New Statistical Analytical Tools for Big Data

Zhengjun Zhang

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

2018 Summer Short Course
School of Data Science, Fudan University

Outline

References

- Fan, J., Han, F., and Liu, H. (2013), Challenges of big data analysis.
- Manyika, J., Chui M., Brown. B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers A. H. (2011), Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute
- Malinowski, A., Schlather, M., and Zhang, Z. (2015), Marked point process adjusted tail dependence analysis for high-frequency financial data.
- Yang, X., Frees, J., and Zhang, Z. (2011), A generalized beta copula with applications in modeling multivariate long-tailed data. *Insurance: Mathemtaics and Economics*. **49**, 265-284.
- Zhang, Z., Zhang, C., and Cui, Q. (2016), Random threshold driven tail dependence measures with application to precipitation data.
- Zhang, Z. and Zhu, B. (2016), Copula structured M4 processes with application to high-frequency financial data.
- More to be added .....

What do we mean by “big data"?

Ways to define “big data"

- The size of the datasets is too big to capture, store, manage, and analyze.
- Definitions are subjective and ‘moving.’
- Vary by sectors.

Where are the data collected from

- Financial market (global)
- Insurance industry (global)
- Health care (US)
- Public sector administration
- Retail
- Manufacturing
- Personal location data

What do we mean by “big data"?

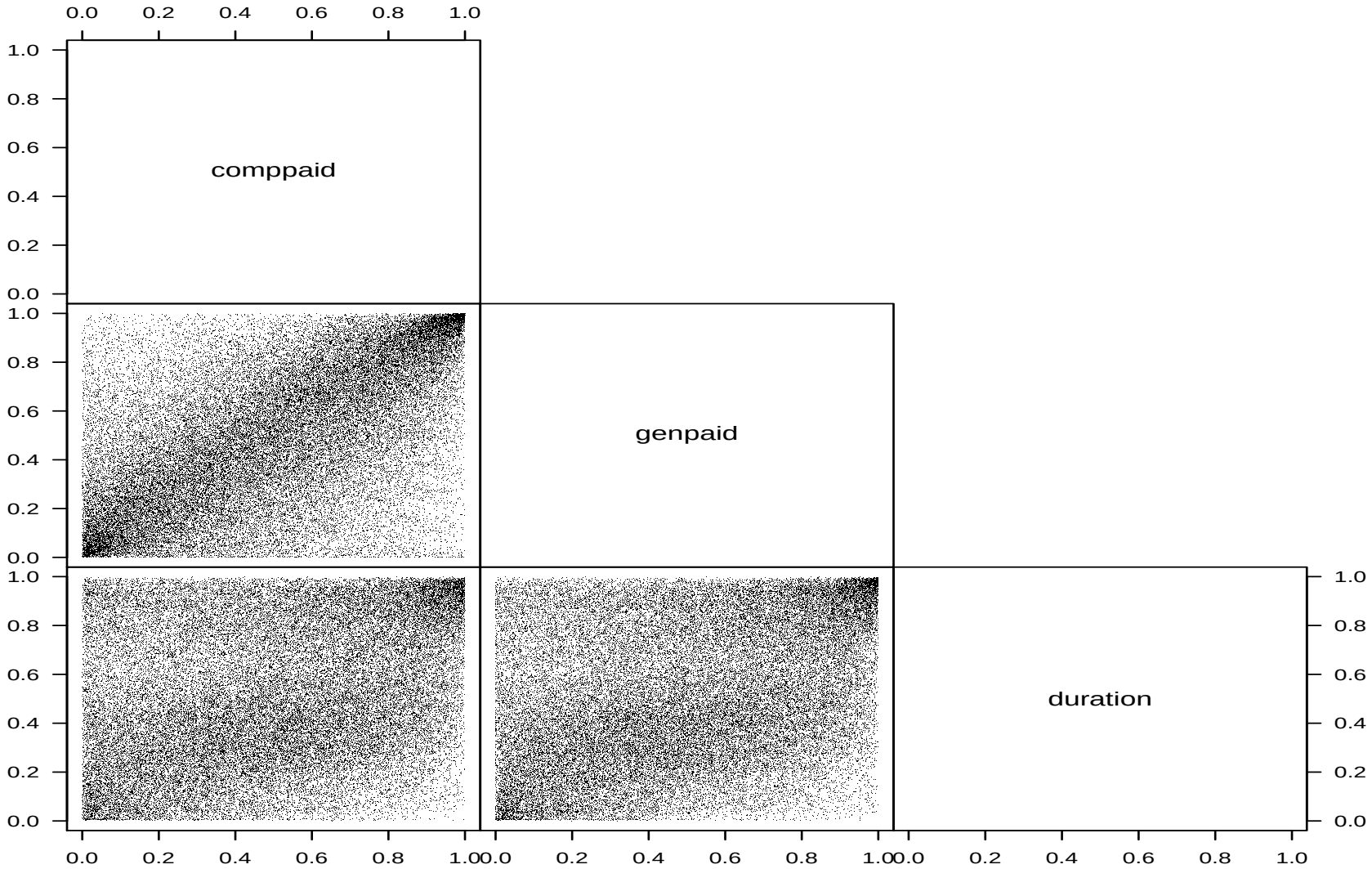
Ways to define “big data"

- The size of the datasets is too big to capture, store, manage, and analyze.
- Definitions are subjective and ‘moving.’
- Vary by sectors.

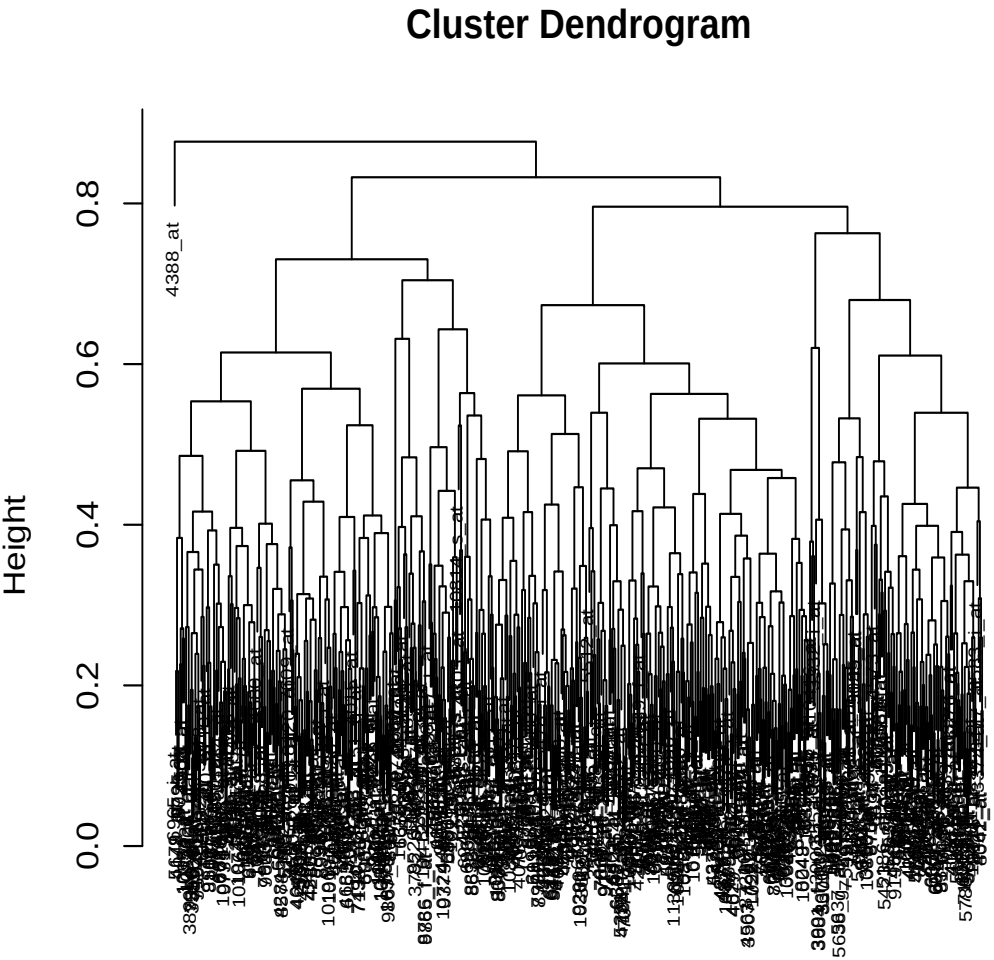
Where are the data collected from

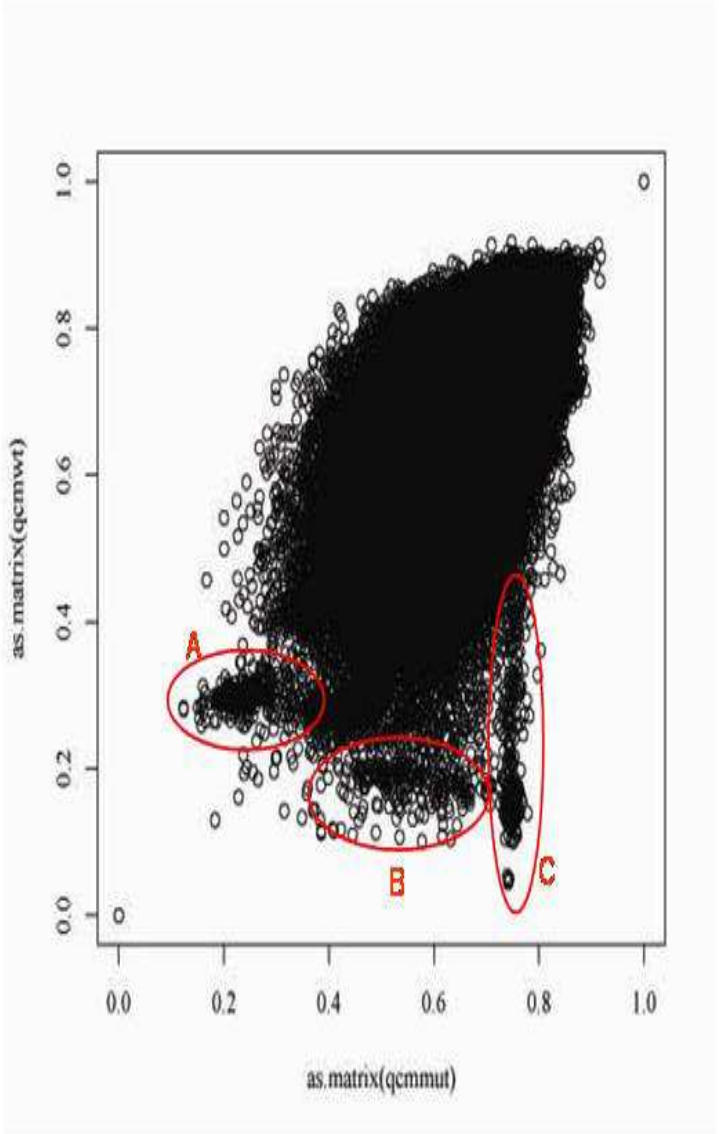
- Financial market (global)
- Insurance industry (global)
- Health care (US)
- Public sector administration
- Retail
- Manufacturing
- Personal location data

My ‘Big Data’ examples



My ‘Big Data’ examples





My ‘Big Data’ examples

DSDR | DATA SHARING FOR DEMOGRAPHIC RESEARCH

ICPSR 27021

National Longitudinal Study of Adolescent Health (Add Health), 1994-2008: Core Files [Restricted Use]

Kathleen Mullan Harris
University of North Carolina-Chapel Hill

J. Richard Udry
University of North Carolina-Chapel Hill

Original Add Health Wave I In-Home Interview Data Codebook

Inter-university Consortium for Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

Outline

Descriptive statistics

- Sample mean
- Sample variance
- Sample covariance

'Intermediate' statistical analysis

- Classification.
- Cluster analysis.
- Regression.
- Data mining.

Descriptive statistics

- Sample mean
- Sample variance
- Sample covariance

'Intermediate' statistical analysis

- Classification.
- Cluster analysis.
- Regression.
- Data mining.

Outline

A new look of a twice-told tale:

$$\mathrm{var}(X) = \mathrm{var}\big(E(X|Y)\big) + E\big(\mathrm{var}(X|Y)\big) \tag{1}$$

- $\mathrm{var}\big(E(X|Y)\big)$ measures the spread of the conditional mean (center) of X given Y
- $\mathrm{var}\big(E(X|Y)\big)/\mathrm{var}(X)$ can certainly be interpreted as the explained variance of X by Y.

$$\frac{\mathrm{var}\big(E(X|Y)\big)}{\mathrm{var}(X)} = 1 - \frac{E\big(\mathrm{var}(X|Y)\big)}{\mathrm{var}(X)} = 1 - \frac{E[\{X - E(X|Y)\}^2]}{\mathrm{var}(X)}.$$

Correlation ratios: Kendall and Stuart (1979), Doksum and Samarov (1995), Wang (2001)

The generalized measures of correlation (GMC)

$$\big\{\mathrm{GMC}(Y|X),\ \mathrm{GMC}(X|Y)\big\} = \Big\{1 - \frac{E[\{Y - E(Y|X)\}^2]}{\mathrm{var}(Y)},\ 1 - \frac{E[\{X - E(X|Y)\}^2]}{\mathrm{var}(X)}\Big\}. \tag{2}$$

Zheng, Shi, and Zhang (2012)

A new look of a twice-told tale:

$$\mathrm{var}(X) = \mathrm{var}\big(E(X|Y)\big) + E\big(\mathrm{var}(X|Y)\big) \tag{1}$$

- $\mathrm{var}\big(E(X|Y)\big)$ measures the spread of the conditional mean (center) of X given Y
- $\mathrm{var}\big(E(X|Y)\big)/\mathrm{var}(X)$ can certainly be interpreted as the explained variance of X by Y.

$$\frac{\mathrm{var}\big(E(X|Y)\big)}{\mathrm{var}(X)} = 1 - \frac{E\big(\mathrm{var}(X|Y)\big)}{\mathrm{var}(X)} = 1 - \frac{E[\{X - E(X|Y)\}^2]}{\mathrm{var}(X)}.$$

Correlation ratios: Kendall and Stuart (1979), Doksum and Samarov (1995), Wang (2001)

The generalized measures of correlation (GMC)

$$\{\mathrm{GMC}(Y|X),\ \mathrm{GMC}(X|Y)\} = \Big\{1 - \frac{E[\{Y - E(Y|X)\}^2]}{\mathrm{var}(Y)},\ 1 - \frac{E[\{X - E(X|Y)\}^2]}{\mathrm{var}(X)}\Big\}. \tag{2}$$

Zheng, Shi, and Zhang (2012)

Properties of generalized measures of correlation

Suppose both $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then

(i) $0 \leq \mathrm{GMC}(Y|X),\ \ \mathrm{GMC}(X|Y) \leq 1$,

- and if X and Y are independent, then $\mathrm{GMC}(Y|X) = 0,\ \mathrm{GMC}(X|Y) = 0$.

