Goodness-of-Fit Testing with Empirical Copulas

Sami Umut Can
John Einmahl
Roger Laeven

EURANDOM

August 29, 2011

Overview of Copulas

Overview of Copulas

- A bivariate copula C is a bivariate cdf defined on $[0,1]^2$ with uniform marginal distributions on $[0,1]$.

Overview of Copulas

- A bivariate copula C is a bivariate cdf defined on $[0,1]^2$ with uniform marginal distributions on $[0,1]$.
- More precisely, a function $C : [0,1]^2 \to [0,1]$ is called a bivariate copula if
  - $C(x,0) = C(0,y) = 0$ for any $x, y \in [0,1]$
  - $C(x,1) = x, C(1,y) = y$ for any $x, y \in [0,1]$
  - $C(x_2,y_2) - C(x_1,y_2) - C(x_2,y_1) + C(x_1,y_1) \geq 0$ for any $x_1, x_2, y_1, y_2 \in [0,1]$ with $x_1 \leq x_2$ and $y_1 \leq y_2$

- Sklar's Theorem:
  Let H be a bivariate cdf with continuous marginal cdf's

$$H(x, \infty) = F(x), \quad H(\infty, y) = G(y).$$

Then there exists a unique copula C such that

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

Conversely, for any univariate cdf's F and G and any copula C, (1) defines a bivariate cdf H with marginals F and G.

- Sklar's Theorem:
  Let H be a bivariate cdf with continuous marginal cdf's

$$H(x, \infty) = F(x), \quad H(\infty, y) = G(y).$$

  Then there exists a unique copula C such that

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

  Conversely, for any univariate cdf's F and G and any copula C, (1) defines a bivariate cdf H with marginals F and G.

- C captures the dependence structure of two random variables. It is used for dependence modeling in finance and actuarial science.

Goodness-of-Fit Testing

Goodness-of-Fit Testing

- Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an unknown bivariate distribution H, with unknown continuous marginal distributions F and G, and a corresponding copula C, how can we decide if a given copula C_0 or a given parametric family of copulas $\{C_\theta, \theta \in \Theta\}$ is a good fit for the sample?

Goodness-of-Fit Testing

- Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an unknown bivariate distribution H, with unknown continuous marginal distributions F and G, and a corresponding copula C, how can we decide if a given copula C_0 or a given parametric family of copulas $\{C_\theta, \theta \in \Theta\}$ is a good fit for the sample?
- In other words, we would like to perform a hypothesis test about C, with a null hypothesis of the form $C = C_0$ or $C \in \{C_\theta, \theta \in \Theta\}$. For now, we consider the simple hypothesis ($C = C_0$) only.

Goodness-of-Fit Testing

- Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an unknown bivariate distribution H, with unknown continuous marginal distributions F and G, and a corresponding copula C, how can we decide if a given copula C_0 or a given parametric family of copulas $\{C_\theta, \theta \in \Theta\}$ is a good fit for the sample?
- In other words, we would like to perform a hypothesis test about C, with a null hypothesis of the form $C = C_0$ or $C \in \{C_\theta, \theta \in \Theta\}$. For now, we consider the simple hypothesis ($C = C_0$) only.
- A natural starting point for constructing goodness-of-fit tests is the so-called *empirical copula*.

- Note that we can write

$$C(x, y) = H(F^{-1}(x), G^{-1}(y)), \quad (x, y) \in [0, 1]^2,$$

with $F^{-1}(x) = \inf\{t \in \mathbb{R} : F(t) \geq x\}$, and similarly for G^{-1}.

- Note that we can write

$$C(x, y) = H(F^{-1}(x), G^{-1}(y)), \quad (x, y) \in [0, 1]^2,$$

with $F^{-1}(x) = \inf\{t \in \mathbb{R} : F(t) \geq x\}$, and similarly for G^{-1}.

