Class #8 Wednesday 9 February 2011

What did we cover last time?

- Description & Inference
- Robustness & Resistance
- Median & Quartiles
- Location, Spread and Symmetry (parallels from classical statistics: Mean, Standard Dev., Skewness)
    - Location (Median, Trimean, Trimmed mean)
    - Spread (IQR, MAD, Trimmed variance)
    - Symmetry (Yule-Kendall index)

- Graphical Techniques
    - Stem and Leaf
    - Box plot
    - Histograms
    - Cumulative Frequency Distributions

2.2.3 Reexpression (Ref: Wilkes 3.4)

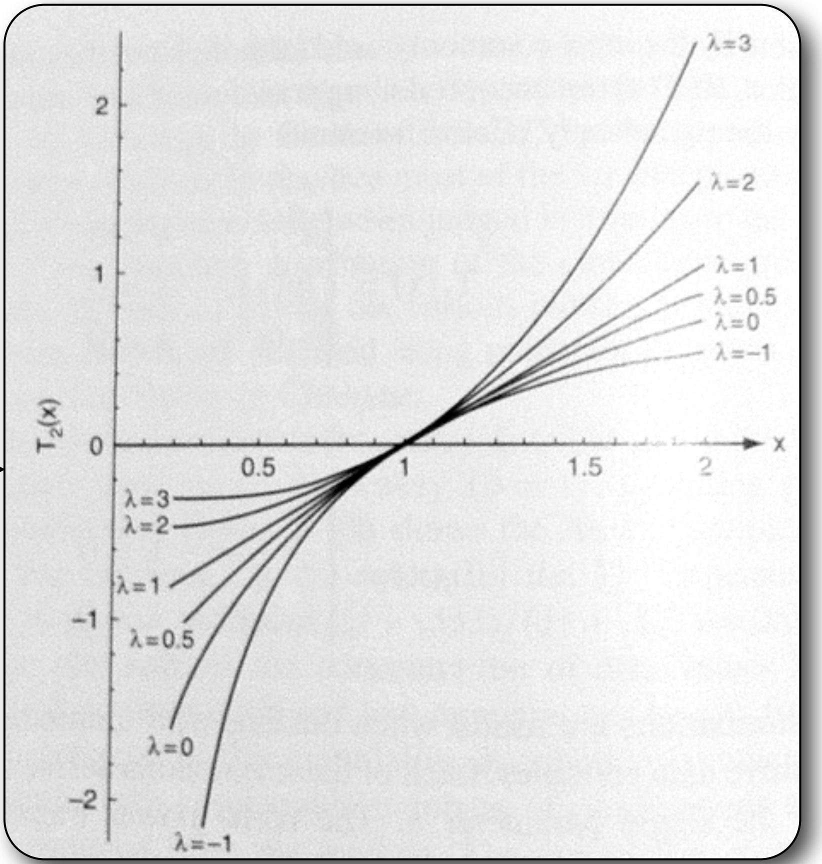
Transform data to:
- Reveal data features
- Adjust the distribution of data
- Variance stabilizing (reduce dependence of one variable on another)

1. Power Transformations
2. Standardization

1. Power Transformations

$$T_1(x) = \left\{ \begin{array}{c} x^\lambda, \lambda < 0 \\ \ln(x), \lambda = 0 \\ -\left(x^\lambda\right), \lambda > 0 \end{array} \right\} \quad 3.18a$$

$$T_2(x) = \left\{ \begin{array}{c} \dfrac{x^\lambda - 1}{\lambda}, \lambda \neq 0 \\ \ln(x), \lambda = 0 \end{array} \right\} \quad 3.18b$$



