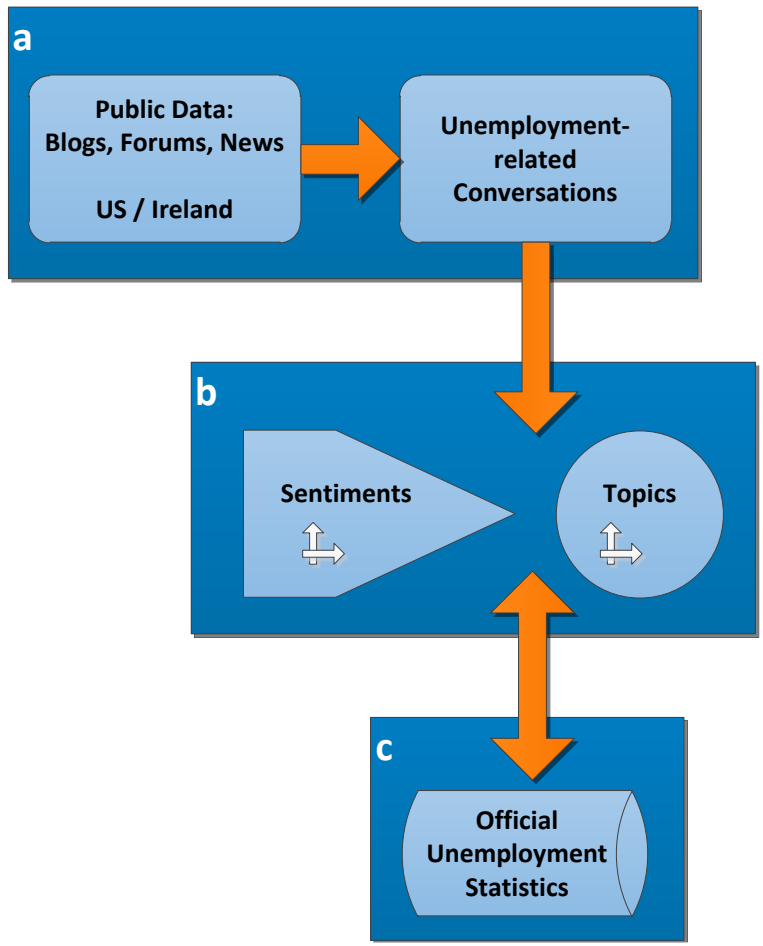
Ch 11 Part 1: Cross-correlation and dynamic regression models

- Motivating example (a UN Global Pulse project):
  "Using social media and online conversations to add depth to unemployment statistics".
  - Source: http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics
- Research questions:
  - Can online conversations provide an early indicator of impending job losses?
  - Can these conversations help policy makers enrich their understanding of the type and sequence of coping strategies employed by individuals?

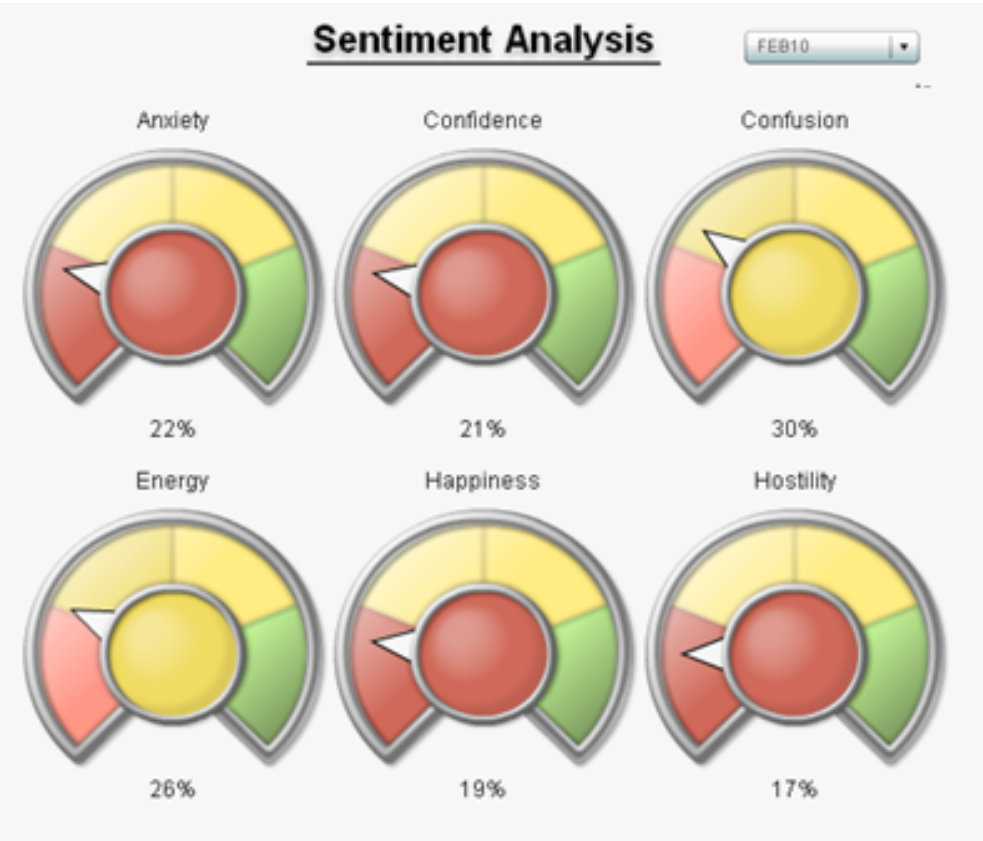
Using social media and online conversations to add depth to unemployment statistics

