The multivariate probability integral transform

Fabrizio Durante

Faculty of Economics and Management
Free University of Bozen-Bolzano (Italy)
fabrizio.durante@unibz.it
http://sites.google.com/site/fbdurante

Austrian Statistical Days, 21–23 October 2015, Vienna
Outline

1. Univariate Probability Integral Transform
2. Multivariate Probability Integral Transform
3. Extreme Risks and Hazard scenarios
4. Clustering extreme events
5. Conclusions
Probability integral transform

Let $X$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution function $F$.

If $F$ is continuous, then $F(X)$ is uniformly distributed on $[0, 1]$.
The r.v. $F(X)$ is called **Probability Integral Transform** (shortly, PIT) of $X$.

Histograms of 1000 points from $Z \sim N(0, 1)$ (left) and the PIT of $Z$ (right).
Application of PIT

Suppose that \( X_1, \ldots, X_n \), is a random sample from a continuous d.f. \( F \), and suppose that one is interested to test

\[
H_0: F = F_0 \quad \text{vs} \quad H_1: F \neq F_0
\]

for some completely known continuous d.f. \( F_0 \).

Under the null hypothesis,

\[
F_0(X_{(1)}) < F_0(X_{(2)}) < \cdots < F_0(X_{(n)}),
\]

are distributed like the ordered statistics from a random sample of size \( n \) from the uniform distribution on \([0, 1] \).

Since the \( i \)--th smallest ordered value from a sample of size \( n \) from the distribution of \( U(0, 1) \) has expectation \( i/(n+1) \), a plot of the points \( (i/(n+1), F_0(X_{(i)})) \), \( i = 1, \ldots, n \), should lie roughly along a straight line of slope 1, if \( F = F_0 \).

Application: evaluating forecast accuracy (Diebold et al., 1998).
Quantile function theorem

Let $F$ be a univariate distribution function.

We call quantile inverse of $F$ the function $F^{(-1)} : (0, 1) \rightarrow (-\infty, \infty)$ given by

$$F^{(-1)}(t) := \inf \{ x \in \mathbb{R} : F(x) \geq t \}$$

with the convention $\inf \emptyset = +\infty$.

If $U$ is a random variable that is uniformly distributed on $[0, 1]$, then $F^{-1}(U)$ has distribution function equal to $F$.

The previous result gives a procedure for simulating a random sample from a given d.f. $F$. 
Quantile function

An increasing function (left) and its corresponding quantile inverse (right).
Uniform representation of random vectors

For r.v.’s $X_1, X_2$ with continuous d.f.’s $F_1, F_2$, $U_1 := F_1(X_1)$ and $U_2 := F_2(X_2)$ are uniformly distributed on $[0, 1]$. Thus,

\[ F(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) \]
\[ = \mathbb{P}(F_1(X_1) \leq F_1(x_1), F_2(X_2) \leq F_2(x_2)) \]
\[ = \mathbb{P}(U_1 \leq F_1(x_1), U_2 \leq F_2(x_2)) \]
\[ = C(F_1(x_1), F_2(x_2)), \]

where $C$ is the d.f. of $(U_1, U_2) := (F_1(X_1), F_2(X_2))$.

Since the d.f. $C$ joins or “couples” a multivariate d.f. to its one-dimensional marginal d.f.’s, we call it copula.
Sklar’s Theorem

Let \((X_1, \ldots, X_d)\) be a r.v. with joint d.f. \(F\) and univariate marginals \(F_1, F_2, \ldots, F_d\). Then there exists a copula (=joint d.f. with uniform margins) \(C\) such that, for all \(x \in \mathbb{R}^d\),

\[
F(x_1, x_2, \ldots, x_d) = C\left(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\right).
\]

\(C\) is uniquely determined on \(\text{Range}(F_1) \times \cdots \times \text{Range}(F_d)\) and, hence, it is unique when \(F_1, \ldots, F_d\) are continuous.