(ii) The relation of GMC and Pearson's correlation coefficient ρ_{XY} satisfies:

- If $\rho_{XY} = \pm 1$, then $\mathrm{GMC}(Y|X) = 1$ and $\mathrm{GMC}(X|Y) = 1$.
- If $\rho_{XY} \neq 0$, then $\mathrm{GMC}(X|Y) \neq 0$ and $\mathrm{GMC}(Y|X) \neq 0$.
- If $\mathrm{GMC}(Y|X) = 0$ and/or $\mathrm{GMC}(X|Y) = 0$, then $\rho_{XY} = 0$.
- $\mathrm{GMC}(Y|X) \geq \rho_{XY}^2,\ \mathrm{GMC}(X|Y) \geq \rho_{XY}^2$

Properties of generalized measures of correlation

(iii) Suppose $Y = g(X) + \varepsilon$, X and ε are independent, and both $E(g^2(X)) < \infty$ and $E(\varepsilon^2) < \infty$. Then

$$\mathrm{GMC}(Y|g(X)) = \frac{\mathrm{var}(g(X))}{\mathrm{var}(g(X)) + \mathrm{var}(\varepsilon)}.$$

Particularly, if $g(x) = ax + b$ for $a \neq 0$ and b being constants, we have

$$\mathrm{GMC}(Y|X) = \rho_{XY}^2.$$

For the extreme values of GMC, we have

$$\mathrm{GMC}(Y|X) = 1 \Longleftrightarrow Y = g(X) \quad a.s.$$

- If g is a one to one measurable function, then $\mathrm{GMC}(Y|X) = \mathrm{GMC}(X|Y) = 1$;
- If g is not one to one, then $\mathrm{GMC}(Y|X) = 1 > \mathrm{GMC}(X|Y) \geq 0$.

Properties of generalized measures of correlation

(iv) Suppose $Y_1 = g_1(X) + \varepsilon_1$ and $Y_2 = g_2(X) + \varepsilon_2$, where ε_1 and ε_2 are independent of X, $g_1(\cdot)$ and $g_2(\cdot)$ are linear or nonlinear measurable functions. If either
1) $\mathrm{var}\big(g_1(X)\big) = \mathrm{var}\big(g_2(X)\big)$, $\mathrm{var}(\varepsilon_1) < \mathrm{var}(\varepsilon_2)$;
or 2) $\mathrm{var}\big(g_1(X)\big) > \mathrm{var}\big(g_2(X)\big)$, $\mathrm{var}(\varepsilon_1) = \mathrm{var}(\varepsilon_2)$,
we have

$$\mathrm{GMC}(Y_1|X) > \mathrm{GMC}(Y_2|X).$$

(v) If $\mathrm{var}(Y_1) = \mathrm{var}(Y_2)$ and $\inf_f E[\{Y_1 - f(X)\}^2] = \inf_g E[\{Y_2 - g(X)\}^2]$, where f, g are measurable functions, then

$$\mathrm{GMC}(Y_1|X) = \mathrm{GMC}(Y_2|X).$$

If $\mathrm{var}(Y_1) = \mathrm{var}(Y_2)$ and $\inf_f E[\{Y_1 - f(X)\}^2] < \inf_g E[\{Y_2 - g(X)\}^2]$, then we have

$$\mathrm{GMC}(Y_1|X) > \mathrm{GMC}(Y_2|X).$$

Example 1

- $Y = X^2$, $X \sim N(0,1)$.
  - $\mathrm{GMC}(Y|X) = 1$, $\mathrm{GMC}(X|Y) = 0$,
  - $\rho_{XY} = 0$.

Proposition 2.2: Bivariate normal

$$\mathrm{GMC}(Y|X) = \mathrm{GMC}(X|Y) = \rho^2_{XY}.$$

GMCs with marginal distributions being uniform on [0,1]

$$\mathrm{GMC}\big(F_Y(Y)|X\big) = 12E\left(\{E(F_Y(Y)|X)\}^2\right) - 3,$$

$$\mathrm{GMC}\big(F_X(X)|Y\big) = 12E\left(\{E(F_X(X)|Y)\}^2\right) - 3.$$

Compared with Spearman's correlation

$$\rho_S(X,Y) = \mathrm{cor}\big(F_X(X), F_Y(Y)\big) = 12E\big(F_X(X)F_Y(Y)\big) - 3$$

Example 1

- $Y = X^2$, $X \sim N(0,1)$.
  - $\mathrm{GMC}(Y|X) = 1$, $\mathrm{GMC}(X|Y) = 0$,
  - $\rho_{XY} = 0$.

Proposition 2.2: Bivariate normal

$$\mathrm{GMC}(Y|X) = \mathrm{GMC}(X|Y) = \rho_{XY}^2.$$

GMCs with marginal distributions being uniform on [0,1]

$$\mathrm{GMC}\big(F_Y(Y)|X\big) = 12E\left(\{E(F_Y(Y)|X)\}^2\right) - 3,$$

$$\mathrm{GMC}\big(F_X(X)|Y\big) = 12E\left(\{E(F_X(X)|Y)\}^2\right) - 3.$$

Compared with Spearman's correlation

$$\rho_S(X,Y) = \mathrm{cor}\big(F_X(X), F_Y(Y)\big) = 12E\big(F_X(X)F_Y(Y)\big) - 3$$

Example 1

- $Y = X^2$, $X \sim N(0,1)$.
  - $\mathrm{GMC}(Y|X) = 1$, $\mathrm{GMC}(X|Y) = 0$,
  - $\rho_{XY} = 0$.

Proposition 2.2: Bivariate normal

$$\mathrm{GMC}(Y|X) = \mathrm{GMC}(X|Y) = \rho_{XY}^2.$$

GMCs with marginal distributions being uniform on [0,1]

$$\mathrm{GMC}\big(F_Y(Y)|X\big) = 12E\left(\{E(F_Y(Y)|X)\}^2\right) - 3,$$

$$\mathrm{GMC}\big(F_X(X)|Y\big) = 12E\left(\{E(F_X(X)|Y)\}^2\right) - 3.$$

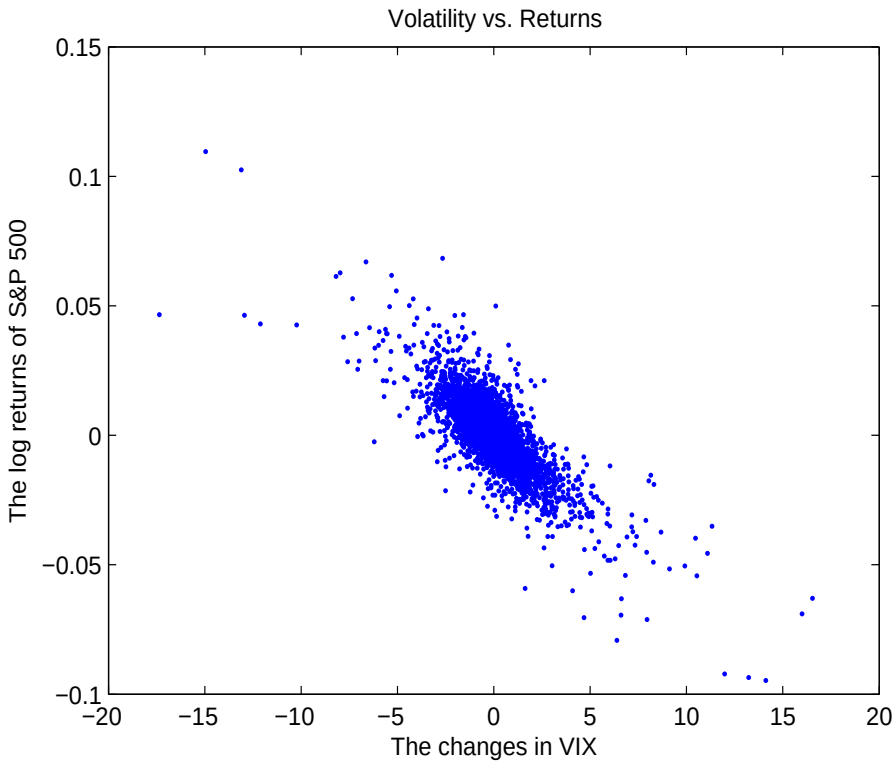
Compared with Spearman's correlation

$$\rho_S(X,Y) = \mathrm{cor}\big(F_X(X), F_Y(Y)\big) = 12E\left(F_X(X)F_Y(Y)\right) - 3$$

Outline

VIX and S&P 500 return

Some stochastic volatility models are often motivated from jointly modeling VIX and S&P 500 return together.



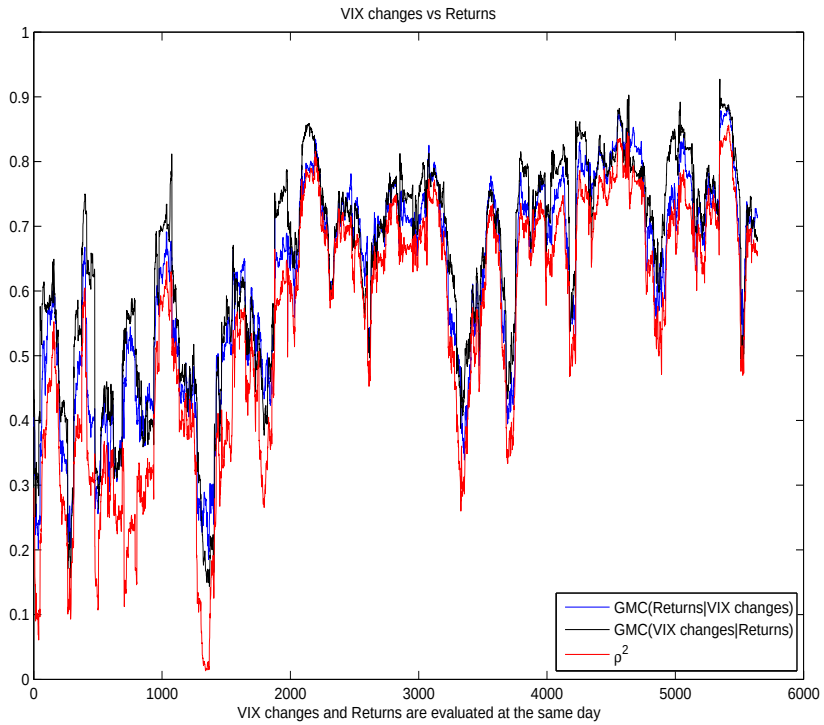


Figure 1: The demonstration of VIX and Return.

- Observation: Symmetric and linear relationship.

S&P 500 return and realized volatility

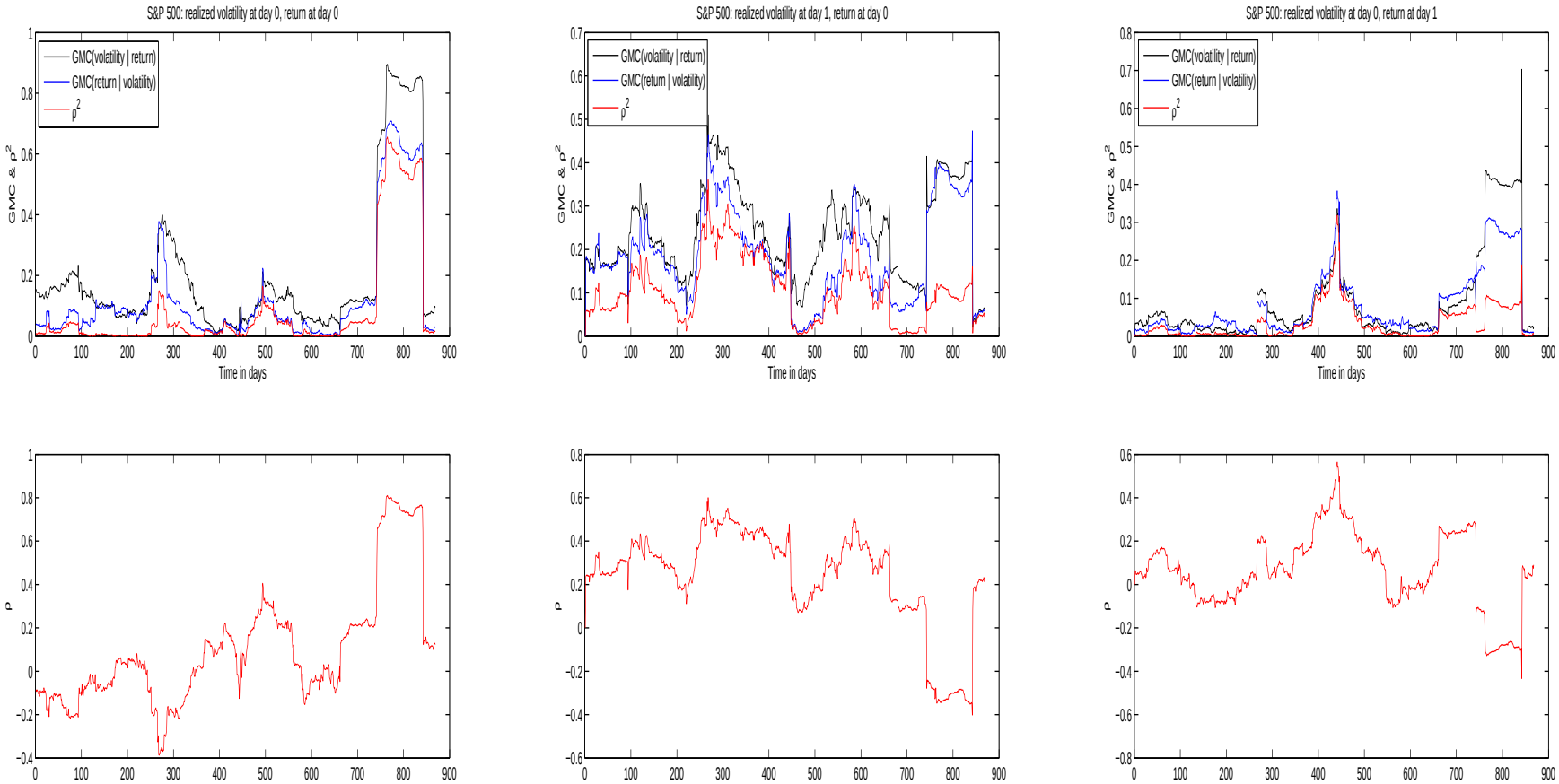


Figure 2: The demonstration of realized volatility and Return.