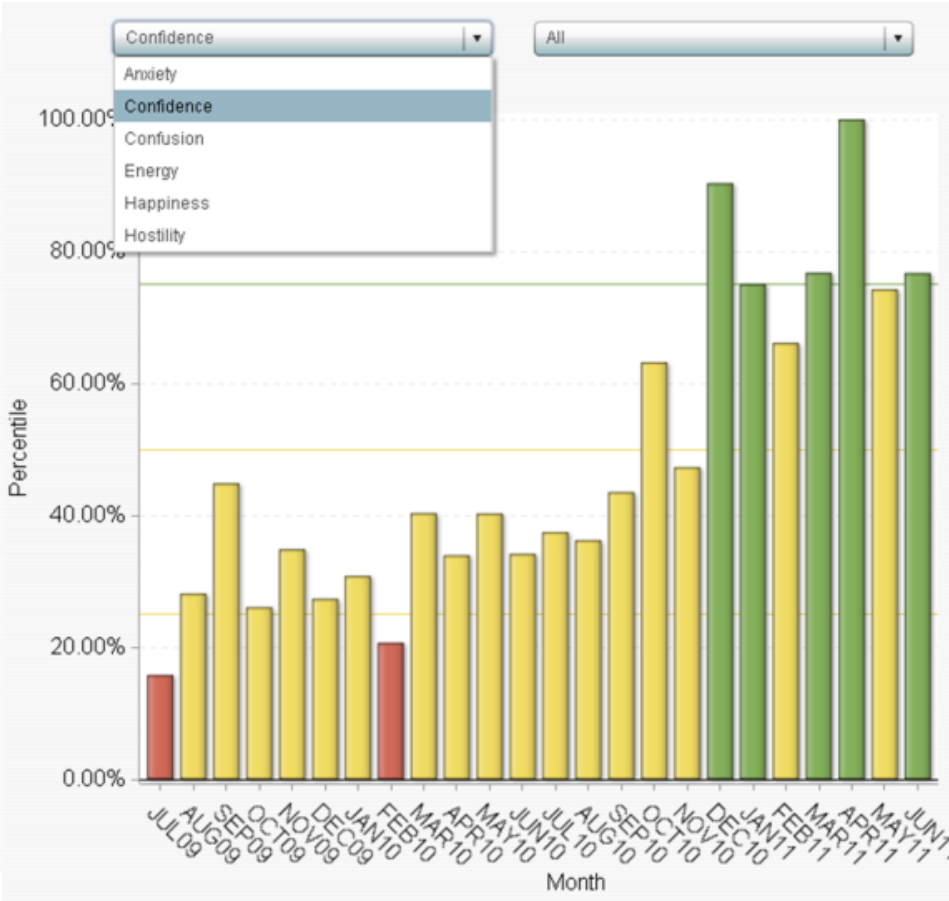
- Context: USA and Ireland, June 2009 - June 2011
- Data:
  - Unemployment rate (high for both countries during this period).
  - Job-related conversations in social media (from facebook, twitter, blogs, forums, ...)
- Social media data were summarized into quantitative values relating to
  - mood/sentiments
  - amount of chatter on specific topics (public transportation, having to downgrade/sell house, spending on travel, entertainment, ...)



Using social media and online conversations to add depth to unemployment statistics

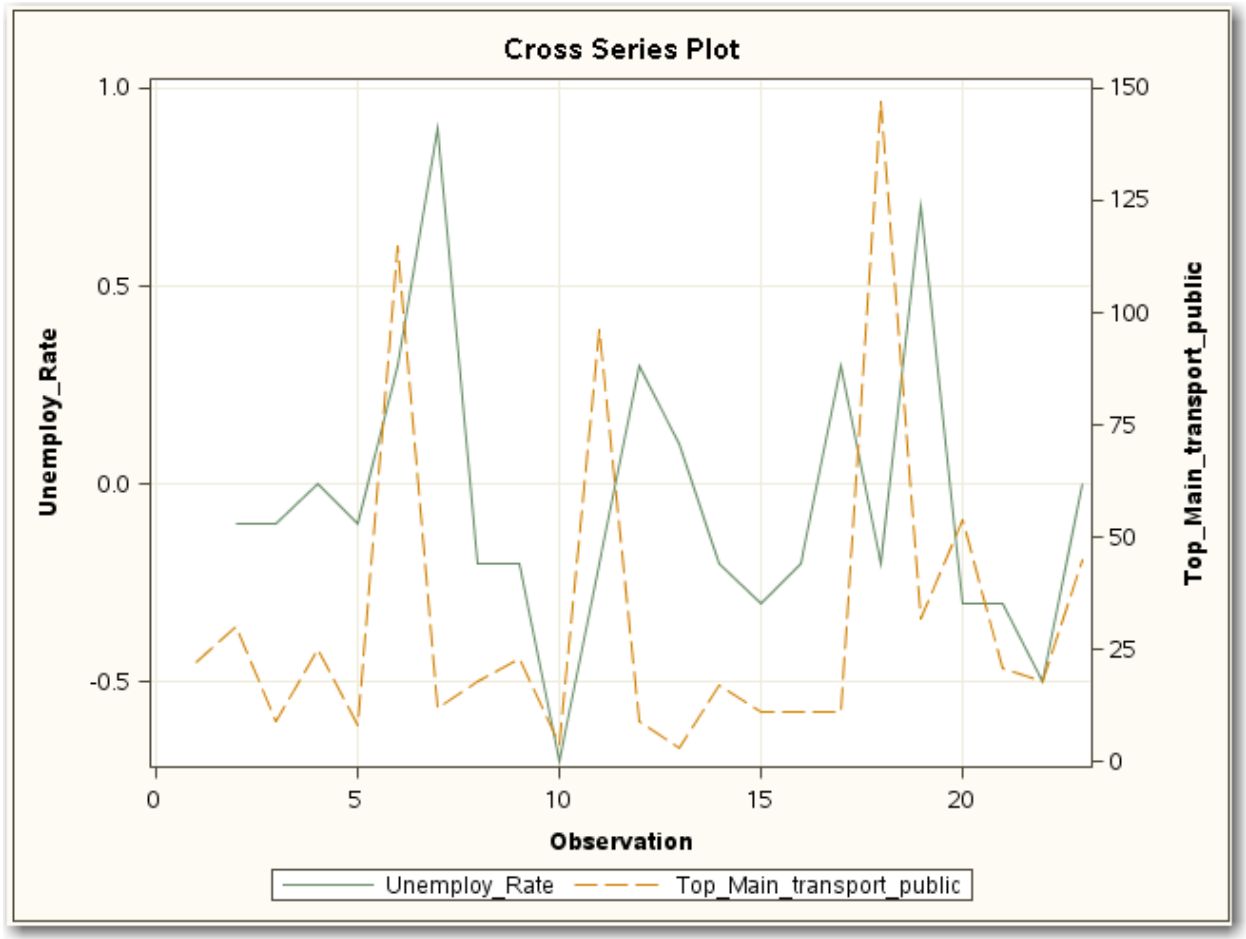
Information on sentiments





Using social media and online conversations to add depth to unemployment statistics

Information on employment and how often the topic "transportion" was mentioned



(They did not use R in this project...)

And??

- Can online conversations provide an early indicator of impending job losses and help policy makers enrich their understanding of the type and sequence of coping strategies employed by individuals?