Every known multivariate d.f. is associated with a copula

The copula \(C\) associated with a joint d.f. \(F\) is given, for all \(u \in [0, 1]^d\) by

\[
C(u_1, \ldots, u_d) = F\left(F_1^{(-1)}(u_1), \ldots, F_d^{(-1)}(u_d)\right).
\]
Basic examples of copulas are:

- the **independence copula** $\Pi_d(u) = u_1 u_2 \cdots u_d$ associated with a r.v. $X$ whose components are independent;

- the **comonotonicity copula** $M_d(u) = \min\{u_1, u_2, \ldots, u_d\}$ associated with a r.v. $X$ such that $X \overset{d}{=} (T_1(Z), \ldots, T_d(Z))$ for some random variable $Z$ and increasing functions $T_1, \ldots, T_d$;

- the **countermonotonicity copula** $W_2(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}$ associated with a r.v. $(X_1, X_2)$ such that $X_2 \overset{d}{=} T(X_1)$ for some decreasing function $T$. 
Sklar’s recipe

Thanks to copulas, we may construct (and fit) parametric statistical models via a two-stage procedure:

- first, choose the univariate marginals \( F_1^{\alpha_1}, \ldots, F_d^{\alpha_d} \),
- then, choose our favorite copula \( C_\theta \),
- mix the two ingredients and obtain the model

\[
F_{\theta,\alpha}(x) = C_\theta \left( F_1^{\alpha_1}(x_1), \ldots, F_d^{\alpha_d}(x_d) \right).
\]

This is potentially very useful for risk management and sensitivity analysis since it allows for testing several scenarios with different kinds of dependence between the assets, keeping the marginals fixed.
Tail behaviour under different scenarios

Bivariate sample clouds of 2500 points from the d.f. $F = C(F_1, F_2)$ where $F_1, F_2 \sim N(0, 1)$, the Spearman’s $\rho$ is equal to 0.75, and $C$ is a Gaussian copula (left) or a Gumbel copula (right).
Outline

1. Univariate Probability Integral Transform
2. Multivariate Probability Integral Transform
3. Extreme Risks and Hazard scenarios
4. Clustering extreme events
5. Conclusions
Kendall d.f.

Let $X$ be a continuous random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose distribution function is equal to $H$. Then we say that:

- The PIT of $X$ is the random variable $V = H(X)$.

- The distribution function $K_X$ of $V$ is called **Kendall distribution function** associated with $X$.

Notice that, in contrast to the univariate case, it is not generally true that $K$ is uniform on $[0, 1]$. 

(Genest and Rivest, 2001)
Kendall d.f.

Empirical Kendall d.f. obtained from a random sample of 1000 points from a Gumbel copula with \( \tau = 0.5 \).
Calculation of Kendall d.f.

It is neither possible to characterize \( H(x) = \mathbb{P}(X \leq x) \) nor reconstruct it from the knowledge of \( K \) alone. In fact, the calculation of \( K \) depends only on the copula \( C \) of \( X \) and does not involve the knowledge the marginal distributions.

Specifically, for every \( t \in [0, 1] \), we have

\[
K(t) = \mathbb{P}(H(X) \leq t) = \mu_H(\{x \in \mathbb{R}^d : H(x) \leq t\}) = \mu_C(\{u \in [0, 1]^d : C(u) \leq t\})
\]

where \( \mu_F \) and \( \mu_C \) are, respectively, the measures induced by the d.f. \( H \) and the copula \( C \) on \( \mathbb{R}^d \).
Properties of Kendall d.f.

- $K(0^-) = 0$
- $t \leq K(t) \leq 1$ for all $t \in [0, 1]$
- $K(t) = t$ if and only if $C = M_d$
- In the bivariate case, $K(t) = 1$ if, and only if, $C = W_2$

The Kendall d.f. is related to the population value of Kendall’s $\tau$ rank correlation coefficient for a pair of random variables $(X, Y) \sim H$ via the formula:

$$\tau(X, Y) = 3 - \int_0^1 K(t) dt.$$
Kendall d.f. and associative copula

Notice that two (different) copulas may be associated with the same Kendall d.f.. In particular, Kendall d.f.’s may be used to define an equivalence relation $\equiv_K$ on the class of copulas.

