Motivating Complex Dependence Structures in Data Mining: A Case Study with Anomaly Detection in Climate

Shih-Chieh Kao[1], Auroop R. Ganguly[1,¶], and Karsten Steinhaeuser[1,2]

[1] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[2] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

Introduction

- The need for handling massive and high dimensional datasets with a focused consideration on preserving the dependence structure among variables.
- Spatio-temporal & inter-variable dependence
  - Temporal: Autoregression model (AR), Markov chain
  - Spatial: Geostatistics (Kriging method)
  - Inter-variable: Bayesian approach
- Lack of an effective mathematical tool to characterize the dependence structure when variables are
  - Multidimensional and non-Gaussian
  - Combination of various marginal distributions
  - Non-linear physical processes
- Risk, statistical similarity, extreme value and frequency, anomaly detection, and data simulation

Concept of Correlation and Dependence Structure

- From moment-based statistics (mean, standard deviation, coefficients of skewness and kurtosis) to probability density function (PDF).
  - Can the cross-moment (e.g., Pearson's correlation coefficient ρ) be extended in a similar manner?
  - A two-dimensional mathematical formula rather than a single measurement

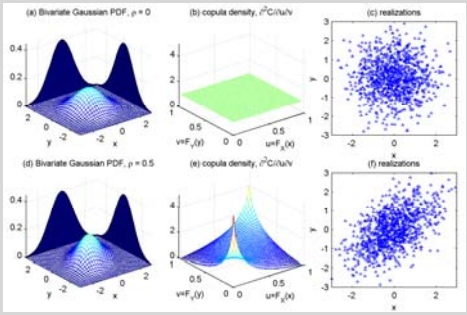

Figure 1 - Bivariate Gaussian distributions (ρ = 0 & ρ = 0.5) and the corresponding copula densities and realizations

Copulas

- Transformation of joint cumulative distribution
  - Univariate marginals: $u_i = F_{Xi}(x_i)$
  - $H_{X1,\ldots Xn}(x_1,\ldots,x_n) = C_{U1,\ldots,Un}(u_1,\ldots,u_n)$
- Archimedean copulas

Frank family of Archimedean copulas:

$$\begin{cases}\varphi(t) = -\ln((e^{-\theta t}-1)/(e^{-\theta}-1)) \\ C(u,v) = -\frac{1}{\theta}\ln(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}) \\ \theta \in (-\infty,0)\cup(0,\infty)\end{cases}$$

Clayton family of Archimedean copulas:

$$\begin{cases}\varphi(t) = (t^{-\theta}-1)/\theta \\ C(u,v) = (\max(u^{-\theta}+v^{-\theta}-1,0))^{-1/\theta} \\ \theta \in [-1,0)\cup(0,\infty)\end{cases}$$

  - Maximum-likelihood estimator versus non-parametric estimator through Kendall's tau
  - Goodness-of-fit of copulas

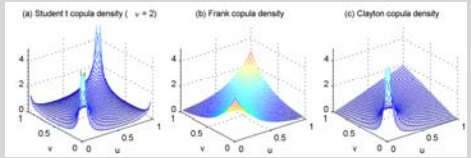

Figure 2 - Illustration of (a) Student t copulas (degree of freedom v = 2), (b) Frank copulas, and (c) Clayton copulas. All three copulas have the same Pearson's correlation coefficient ρ = 0.5

- Flexibility in modeling various types of joint distributions
- Strength in generating correlated random variables

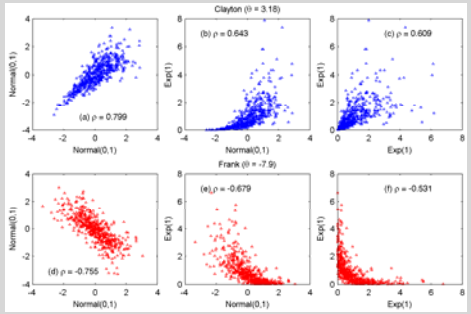

Figure 3 - The capability of copulas in random number generation. All simulated patterns (500 data points in each panel) are combinations of Normal(0,1) and Exp(1) marginals, and Clayton(3.18) and Frank(-7.9) copulas

Dependence in Climate Data

- Climate data contains multiple variables (e.g., temperature, pressure, wind speed, humidity, precipitable water and precipitation). Each variable has its own type of distribution, seasonal variability, and long-term non-stationary trend.
- Most of the hydro-meteorological variables are governed by non-linear processes with non-intuitive mechanisms.
- Data used in this study
  - NCEP2 reanalysis: 1.9°x1.9° spatial resolution (total of 18,048 grid cells), beginning from 1979 until present.
  - Monthly temperature (X), precipitation (Y), precipitable water (Z)
  - Z-score transformation to remove seasonal variability
- Global correlation between normalized variables
  - Category I: ρ < -0.3, negatively dependent
  - Category II: -0.3 < ρ < 0.3, near independent
  - Category III: ρ > 0.3, positively dependent