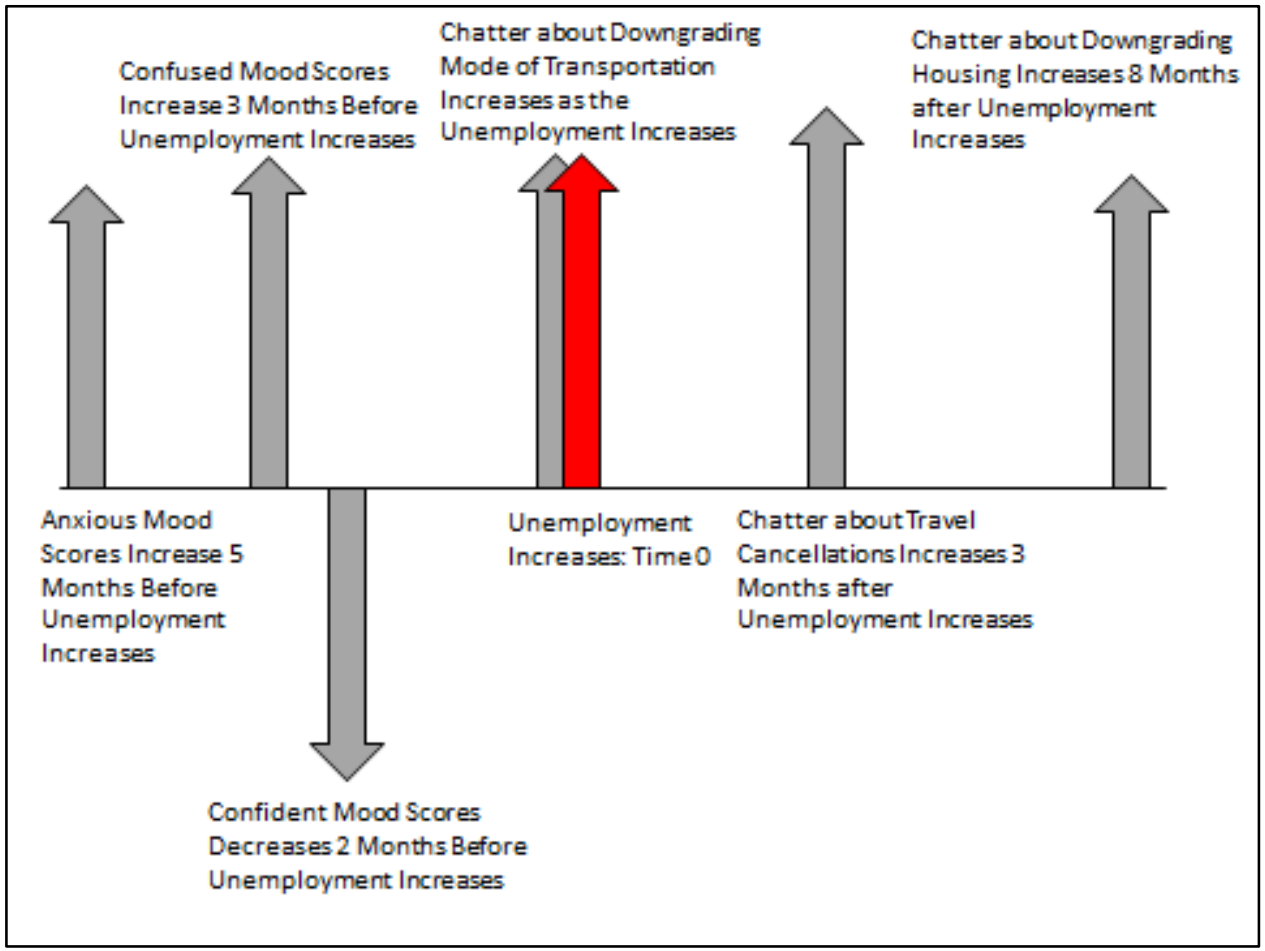


Figure 9: Ireland Chatter

Cross-correlation and dynamic regression models

- The motivating example boils down to the following type of analysis:
  - Suppose we have a time series of interest $Y_1, Y_2, \ldots, Y_t$ (e.g. changes in the unemployment rate), here denoted by Y, and we want to explore whether/how another time series $X_1, X_2, \ldots, X_t$, here denoted by X (e.g. depression measured in employment-related social media output), relates to Y.
  - E.g. does depression increase before or after unemployment increases? Or is there no relation at all?
- Other examples: price and sales of an item, weather/climate and dengue outbreaks, ...
- Topics to discuss:
  - Summarizing the correlation between X and Y using the (sample) cross-correlation function
  - Modeling Y using X while accounting for autocorrelation in Y: dynamic regression models
- Material: Ch 11.3 + 11.4

Summarizing the correlation between X and Y

- We can summarize the correlation between time series X and Y, for any pair of times t and s, using the **cross-correlation function**, CCF, $\rho_{t,s}(X, Y)$:

$$\rho_{t,s}(X, Y) = Corr(X_t, Y_s) = \frac{Cov(X_t, Y_s)}{\sqrt{Var(X_t)Var(Y_s)}}.$$

- This function simplifies a bit when X and Y are jointly (weakly) stationary, which holds true if
  - both processes are (weakly) stationary *(what did that mean again?)*,
  - AND $\rho_{t,s}(X, Y)$ depends on $(t - s)$ only.
- For jointly stationary X and Y, we define the cross-correlation function $\rho_k(X, Y) = \rho_{t+k,t}(X, Y)$:

$$\rho_k(X, Y) = \rho_{t+k,t}(X, Y) = Corr(X_{t+k}, Y_t) = Corr(X_t, Y_{t-k}).$$

Cross-correlation function: Example

- $Y_t = \beta_0 + \beta_1 X_{t-m} + e_t$, where the X_t's are white noise with $Var(X_t) = \sigma_X^2$, independent of e_t.
  - $m > 0$ is referred to as X leading Y.

$$\rho_{-m}(X,Y) = \frac{Cov(X_{t-m}, Y_t)}{\sqrt{Var(X_t)Var(Y_t)}},$$

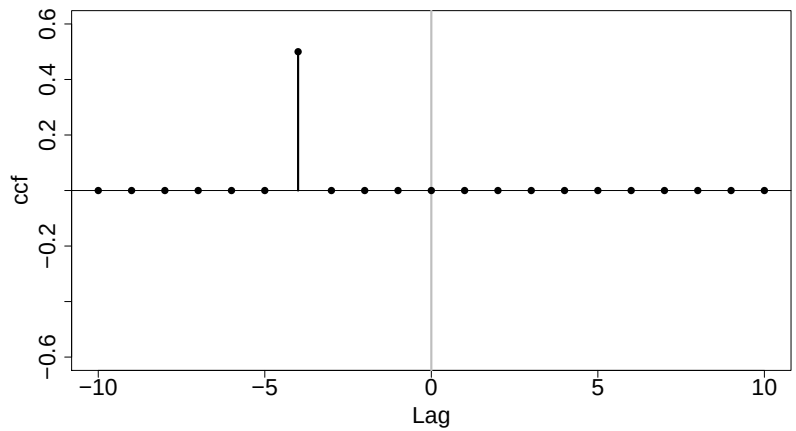
$$= \frac{Cov(X_{t-m}, \beta_0 + \beta_1 X_{t-m} + e_t)}{\sqrt{\sigma_x^2}\sqrt{\beta_1^2\sigma_X^2 + \sigma_e^2}},$$

$$= \frac{\beta_1\sigma_X^2}{\sigma_x\sqrt{\beta_1^2\sigma_X^2 + \sigma_e^2}},$$

$$= \frac{\beta_1\sigma_X}{\sqrt{\beta_1^2\sigma_X^2 + \sigma_e^2}},$$

and $\rho_k(X,Y) = 0$ for $k \neq -m$.