- So a natural way of estimating the copula C is using the *empirical copula*

$$C_n(x, y) = H_n(F_n^{-1}(x), G_n^{-1}(y)), \quad (x, y) \in [0, 1]^2,$$

with

$$H_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x, Y_i \leq y\},$$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}, \quad G_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$$

- It is known that the *empirical copula process*

$$D_n(x,y) = \sqrt{n}(C_n(x,y) - C(x,y)), \quad (x,y) \in [0,1]^2$$

converges weakly in $\ell^\infty([0,1]^2)$ to a C-Brownian pillow, under the assumption that

$$C^x(x,y) \text{ is continuous on } \{(x,y) \in [0,1]^2 : 0 < x < 1\},$$
$$C^y(x,y) \text{ is continuous on } \{(x,y) \in [0,1]^2 : 0 < y < 1\}.$$

- It is known that the *empirical copula process*

$$D_n(x, y) = \sqrt{n}(C_n(x, y) - C(x, y)), \quad (x, y) \in [0, 1]^2$$

converges weakly in $\ell^\infty([0, 1]^2)$ to a C-Brownian pillow, under the assumption that

$$C^x(x, y) \text{ is continuous on } \{(x, y) \in [0, 1]^2 : 0 < x < 1\},$$
$$C^y(x, y) \text{ is continuous on } \{(x, y) \in [0, 1]^2 : 0 < y < 1\}.$$

- A C-Brownian sheet $W(x, y)$ is a mean zero Gaussian process with covariance function

$$\mathrm{Cov}[W(x, y), W(x', y')] = C(x \wedge x', y \wedge y'), \quad x, x', y, y' \in [0, 1].$$

- It is known that the *empirical copula process*

$$D_n(x, y) = \sqrt{n}(C_n(x, y) - C(x, y)), \quad (x, y) \in [0, 1]^2$$

converges weakly in $\ell^\infty([0, 1]^2)$ to a C-Brownian pillow, under the assumption that

$C^x(x, y)$ is continuous on $\{(x, y) \in [0, 1]^2 : 0 < x < 1\}$,
$C^y(x, y)$ is continuous on $\{(x, y) \in [0, 1]^2 : 0 < y < 1\}$.

- A C-Brownian sheet $W(x, y)$ is a mean zero Gaussian process with covariance function

$$\mathrm{Cov}[W(x, y), W(x', y')] = C(x \wedge x', y \wedge y'), \quad x, x', y, y' \in [0, 1].$$

- A C-Brownian pillow $D(x, y)$ is a mean zero Gaussian process that is equal in distribution to the C-Brownian sheet W, conditioned on $W(x, y) = 0$ for any $(x, y) \in [0, 1]^2 \setminus (0, 1)^2$.

- We have

$$D(x,y) = W(x,y) - C^x(x,y)W(x,1) - C^y(x,y)W(1,y) \\ - (C(x,y) - xC^x(x,y) - yC^y(x,y))W(1,1).$$

- We have

$$D(x,y) = W(x,y) - C^x(x,y)W(x,1) - C^y(x,y)W(1,y) \\ - (C(x,y) - xC^x(x,y) - yC^y(x,y))W(1,1).$$

- So we know the asymptotic distribution of the empirical copula process

$$D_n(x,y) = \sqrt{n}(C_n(x,y) - C(x,y)),$$

and we can take a functional of D_n (such as the sup over $[0,1]^2$ or an appropriate integral) as a test statistic for a goodness-of-fit test.

- We have

$$\begin{aligned} D(x,y) = W(x,y) - C^x(x,y)W(x,1) - C^y(x,y)W(1,y) \\ - (C(x,y) - xC^x(x,y) - yC^y(x,y))W(1,1). \end{aligned}$$

- So we know the asymptotic distribution of the empirical copula process

$$D_n(x,y) = \sqrt{n}(C_n(x,y) - C(x,y)),$$

and we can take a functional of D_n (such as the sup over $[0,1]^2$ or an appropriate integral) as a test statistic for a goodness-of-fit test.

- Problem: The asymptotic distribution of D_n, and that of the test statistic, depends on C. We would like to have a *distribution-free* goodness-of-fit test.

Scanning

Scanning

- Idea: Transform D_n into another process, say Z_n, whose asymptotic distribution is independent of C. Use an appropriate functional of the new process Z_n as a test statistic for goodness-of-fit tests.