If $C_1$ and $C_2$ are copulas with Kendall d.f.’s $K_1$ and $K_2$, respectively, then

$$C_1 \equiv_K C_2 \text{ if and only if } K_1 = K_2$$

**Theorem**

*In dimension 2, every equivalence class given by the relation $\equiv_K$ contains one and only one associative copula.*

*(Nelsen, Quesada-Molina, Rodríguez-Lallena, Úbeda-Flores, 2009)*
A copula $C$ is Archimedean if it can be expressed in the form

$$C(u) = \phi^{-1}(\phi(u_1) + \cdots + \phi(u_d))$$

for a suitable generator $\phi$.

**Theorem**

*Under mild regularity conditions, any Kendall d.f. $K$ is associated with a unique Archimedean copula generated by*

$$\phi(t) = \exp \left( \int_{t_0}^{t} \frac{dw}{w - K(w)} \right)$$

*for some $t_0 \in (0, 1)$.*

*(Genest and Rivest, 2001; Genest, Neslehova and Ziegel, 2011)*
Estimation of the Kendall distribution

Suppose that \((X_{11}, X_{12}), \ldots, (X_{T1}, X_{T2})\) is a random sample from a distribution \(H\) with copula \(C\).

The empirical Kendall distribution function \(K_T\) is given, for all \(q \in [0, 1]\), by

\[
K_T(q) = \frac{1}{T} \sum_{j=1}^{T} 1(W_j \leq q),
\]

where, for each \(j \in \{1, \ldots, T\}\),

\[
W_j = \frac{1}{T + 1} \sum_{t=1}^{T} 1(X_{t1} < X_{j1}, X_{t2} < X_{j2}).
\]

The empirical process \(\sqrt{T}(K_T - K)\) convergence in law to a centered Gaussian limit under mild regularity conditions.

(Barbe et al., 1996; Genest, Neslehova and Ziegel, 2011)
Approximation of copulas via Kendall distribution

Given the correspondence between Kendall d.f.’s and Archimedean copulas, we can consider a general approximation of the dependence structure as follows:

- Consider a sequence of iid observations $X_1, X_2, \ldots$ from a continuous $X \sim F = C(F_1, \ldots, F_d)$.
- Choose an estimator $K_n$ of the corresponding Kendall d.f. $K_C$.
- Provided that $K_n$ is a bona fide Kendall d.f., construct the related Archimedean copula.

The procedure works well even if we restrict to consider $K_n$ from specific classes (e.g. piecewise linear).
Example: Gumbel copula

Comparison between Kendall’s d.f. of the reference Gumbel copula, its empirical estimate, and the corresponding piecewise linear approximation. See (Salvadori, Durante and Perrone, 2013).
Application: structural risk

We consider an application in coastal engineering related to the preliminary design of a rubble mound breakwater.

The target is to compute the quantiles associated to the weight $W$ of a concrete cube element forming the breakwater structure, assuming that the environmental load is given by

- $H$, the significant wave height (in meters),
- $D$, the sea storm duration (in hours),

and the existence of a structure function $\Psi$ given by

$$W = \Psi(H, D) = \rho_s \cdot \left[ H \left( \frac{2 \pi H}{g \left[ 4.597 \cdot H^{0.328} \right]^2} \right)^{0.1} \right]^3 / \left[ \left( \frac{\rho_s}{\rho_W} - 1 \right) \cdot \left( 1 + \frac{6.7 \cdot N_d^{0.4}}{(3600 D / \left[ 4.597 \cdot H^{0.328} \right])^{0.3}} \right) \right]^3$$

For more details, see (Salvadori, Tomsicchio and D’Alessandro, 2014).
Application: structural risk
Behavior of the approximations of the cube weight for four different design quantiles and three different Sea Storm Sample Sizes (SSSS). Reference copula: Gumbel with $\tau = 0.5$. See (Pappadà et al., 2015).
Outline

