Stat 521, Graphical Models and Parsimonious Dependence Prof Harry Joe, Fall 2015, 2015.10.26–2015.12.05

Examples to show why graphical models should be considered on whether they extend from Gaussian to non-Gaussian.

A. Probability integral transform (to normal/Gaussian)

B. Data sets

1. Abalone data; abalone is a sea animal and a mollusk, from archive.ics.uci.edu/ml/datasets.html

2. Exchange rate log returns for a few international currencies relative to USD. Log return for day t means $\log(x_t/x_{t-1})$ where x_t is the exchange rate on day t.

3. Market log returns for a few European stock markets. Log return for day t means $\log(P_t/P_{t-1})$ where P_t is the value of the stock market index on day t.

Main result for simulation of random variables and copula models

If $U \sim U(0,1)$ and F is a univariate cdf, then $F^{-1}(U) \sim F$; F^{-1} is the (generalized) function inverse of F.

If $X \sim F_X$ and F_X is continuous, then $F_X(X) \sim U(0,1)$.

Transform from a continuous random variables $X \sim F_X$ to $N(0,1)$: Let $\Phi(\cdot)$ be the standard normal cdf with function inverse Φ^{-1}. Then from the above, $F_X(X) \sim U(0,1)$ and $\Phi^{-1}[F_X(X)] = [\Phi^{-1} \circ F_X](X) \sim N(0,1)$.

So if data variables are not normally distributed, they can be transformed to $N(0,1)$, and then one can consider whether a multivariate normal/Gaussian model is suitable for the transformed variables.

Pay attention to notation:

Random variables appear as upper case letters in subscripts of cdfs F and density functions f.

The following has incorrect notation.

X, Y are independent random variables if $f(x,y) = f(x)\, f(y)$. [This overloads f, what does $f(2)$ refer to?]

Proper writing: X, Y are independent random variables if $f_{X,Y} = f_X f_Y$ (when domain of density functions are understood), or

X, Y are independent random variables if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all $x, y \in \Re$.

Let X and Y be continuous random variables with respective cdfs F_X, F_Y, and respective density functions f_X, f_Y. Next is a derivation for a transform based on a differentiable, monotone increasing, real-valued function g.

Let $Y = g(X)$: express the cdf and density of Y in terms of those of X.

$$F_Y(y) = \Pr(Y \le y) = \Pr(X \le g^{-1}(y)) = F_X(g^{-1}(y)),$$

$$f_Y(y) = F_Y'(y) = f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}.$$

When making derivations involving transformed variables, the arguments of cdfs and density functions are not necessarily dummy variable version of the random variables.

Normal scores transform for continuous variables.

Data $(y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$.

(a) Rank transform to normal scores

Rank the jth variable vector $(y_{1j}, \ldots, y_{nj})$ in increasing order to get ranks $R_{1j}, \ldots, R_{nj}$ (permutation of $1, \ldots, n$ if no ties).

$\hat{z}_{ij} = \Phi^{-1}((R_{ij} - 0.5)/n)$ for $i = 1, \ldots, n$, consist of normal scores transform for jth variable.

(b) Parametric transform to normal scores

jth variable univariate model $F_j(\cdot; \eta_j)$ is fitted to $(y_{1j}, \ldots, y_{nj})$.

$\hat{u}_{ij} = F_j(y_{ij}; \hat{\eta}_j)$; $\hat{z}_{ij} = \Phi^{-1}(\hat{u}_{ij})$ for normal scores transform.

Bivariate plots: look for deviations from elliptical shape clouds.

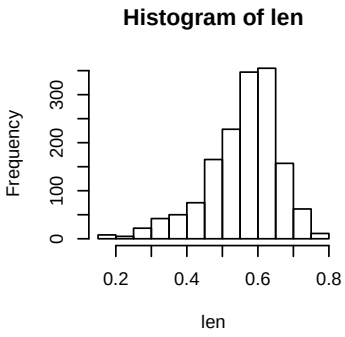
For $j \ne k$, plot $(\hat{z}_{ij}, \hat{z}_{ik})$, $i = 1, \ldots, n$.

