this example. For details see Section 4.*

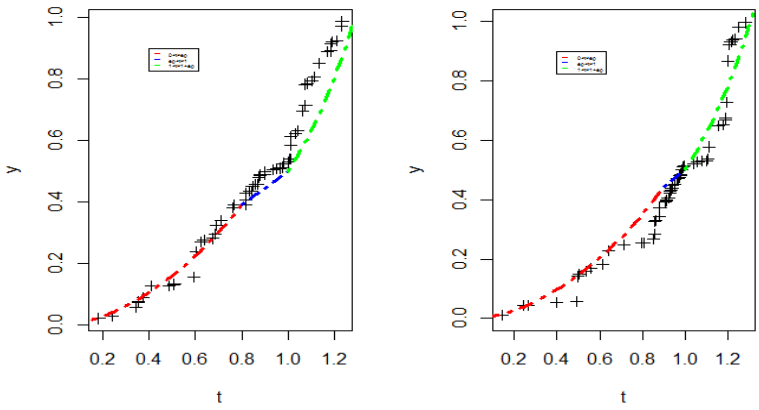


Figure 3: Berkson measurement error regression equations and the simulated data for example 3.2. The "+" symbols are depicting the simulated data and the colored lines are their corresponding fitted curves.

Example 3.3. *Let* X *and* Y *be random variables uniformly distributed on* $[0, 1]$ *and random variable* Z *have uniform distribution on* $[0, \varepsilon]$, $0 < \varepsilon \leq 1$ *and* $C_{X,Y,Z}(u, v, w) = uM_2(v, w)$, *then*

$$r_{Y|T}(t) = \begin{cases} \frac{t}{2\varepsilon} & 0 < t < \varepsilon \\ \frac{1}{2} & \varepsilon < t < 1 \\ \frac{\varepsilon+t-1}{2\varepsilon} & 1 < t < 1 + \varepsilon. \end{cases} \tag{22}$$

Example 3.4. *Let* X, Y *and* Z *are random variables uniformly distributed on* $[0, 1]$ *and* X, Z *are connected by the Fréchet-Mardia copula* (3) *with parameter* $\alpha \in [0, 1]$ *and* Y, Z *associated to* M_2, *then*

$$r_{Y|T}(t) = \begin{cases} \frac{(1-\alpha)t^2}{2t(1-\alpha)+\alpha} & 0 \leqslant t \leqslant 1 \\ \frac{(1-\alpha)(2t-t^2)}{1+(\alpha-1)(2t-3)} & 1 \leqslant t \leqslant 2. \end{cases} \tag{23}$$

4 Simulation results

In this section, a simulation study is conducted to evaluate the results of the previous section. In order to investigate the results of Example 3.2, we consider that X and Y are uniformly distributed on $[0, 1]$ and that the random variable Z has a uniform distribution on $[0, \varepsilon]$, for values of $\varepsilon = 0.8$ and 0.9. Also, we assume that the Farlie-Gumbel-Morgenstern copula couples Y and Z with two values of $\theta = 0.3$ and $\theta = 0.4$. We take a sample size of 200 from these distributions and repeat this simulation ten times and calculate the average of the simulated data. Figure 3 depicts Berkson measurement error regression equations for some values of ε and θ in this example. The Left panel shows the simulated data, which are denoted by "+" symbols, as well as the fitted regression function (21) when the random variable Z is simulated from $U[0, \varepsilon]$ with $\varepsilon = 0.8$ and Y and Z are coupled by a Farlie-Gumbel-Morgenstern copula with $\theta = 0.3$. Similarly, the right panel, is for $\varepsilon = 0.9$ and $\theta = 0.4$.