Scanning

- Idea: Transform D_n into another process, say Z_n, whose asymptotic distribution is independent of C. Use an appropriate functional of the new process Z_n as a test statistic for goodness-of-fit tests.
- We use E. Khmaladze's "scanning" idea to transform D into a standard two-parameter Wiener process Z defined on $[0, 1]^2$. The same transformation applied to D_n will then produce a process Z_n that will, hopefully, converge to Z.

Scanning

- Idea: Transform D_n into another process, say Z_n, whose asymptotic distribution is independent of C. Use an appropriate functional of the new process Z_n as a test statistic for goodness-of-fit tests.

- We use E. Khmaladze's "scanning" idea to transform D into a standard two-parameter Wiener process Z defined on $[0,1]^2$. The same transformation applied to D_n will then produce a process Z_n that will, hopefully, converge to Z.

- Assumptions on C: Continuous first-order partial derivatives on $[0,1]^2 \setminus \{(0,0),(0,1),(1,0),(1,1)\}$, continuous second-order partial derivatives on $(0,1)^2$, strictly positive mixed partial C^{xy} on $(0,1)^2$, and more (to be determined).

- Define a grid $\{(x_i, y_j) : 0 \leq i, j \leq N\}$ on $[0,1]^2$ such that

$$0 = x_0 < x_1 < \ldots < x_N = 1$$
$$0 = y_0 < y_1 < \ldots < y_N = 1$$

  and define filtrations

$$\mathcal{F}_x(x_i) = \sigma\{D(x_h, y_k) : 0 \leq h \leq i, 0 \leq k \leq N\}, \quad 0 \leq i \leq N$$
$$\mathcal{F}_y(y_j) = \sigma\{D(x_h, y_k) : 0 \leq h \leq N, 0 \leq k \leq j\}, \quad 0 \leq j \leq N$$

- Define a grid $\{(x_i, y_j) : 0 \leq i, j \leq N\}$ on $[0, 1]^2$ such that

$$0 = x_0 < x_1 < \ldots < x_N = 1$$
$$0 = y_0 < y_1 < \ldots < y_N = 1$$

  and define filtrations

$$\mathcal{F}_x(x_i) = \sigma\{D(x_h, y_k) : 0 \leq h \leq i, 0 \leq k \leq N\}, \quad 0 \leq i \leq N$$
$$\mathcal{F}_y(y_j) = \sigma\{D(x_h, y_k) : 0 \leq h \leq N, 0 \leq k \leq j\}, \quad 0 \leq j \leq N$$

- "Scan" the process D with respect to the filtration $\{\mathcal{F}_x\}$:

$$K_1^{(N)}(x_i, y_j) = \sum_{h=0}^{i-1} \Big(D(x_{h+1}, y_j) - D(x_h, y_j) \\ - E[D(x_{h+1}, y_j) - D(x_h, y_j) | \mathcal{F}_x(x_h)] \Big)$$

- We compute

$$
\begin{aligned}
&K_1^{(N)}(x_i, y_j) \\
&= D(x_i, y_j) - \sum_{h=0}^{i-1} D(x_h, y_j) \left(\frac{E[D(x_h, y_j) D(x_{h+1}, y_j)]}{E[D(x_h, y_j)^2]} - 1 \right)
\end{aligned}
$$

- We compute

$$K_1^{(N)}(x_i, y_j)$$
$$= D(x_i, y_j) - \sum_{h=0}^{i-1} D(x_h, y_j) \left(\frac{E[D(x_h, y_j)D(x_{h+1}, y_j)]}{E[D(x_h, y_j)^2]} - 1 \right)$$

- Making the x-partitioning finer and finer, we obtain

$$K_1(x, y_j) = D(x, y_j) - \int_0^x D(s, y_j)\xi_1(ds, y_j)$$

as a limit in probability, where ξ_1 is an absolutely continuous measure whose density is determined by C and its first- and second-order derivatives.

- Next, we scan K_1 with respect to the filtration $\{\mathcal{F}_y\}$ and take the limit as the y-partition gets finer and finer:

$$K(x,y) = D(x,y) - \int_0^x D(s,y)\xi_1(ds,y) - \int_0^y D(x,t)\xi_2(x,dt) + \int_0^x \int_0^y D(s,t)\xi_1(ds,t)\xi_2(s,dt),$$

where ξ_2 is another absolutely continuous measure whose density is determined by C and its first- and second-order derivatives.