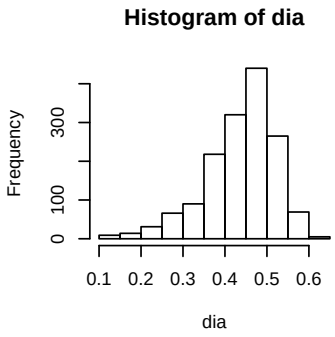
Correlation of normal scores (vander Waarden coefficient):
$\hat{\rho}_{N,jk}$ is the sample correlation of $(\hat{z}_{ij}, \hat{z}_{ik})$.

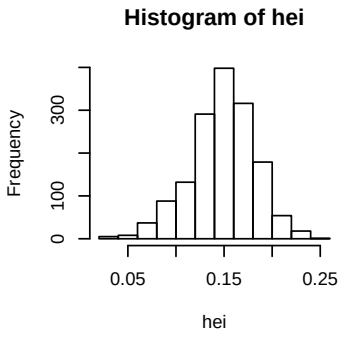
Abalone data (subset of males)

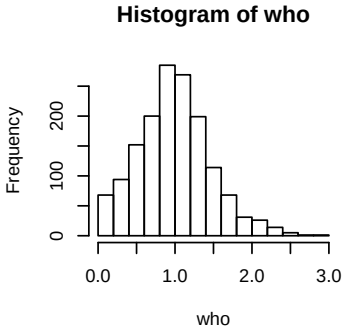
```
1. Length continuous mm Longest shell measurement
2. Diameter continuous mm perpendicular to length
3. Height continuous mm with meat in shell
4. Whole weight continuous grams whole abalone
5. Shucked weight continuous grams weight of meat
6. Viscera weight continuous grams gut weight (after bleeding)
7. Shell weight continuous grams after being dried
8. Rings integer +1.5 gives the age in years
```

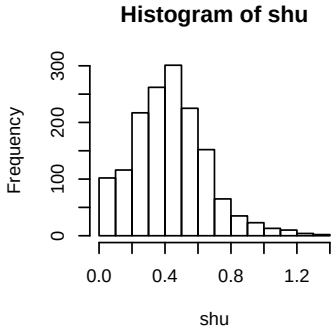
1. univariate histograms (skewed): need 3-parameters parametric densities such as skew-normal or generalized gamma.
2. bivariate normal scores plot: non-elliptical shape implies dependence is not Gaussian even after transforming univariate margins to N(0,1).