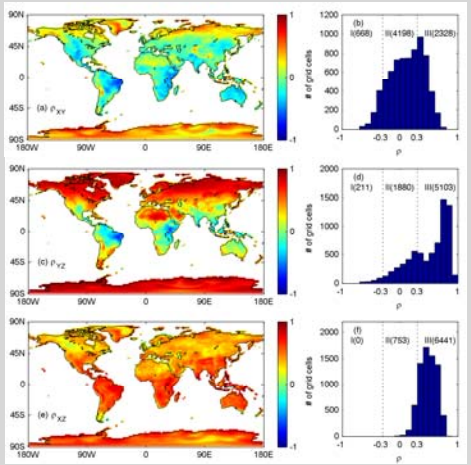

Figure 4 - Global maps and histograms of correlation coefficient ρ for (Plots a & b) normalized temperature (X) versus normalized precipitation (Y), (Plots c & d) Y versus normalized precipitable water (Z), and (3) X versus Z. Number of grid points in Region I (ρ < -0.3), Region II (-0.3 < ρ < 0.3), and Region III (ρ > 0.3) are also marked on the histograms.

- Highly positive dependence between normalized temperature and precipitable water.
- Dependence level between normalized temperature and precipitation varies widely.
- Reflecting regional weather characteristics

- Example: grid point near Miami, FL (79.67W~81.56W, 24.76N~26.67N)
  - Marginal distribution: kernel density function
  - Dependence structure: Frank Archimedean copulas

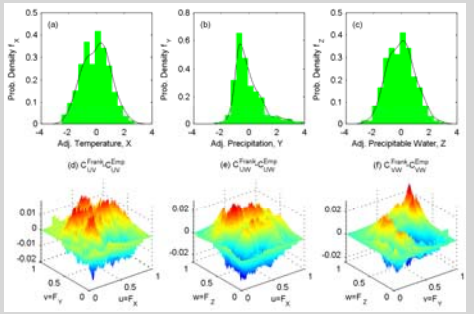

Figure 5 - Taking the grid cell containing Miami as an example, (a-c) show the histograms and kernel density fitting of adjusted temperature (X), precipitation (Y), precipitable water (Z), and (d-f) show the differences between fitted Frank copulas and empirical copulas of each pair of variables.

Climatic Anomaly Detection

- Bivariate abnormal climate events:
  - Months which are hot and have low precipitation, have low precipitation and dry, or hot and dry
- Fixed versus adjusted thresholds (from the theory of copulas) in detecting abnormal months

Table 1 – Summary of bivariate climatic anomaly detection

	Regions		
	I	II	III
Case 1, fixed threshold			
$\{X^{(i,j)} \geq x^{(i,j)}_{80\%}, Y^{(i,j)} \leq y^{(i,j)}_{20\%}\}$	8.41%	4.39%	1.43%
$\{X^{(i,j)} \geq x^{(i,j)}_{80\%}, Z^{(i,j)} \leq z^{(i,j)}_{20\%}\}$	8.38%	3.67%	0.52%
$\{Y^{(i,j)} \leq y^{(i,j)}_{20\%}, Z^{(i,j)} \leq z^{(i,j)}_{20\%}\}$	-	5.98%	8.96%
Case 2, adjusted threshold			
$\{X^{(i,j)} \geq x^{(i,j)}_{q1}, Y^{(i,j)} \leq y^{(i,j)}_{q1}\}$	4.81%	4.48%	4.47%
$\{X^{(i,j)} \geq x^{(i,j)}_{q2}, Z^{(i,j)} \leq z^{(i,j)}_{q2}\}$	4.52%	4.24%	4.85%
$\{Y^{(i,j)} \leq y^{(i,j)}_{q3}, Z^{(i,j)} \leq z^{(i,j)}_{q3}\}$	-	4.15%	4.45%

- The number of abnormal months is affected by the dependence levels between variables.
- By considering the dependence levels between variables, the adjusted ratio approach results in similar amount of anomalies.

References

G. Kuhn, S. Khan, A. R. Ganguly and M. L. Branstetter, Geospatial-temporal Dependence among Weekly Precipitation Extremes with Applications to Observations and Climate Model Simulations in South America, Adv. Water Resour., 30(12), pages 2401-2423, 2007.

S.-C. Kao and R. S. Govindaraju, Trivariate Statistical Analysis of Extreme Rainfall Events via Plackett Family of Copulas, Water Resour. Res., 44, W02415, 2008.

¶ Point of Contact

Auroop R. Ganguly, Ph.D.
Email: gangulyar@ornl.gov
http://www.ornl.gov/knowledgediscovery/ClimateExtremes

Geo. Information Sci. & Tech. Group
Computational Sci. & Eng. Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6017
Tel.: +1 (865) 241-1305
Fax: +1 (865) 241- 6261

This research was funded by the Laboratory Directed Research and Development (LDRD) Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.