Power Transformations

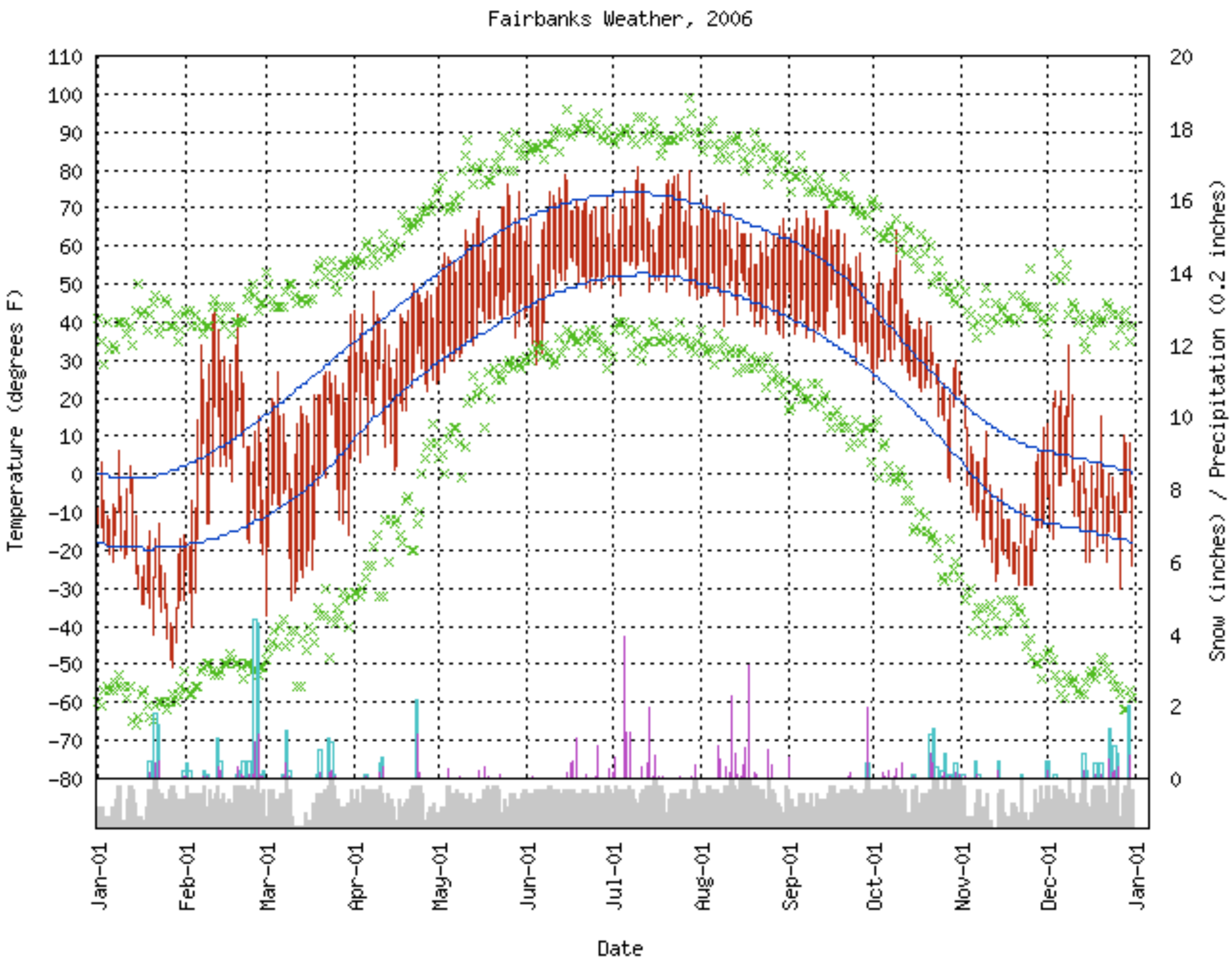
- Use for unimodal data
- Make data more symmetric
- ‘Order statistics’ will have one-to-one correspondence