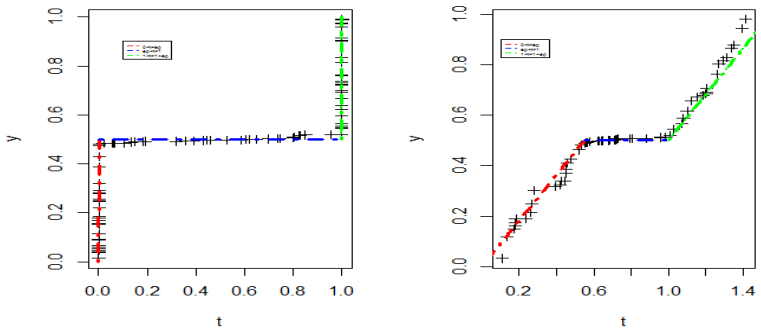


Figure 4: Berkson measurement error regression equations and the simulated data for example 3.3. The "+" symbols are depicting the simulated data and the colored lines are their corresponding fitted curves.

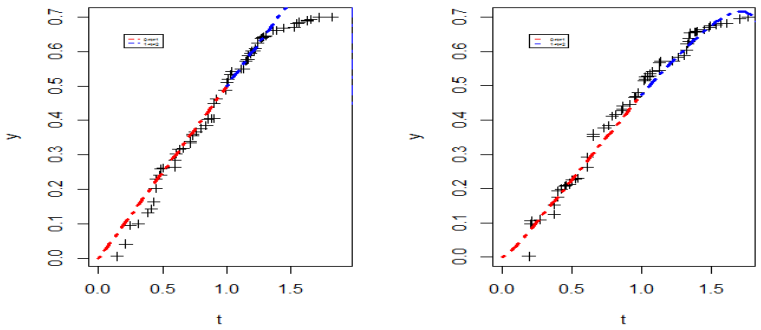


Figure 5: Berkson measurement error regression equations and the simulated data for Example 3.4. The "+" symbols are depicting the simulated data and the colored lines are their corresponding fitted curves.

A similar strategy is done to assess and visualize the results of Examples 3.3 and 3.4. The left panel of figure 4 present the simulated data, which are denoted by "+" symbols, as well as the fitted regression function (22) of Example 3.3 with the value of $\varepsilon = 0.001$ and the right panel, is for $\varepsilon = 0.55$. Finally, the left panel of Figure 5 presents the simulated data, which are denoted by "+" symbols as well as the fitted regression function (23) of Example 3.4 with values of $\alpha = 0.001$ and $\alpha = 0.1$ for the right one.

5 Conclusion

We introduced a copula-based Berkson measurement error model by considering the function $C_{\boldsymbol{X},Y,Z}$ as a copula of the random vector $(\boldsymbol{X}, Y, Z)$. By presenting a general form, we investigated some of its special cases and obtained the bivariate distribution function as well as the copula of the random vector $(\boldsymbol{X} + Z\boldsymbol{e}, Y)$. These two functions led us to carry out a Berkson measurement error model.

The most important advantage of this approach is the implanting any form of linear and nonlinear relations between explanatory variables and the measurement error variables, which can be captured by various forms of copulas, besides its computational complexities.

The idea of the present work can be extended in many cases. In our examples, we assumed uniform marginals while other statistical distributions may be assumed elsewhere. Also, considering other connecting copulas such as Archimedean copulas as well as the elliptical copulas, especially the Gaussian copula, would be of interest. In this work, we have considered one of the explanatory variables being prone to measurement error. Other cases may arise when more than one of the explanatory variables are measured with error. Also, performing a measurement error analysis when one of the regressors is associated with the classical measurement error, and one of the other regressors is related to the Berkson-type measurement error will open interesting fields in this topic [4, 6].

References

[1] P. Arbenz, P. Embrechts, G. Puccetti, *The AEP algorithm for the fast computation of the distribution of the sum of dependent random variables*, Bernoulli, **17**(2) (2011), 562-591.

[2] J. Berkson, *Are there two regressions?*, Journal of the American Statistical Association, **45**(250) (1950), 164-180.

[3] J. Buonaccorsi, *Measurement error: Models, methods, and applications*, CRC press, 2010.

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, *Measurement error in nonlinear models: A modern perspective*, CRC press, 2006.

[5] X. Chen, H. Hong, D. Nekipelov, *Nonlinear models of measurement errors*, Journal of Economic Literature, **49**(4) (2011), 901-937.