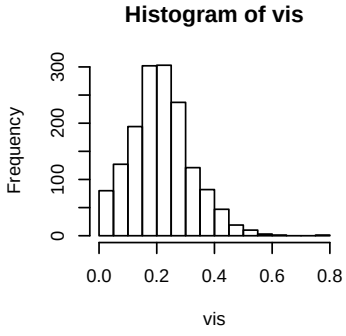

Histogram of len
Frequency
0
100
200
300
0.2
0.4
0.6
0.8
len

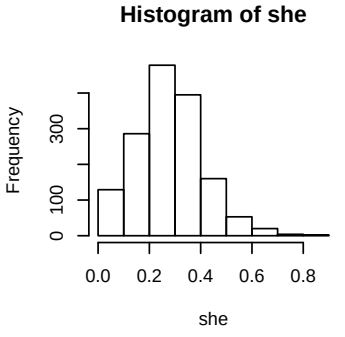

Histogram of dia
Frequency
0
100
300
0.1
0.2
0.3
0.4
0.5
0.6
dia

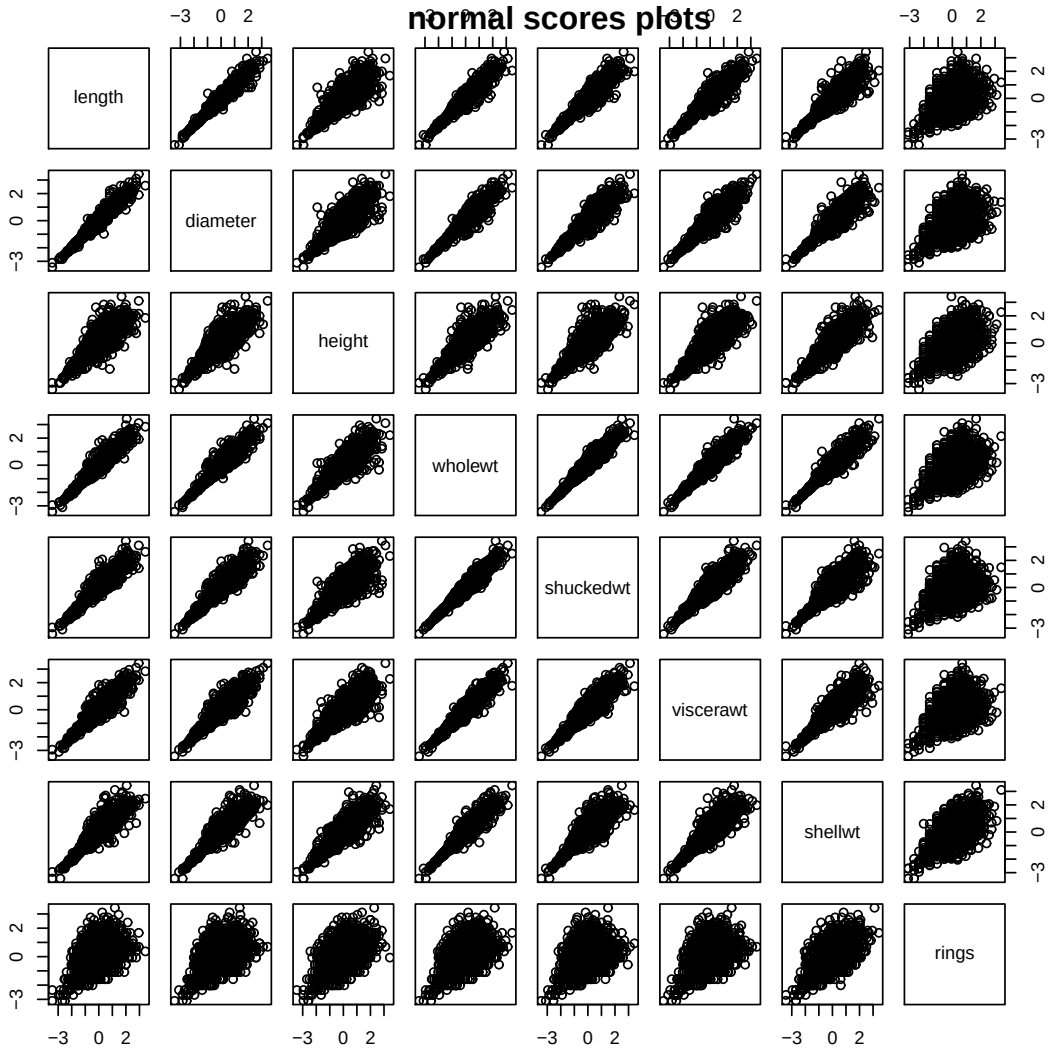

Histogram of hei
Frequency
0
100
300
0.05
0.15
0.25
hei


Histogram of who
Frequency
0
100
200
0.0
1.0
2.0
3.0
who

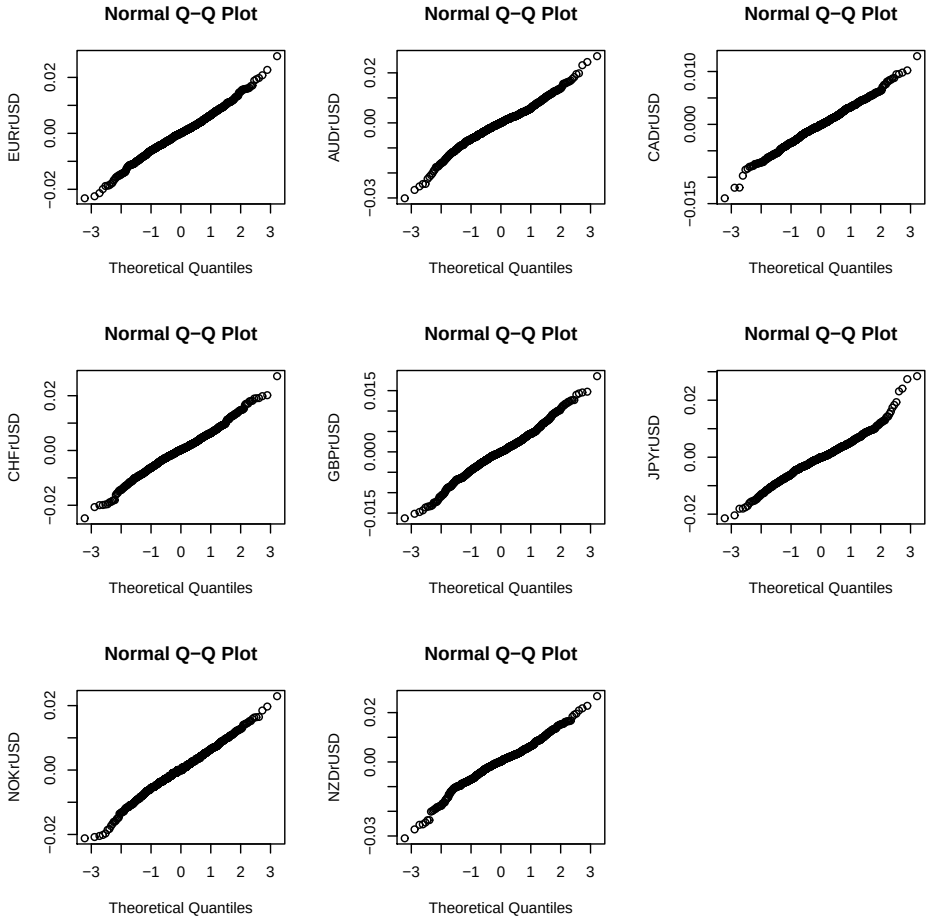