- Observation: ???.

Will GARCH(p,q) type models be suitable?

$$r_t = \sigma_t \varepsilon_t,\ \sigma_t^2 = K + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j r_{t-j}^2.$$

$$K > 0,\ \alpha_i \geq 0,\ \beta_j \geq 0, \sum \alpha_i + \sum \beta_j < 1.$$

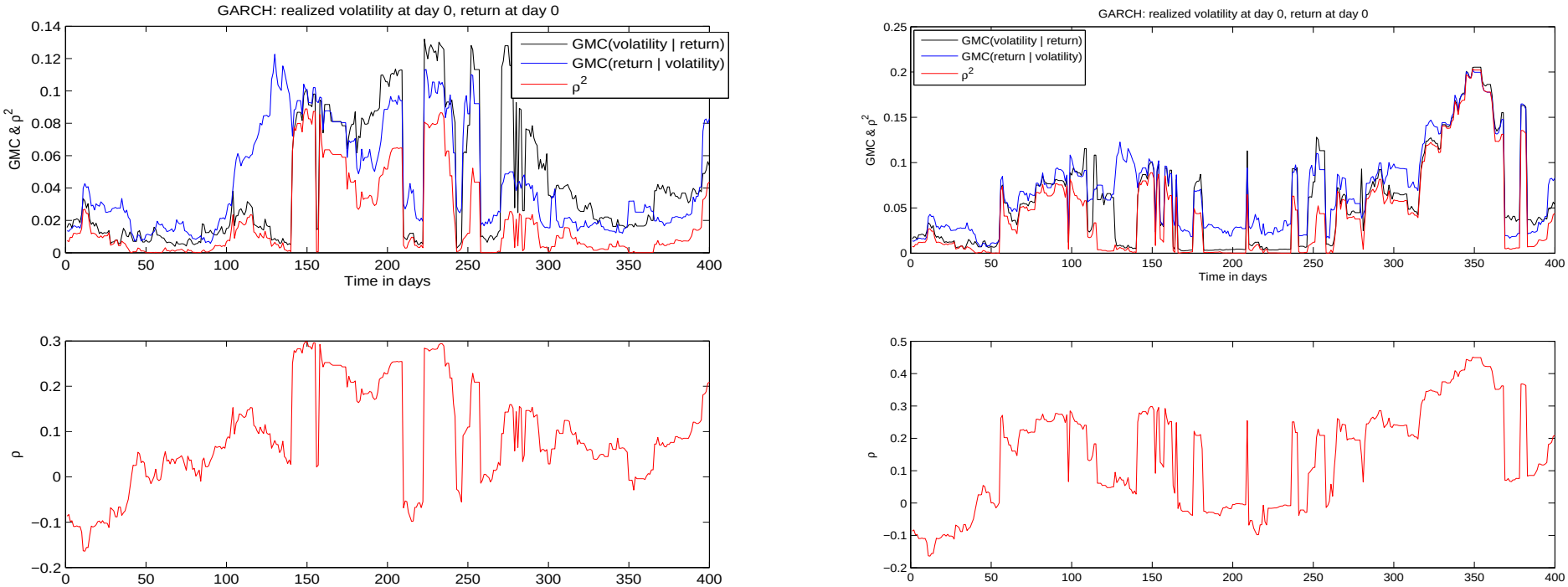
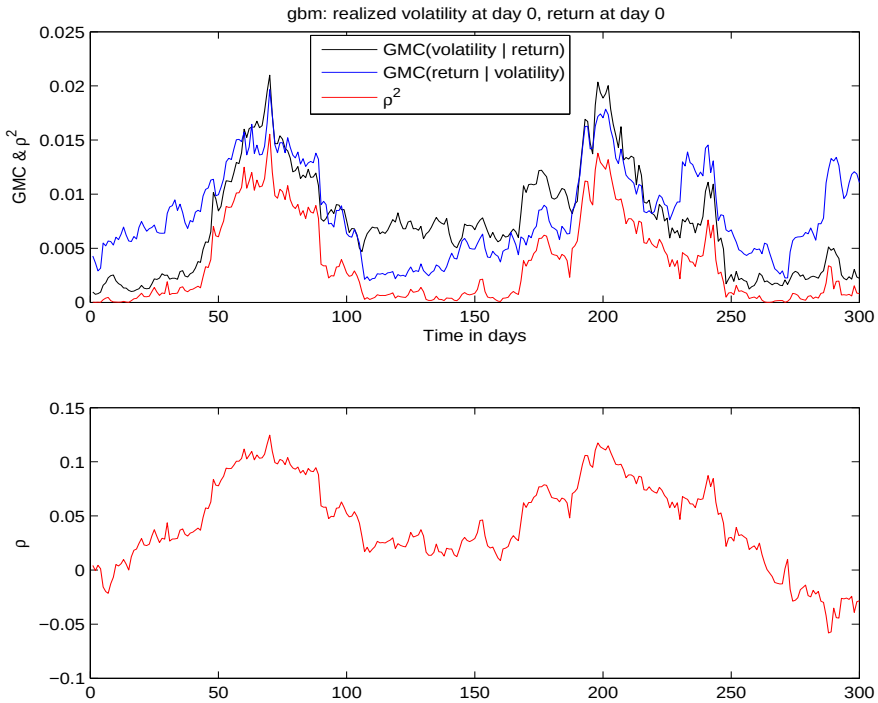


Figure 3: The demonstration of realized volatility and Return.

- Observation: ???.

Will GBM type models be suitable?

$$dX_t = \mu(t)X_t dt + D(t, X_t)V(t)dW_t.$$



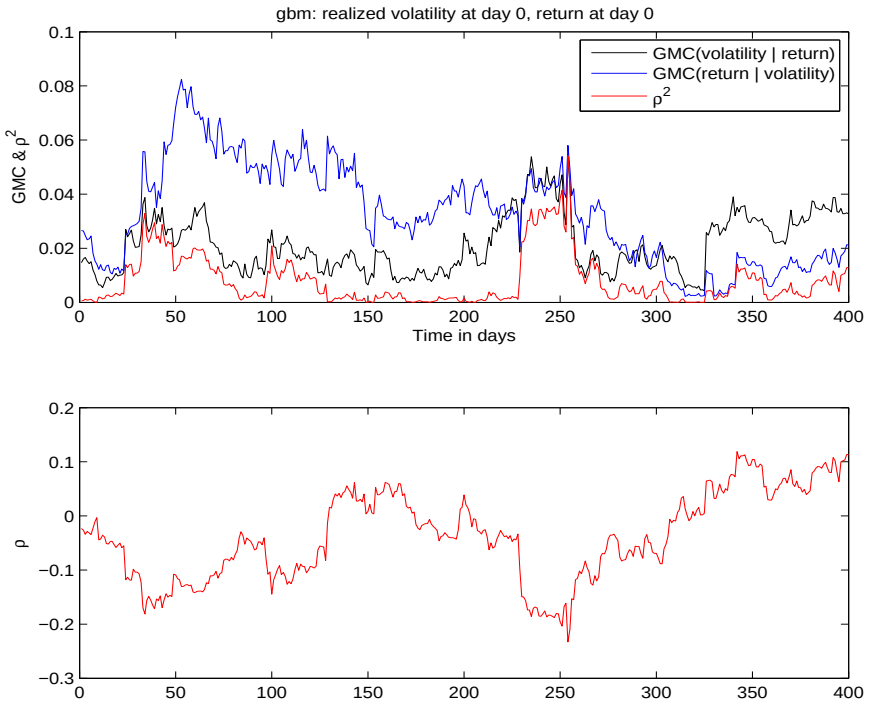
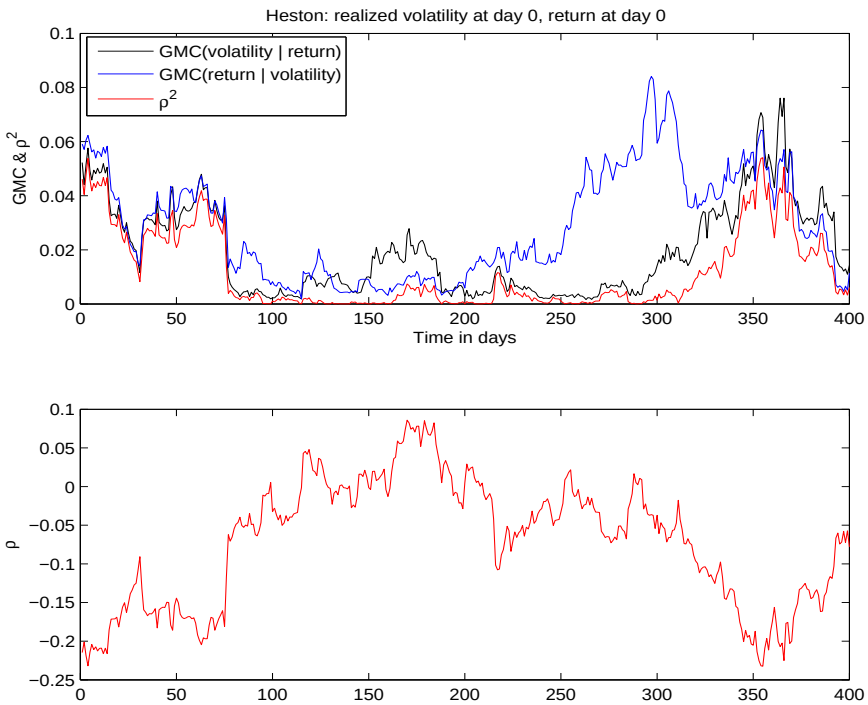


Figure 4: The demonstration of realized volatility and Return.

- Observation: ???.

Will Heston type models be suitable?

$$dX_{1t} = B(t)X_{1t}dt + \sqrt{X_{2t}}X_{1t}dW_{1t},$$
$$dX_{2t} = S(t)[L(t) - X_{2t}]dt + V(t)\sqrt{X_{2t}}dW_{2t}.$$



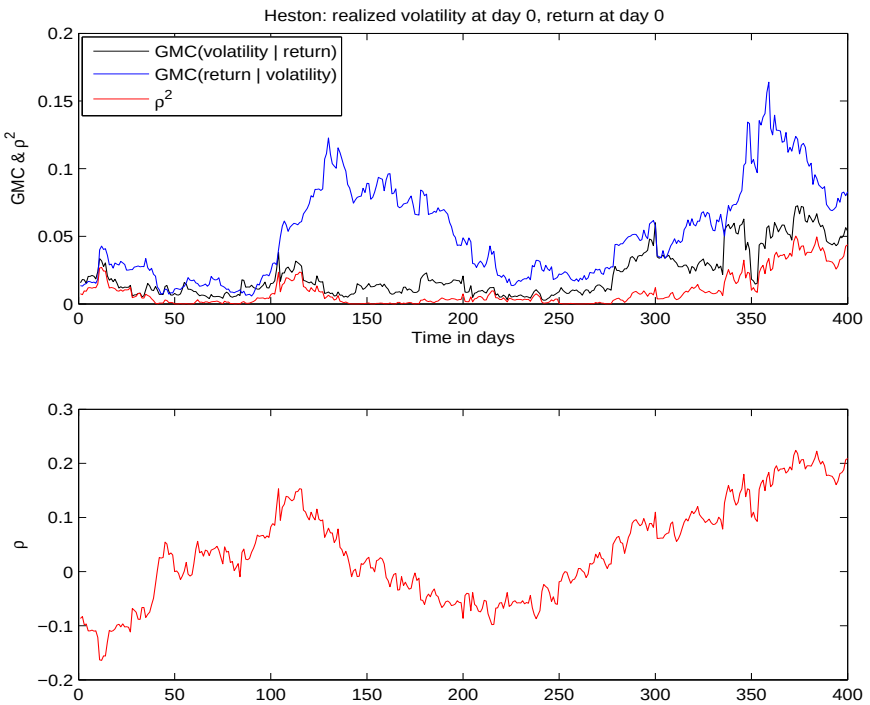
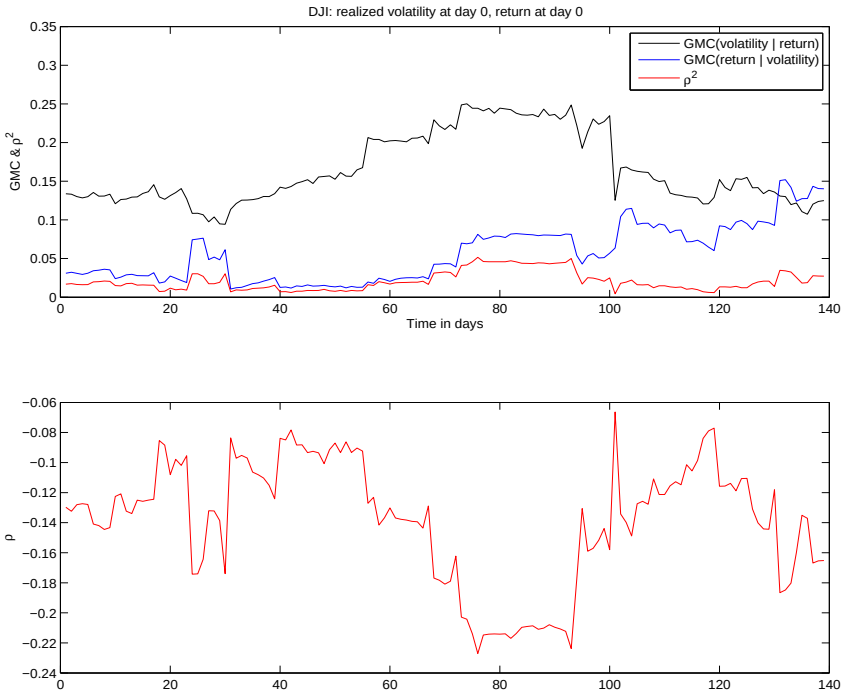


Figure 5: The demonstration of realized volatility and Return.