1. Univariate Probability Integral Transform
2. Multivariate Probability Integral Transform
3. Extreme Risks and Hazard scenarios
4. Clustering extreme events
5. Conclusions
Global economic losses from extreme weather

The global insured and uninsured economic losses from the two biggest categories of weather related extreme events (category 6 “great natural catastrophes” and category 5 “devastating catastrophes”) over the last 30 years from the Munich Re NatCatSE RVICE database.
Motivation

Climate Extreme (extreme weather or climate event): The occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends of the range of observed values of the variable.

Much of the analysis of changes of extremes has, up to now, focused on individual extremes of a single variable. However, recent literature in climate research is starting to consider compound events and explore appropriate methods for their analysis.

(IPCC Report, 2012)
Motivation

The flood risk management should require the implementation of suitable flood hazard maps covering the geographical areas which could be flooded according to the following scenarios:

(a) floods with a low probability, or extreme event scenarios;
(b) floods with a medium probability (likely return period $\geq 100$ years);
(c) floods with a high probability, where appropriate.

For each flood scenario, the following quantities should be considered:

(a) the flood extent;
(b) water depths or water level, as appropriate;
(c) where appropriate, the flow velocity or the relevant water flow.
Hazard scenario

Let $X$ be a random vector describing the phenomenon of interest. A Hazard Scenario (hereinafter, HS) of level $\alpha \in (0, 1)$ is any Borel upper set $S \subseteq \mathbb{R}^d$ such that $P(X \in S) = \alpha$.

If $S$ is an upper set, it follows that, for all $x \in S$, $S$ also contains all the occurrences $y \geq x$ componentwise.

As will be shown below, given a realization $x \in \mathbb{R}^d$, there exist several ways to associate $x$ with a suitable HS, occasionally denoted by $S_x \subseteq \mathbb{R}^d$.
Hazard scenario: OR

For every \( \mathbf{x} \in \mathbb{R}^d \), OR HS is given by

\[
S^\vee_{\mathbf{x}} = \bigcup_{i=1}^{d} (\mathbb{R} \times \cdots \times \mathbb{R} \times [x_i, +\infty[ \times \mathbb{R} \times \cdots \times \mathbb{R}) .
\]

For the realization of the event \( \{ \mathbf{X} \in S^\vee_{\mathbf{x}} \} \) it is enough that one of the variables \( X_i \)'s exceeds the corresponding threshold \( x_i \).

In turn,

\[
\alpha^\vee_{\mathbf{x}} = P (\mathbf{X} \in S^\vee_{\mathbf{x}}) = 1 - C(F_1(x_1), \ldots, F_d(x_d))
\]

\[
= 1 - C(u_1, \ldots, u_d),
\]

where \( (u_1, \ldots, u_d) = (F_1(x_1), \ldots, F_d(x_d)) \).
Hazard scenario: OR

\[ \text{OR} \]

\[ \begin{array}{c}
\text{U} \\
\text{V}
\end{array} \]

\[ S \]

\[ \text{u} \quad \text{v} \]

\[ 0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1 \\
0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.6 \quad 0.7 \quad 0.8 \quad 0.9 \quad 1 \]
Evaluation of dangerous flood events at the confluence of two rivers, where the threatening occurrence can be due to the contribution of one river, or the other, or both (Favre et al., 2004)
Hazard scenario: AND

For every $x \in \mathbb{R}^d$ AND HS is given by

$$S_x^\wedge = \bigcap_{i=1}^{d} (\mathbb{R} \times \cdots \times \mathbb{R} \times [x_i, +\infty[ \times \mathbb{R} \times \cdots \times \mathbb{R}).$$

For the realization of the event $\{X \in S_x^\wedge\}$ it is necessary that all the variables $X_i$’s exceed the corresponding thresholds $x_i$’s.