Histogram of shu
Frequency
0
100
200
300
0.0
0.4
0.8
1.2
shu

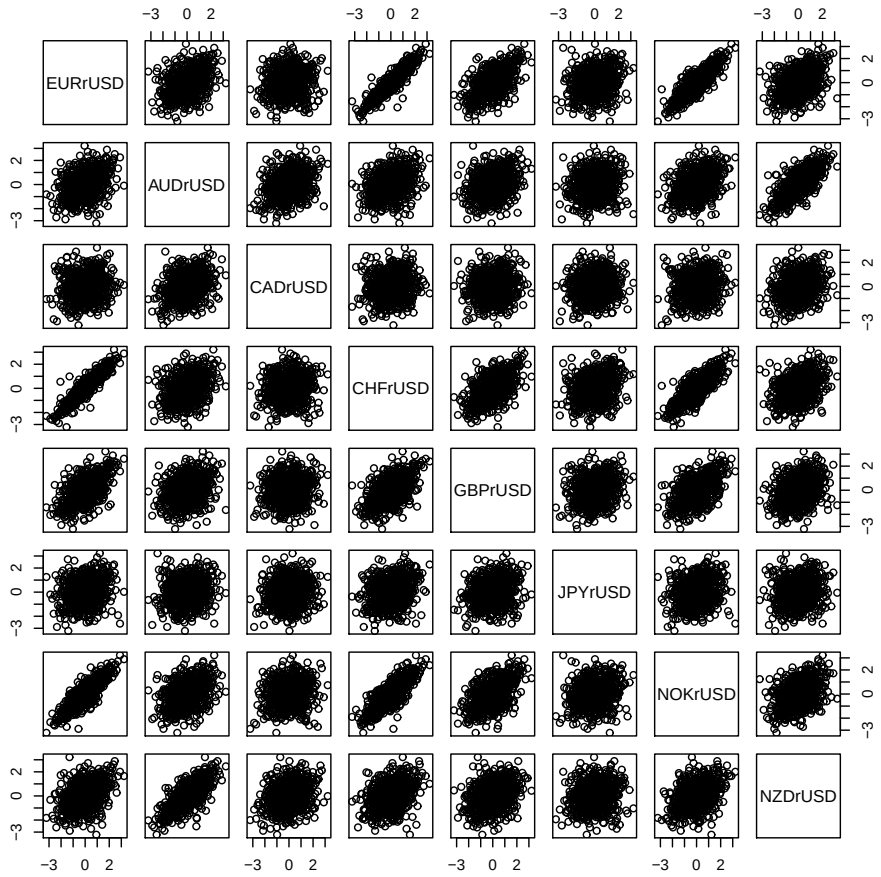

Histogram of vis
Frequency
0
100
200
300
0.0
0.2
0.4
0.6
0.8
vis


Histogram of she
Frequency
0
100
300
0.0
0.2
0.4
0.6
0.8
she



Exchange rate data 2001-2003: univariate normal probability plots, bivariate normal scores plot. These can look close to Gaussian if time period for data is 3 years.