[6] V. Deffner, H. Küchenhoff, S. Breitner, A. Schneider, J. Cyrys, A. Peters, *Mixtures of Berkson and classical covariate measurement error in the linear mixed model: Bias analysis and application to a study on ultra-fine particles*, Biometrical Journal, **60**(3) (2018), 480-497.

[7] A. Delaigle, P. Hall, P. Qiu, *Nonparametric methods for solving the Berkson errors-in-variables problem*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **68**(2) (2006), 201-220.

[8] F. Durante, C. Sempi, *Principles of copula theory*, CRC press, 2015.

[9] M. Fréchet, *Au sujet de la note précédente*, Comptes Rendus Hebdomadires des Séances de l'Académie des Sciences, **246**(19) (1958), 2719-2720.

[10] W. A. Fuller, *Measurement error models*, John Wiley and Sons, 2009.

[11] R. Mesiar, A. Sheikhi, M. Komorníková, *Random noise and perturbation of copulas*, Kybernetika, **55**(2) (2019), 422-434.

[12] R. B. Nelson, *An introduction to copulas*, Springer Science and Business Media, 2007.

[13] S. M. Schennach, *Regressions with Berkson errors in covariates-a nonparametric approach*, The Annals of Statistics, **41**(3) (2013), 1642-1668.

[14] A. Sheikhi, F. Arad, R. Mesiar, L. Vavríková, *Random noise and perturbation of copula with a copula induced noise*, International Journal of General Systems, **49**(8) (2020), 856-871.

[15] A. Sheikhi, R. Mesiar, *Copula-based measurement error models*, Iranian Journal of Fuzzy Systems, **17**(6) (2020), 29-38.

[16] A. Sklar, *Fonctions de répartition an dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris, **8** (1959), 229-231.

[17] L. Wang, *Estimation of nonlinear models with Berkson measurement errors*, The Annals of Statistics, **32**(6) (2004), 2559-2579.

[18] R. Wang, *Sum of arbitrarily dependent random variables*, Electronic Journal of Probability, **19** (2014), 1-18.

Appendix

Detailed computation for Example 2.5:

$$\begin{aligned} F_{X+Z,Y}(t,v) &= \int\int D_1 C_{Y,Z}(F_Y(u), F_Z(\varepsilon(t-x)))\mathrm{d}u\mathrm{d}F_X(x) \\ &= \int\int (t-x) + \theta(t-x)(1-(t-x))\frac{1}{\varepsilon}(\varepsilon - 2\varepsilon u)\mathrm{d}u\mathrm{d}x \\ &= -\frac{1}{\varepsilon^2}\int\int \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z, \end{aligned}$$

where $\varepsilon u = z$ and $t - x = M$.
Using (13) we have the following cases to obtain the joint CDF of $X + Z, Y$:

1. If $0 < t < \varepsilon$ and $\varepsilon v > t$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon^2}\int_0^t \int_z^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

2. If $0 < t < \varepsilon$ and $\varepsilon v < t$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v} \int_z^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

3. If $\varepsilon < t < 1$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v} \int_z^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

4. If $1 < t < 1 + \varepsilon$ and $\varepsilon v > t - 1$:

$$\begin{aligned} F_{X+Z,Y}(t,v) &= \frac{1}{\varepsilon^2}\int_0^{t-1} \int_{t-1}^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z \\ &\quad + \frac{1}{\varepsilon^2}\int_{t-1}^{\varepsilon v} \int_z^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z. \end{aligned}$$

5. If $1 < t < 1 + \varepsilon$ and $\varepsilon v < t - 1$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v} \int_{t-1}^t \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

Now for computation of $C_{X+Z,Y}$ we have

$$\begin{aligned} C_{X+Z,Y}(u,v) &= \int D_2 C_{X,Z}(F_X(F^{-1}_{X+Z}(u) - F^{-1}_Z(v)), v)\mathrm{d}v \\ &= \frac{1}{\varepsilon^2}\int\int \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z. \end{aligned}$$