In turn,

$$\alpha_x^\wedge = \mathbb{P} (X \in S_x^\wedge) = \hat{C}(F_1(x_1), \ldots, F_d(x_d)),$$

where $F_1, \ldots, F_d$ are the survival functions and $\hat{C}$ is the survival copula associated with $C$. 
Hazard scenario: AND
The Mekong Delta is at risk if both flood peak and flood volume are high. The hydraulic system in the Vietnamese Mekong Delta is characterized by a large number of channels, dikes and control structures such as sluice gates in the dikes connecting channels and floodplains. The presence of the dike system requires certain water levels [...], but as socio-economic and agricultural systems are well adapted to the annual floods, large discharge values exceeding dike levels do not automatically imply a disaster. For a flood event to become a disaster, it needs also a high flood volume. (Dung et al., 2015)
Hazard scenario: Kendall

Given a $d$–dimensional continuous d.f. $F$ with strictly increasing margins and $t \in (0, 1)$ we define:

- **the critical layer**
  \[ \mathcal{L}^F_t = \{ \mathbf{x} \in \mathbb{R}^d : F(\mathbf{x}) = t \} \]

- **the sub–critical region** (lower level set)
  \[ \mathcal{R}^<_t = \{ \mathbf{x} \in \mathbb{R}^d : F(\mathbf{x}) < t \} \]

- **the super–critical region** (upper level set)
  \[ \mathcal{R}^>_t = \{ \mathbf{x} \in \mathbb{R}^d : F(\mathbf{x}) > t \} \]

At any occurrence of the phenomenon, with probability 1, either a realization of $\mathbf{X}$ lies in $\mathcal{R}^>_t$ or in $\mathcal{R}^<_t$.
Hazard scenario: Kendall

Kendall scenario

For every \( x \in \mathbb{R}^d \) with \( F(x) = t \) the Kendall HS is defined as

\[
S^K_t = \{ x \in \mathbb{R}^d : F(x) > t \} = \{ x \in \mathbb{R}^d : C(F_1(x_1), \ldots, F_d(x_d)) > t \},
\]

whose level \( \alpha \) is given by

\[
\alpha^K_u = \alpha^K_x = \alpha^K_t = P(X \in S^K_t) = 1 - K(t),
\]

where \( t = C(u) = F(x) \).

(Salvadori, De Michele and Durante, 2011)
Hazard scenario: Kendall

![Diagram showing Kendall scenario]
Fig. 7. Level curves of storm parameters $H$ and $D$ for storm (solid lines) and erosion (dashed lines) return periods of 100, 50 and 25 yr for $T = 15$ s and $W = 1.0$ m above mean sea level.

For more details, see S. Corbella and D. D. Stretch (2012).
Hazard scenario: Survival Kendall

**Survival Kendall scenario**

For every $\mathbf{x} \in \mathbb{R}^d$ with $\overline{F} (\mathbf{x}) = t$, the Survival Kendall HS is defined as

$$S^K_t = \{ \mathbf{x} \in \mathbb{R}^d : \overline{F} (\mathbf{x}) < t \} = \{ \mathbf{x} \in \mathbb{R}^d : \hat{C} (\overline{F}_1 (x_1), \ldots, \overline{F}_d (x_d)) < t \},$$

whose level $\alpha$ is given by

$$\alpha^K_x = \alpha^K_t = \mathbb{P} \left( \mathbf{X} \in S^K_t \right) = 1 - \hat{K} (t) = \hat{K} (t),$$

where $t = \hat{C} (1 - u) = \overline{F} (\mathbf{x})$ and

$$\hat{K} (t) = \mathbb{P} \left( \overline{F} (X_1, \ldots, X_d) \leq t \right) = \mathbb{P} \left( \hat{C} (\overline{F}_1 (X_1), \ldots, \overline{F}_d (X_d)) \leq t \right).$$