2. Standardized Anomalies

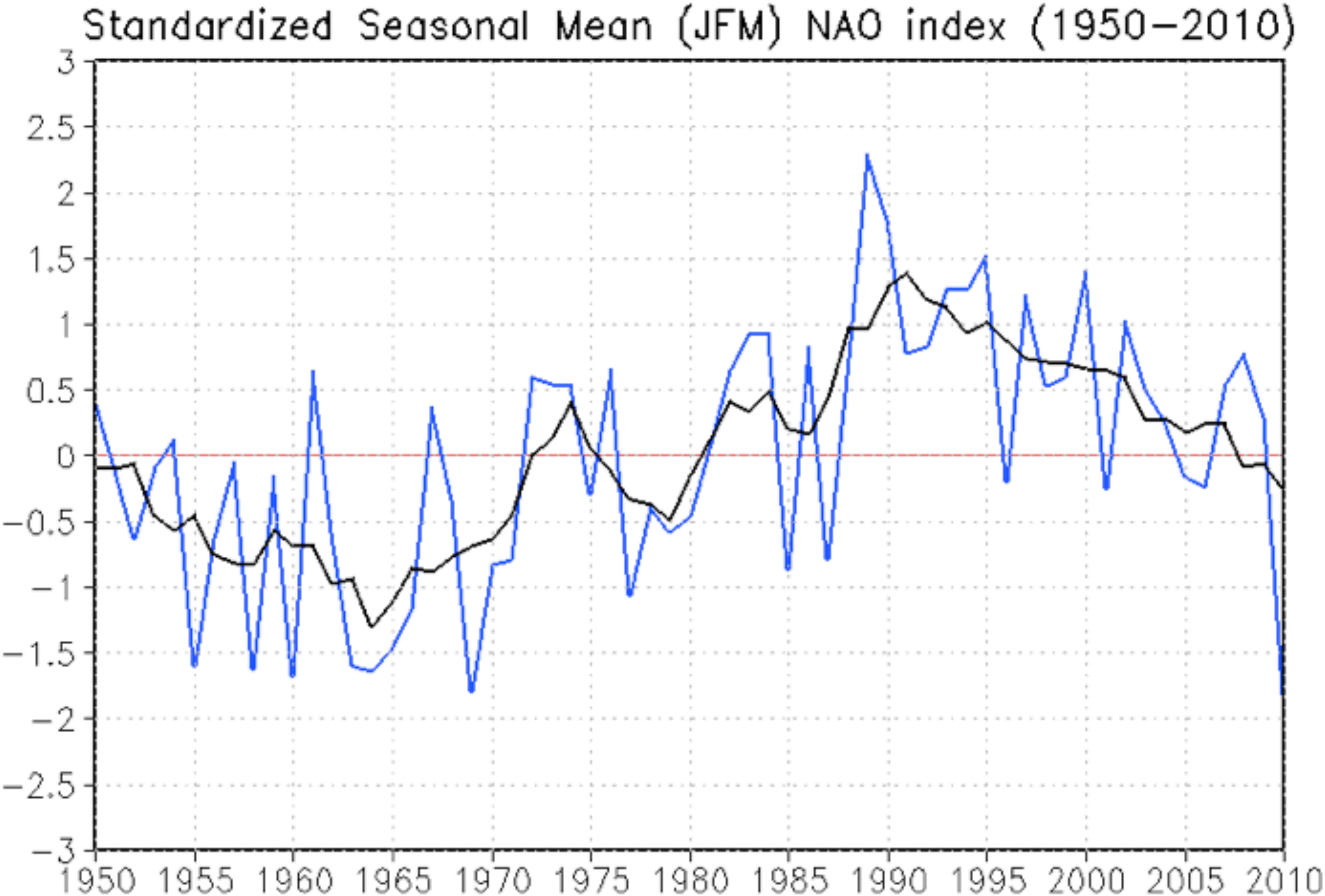
Used to work with two types of data which have very different variability