Example: $m = 4$



Estimating the cross-correlation function

- The sample ccf, based on pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, is given by:

$$r_k(X, Y) = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2}\sqrt{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}}$$

  - Compare to $r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$.
- Large sample distribution:
  - If X and Y are white noise processes, then for k with $\rho_k(X, Y) = 0$, approximately $r_k \sim N(0, 1/n)$.
  - So we can use critical values $1.96 \pm 1/\sqrt{n}$ (again) for finding significant CCFs.
  - Q: How many "false positives" with
    - $|r_k(X, Y)| > 1.96\sqrt{1/n}$ while $\rho_k(X, Y) = 0$

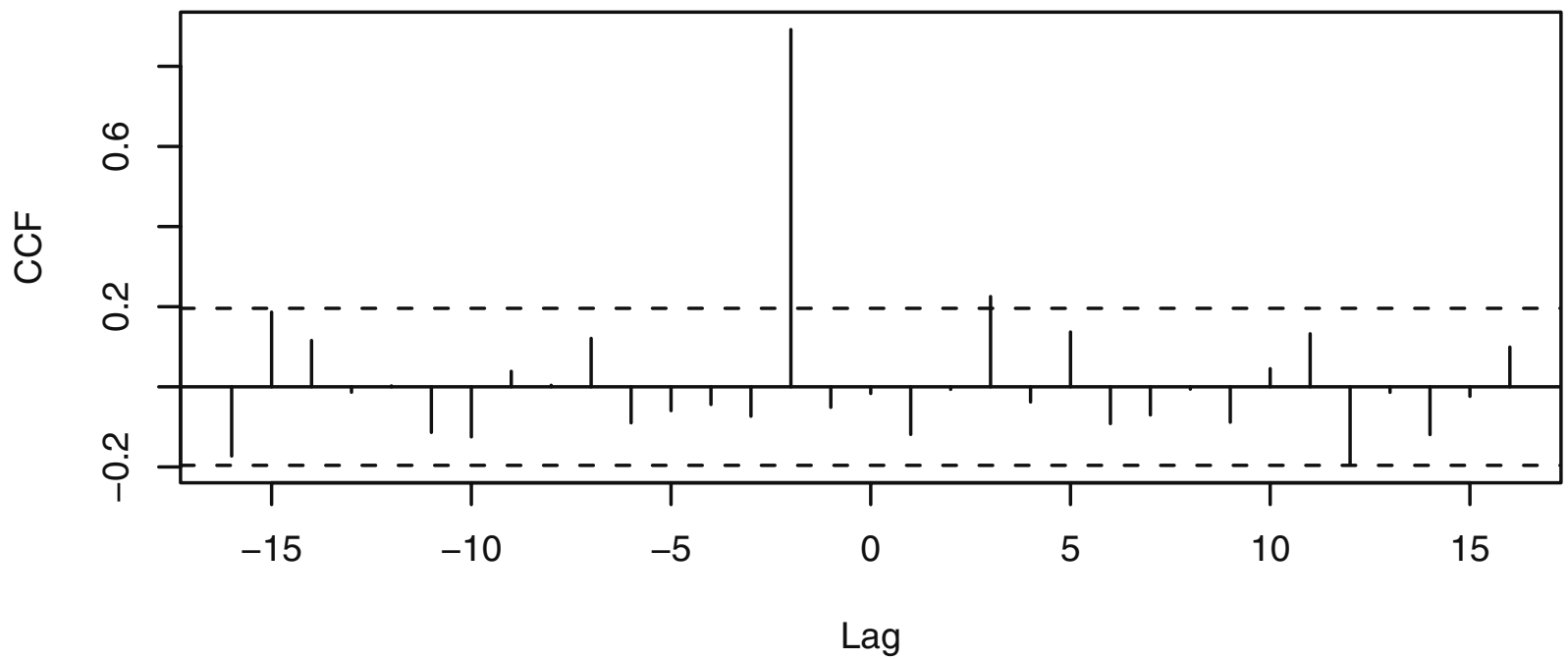
    do we expect if we calculate say 20 sample CCFs?

Simple example

- $Y_t = \beta_0 + \beta_1 X_{t-m} + e_t$, for $t = 1, 2, \ldots, 100$ where $X_t \sim N(0, 1)$ and $e_t \sim N(0, 0.5^2)$ (all independent), $m = 2, \beta_0 = 0, \beta_1 = 1$.
- Then the critical value is ± 0.196 and

$$\rho_k(X, Y) = \begin{cases} \frac{\beta_1 \sigma_X}{\sqrt{\beta_1^2 \sigma_X^2 + \sigma_e^2}} = 1/\sqrt{(1 + 0.5^2)} \approx 0.9, & \text{if } k = -m, \\ 0 & \text{otherwise.} \end{cases}$$

- Example



More realistic settings for time series data

- So far, we discussed the relation

$$Y_t = \beta_0 + \beta_1 X_{t-m} + e_t,$$

  where the X_t's and e_t's are both white noise series to explore correlation between Y_t and X_{t-m}.
- The specification for X_t and e_t is not necessarily very realistic when working with time series data.
  - For example, in the research project, if we would be interested in the correlation between unemployment (Y_t) and measuring "chatter about travel cancelation" (X_t), then most likely, both variables are some time series process instead of white noise!
- Let's investigate the CCF for the following model:

$$Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t,$$

  where X_t and Z_t are independent from each other but not necessarily white noise.
- It turns out that things get a bit more complicated...

The CCF $\rho_k(X, Y)$ when X_t's are autocorrelated

- Do we still find that $\rho_k(X, Y) = 0$ for $k \neq -m$ in the model

$$Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t,$$

  where X_t and Z_t are independent from each other, but where X_t is not necessarily white noise?