(Salvadori, De Michele and Durante, 2014)
Hazard scenario: Survival Kendall
Concurrent temperature and precipitation extremes return period based on November-April data from 1896 to 2014. The blue dots represent historical observations, and the isolines show the return periods. For more details, see AgaKouchak et al. (2014).
Univariate empirical return period of extreme droughts in California and their corresponding concurrent extreme (red text) return periods. The latter includes November-April 1896-2014 precipitation and temperature data, whereas the former is solely based on precipitation in the same period. For more details, see AgaKouchak et al. (2014).
Example: Ceppo Morelli dam (Italy)

Variables of interest: $Q$, maximum annual flood peak; $V$ maximum annual volume. See (Salvadori et al., 2015).
Example: Ceppo Morelli dam (Italy)

Comparison of the Failure Probabilities (probability of occurrence of at least one extreme event under a given time period considering different Hazard Scenarios and samples. See (Salvadori et al., 2015).
Outline

1. Univariate Probability Integral Transform
2. Multivariate Probability Integral Transform
3. Extreme Risks and Hazard scenarios
4. Clustering extreme events
5. Conclusions
Motivation

- The requirement for projections of extreme events has motivated the development of **regionalization techniques** to simulate weather and climate at finer spatial resolutions.

- The identification of different groups in a set of climate time series is relevant to identify subgroups characterized by **similar behavior** in order to adopt specific risk management strategies.

- Moreover, management of environmental risk often requires the analysis of spatial rainfall extremes which typically exhibit **joint tail dependence**.
The goal

We aim at creating clusters of climate time series that are homogeneous in the sense that the elements of each cluster tend to comove, while elements from different clusters are characterized by some weak dependence.

The procedure is based on the introduction of a suitable dissimilarity matrix $\Delta = (\delta_{ij})$ that describes the pairwise association and it is tailored to some risk measure currently adopted in environmental science.

Once the dissimilarity matrix is obtained, we may apply standard clustering techniques like hierarchical clustering (R: hclust) or “fuzzy” clustering (R: fanny).
The clustering procedure

Consider an iid sample $X^t_1, \ldots, X^t_n$ from a given r.v. $X$ corresponding to $n$ different measurements collected at time $t \in \{1, \ldots, T\}$.

The procedure to calculate the dissimilarity matrix $\Delta = (\delta_{ij})$ consists of the following steps:

- Calculate the Kendall d.f. $K_{ij}$ for each pair $(X_i, X_j)$.
- Define a dissimilarity matrix among the time series such that the dissimilarity between $X_i$ and $X_j$ is given by:

  $$\delta_{ij} = \int_0^1 (q - K_{ij}(q))^2 dq,$$

  where $K(t) = t$ is the Kendall d.f. of the comonotone case.
- Apply a suitable cluster algorithm.

(Durante and Pappadà, 2015)
The data

The data

We consider daily rainfall measurements recorded at 18 gauge stations spread across the province of Bolzano-Bozen in the North-Eastern Italy.

From these time series, we extracted annual maxima at each spatial location resulting in a $50 \times 18$ matrix of time series observations.

The selection of annual maxima transforms strong seasonal time series into data that are assumed to be iid.
The data

Box plots of annual maxima (in mm) at each station from 1961 to 2010.
The cluster dendrogram

Dendrogram for the 18 rainfall measurement stations based on the complete linkage method.
Map of the rainfall measurement stations marked according the 4-clusters solution in the province of Bolzano–Bozen (North-Eastern, Italy).

See (Durante and Pappadà, 2015).
Outline

1. Univariate Probability Integral Transform
2. Multivariate Probability Integral Transform
3. Extreme Risks and Hazard scenarios
4. Clustering extreme events
5. Conclusions
Conclusions

- We have presented the univariate PIT and shown how it is can be used to introduce the concept of copula.
- We have considered a multivariate PIT and outlined how it can be used to address the problem of identifying hazard scenarios, when several dependent variables are involved.
- Several studies are hence illustrated about their use in hydrology.
Bibliography

Questions? Comments?

Thanks for your attention!

More information about this talk:

- visit my home-page
  http://sites.google.com/site/fbdurante