- Example: Seasonality in data. Temperature variability is larger in winter than summer.
- Example: Perform cluster analysis on Temperature and Precipitation data to determine climate divisions.
- Example: North Atlantic Oscillation
- **Standardize or Normalize of Anomalies** to remove influence of location and spread.

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x} \quad 3.21$$



Fairbanks Weather, 2006
Temperature (degrees F)
110
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
-80
Snow (inches) / Precipitation (0.2 inches)
20
18
16
14
12
10
8
6
4
2
0
Jan-01
Feb-01
Mar-01
Apr-01
May-01
Jun-01
Jul-01
Aug-01
Sep-01
Oct-01
Nov-01
Dec-01
Jan-01
Date



http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/season.JFM.nao.gif

2.2.4 EDA for Paired Data (Ref: Wilkes 3.5)

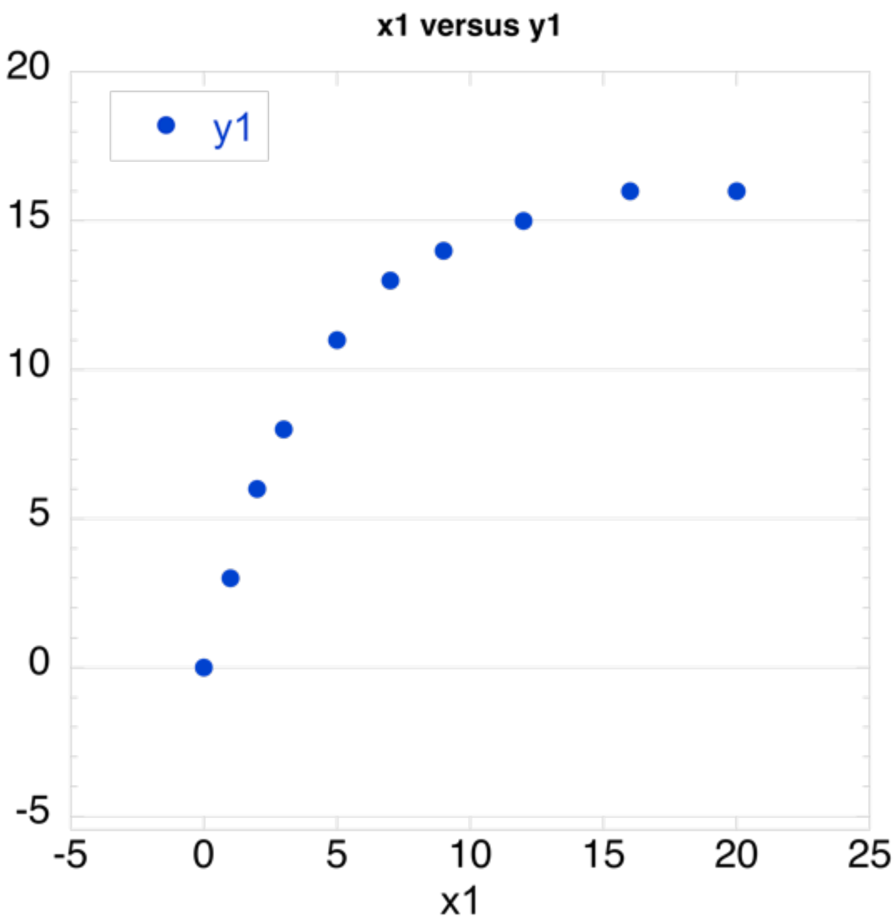
Table3.4

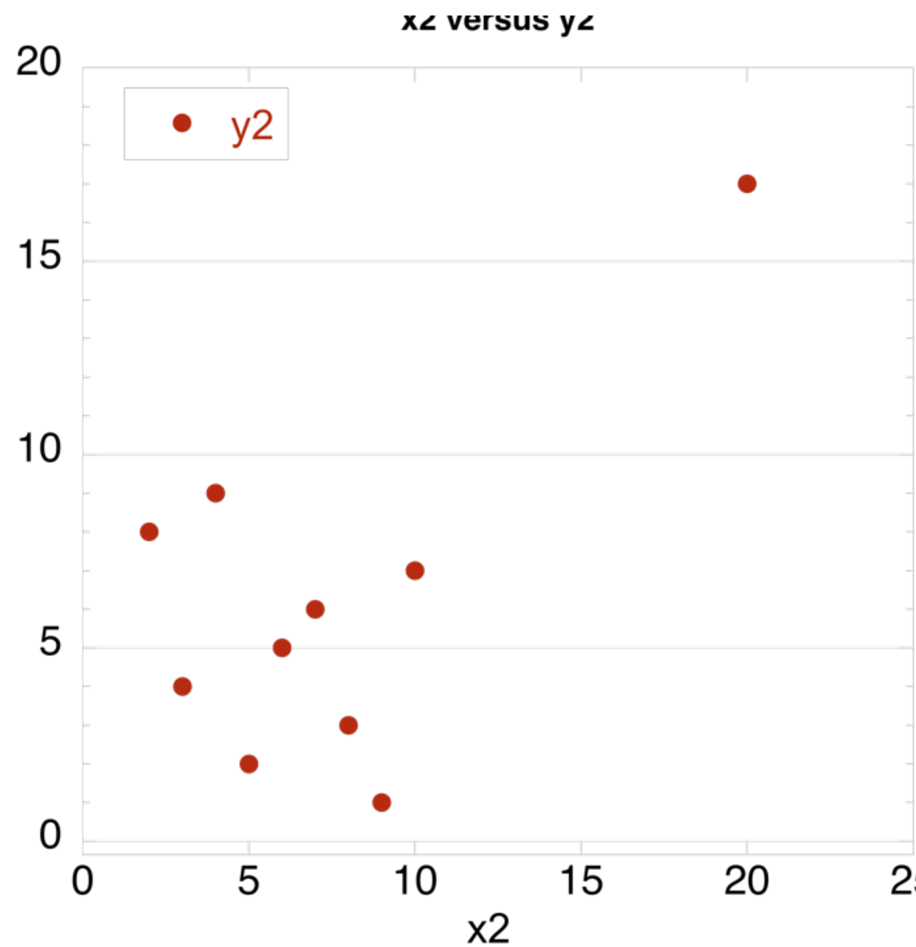
	x1	y1	x2	y2	E
0	0	0	2	8	
1	1	3	3	4	
2	2	6	4	9	
3	3	8	5	2	
4	5	11	6	5	
5	7	13	7	6	
6	9	14	8	3	
7	12	15	9	1	
8	16	16	10	7	
9	20	16	20	17	
10					
11					
12					
13					
14					
15					
16					
17					

Row : 13 Column : 2

Scatter Plots

Pairs of data are plotted against each other.
Useful to see relationship between variables.