- Observation: ???.

How about applicability in Dow Jones and NASDAQ



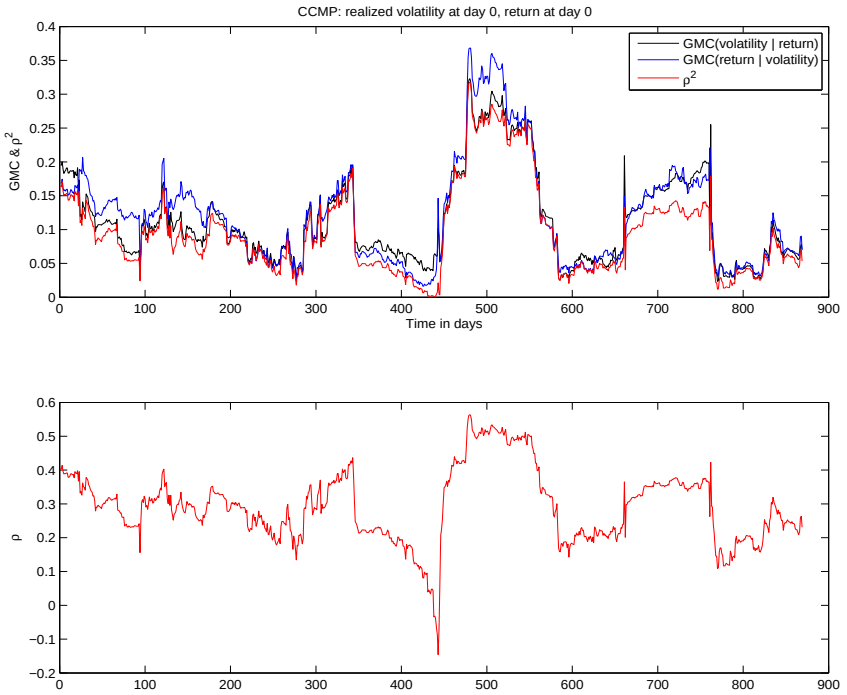


Figure 6: The demonstration of realized volatility and Return.

- Observation: ???.

Outline

List of areas of GMCs applications

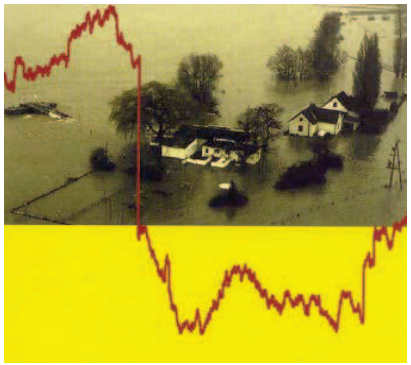
- Whenever Pearson correlation coefficients are applied, GMCs are applicable.
- Causal inference.
- Nonlinear and asymmetric inference.
- Graphical models.
- Asymmetric game theory.
- .....

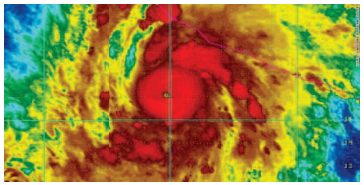
Future work

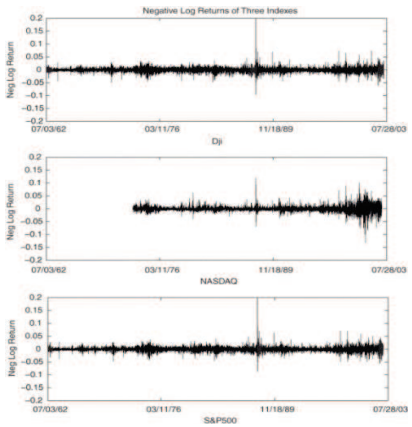
- Discrete GMC
- Time series GMC
- Graphical models
- Variable selection
- .....

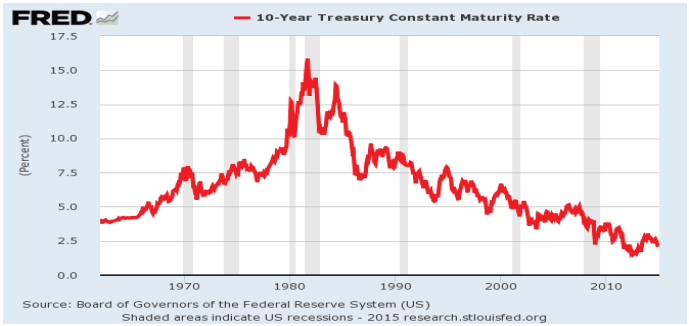
Outline

Big Data in Extreme Values



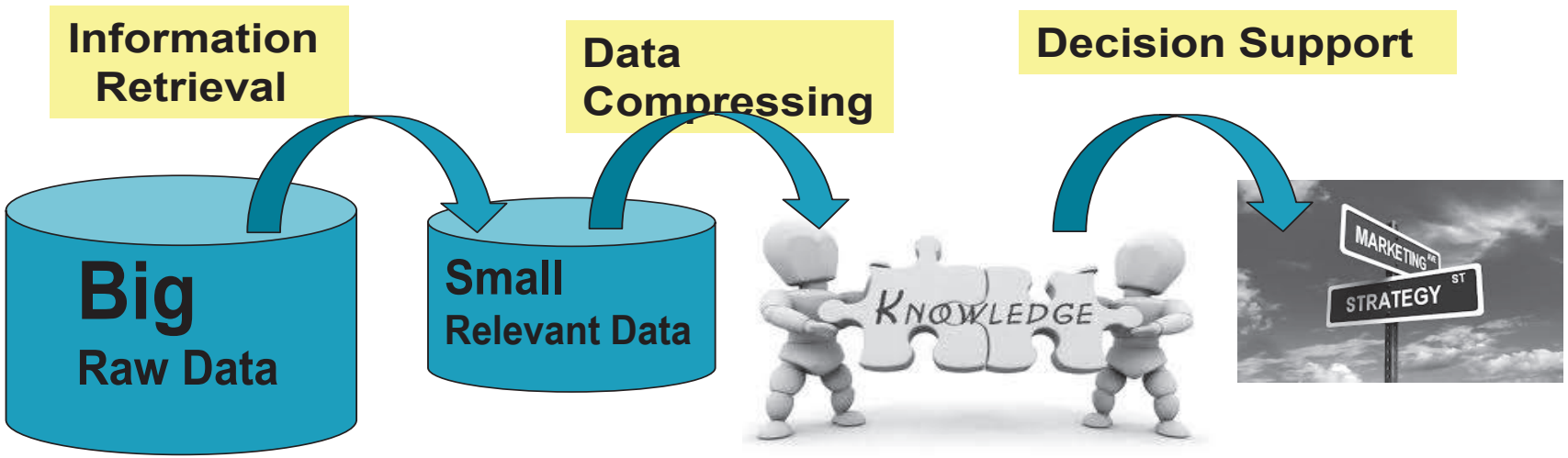






My ‘Big Data’ examples

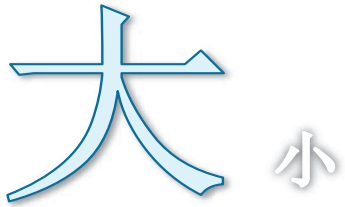
Extreme value analytics



Flow chat borrowed from Chengxiang Zhai

Extreme Value Analytics for Big Data

比



Zhengjun Zhang

Department of Statistics

University of Wisconsin

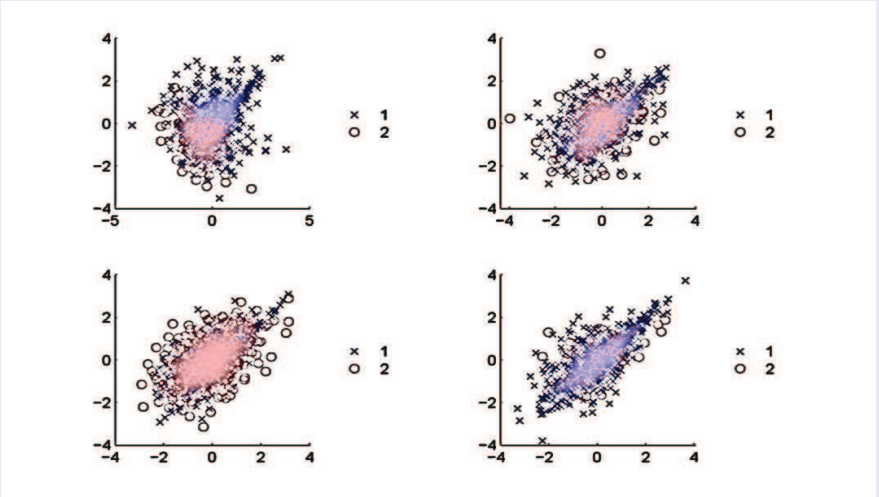
Madison, WI 53706, USA

Outline

Beyond normal distributions

John Tukey:

"As I am sure almost every geophysicist knows, distributions of actual errors and fluctuations have much more straggling extreme values than would correspond to the magic bell-shaped distribution of Gauss and Laplace."



WASHINGTON, Nov. 23 (UPI) -- According to the World Bank, climate change is getting to the point that extreme weather should be seen as the new normal.

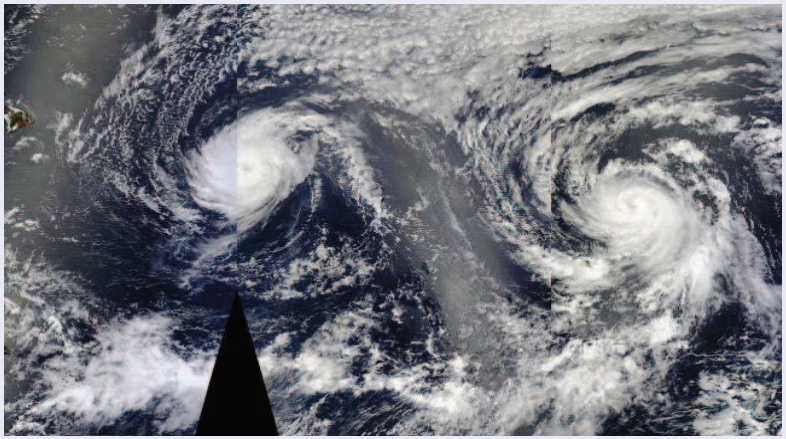
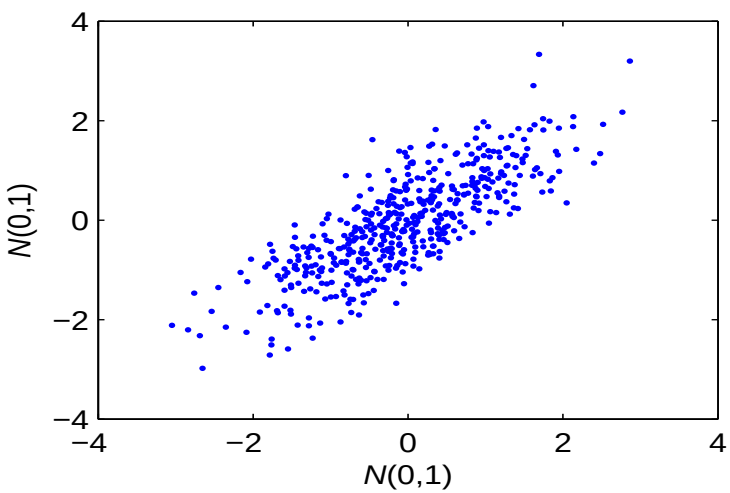
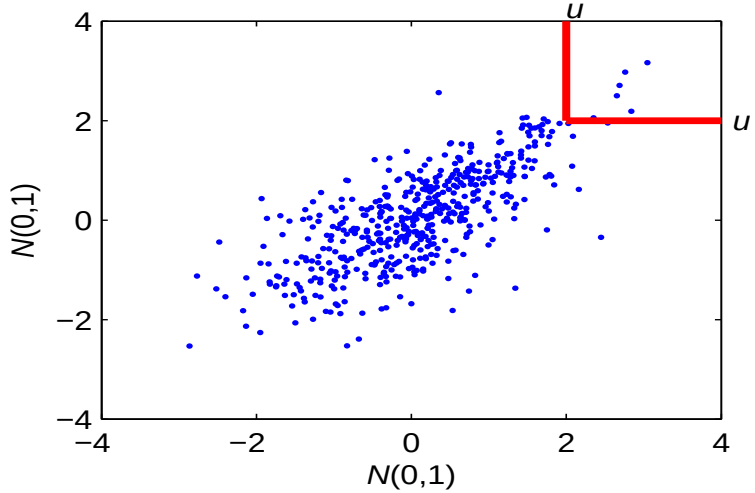
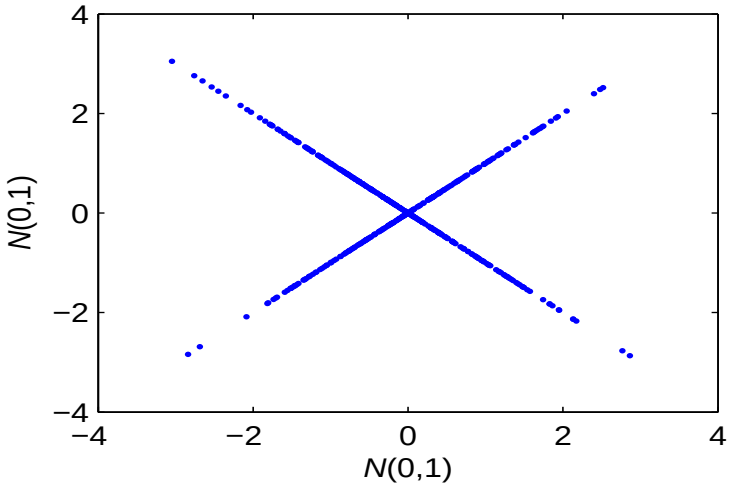
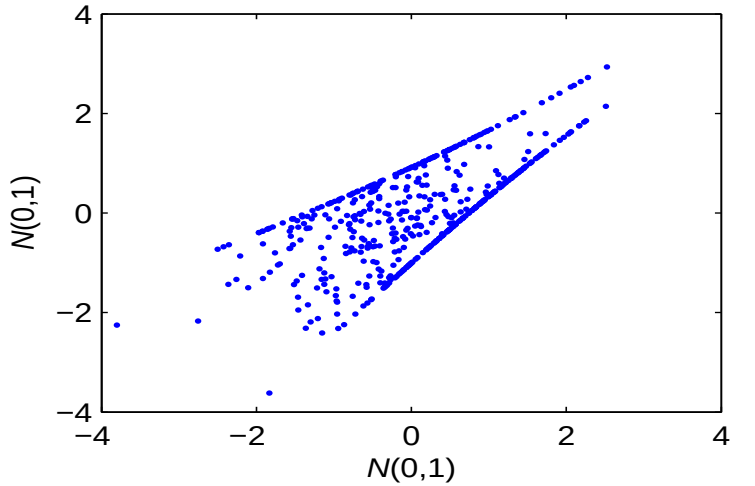


Illustration of bivariate tail (in)dependence









Tail (in)dependence definition: Sibuya (1960), Zhang (2005, 2008)

Two identically distributed r.v. X and Y are called *tail independent* if

$$\lambda = \lim_{u \to x_F} P(Y > u \,|\, X > u) \tag{3}$$

exists and equals 0, where $x_F = \sup\{x \in \mathbb{R} : \ P(X \leq x) < 1\}$. The quantity λ, if exists, is called the bivariate tail dependence index. If $\lambda > 0$, then (X, Y) is called *tail dependent* and we say there are extreme co-movements between X and Y.