- **Theorem:** K is a C-Brownian sheet.

- **Theorem:** K is a C-Brownian sheet.
- **Proof:**

- **Theorem:** K is a C-Brownian sheet.
- **Proof:**
  - K is a mean zero Gaussian process since D is, so it remains to show that K has the covariance structure of a C-Brownian sheet.

- **Theorem:** K is a C-Brownian sheet.
- **Proof:**
  - K is a mean zero Gaussian process since D is, so it remains to show that K has the covariance structure of a C-Brownian sheet.
  - K has independent (rectangle) increments by construction, so it will suffice to show that $\mathrm{Var}[K(x,y)] = C(x,y)$ for all $(x,y) \in [0,1]^2$.

- **Theorem:** K is a C-Brownian sheet.
- **Proof:**
    - K is a mean zero Gaussian process since D is, so it remains to show that K has the covariance structure of a C-Brownian sheet.
    - K has independent (rectangle) increments by construction, so it will suffice to show that $\mathrm{Var}[K(x, y)] = C(x, y)$ for all $(x, y) \in [0, 1]^2$.
    - The variance of the K-increment over a small rectangle is “close” to the variance of the D-increment over the same rectangle, which is in turn “close” to the W-increment over the same rectangle.

- **Corollary:** The process

$$Z(x,y) = \int_0^x \int_0^y \frac{1}{\sqrt{C^{xy}(s,t)}} dK(s,t), \quad (x,y) \in [0,1]^2$$

is a standard two-parameter Wiener process.

- **Corollary:** The process

$$Z(x,y) = \int_0^x \int_0^y \frac{1}{\sqrt{C^{xy}(s,t)}} dK(s,t), \quad (x,y) \in [0,1]^2$$

is a standard two-parameter Wiener process.

- We have thus transformed the C-Brownian pillow D into a standard two-parameter Wiener process Z, through a two-step transformation:

$$D \mapsto K \mapsto Z.$$

- We apply the same two-step transformation to D_n, i.e. we define

$$
\begin{aligned}
K_n(x,y) = D_n(x,y) &- \int_0^x D_n(s,y)\xi_1(ds,y) \\
&- \int_0^y D_n(x,t)\xi_2(x,dt) \\
&+ \int_0^x \int_0^y D_n(s,t)\xi_1(ds,t)\xi_2(s,dt), \\
Z_n(x,y) = \int_0^x \int_0^y & \frac{1}{\sqrt{C^{xy}(s,t)}} dK_n(s,t)
\end{aligned}
$$

for $(x,y) \in [0,1]^2$.

- **Theorem:** Z_n converges weakly to Z.

- **Theorem:** Z_n converges weakly to Z.
- As an intermediate step, we need to show that K_n converges weakly to K.

- **Theorem:** Z_n converges weakly to Z.
- As an intermediate step, we need to show that K_n converges weakly to K.
- Thus the asymptotical distribution of Z_n is independent of C, and functionals of Z_n can be used for goodness-of-fit tests.

- **Theorem:** Z_n converges weakly to Z.

- As an intermediate step, we need to show that K_n converges weakly to K.

- Thus the asymptotical distribution of Z_n is independent of C, and functionals of Z_n can be used for goodness-of-fit tests.

- Future: Construct actual test statistics and procedures for goodness-of-fit tests. Consider composite null hypotheses of the form $C \in \{C_\theta : \theta \in \Theta\}$ and consider m-dimensional copulas with $m > 2$.

- **Theorem:** Z_n converges weakly to Z.
- As an intermediate step, we need to show that K_n converges weakly to K.
- Thus the asymptotical distribution of Z_n is independent of C, and functionals of Z_n can be used for goodness-of-fit tests.
- Future: Construct actual test statistics and procedures for goodness-of-fit tests. Consider composite null hypotheses of the form $C \in \{C_\theta : \theta \in \Theta\}$ and consider m-dimensional copulas with $m > 2$.
- Thank you for listening!