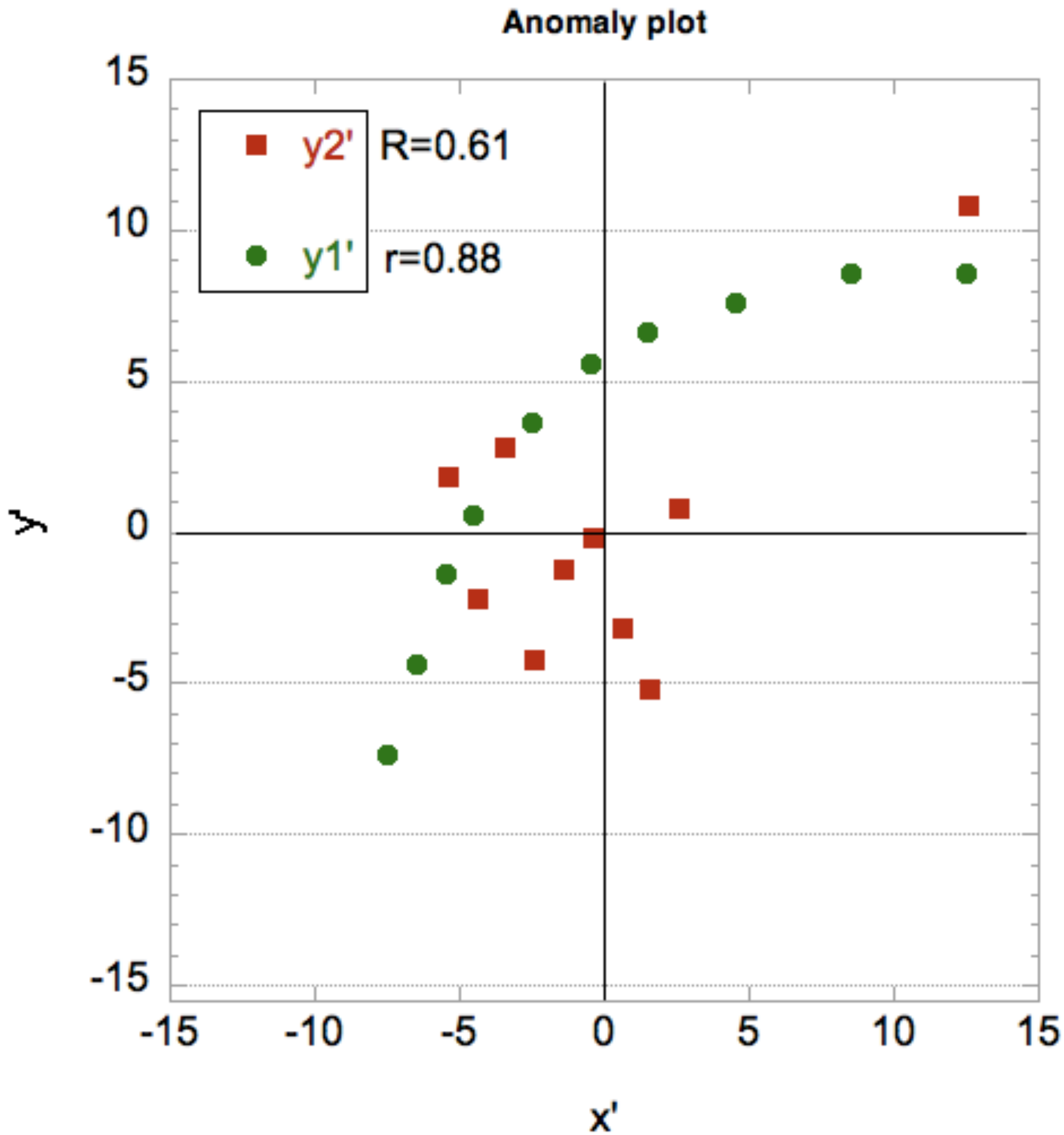




Pearson (Ordinary) Correlation Coefficient

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}\left[(x_i-\overline{x})(y_i-\overline{y})\right]}{\left[\frac{1}{(n-1)}\sum_{i=1}^{n}\left[(x_i-\overline{x})\right]^2\right]^{1/2}\left[\frac{1}{(n-1)}\sum_{i=1}^{n}\left[(y_i-\overline{y})\right]^2\right]^{1/2}} = \frac{\sum_{i=1}^{n}\left[x_i^{'}y_i^{'}\right]}{\left[\sum_{i=1}^{n}\left[x_i^{'}\right]^2\right]\left[\sum_{i=1}^{n}\left[y_i^{'}\right]^2\right]} \quad 3.22$$

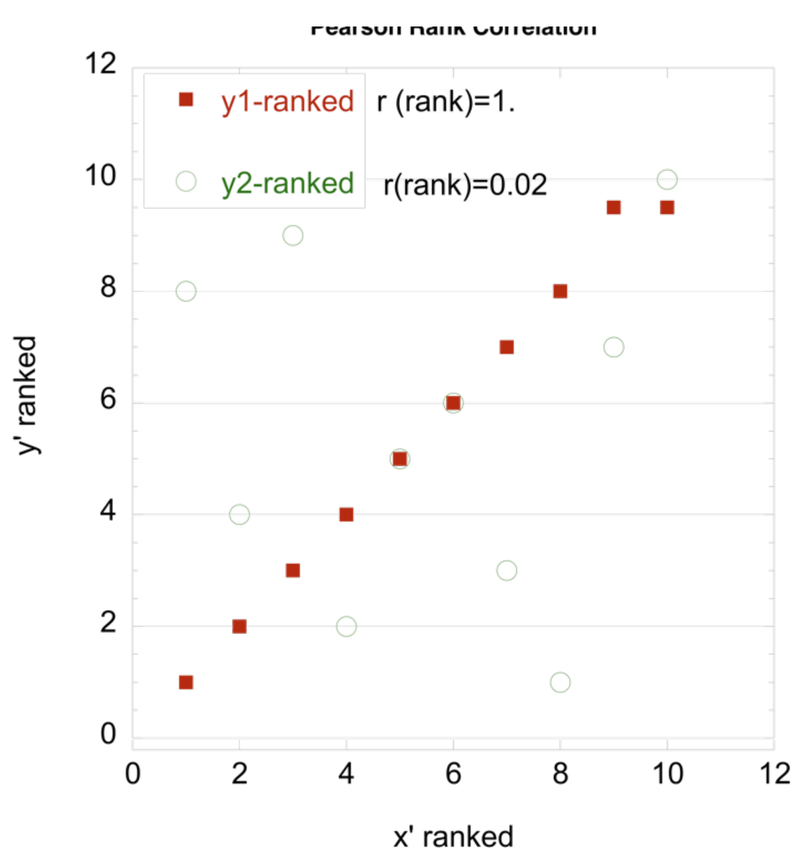
- Not Robust (possibly nonlinear relationships)
- Not Resistant since sensitive to outliers
- Properties (between -1 and 1, Square of coefficient explains proportion of variability, does not give physical causality)