Using (14) we have the following cases to obtain the conecting copula of $X + Z, Y$:

1. If $0 < u < \frac{\varepsilon}{2}$ and $\varepsilon v > A$ where $A = \sqrt{2\varepsilon u}$:

$$C_{X+Z,Y}(u,v) = \frac{1}{\varepsilon^2}\int_0^A \int_z^A \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

2. If $0 < u < \frac{\varepsilon}{2}$ and $\varepsilon v < A$ where $A = \sqrt{2\varepsilon u}$:

$$C_{X+Z,Y}(u,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v} \int_z^A \varepsilon M + \theta(M-M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

3. If $\frac{\varepsilon}{2} < u < 1 - \frac{\varepsilon}{2}$ where $B = u + \frac{\varepsilon}{2}$:

$$C_{X+Z,Y}(u,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v}\int_z^B \varepsilon M + \theta(M - M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

4. If $1 - \frac{\varepsilon}{2} < u < 1$ and $\varepsilon v > C - 1$ where $C = \varepsilon + 1 - \sqrt{2\varepsilon(1-u)}$:

$$\begin{aligned} C_{X+Z,Y}(u,v) = &\frac{1}{\varepsilon^2}\int_0^{C-1}\int_{C-1}^{C} \varepsilon M + \theta(M - M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z \\ &+ \frac{1}{\varepsilon^2}\int_{C-1}^{\varepsilon v}\int_z^{C} \varepsilon M + \theta(M - M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z. \end{aligned}$$

5. If $1 - \frac{\varepsilon}{2} < u < 1$ and $\varepsilon v < C - 1$ where $C = \varepsilon + 1 - \sqrt{2\varepsilon(1-u)}$:

$$C_{X+Z,Y}(u,v) = \frac{1}{\varepsilon^2}\int_0^{\varepsilon v}\int_{C-1}^{C} \varepsilon M + \theta(M - M^2)(\varepsilon - 2z)\mathrm{d}M\mathrm{d}z.$$

Detailed computation for Example 2.6:
We have $D_1 C_{Y,Z}(y, F_Z(t-x)) = 1$, see e.g., [14], then by using (13) we have the following cases to obtain the joint CDF of $X + Z, Y$:

$$F_{X+Z,Y}(t,v) = \int\int D_1 C_{Y,Z}(y, F_Z(t-x))\mathrm{d}y\mathrm{d}x = \frac{1}{\varepsilon}\int\int \mathrm{d}z\mathrm{d}x.$$

1. If $0 < t < \varepsilon$ and $\varepsilon v > t$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon}\int_0^t\int_0^{t-z} \mathrm{d}x\ \mathrm{d}z = \frac{t^2}{2\varepsilon}.$$

2. If $0 < t < \varepsilon$ and $\varepsilon v < t$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon}\int_0^{\varepsilon v}\int_0^{t-z} \mathrm{d}x\ \mathrm{d}z = vt - \frac{1}{2}\varepsilon v^2.$$

3. If $\varepsilon < t < 1$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon}\int_0^{\varepsilon v}\int_0^{t-z} \mathrm{d}x\ \mathrm{d}z = vt - \frac{1}{2}\varepsilon v^2.$$

4. If $1 < t < 1 + \varepsilon$ and $\varepsilon v < t - 1$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon}\int_0^{\varepsilon v}\int_0^{1} \mathrm{d}x\ \mathrm{d}z = v.$$

5. If $1 < t < 1 + \varepsilon$ and $\varepsilon v > t - 1$:

$$F_{X+Z,Y}(t,v) = \frac{1}{\varepsilon}\int_0^{t-1}\int_0^{1} \mathrm{d}x\mathrm{d}z + \frac{1}{\varepsilon}\int_{t-1}^{\varepsilon v}\int_0^{t-z} \mathrm{d}x\ \mathrm{d}z = \frac{1}{\varepsilon}\left[-\frac{1}{2} + \varepsilon tv - \frac{1}{2}\varepsilon^2 v^2 - \frac{1}{2}t^2 + t\right].$$