- Let's check:

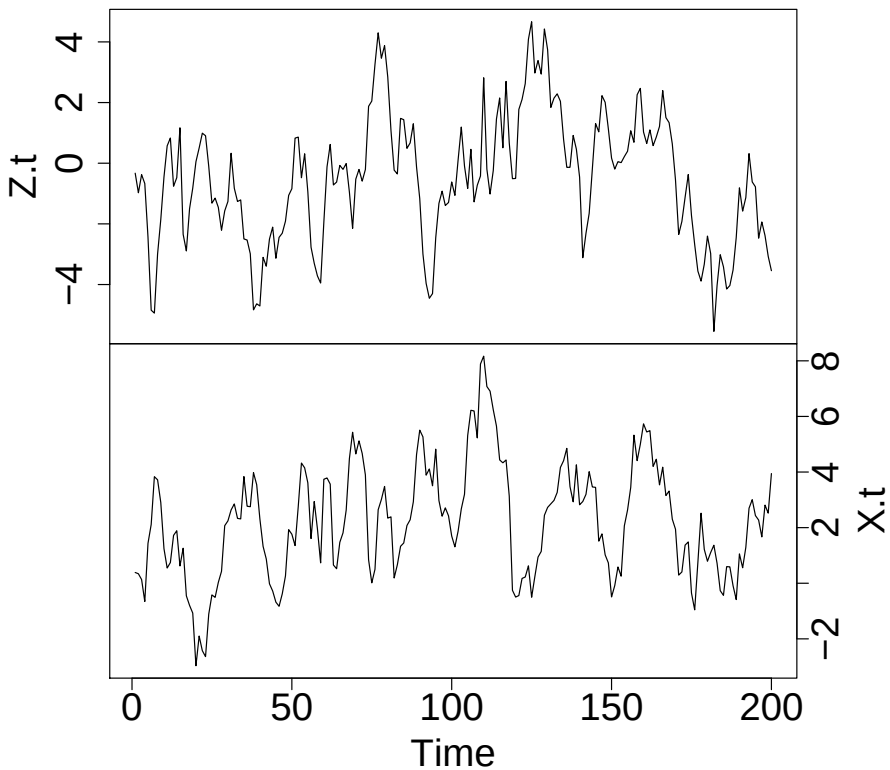
$$\rho_k(X, Y) = \frac{Cov(X_{t+k}, Y_t)}{\sqrt{Var(X_t)Var(Y_t)}} = \frac{Cov(X_{t+k}, \beta_0 + \beta_1 X_{t-m} + Z_t)}{\sigma_X \sigma_Y}$$

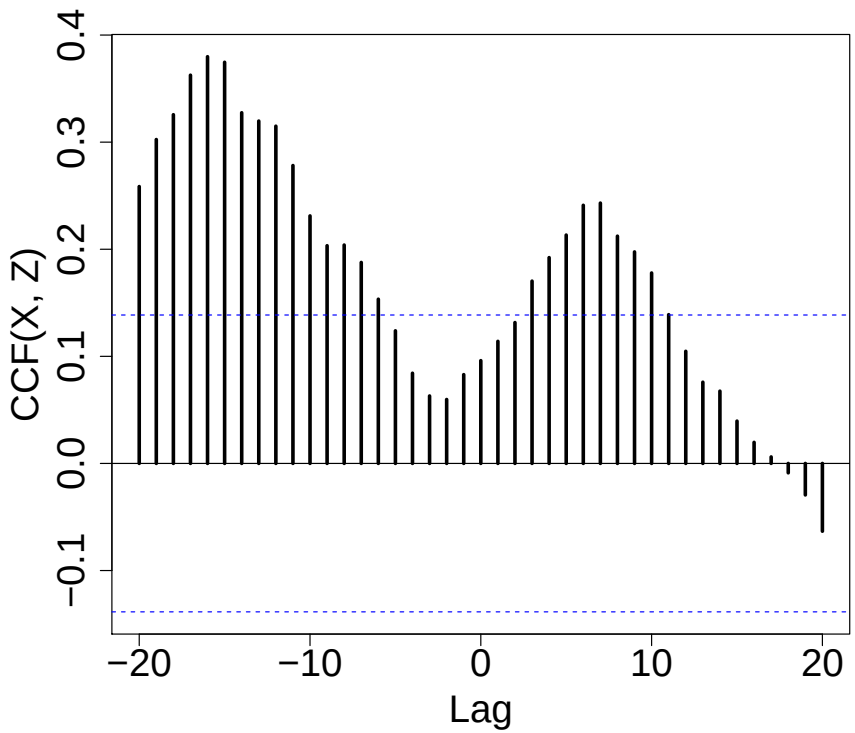
  $\rho_k(X, Y)$ can be non-zero for $k \neq -m$ if the X_t's are autocorrelated!

- So how to figure out which lag m is important?

Another issue when trying to figure out what lags to focus on...

Simulation example for the sample CCF when $Z_t \sim AR(1)$, $X_t \sim AR(1)$ (independent of Z_t) and $n = 200$.





- What's going on?

What's going on?

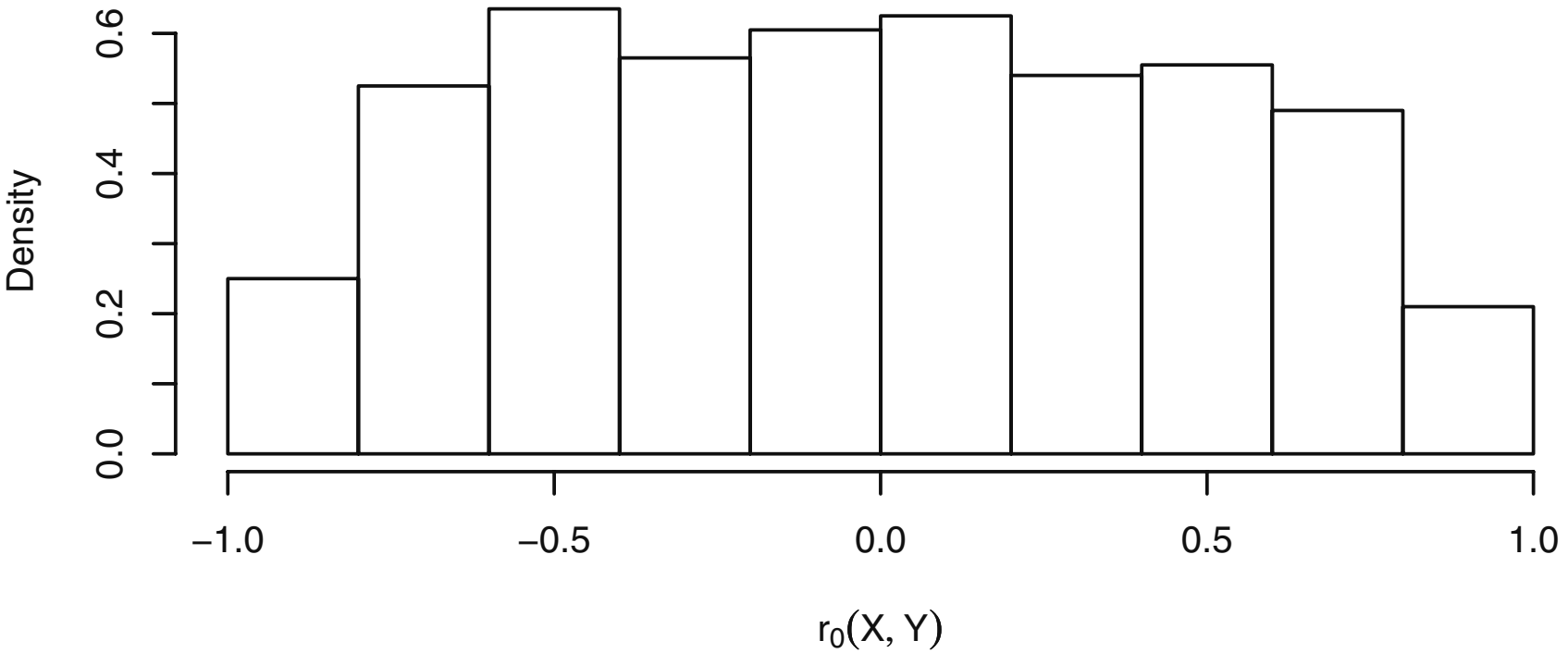
- In our last example: $X_t \sim AR(1)$ and $Z_t \sim AR(1)$ were independent, why does the sample CCF $r_k(X, Z)$ show "large" values?
- It turns out that we can get large values for $r_k(X, Y)$ for two series X and Y by chance if the series are autocorrelated:
  - The approximate sampling variance of $r_k(X, Y)$ is $1/n$ for white noise processes X and Y for k with $\rho_k(X, Y) = 0$.
  - But for stationary processes X and Y, the approximate sampling variance of $r_k(X, Y)$ for k with $\rho_k(X, Y) = 0$ is given by:

$$1/n(1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)),$$

    where $\rho_k(X)$ is ACF for X_t and $\rho_k(Y)$ the ACF for Y_t.
  - In our last example:
    - $n = 200$ thus $\sqrt{1/n} \approx 0.07$ but the approximate SE for $r_k(X, Y)$ is much larger (0.3).
    - The critical bounds $\pm 1.96\sqrt{1/n}$ are no longer valid!
- Things get even more problematic for non-stationary Y_t and X_t: distribution of $r_k(X, Y)$ may no longer be approximately normal!

Simulation of sampling distribution for $r_0(X, Y)$ if both are IMA(1,1) processes (independent)

Exhibit 11.13 Histogram of 1000 Sample Lag Zero Cross-Correlations of Two Independent IMA(1,1) Processes Each of Size 500



Summary

- Suppose $Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t$, with X_t independent of Z_t but where Z_t and X_t are time series processes with $\rho_k(X)$ and $\rho_k(Z)$ non-zero for $k \neq 0$.
- Issues that complicate an analysis of cross-correlation:
  - The CCF is non-zero for various lags, not just $k = -m$.
  - If $\beta_1 = 0$ (no relation at all between X and Y), then the approximate variance of $r_k(X, Y)$ is given by

$$1/n(1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)),$$

    instead of $1/n$, which may be substantial.
  - For non-stationary series, distribution of r_k may no longer be approximately normal.
- Conclusion: the sample CCF $r_k(X, Y)$ is not useful for identifying true correlation between two processes X_t and Y_t if autocorrelation in X_t and Z_t is present.
  - We will find spurious (nonsense) correlation.

Huh... how were these results obtained?