Importance of Studying Tail (In)dependence

In theory

- Suppose $\{(X_i, Y_i), i = 1, \ldots, n\}$ is a random sample of (X, Y).
- If $\lambda = 0$, then the limit joint bivariate extreme value distribution is the product of the univariate limit distributions, i.e.

$$\lim_{n\to\infty} P\left\{a_n(\max_i X_i - b_n) < x,\ c_n(\max_i Y_i - d_n) < y\right\}$$
$$= \lim_{n\to\infty} P\left\{a_n(\max_i X_i - b_n) < x\right\} \lim_{n\to\infty} P\left\{c_n(\max_i Y_i - d_n) < y\right\}.$$

- When $\lambda = 0$, the limit theory of the joint maxima is simple and easy!
  - Example: a bivariate normal random variable with correlation coefficient $\rho \neq 1$.
- When $\lambda > 0$, the limit theory of the joint maxima does not show a unified parametric form!!
  - Example: a bivariate t random variable with correlation coefficient $\rho > 0$.

Importance of Studying Tail (In)dependence: In extremal climatic conditions.

Notes

- Global warming causes severe storms.
- Increased ocean temperatures cause increasingly intense hurricanes. Hurricane Sandy forced Wall Street to close first time since 2001.
- “Hurricane Sandy’s effect on the politics of climate change was immediate." Source: TINA ROSENBERG
- “Will Storm’s Wall Street Impact Influence U.S. Carbon Policy?" Source: ANDREW C. REVKIN.
- There are increasing concerns about the reliability of climate models.

Importance of Studying Tail (In)dependence: In extremal climatic conditions.

Notes

- Mexico hunkers down for Patricia, 'the most dangerous storm in history'
- 2015 might have seen hottest summer in 4,000 years
- CHARLESTON, S.C. (Reuters) - Rainfall in South Carolina over the last few days reached historic in parts of the state that are expected to be seen once in 1,000 years, Governor Nikki Haley announced at a press conference on Sunday.

The specific question for forecasting future extremal climatic conditions:

- How to account for historical records.

Two fundamental questions:

- How to identify tail dependencies and nonlinear dependencies between variables.
- How to develop statistical models dealing with tail dependencies and nonlinear dependencies.

Our focus:

- The first question.

The specific question for forecasting future extremal climatic conditions:

- How to account for historical records.

Two fundamental questions:

- How to identify tail dependencies and nonlinear dependencies between variables.
- How to develop statistical models dealing with tail dependencies and nonlinear dependencies.

Our focus:

- The first question.

The specific question for forecasting future extremal climatic conditions:

- How to account for historical records.

Two fundamental questions:

- How to identify tail dependencies and nonlinear dependencies between variables.
- How to develop statistical models dealing with tail dependencies and nonlinear dependencies.

Our focus:

- The first question.

Outline

Analog definitions of linear and nonlinear (tail) correlation coefficients under different stable law

Under stable law	Under max-stable law
The simplest factor model	The simplest max-linear factor model
$X = a_1 Z_1 + a_2 Z_2$ $Y = b_1 Z_1 + b_2 Z_2$	$X = \max(c_1 Z_1, c_2 Z_2)$, $c_i \geq 0$ $Y = \max(d_1 Z_1, d_2 Z_2)$, $d_i \geq 0$
The 'fourteenth' example	The first example
Pearson's linear correlation coefficient	sample based quotient correlation coefficient
$a_1^2 + a_2^2 = 1$, $b_1^2 + b_2^2 = 1$ $\rho = a_1 b_1 + a_2 b_2$	$c_1 + c_2 = 1$, $d_1 + d_2 = 1$ $\lambda = \min(c_1 + d_2,\ c_2 + d_1)$

Elementary facts

- **Ways to measure relative positions:**
  - the difference $X - Y$;
  - the quotient X/Y for positive variables X and Y.
    - $X/Y = 0$ 'means' no relation.
    - $X/Y = 1$ means they are identical.
- **In a 'Normal' world:** Pearson correlation coefficient is the sum of the products of the Z scores,

$$r_n = \frac{1}{n}\sum Z_{x_i} Z_{y_i}, \quad r_n \xrightarrow{\mathscr{P}} \rho.$$

- Quotient correlation coefficients are based on the maxima of the quotients of the Fréchet scores,

$$\mathrm{q}_n = \frac{\max_{i\le n}\{Y_i/X_i\} + \max_{i\le n}\{X_i/Y_i\} - 2}{\max_{i\le n}\{Y_i/X_i\} \times \max_{i\le n}\{X_i/Y_i\} - 1}, \ \mathrm{q}_n \xrightarrow{\mathscr{P}} \lambda(?). \qquad (4)$$

Motivations:

- In view of (3), the tail dependence index λ is mainly relying on a high threshold value u and the dependence between tails of two random variables.

A generalized tail dependence measure: Tail quotient correlation coefficient (TQCC)

Suppose now X_i and Y_i are two dependent unit Fréchet random variables. Define a sample based tail dependence measure by

$$q_{u_n} = \frac{\max\limits_{1\le i\le n}\left\{\frac{\max(X_i,u_n)}{\max(Y_i,u_n)}\right\} + \max\limits_{1\le i\le n}\left\{\frac{\max(Y_i,u_n)}{\max(X_i,u_n)}\right\} - 2}{\max\limits_{1\le i\le n}\left\{\frac{\max(X_i,u_n)}{\max(Y_i,u_n)}\right\} \times \max\limits_{1\le i\le n}\left\{\frac{\max(Y_i,u_n)}{\max(X_i,u_n)}\right\} - 1}. \tag{5}$$

- In the particular case of $u_n = u$ (a constant), definition (5) coincides with the one defined in Zhang (2008). In this talk, u_n is random, and it is allowed to diverge to infinity. Engle (2005).

Hypotheses of tail (in)dependence

$$H_0 : X \text{ and } Y \text{ are tail independent}$$
$$\leftrightarrow H_1 : X \text{ and } Y \text{ are tail dependent},$$

which can also be written as

$$H_0 : \lambda = 0 \longleftrightarrow H_1 : \lambda > 0. \tag{6}$$

In the remaining of the talk, we discuss how to test the null of (6) and how to estimate λ under the alternative hypothesis.

Remarks

- The null and alternative hypotheses in Ledford and Tawn (1996, 1997) are reversed in this talk, see also Peng (1999), Draisma et al. (2004), and others.
- Other significant tests include Falk and Michel (2006), Hüsler and Li (2009), Bacro, Bel, and Lantuéjoul (2010) etc.

Hypotheses of tail (in)dependence

$$H_0 : X \text{ and } Y \text{ are tail independent}$$
$$\leftrightarrow H_1 : X \text{ and } Y \text{ are tail dependent},$$

which can also be written as

$$H_0 : \lambda = 0 \longleftrightarrow H_1 : \lambda > 0. \tag{6}$$

In the remaining of the talk, we discuss how to test the null of (6) and how to estimate λ under the alternative hypothesis.

Remarks

- The null and alternative hypotheses in Ledford and Tawn (1996, 1997) are reversed in this talk, see also Peng (1999), Draisma et al. (2004), and others.
- Other significant tests include Falk and Michel (2006), Hüsler and Li (2009), Bacro, Bel, and Lantuéjoul (2010) etc.

Limit distribution and tail independence test

Assuming $\lambda = 0$. Define

$$\mathrm{q}_{n,t} = \frac{\max_{i\le n} \frac{\max(X_i, T_{n,t})}{\max(Y_i, T_{n,t})} + \max_{i\le n} \frac{\max(Y_i, T_{n,t})}{\max(X_i, T_{n,t})} - 2}{\max_{i\le n} \frac{\max(X_i, T_{n,t})}{\max(Y_i, T_{n,t})} \times \max_{i\le n} \frac{\max(Y_i, T_{n,t})}{\max(X_i, T_{n,t})} - 1},$$

where X_i and Y_i are unit Fréchet random variables, $T_{n,t}$ is distributed as e^{-n/w^t}, for $w > 0$ and $t > 1$. Then

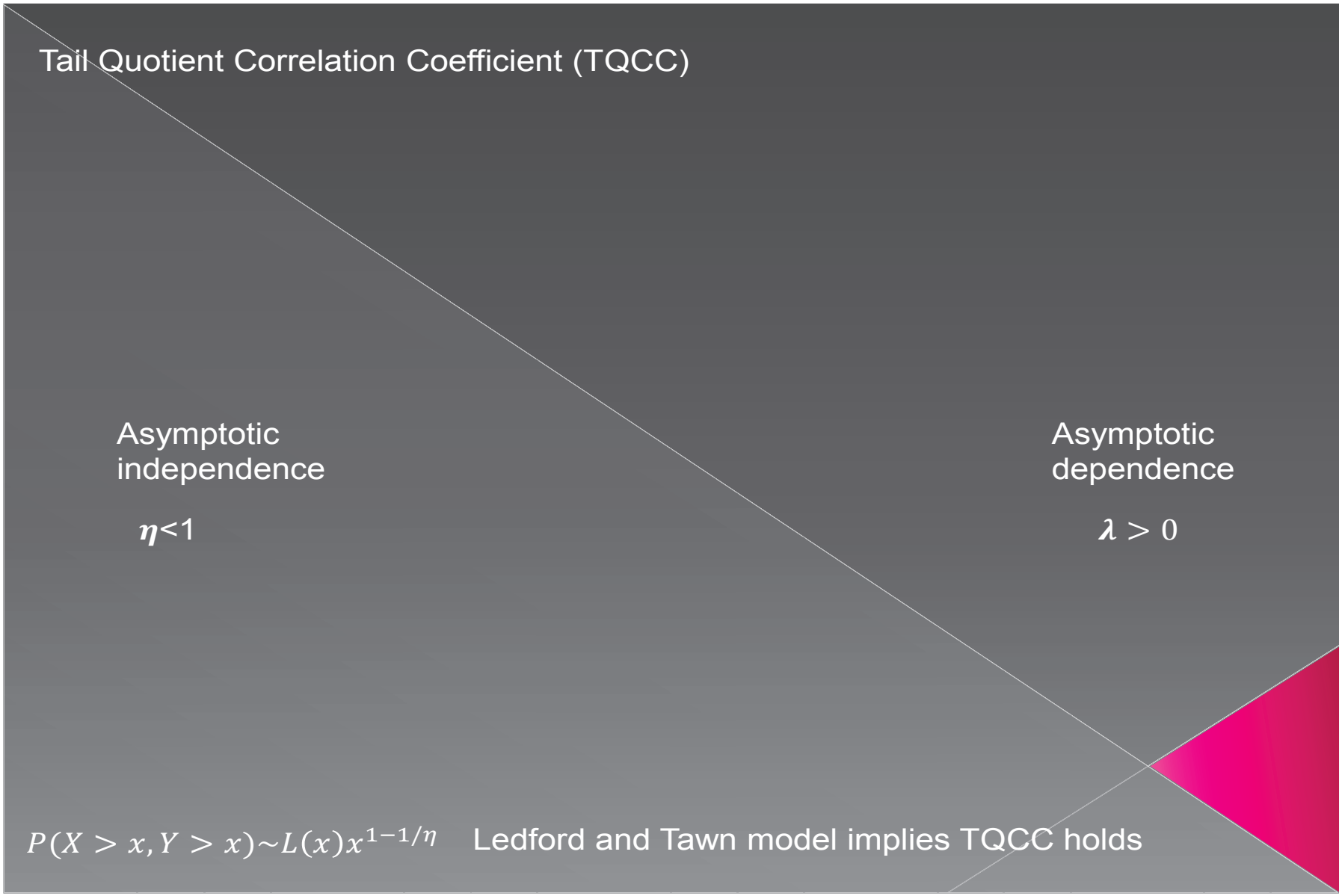
$$2n\{1 - e^{-1/T_{n,t}}\}\mathrm{q}_{n,t} \xrightarrow{\mathscr{L}} \chi^2_4.$$

Hypotheses which are connected to (6)

Testing procedure

- For a given significance level α and an appropriate chosen u_n, if $2n\{1-\exp(-1/u_n)\}\mathrm{q}_{u_n} > \chi^2_{4;\alpha}$, H_0 of (6) is rejected, and we conclude there exists tail dependence between two random variables of interest. Here $\chi^2_{4;\alpha}$ is the upper α percentile of a χ^2 distributed random variable with 4 degrees of freedom.
- If H_0 of (6) is rejected, (5) is an estimate of λ.

