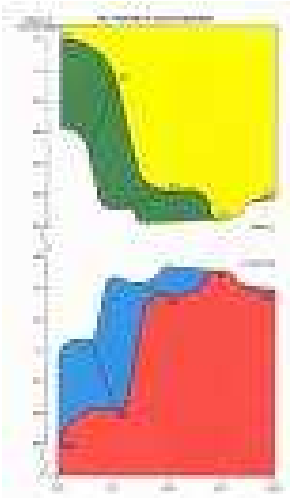
Lecture 14: Correlation and Autocorrelation

Steven Skiena

Department of Computer Science
State University of New York
Stony Brook, NY 11794–4400

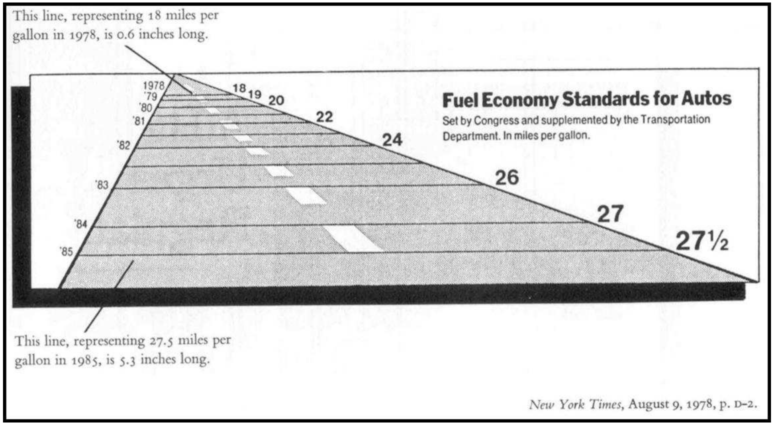
http://www.cs.sunysb.edu/∼skiena

Overuse of Color, Dimensionality, and Plots



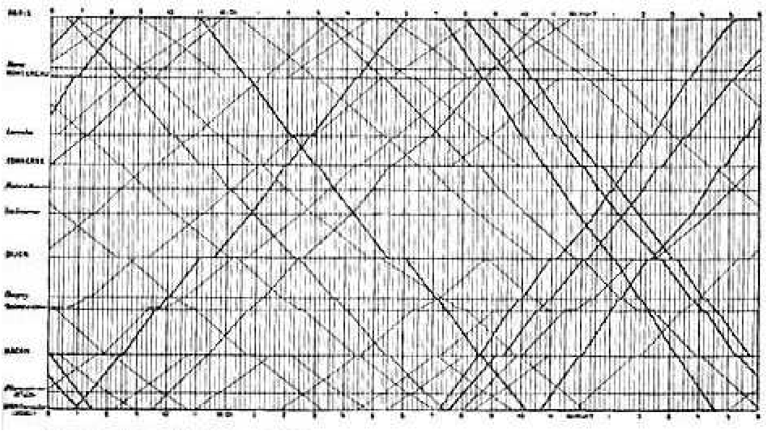
Four colors, three dimensions, and two plots to visualize five data points!

Misleading Scales



Neither the time dimension nor the data magnitudes are represented faithfully.

Railroad Schedules as Time Series



Which trains are fastest? Which trains stop moving? When do you see a passing train out the window?

Variance and Covariance

The *variance* of a random variable X is defined

$$Var(X) = \sigma^2 = \sum(X - \mu_x)^2/N] = E[(X - \mu_x)]$$

Dividing by $N - 1$ provides an unbiased estimate of σ^2 on sampled data, compensating for the difference between the sample mean and the population mean.

The *covariance* of random variables X and Y, is defined

$$\mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

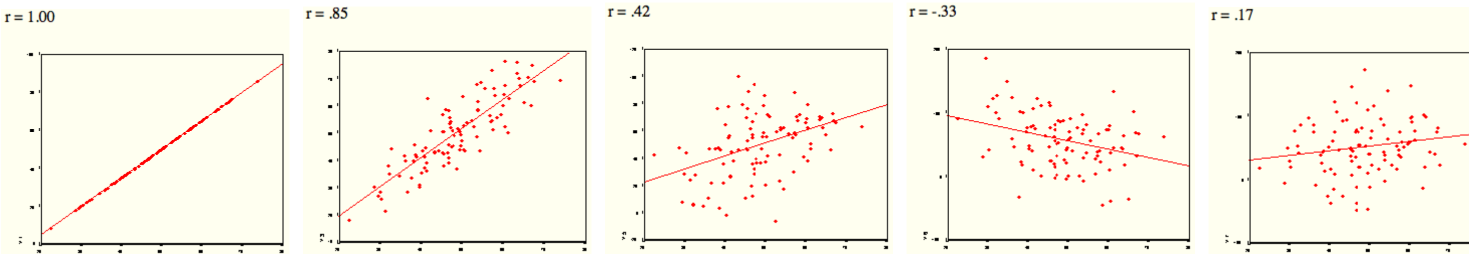
If X and Y are "in sync" the covariance will be high; if they are independent, positive and negative terms should cancel out to give a score around zero.