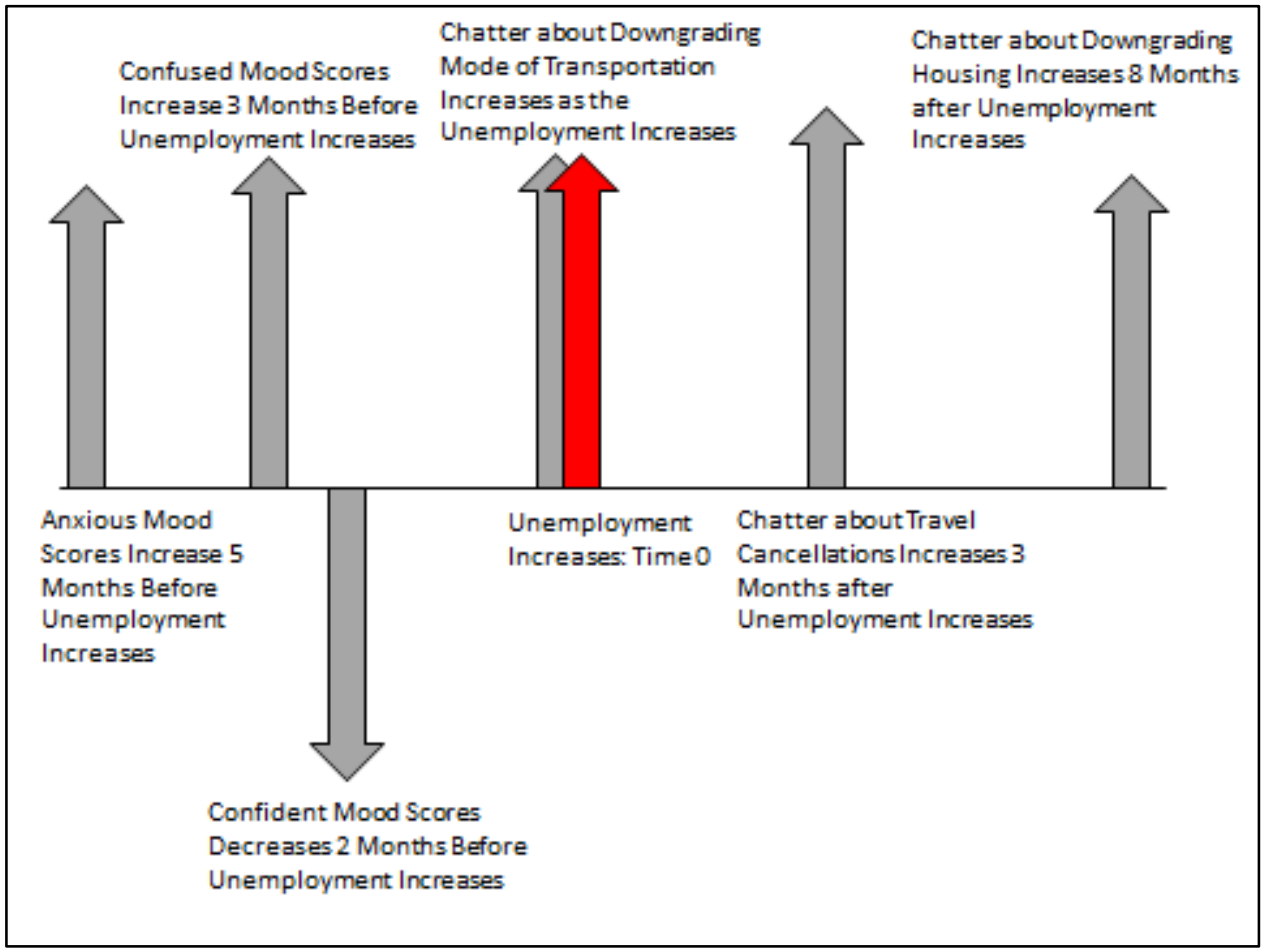
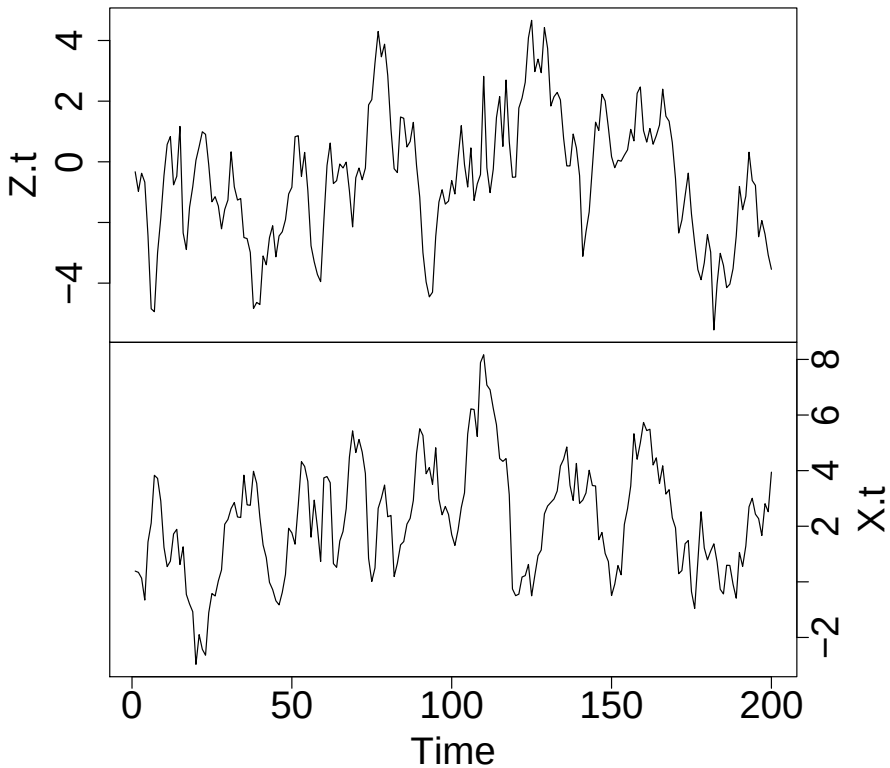
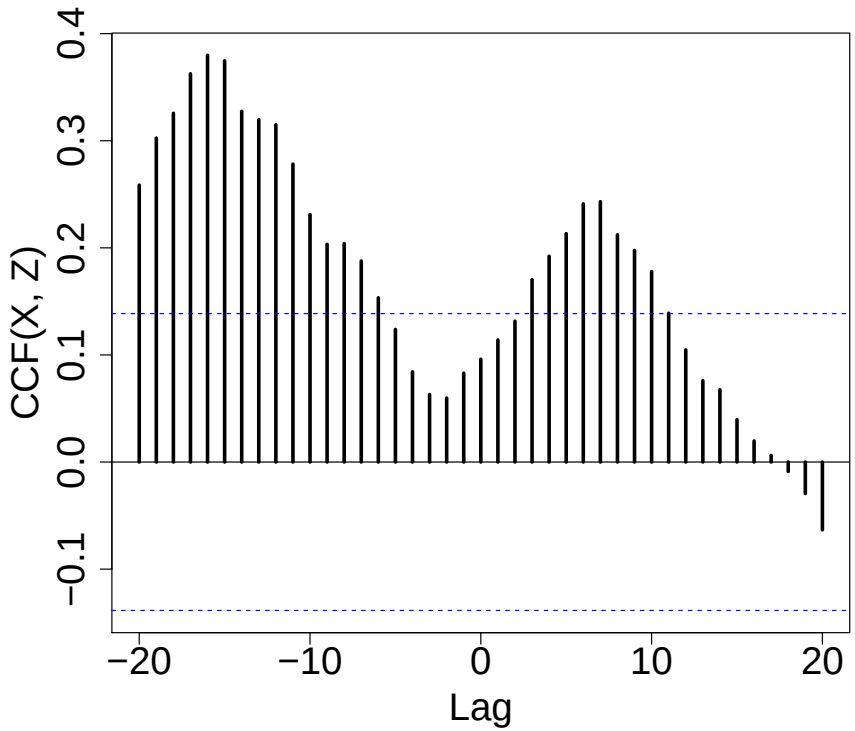


Figure 9: Ireland Chatter

Answer: using prewhitening!

And how to find out that these two series are uncorrelated?





Answer: using prewhitening!

How to figure out whether some X_{t-m}'s are related to Y_t?

- Strategy: examine the correlation between transformed versions of Y_t and X_t, denoted by $\tilde{Y}_t$ and $\tilde{X}_t$ whereby $\tilde{X}_t$ is approximately white noise.
- Let's start with an example, suppose $Y_t = \beta_1 X_{t-m} + Z_t$, where $X \sim AR(1)$ and Z is some other ARIMA process.
- Remember that we can rewrite the AR(1) model for X_t to obtain an expression for the white noise terms related to X_t as follows:

$$\begin{aligned} X_t &= \phi X_{t-1} + e_t. \\ e_t &= X_t - \phi X_{t-1} = (1 - \phi B) X_t = \pi(B) X_t. \end{aligned}$$

- Let's define $\tilde{X}_t = \pi(B) X_t$, such that $\tilde{X}_t$ is white noise.
  - This process is called whitening or prewhitening.
- We could also apply the same "filter" $\pi(B)$ to Y_t to obtain:

$$\tilde{Y}_t = \pi(B) Y_t = Y_t - \phi Y_{t-1}.$$

- Suppose that $\tilde{Y}_t$ is stationary, then what is the $\rho_k(\tilde{X}, \tilde{Y})$?

CCF $\rho_k(\tilde{X}, \tilde{Y})$

- When we apply the filter to the model equation, we obtain the following relation between $\tilde{Y}$ and $\tilde{X}$:

$$\begin{aligned} Y_t &= \beta_1 X_{t-m} + Z_t, \\ \pi(B)Y_t &= \beta_1 \pi(B) X_{t-m} + \pi(B) Z_t, \\ \tilde{Y}_t &= \beta_1 \tilde{X}_{t-m} + \tilde{Z}_t, \end{aligned}$$

- The CCF $\rho_k(\tilde{X}, \tilde{Y})$ for $\tilde{X}$ and $\tilde{Y}$ is given by

$$\rho_k(\tilde{X}, \tilde{Y}) = \frac{Cov(\tilde{X}_{t+k}, \tilde{Y}_t)}{\sqrt{Var(\tilde{X}_t) Var(\tilde{Y}_t)}} = \frac{Cov(\tilde{X}_{t+k}, \beta_1 \tilde{X}_{t-m} + \tilde{Z}_t)}{\sigma_{\tilde{X}} \sigma_{\tilde{Y}}},$$

  where $\tilde{X}_t$ is white noise.
- Aha! If $\tilde{X}_t$ is white noise, then $\rho_k(\tilde{X}, \tilde{Y}) \neq 0$ for $k =$???