λ vs. η



Remarks

Assumption T1: Suppose for $1 < t < 1+\delta,\ \delta > 0$, paired tail independent random variables (X_i, Y_i) satisfy

$$\max_{1\le i\le n} \frac{\max(X_i, T_{n,t})}{\max(Y_i, T_{n,t})} \Big/ \max_{1\le i\le n} \frac{\max(X_i, T_{n,t})}{T_{n,t}} = 1 + o_p(1),$$

$$\max_{1\le i\le n} \frac{\max(Y_i, T_{n,t})}{\max(X_i, T_{n,t})} \Big/ \max_{1\le i\le n} \frac{\max(Y_i, T_{n,t})}{T_{n,t}} = 1 + o_p(1).$$

Proposition 1

Suppose $g(u) \sim L(u)u^{-1+1/\eta},\ \eta \in (0,1]$. *Then*

- *Assumption T1 holds for* $t\eta < 1$ *when* $\eta < 1$.
- *Assumption T1 doesn't hold when* $\eta = 1$, *i.e. tail dependent case.*

- Pearson's sample correlation coefficient and TQCC are asymptotically independent. Zhang, Qi, and Ma (2011).
- TQCC with $u_n = 0$ is $\sqrt{n}$ convergence under the alternative hypothesis of bivariate joint Fréchet distributed , Wang (2012).

Outline

The precipitation data

- The data are daily precipitation totals covering period 1950-1999 over 5873 stations in the continental USA (excluding Alaska and Hawaii). The data units are tenths of a millimeter.
- The data are the same as used by Smith, Grady, and Hegerl (2007), and by Shamseldin, Smith, Sain, Mearns, and Cooley (2008).
- The data are first fitted to GEVs, and then transformed to unit Fréchet margins.

$$H(x;\xi,\mu,\psi) = \exp[-\{1+\xi(x-\mu)/\psi\}_+^{-1/\xi}], \tag{7}$$

to local maxima of observations, where μ is a location parameter, $\psi > 0$ is a scale parameter, and ξ is a shape parameter

Tail indecies of precipitations

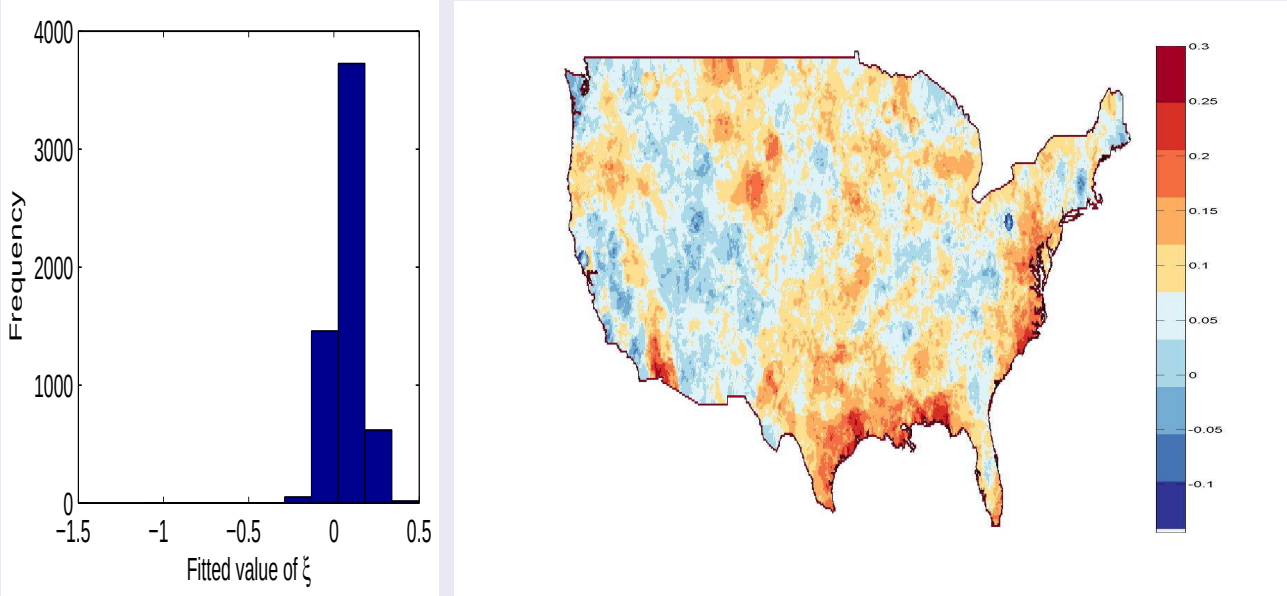


Figure 7: *Fitted tail shape parameter values for all 5873 station time series. The left panel shows the distribution of fitted shape parameter values. The right panel plots fitted shape parameter values to US map using kriging.*

- Precipitations appear to be non-stationarity, spatial clusters, and asymmetry over all stations.
- Precipitations over stations near Mexican bay region and stations near Atlantic ocean and along North Carolina coast have heavier tails then precipitations over other stations have.

Lemma 6.1

Suppose that $X, X_1, X_2, \ldots$ are positive random variables. Then $X_n \xrightarrow{\mathscr{P}} X$ if and only if there are two sequences of positive random variables $\xi_n^{(1)}$ and $\xi_n^{(2)}$ such that $\xi_n^{(1)} \xrightarrow{\mathscr{P}} 1$, $\xi_n^{(2)} \xrightarrow{\mathscr{P}} 1$, and $\xi_n^{(1)} X \leq X_n \leq \xi_n^{(2)} X$, $n = 1, 2, \ldots$.

Lemma 6.2

Suppose that $\{X_i\}_{i=1}^n$ is a random sample from the distribution $F_{\gamma_0}(x) = \exp(-1/x^{\gamma_0})$, with $x > 0$, and the true shape parameter γ_0. Suppose that the estimator of γ_0 is $\widehat{\gamma} = \widehat{\gamma}(X_1, \ldots, X_n)$ satisfying $n^{\alpha}(\widehat{\gamma} - \gamma_0) \xrightarrow{\mathscr{L}} W$, for some $\alpha > 0$ and some random variable W. Then as $n \to \infty$,

$$\max_{1 \le i \le n} \left| \log\{F_{\gamma_0}(X_i)\} / \log\{F_{\widehat{\gamma}}(X_i)\} - 1 \right| \xrightarrow{\mathscr{P}} 0.$$

Limit theorem under estimated marginal transformation

Proposition 2

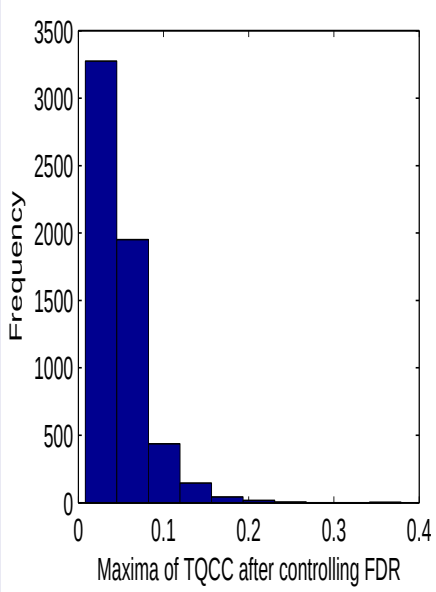
Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ *is a sample of two independent GEV random variables* (X, Y) *with the true shape parameters* $\xi_X = \xi_{0;X}$ *and* $\xi_Y = \xi_{0;Y}$ *respectively. Suppose that estimators of* ξ_X *and* ξ_Y *are* $\widehat{\xi}_X = \xi_{n;X}(X_1, \ldots, X_n)$ *and* $\widehat{\xi}_Y = \xi_{n;Y}(Y_1, \ldots, Y_n)$ *satisfying* $n^{\alpha_X}(\widehat{\xi}_X - \xi_{0;X}) \xrightarrow{\mathscr{L}} W_X$ *and* $n^{\alpha_Y}(\widehat{\xi}_Y - \xi_{0;Y}) \xrightarrow{\mathscr{L}} W_Y$, *where* $\alpha_X > 0$, $\alpha_Y > 0$ *and* W_X *and* W_Y *are random variables. Denote* $\widehat{X}_i = -1/\log\{H(X_i; \widehat{\xi}_X)\}$, $\widehat{Y}_i = -1/\log\{H(Y_i; \widehat{\xi}_Y)\}$ *and set*

$$\widehat{\mathrm{q}}_{u_n} = \frac{\max\limits_{1\le i\le n}\{\max(\widehat{X}_i, u_n)/\max(\widehat{Y}_i, u_n)\} + \max\limits_{1\le i\le n}\{\max(\widehat{Y}_i, u_n)/\max(\widehat{X}_i, u_n)\} - 2}{\max\limits_{1\le i\le n}\{\max(\widehat{X}_i, u_n)/\max(\widehat{Y}_i, u_n)\} \times \max\limits_{1\le i\le n}\{\max(\widehat{Y}_i, u_n)/\max(\widehat{X}_i, u_n)\} - 1}. \tag{8}$$

Then

$$2n\{1 - \exp(-1/u_n)\}\widehat{\mathrm{q}}_{u_n} \xrightarrow{\mathscr{L}} \chi^2_4.$$

Overall tail dependency across all stations



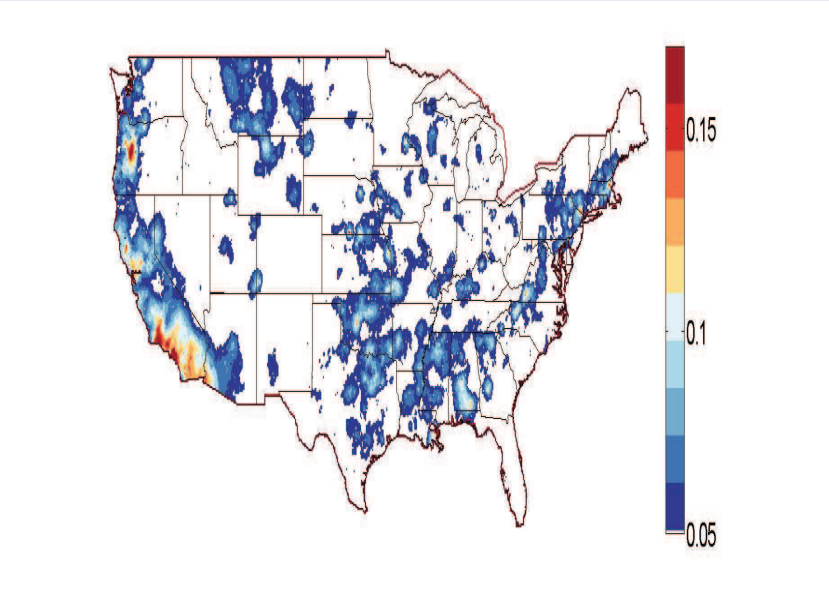


Figure 8: *Maximal extremal precipitation dependencies between one station and the rest of stations on the same day.*

- The tail dependencies are generally clustered. There are some paired distanced stations showing tail dependencies.

Overall tail dependency across all stations

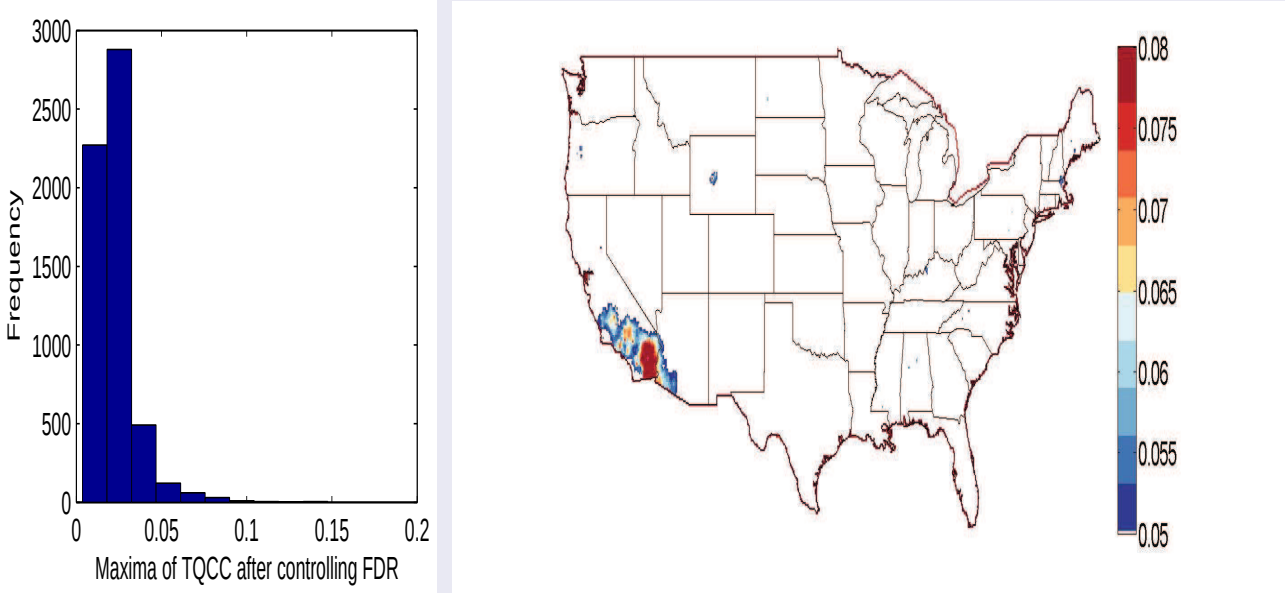
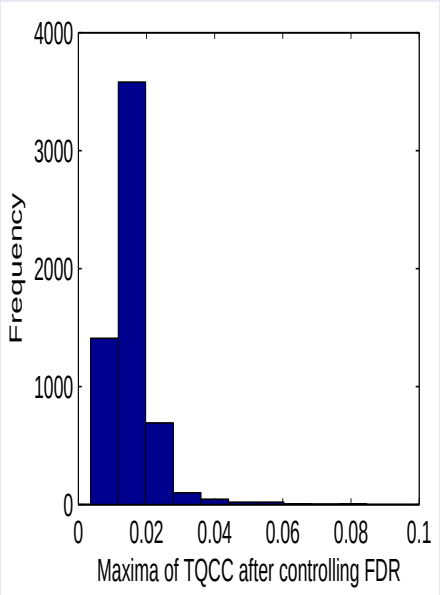


Figure 9: *Maximal extremal precipitation dependencies between one station and the rest of stations on the lagged-1 day.*

Overall tail dependency across all stations



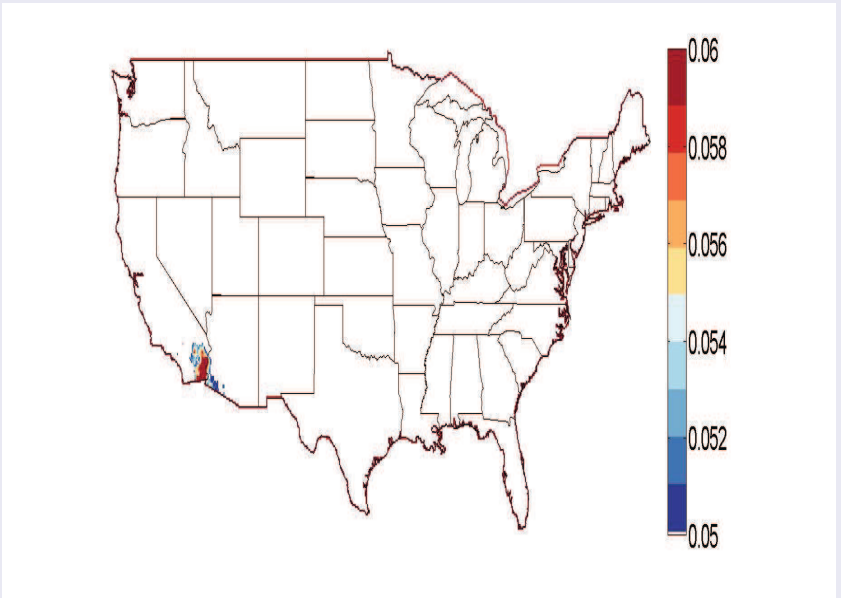


Figure 10: *Maximal extremal precipitation dependencies between one station and the rest of stations on the lagged-7 day.*

- The maximal tail dependencies decay as time goes by.