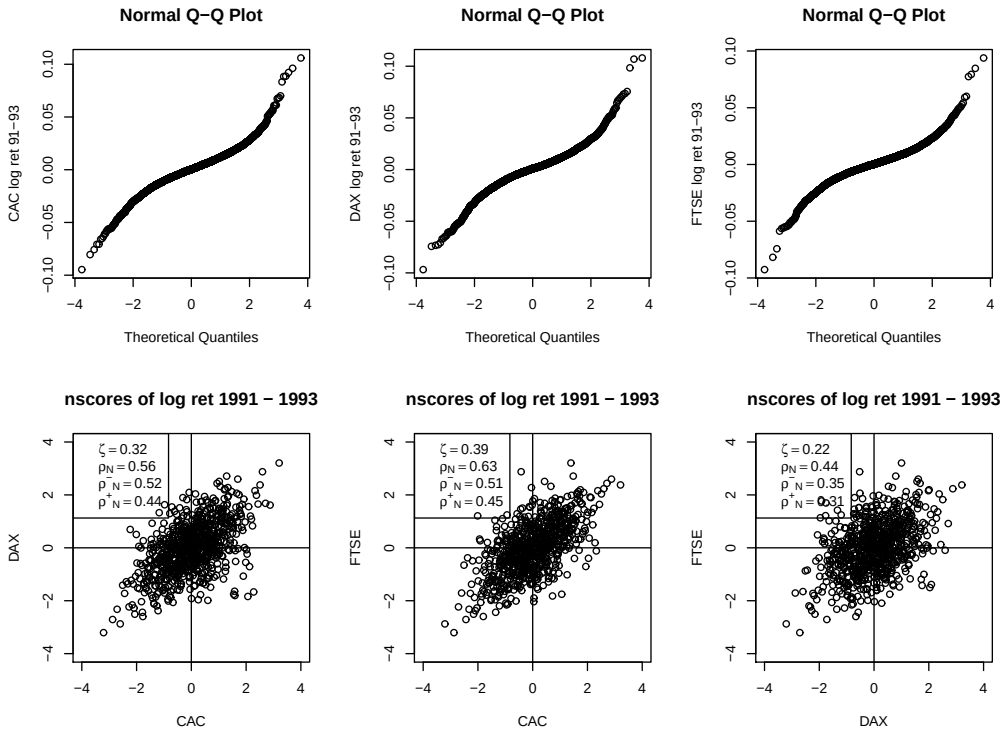

Normal Q−Q Plot
EURrUSD
−0.02
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
AUDrUSD
−0.03
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
CADrUSD
−0.015
0.000
0.010
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
CHFrUSD
−0.02
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
GBPrUSD
−0.015
0.000
0.015
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
JPYrUSD
−0.02
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
NOKrUSD
−0.02
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
NZDrUSD
−0.03
0.00
0.02
−3
−1
0
1
2
3
Theoretical Quantiles



Market return data 1991-1993: CAC (France), DAX (Germany), FTSE (UK)

Even for short periods of 2 or 3 years, univariate distributions are typically heavier tailed than normal and bivariate normal scores plot show non-elliptical shape and joint stronger tails than Gaussian

Plots show semi-correlations ρ_N^-, ρ_N^+ (correlations of points in lower quadrant and in upper quadrant) as well as the normal scores correlation ρ_N and the semi-correlation ζ for a bivariate normal distribution with correlation ρ_N.


Normal Q−Q Plot
CAC log ret 91−93
Theoretical Quantiles
Normal Q−Q Plot
DAX log ret 91−93
Theoretical Quantiles
Normal Q−Q Plot
FTSE log ret 91−93
Theoretical Quantiles
nscores of log ret 1991 − 1993
ζ = 0.32
ρN = 0.56
ρ⁻N = 0.52
ρ⁺N = 0.44
DAX
CAC
nscores of log ret 1991 − 1993
ζ = 0.39
ρN = 0.63
ρ⁻N = 0.51
ρ⁺N = 0.45
FTSE
CAC
nscores of log ret 1991 − 1993
ζ = 0.22
ρN = 0.44
ρ⁻N = 0.35
ρ⁺N = 0.31
FTSE
DAX