Anomaly plot
y2' R=0.61
y1' r=0.88
15
10
5
0
-5
-10
-15
y
-15
-10
-5
0
5
10
15
x'

Spearman Rank Correlation

- More robust than Pearson's correlation and it is calculated using ranked data.
- Represents the strength of the monotone relationship (not linear relationship).

x1-ranked	y1-ranked	x2-ranked	y2-ranked
1	1	1	8
2	2	2	4
3	3	3	9
4	4	4	2
5	5	5	5
6	6	6	6
7	7	7	3
8	8	8	1
9	9.5	9	7
10	9.5	10	10



Kendall Tau test

- ⁕ More robust and resistant than Pearson's correlation
- ⁕ Calculate by determining concordant and discordant pairs from all possible pairs of x_i and y_i, which is n(n-1)/2
- ⁕ The pairs (3,8) and (7,83) are concordant, latter has both larger numbers. The pairs (3,83) and (7,8) are discordant. Identical pairs contribute half to both.

$$\tau = \frac{N_C - N_D}{n(n-1)/2}$$

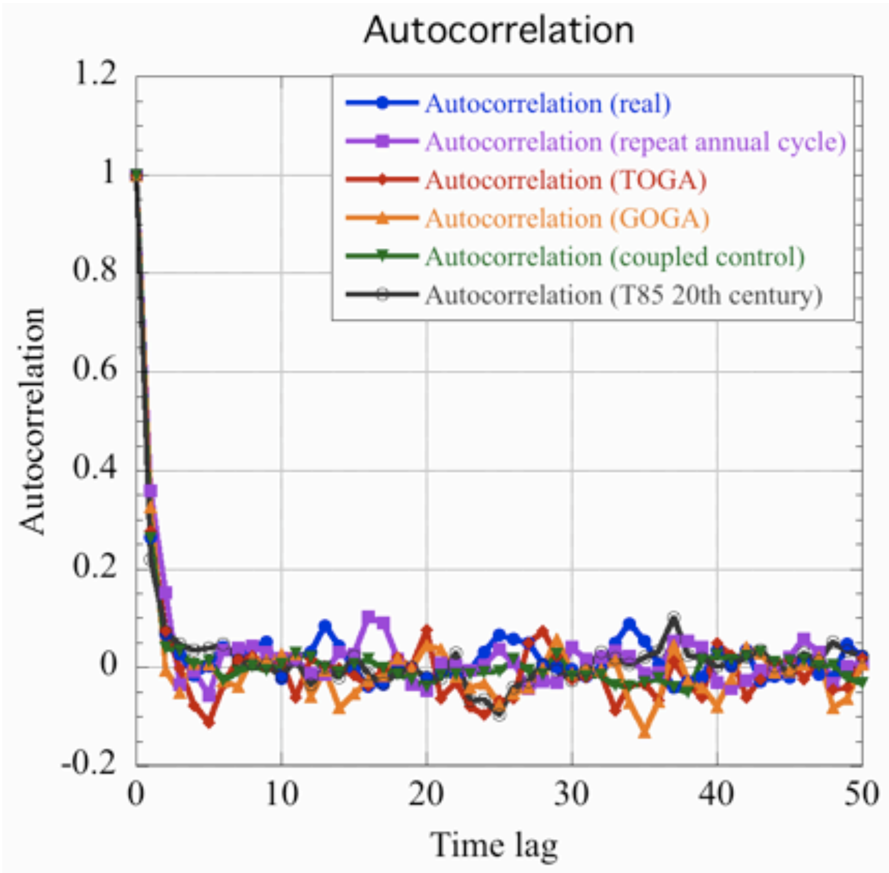
Serial Correlation

- Measure of persistence in a time series! Very important in Meteorology for forecasting.
- Can also be calculated for greater lags