Table 1: *The 10 largest tail quotient correlation coefficients.*

Pair ID	TQCC	Latitude	Longitude	Elevation	Station name
(I)	.3787	33.92	-118.13	34	DOWNEY FIRE STN FC107D
		33.97	-118.02	128	WHITTIER CITY YD FC106C
(II)	.2652	44.40	-122.48	262	CASCADIA
		44.10	-122.68	206	LEABURG 1 SW
(III)	.2324	34.48	-119.50	633	JUNCAL DAM
		34.53	-119.78	312	LOS PRIETOS RANGER STN
(IV)	.2271	40.08	-99.20	610	HARLAN COUNTY LAKE
		40.07	-99.13	573	NAPONEE
(V)	.2208	39.35	-123.12	309	POTTER VALLEY P H
		39.13	-123.20	193	UKIAH
(VI)	.2206	42.48	-71.28	49	BEDFORD
		42.52	-71.13	27	READING
(VII)	.2200	34.52	-119.68	473	GIBRALTAR DAM 2
		34.48	-119.50	633	JUNCAL DAM
(VIII)	.2198	34.08	-117.87	175	COVINA NIGG FC193B
		33.97	-118.02	128	WHITTIER CITY YD FC106C
(IX)	.2128	29.95	-90.13	6	NEW ORLEANS WATER PLT
		29.98	-90.02	3	NEW ORLEANS D P S 5
(X)	.2092	33.53	-117.77	11	LAGUNA BEACH

The most recent flooding areas

Pair ID	TQCC	Latitude	Longitude	Elevation	Station name
(I)	.0223	31.95	-112.80	512	ORGAN PIPE CACTUS N M
		38.62	-122.87	33	HEALDSBURG
(II)	.0239	47.55	-116.17	680	KELLOGG AIRPORT
		47.62	-117.52	718	SPOKANE WSO AIRPORT
(III)	.0363	41.25	-91.37	204	COLUMBUS JUNCT 2 SSW
		41.63	-91.52	195	IOWA CITY
(IV)	.0214	29.98	-90.25	1	NEW ORLEANS WSCMO ARP
		45.52	-89.20	488	SOUTH PELICAN
(V)	.0326	47.93	-97.17	256	GRAND FORKS FAA AP
		47.92	-97.08	253	GRAND FORKS UNIV NWS
(VI)	.0769	33.98	-78.00	6	SOUTHPORT 5 N
		34.32	-77.92	12	WILMINGTON 7 N

Example of far distance tail dependence: spatial tail dependence

Satellite image of China precipitation

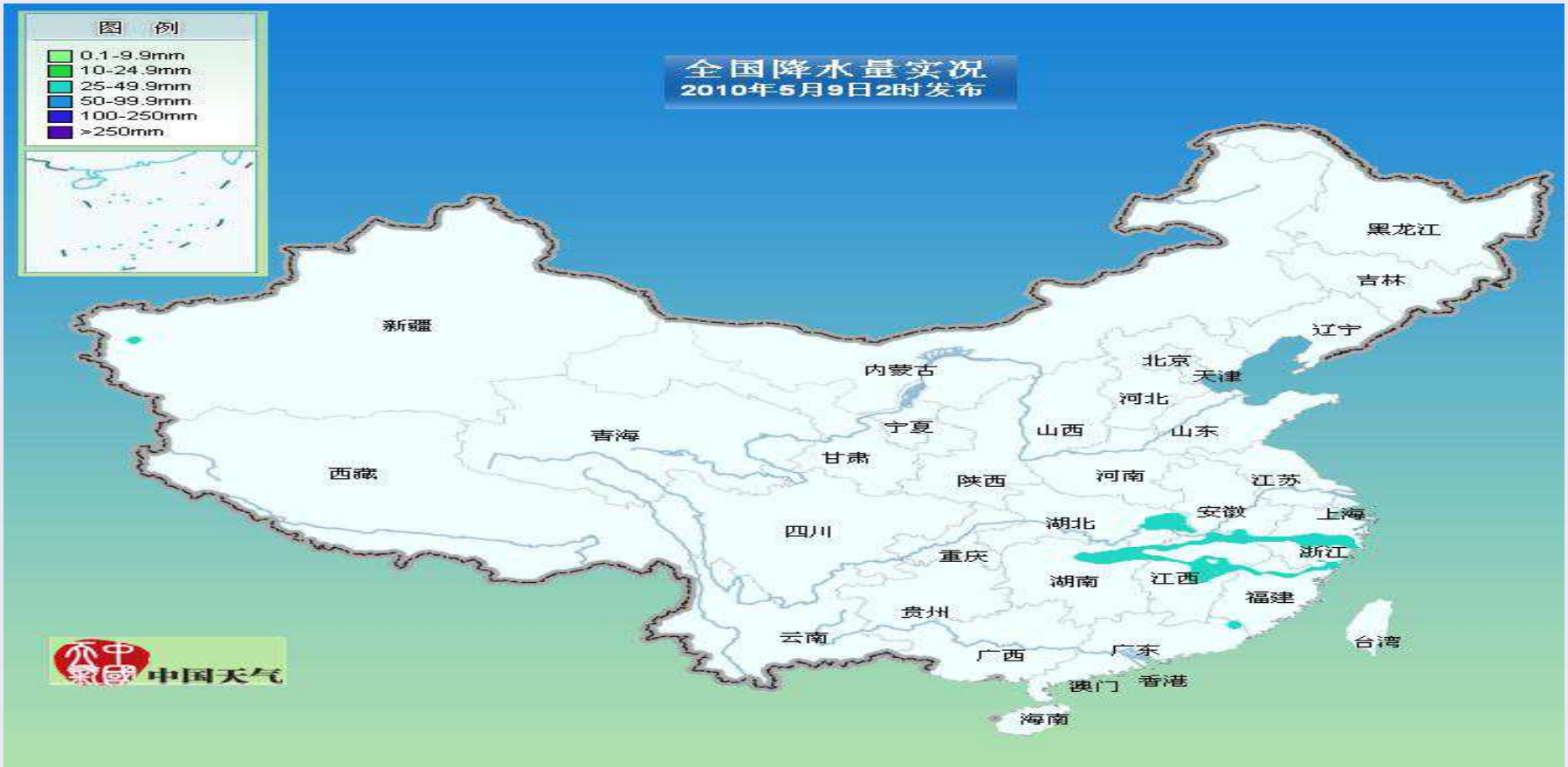


Figure 11: *China heavy rain real time image*

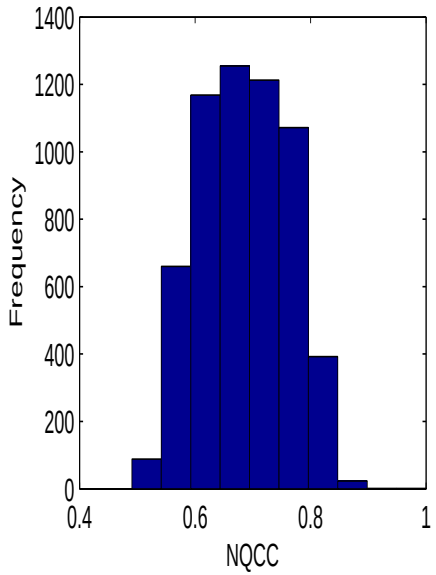
Ecuador and Japan earthquakes: Are they related?

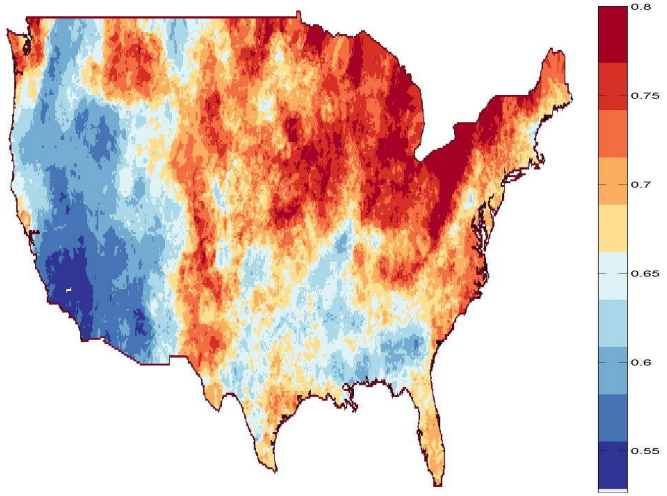
Figure 12: *“Usually we don’t think earthquake are connected across the ocean," Caruso said, but there’s ongoing research in “remote triggering," the idea that a big quake can cause another quake a long distance away.*

Nonlinear quotient correlation for less extremes

$$\mathrm{q}_n(g) = \frac{g\left(\frac{\max(u_n,X_1)}{\max(u_n,Y_1)},\ldots,\frac{\max(u_n,X_n)}{\max(u_n,Y_n)}\right)+g\left(\frac{\max(u_n,Y_1)}{\max(u_n,X_1)},\ldots,\frac{\max(u_n,Y_n)}{\max(u_n,X_n)}\right)-2}{g\left(\frac{\max(u_n,X_1)}{\max(u_n,Y_1)},\ldots,\frac{\max(u_n,X_n)}{\max(u_n,Y_n)}\right)\times g\left(\frac{\max(u_n,Y_1)}{\max(u_n,X_1)},\ldots,\frac{\max(u_n,Y_n)}{\max(u_n,X_n)}\right)-1},$$

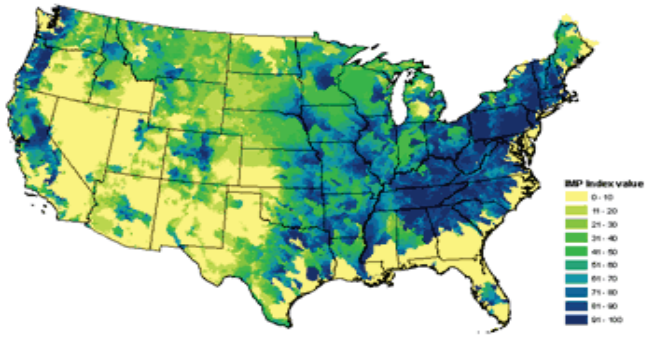
where $g(z_1,\ldots,z_n)$ gives the kth largest value, or the pth percentile, of $\{z_1,\ldots,z_n\}$ such that $g(z_1,\ldots,z_n)\geq 1$.

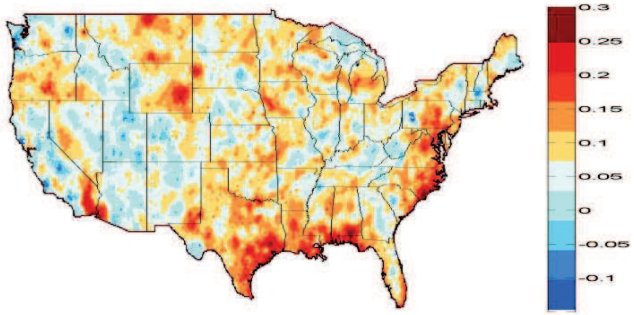


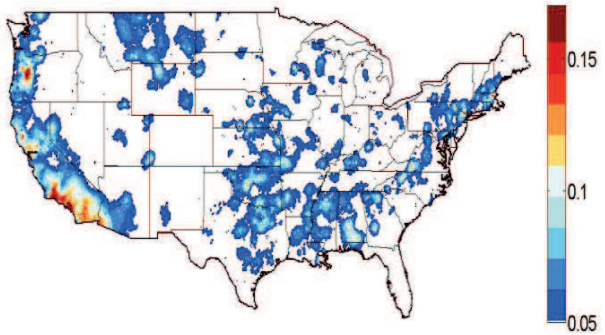


- Note that the shape of nonlinear dependence correlations show a bell shaped curve.