Correlation Coefficent

Pearson's correlation coefficient of random variables X and Y, is defined

$$\rho_{x,y} = \frac{\text{Cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

It measures the strength of linear dependence between X and Y, and lies between -1 and 1.



Correlation and Causation

Note that correlation does not imply causation – the conference of the Super Bowl winner has had amazing success predicting the fate of the stock market that year.

If you investigate the correlation of many pairs of variables (such as in data mining), some are destined to have high correlation by chance.

The meaningfulness of the correlation can be evaluated by considering (1) the number of pairs tested, (2) the number of points in each time series, (3) the sniff test of whether there *should* be a connection, (4) statistical tests.

Significance of Correlation

The squared correlation coefficient (ρ^2) is the proportion of variance in Y that can be accounted for by knowing X, and is a good way to evaluate the strength of a relationship.

If the correlation between height and weight is approximately $\rho = 0.70$, then $\rho^2 = 49\%$ of one's weight is directly accounted for one's height and vice versa.

Thus high correlations are needed for a single factor to have to significant impact on prediction.

Autocorrelation

The *lag-l autocorrelation* ρ_l is the correlation coefficient of r_t and r_{t-l}.

A linear time-series is characterized by its *sample autocorrelation function* $a_l = \rho_l$ for $0 \leq l \leq n$.

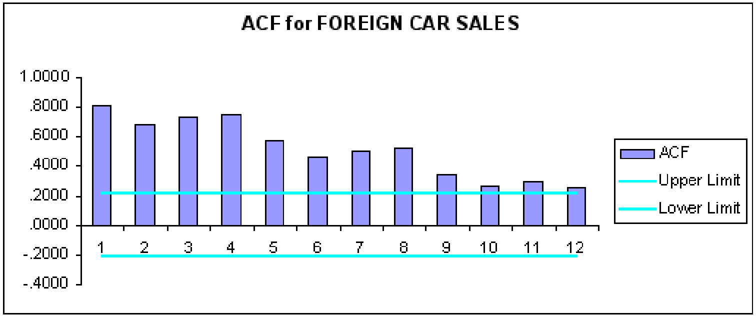
The naive algorithm for computing the autocorrelation function takes $O(n^2)$ time for a series of n terms:

$$A = \forall_{l=0}^{n} \sum_{i=0}^{n-l} r_t r_{t+l}$$

However, fast *convolution* algorithms can compute it in $O(n \log n)$ time.

Interpreting Autocorrelation Functions

A *correlogram* is a plot representing the autocorrelation function.



What does the correlogram of stock *prices* look like? What about stock *returns*?

White Noise

Since stock returns are presumably random, we expect *all* non-trivial lags to show a correlation of around zero.

White noise is a time series consisting of independently distributed, uncorrelated observations which have constant variance.

Thus the autocorrelation function for white noise is $a_l = 0$ for $l > 0$ and $a_0 = 1$.

Gaussian white noise has the additional property that it is normally distributed.

Identifying White Noise

The white noise *residual* terms can be calculated $e_t = r_t - f_t$ for observations r and model f.

Our modeling job is complete when the residuals/errors are Gaussian white noise.

For white noise, distribution of the sample autocorrelation function at lag k is approximately normal with mean 0 and variance $\sigma^2 = 1/n$.

Testing to ensure no residual autocorrelations of magnitude $> 1/\sqrt{n}$ is good way to test if our model is adequate.

Autocorrelation of Sales Data

What is the ACF of the daily gross sales for Walmart?

Today's sales are a good predictor for yesterday's, so we expect high autocorrelations for short lags.

However, there are also day-of-week effects (Sunday is a bigger sales day than Monday) and day-of-year effects (Christmas season is bigger than mid-summer). These show up as lags of 7 and about 365, respectively.

Stationarity

The mathematical tools we apply to the analysis of time series data rest on certain assumptions about the nature of the time series.

A time series $\{r_t\}$ is said to be *weakly stationary* if (1) the mean of r_t, $E(r_t)$, is a constant and (2) the *covariance* $\mathrm{Cov}(r_t, r_{t-l}) = \gamma_l$ depends only upon l.

In a weakly stationary series, the data values fluctuate with constant variation around a constant level.

The financial literature typically assumes that asset returns are weakly stationary, as can be tested empirically.