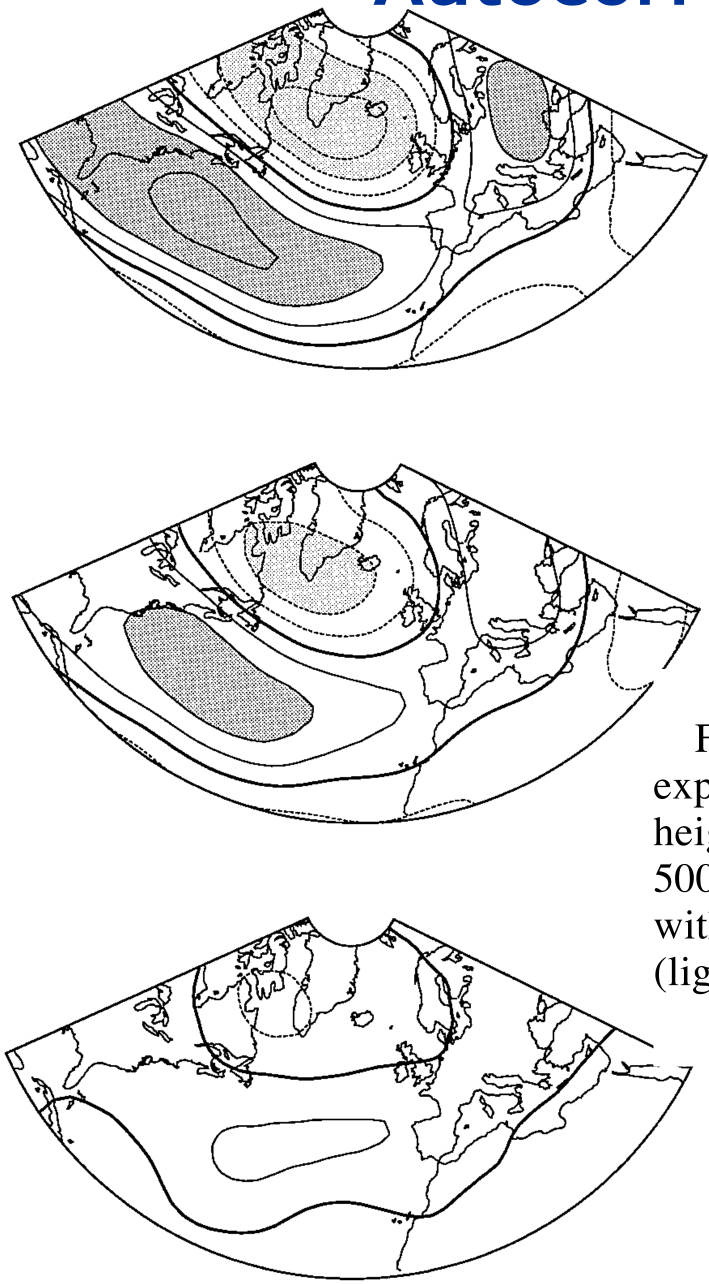
$$r_1 = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x}_-)(x_{i+1} - \bar{x}_+)\right]}{\left(\left[\sum_{i=1}^{n-1}\left[(x_i - \bar{x}_-)\right]^2\right]\left[\sum_{i=2}^{n}\left[x_i - \bar{x}_+\right]^2\right]\right)^{1/2}} \qquad 3.30$$

Autocorrelation Function

❋ The correlations at multiple lags put together constitutes the Autocorrelation function.
❋ Autocovarience is an alternative way to display (construct by multiplying by variance).



Autocorrelation Plots 2-D



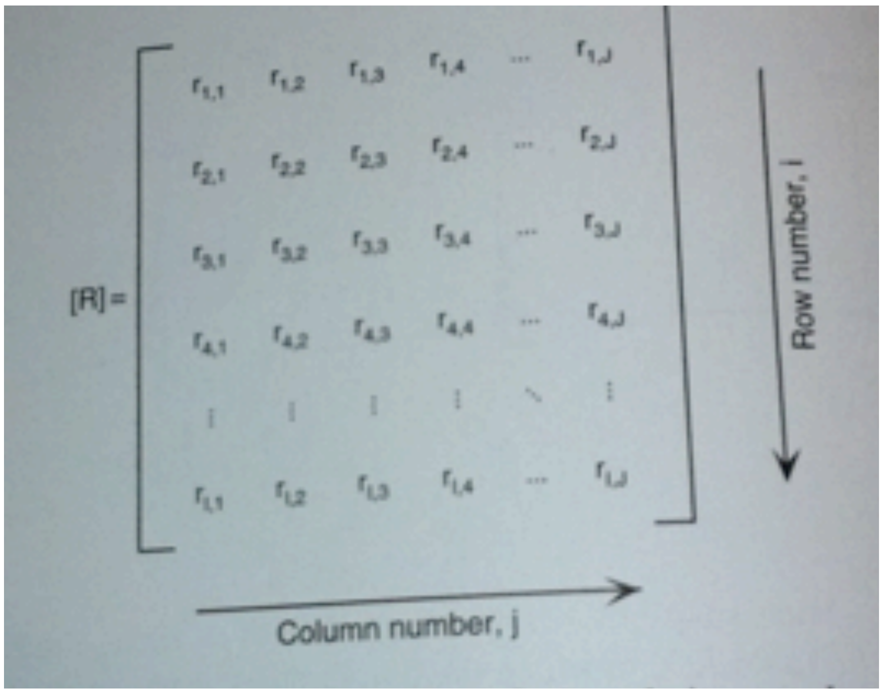
- ❋Useful to try to figure out sequence of events.
- ❋ Ocean lead atmosphere in tropics while atmosphere leads in midlatitudes.