Take-home messages







Outline

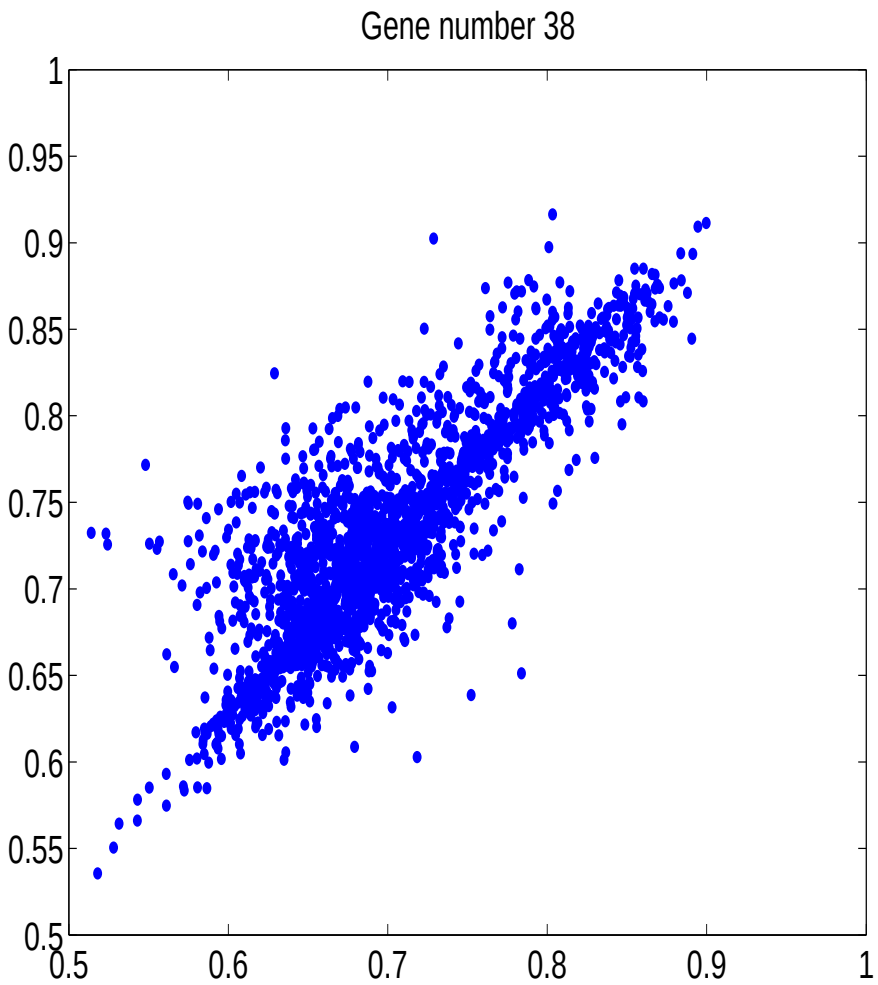
Dimension Reduction from Ultra High to Ultra Low

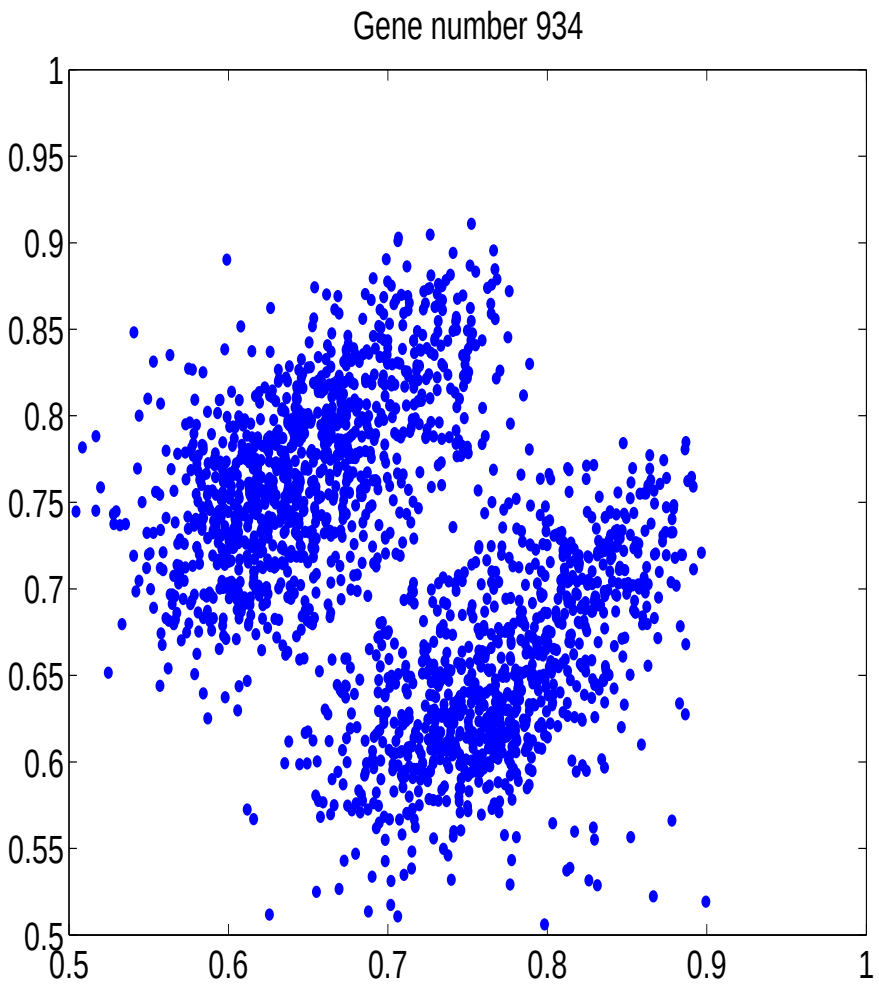
A Three Variable Selector

The Crazy, The Good, and The Ugly

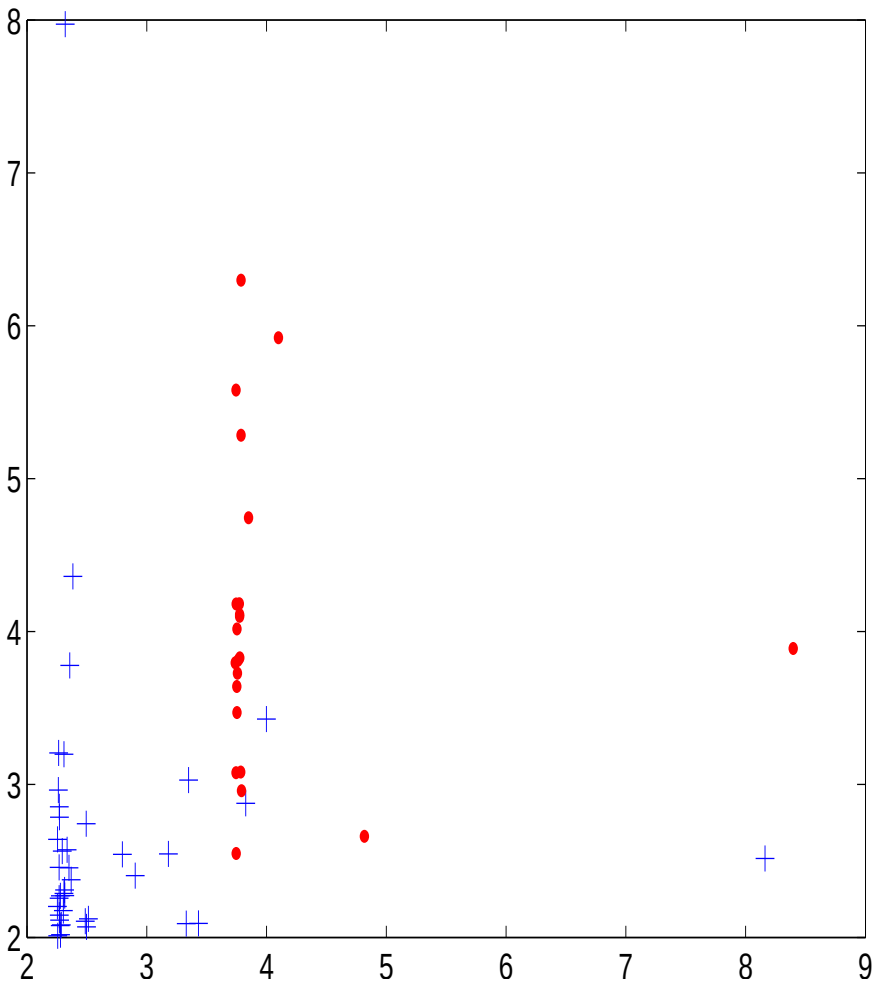
Kunming 2007

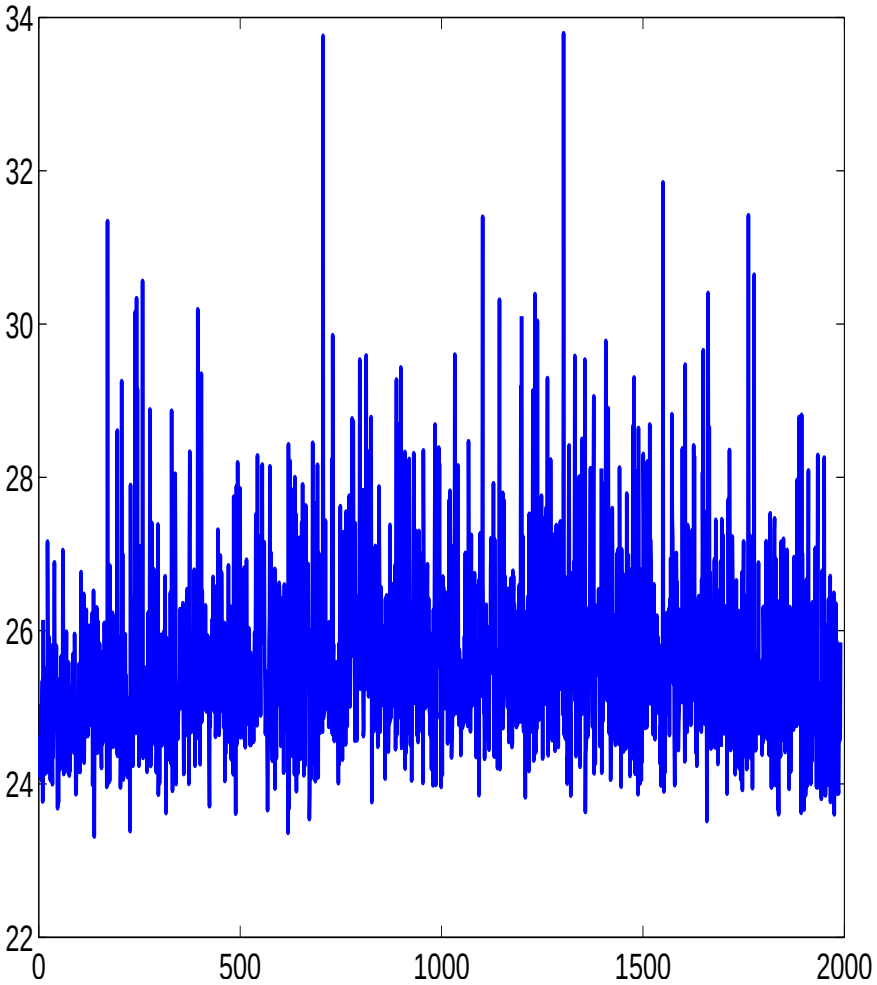
Colon gene expression data analysis



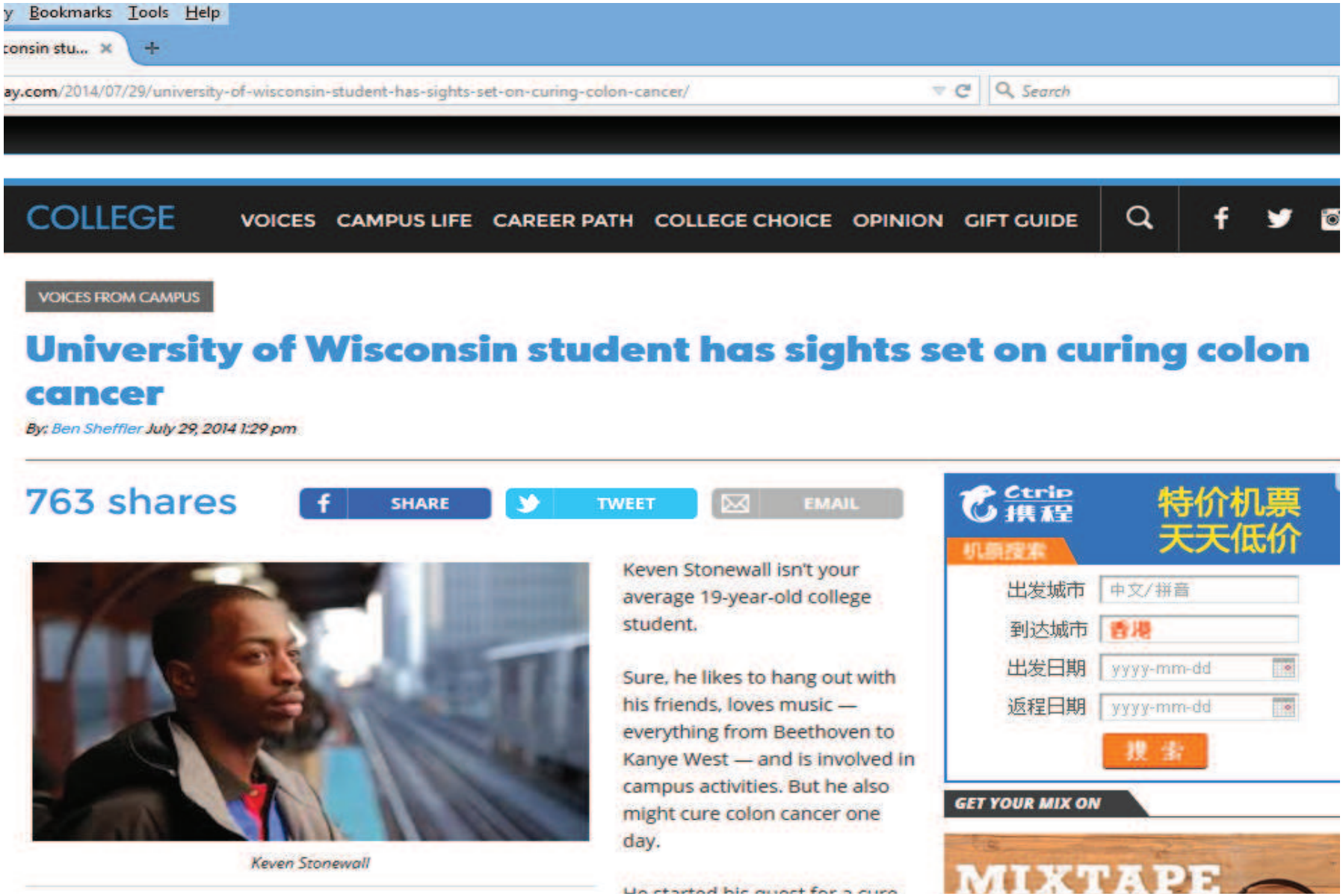


Colon gene expression data analysis





USToday's report



Thank You!