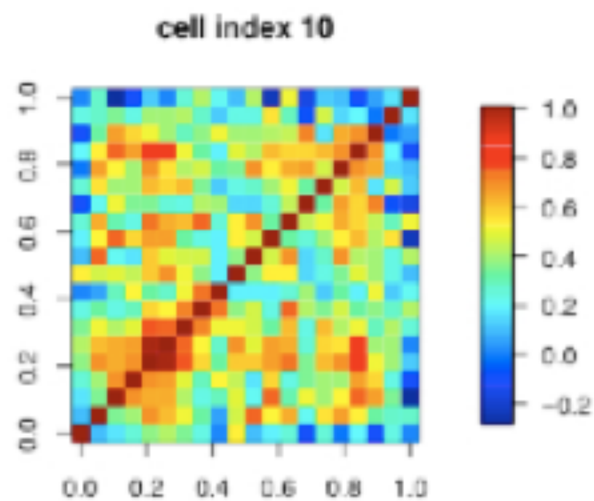
FIG. 3. The 500-mb height correlations with the lag +2 SST SVD expansion coefficient, based on intraseasonal data: (top) 500-mb heights leading SST by 2 weeks, (middle) simultaneous, and (bottom) 500-mb heights lagging SST by 2 weeks. Contour interval = 0.2, with negative contours dashed and the zero contour darkened. Dark (light) shading indicates correlations >0.4 (<−0.4).

Deser and Timlin, 1998

Correlation Matrix

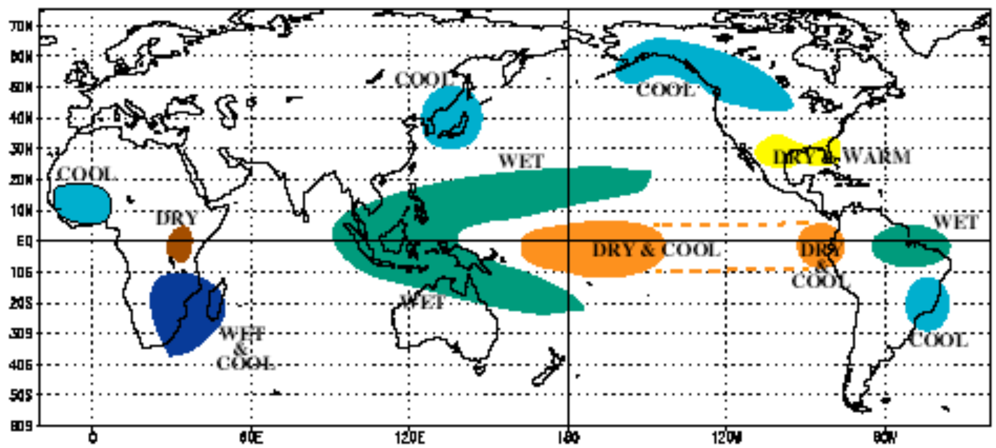
- Contains all possible combinations
- Properties of matrix
- Uses of matrix



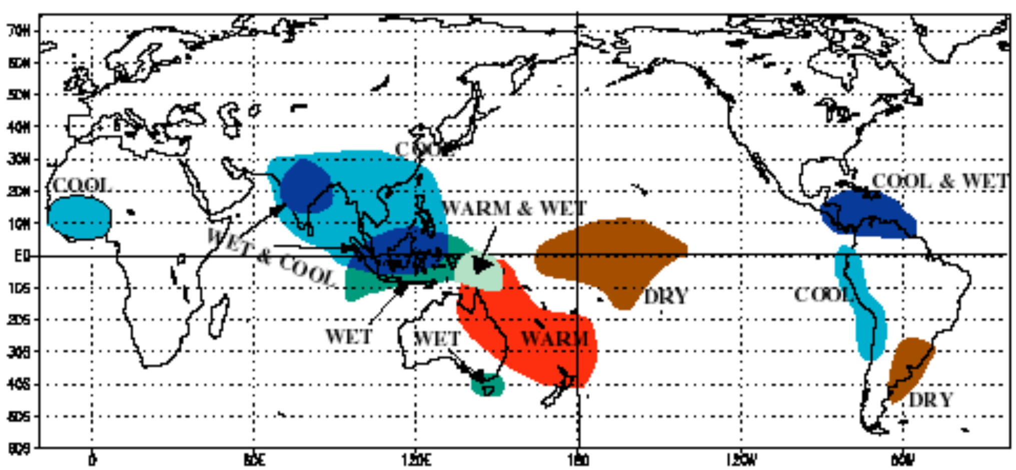


Correlation Map

COLD EPISODE RELATIONSHIPS DECEMBER - FEBRUARY



COLD EPISODE RELATIONSHIPS JUNE - AUGUST



Climate Prediction Center

- ENSO index correlated with temperature and precipitation around the world.