The "promising-looking" CCF $\rho_k(\tilde{X}, \tilde{Y})$

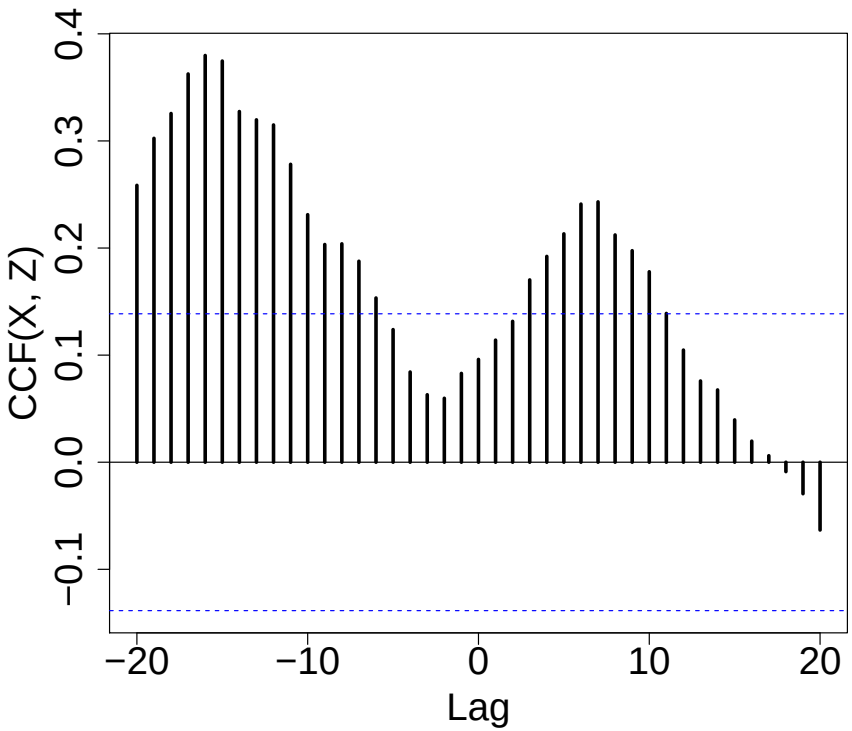
- For the model $Y_t = \beta_1 X_{t-m} + Z_t$, $\rho_k(\tilde{X}, \tilde{Y}) \neq 0$ for $k = -m$ only because $\tilde{X}_t$ is white noise.
- If $\tilde{Y}_t$ is stationary, then approximately for large n, for $k \neq -m$

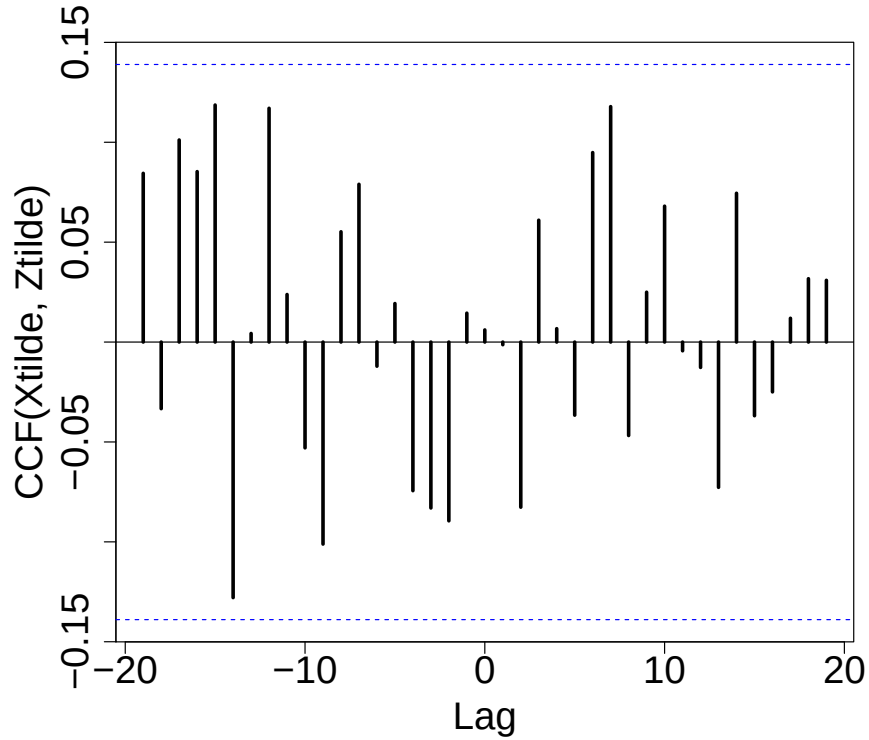
$$\begin{aligned} r_k(\tilde{X}, \tilde{Y}) &\sim N(0, Var(r_k(\tilde{X}, \tilde{Y}))), \\ Var(r_k(\tilde{X}, \tilde{Y})) &= 1/n \left(1 + 2 \sum_{k=1}^{\infty} \rho_k(\tilde{X}) \rho_k(\tilde{Y}) \right) = 1/n. \end{aligned}$$

  because $\rho_k(\tilde{X}) = 0$ for all $k \neq 0$.
- Good news!
  - We can try to use the sample CCF $r_k(\tilde{X}, \tilde{Y})$ to find out which lag $-m$ should be selected!

Simulated example I

- Let $Z \sim AR(1)$, $X \sim AR(1)$, and $n = 200$ as before, with $\phi_X = 0.95$ and $\phi_Z = 0.9$.
- Get $r_k(\tilde{X}, \tilde{Z})$ for $\tilde{X}_t = X_t - \phi_X X_{t-1}$ and $\tilde{Z}_t = Z_t - \phi_X Z_{t-1}$.





Conclusion?

Prewhitening: general approach

- Suppose we want to decide if any X_{t-k}'s are related to Y_t in a model in the form of

$$Y_t = \sum_{k=-\infty}^{\infty} \beta_k X_{t-k} + Z_t,$$

  where X_t and Z_t are time series processes, how can we find out which β_k's are non-zero?
- Approach:
  1. Find filter $\pi(B)$ for X_t such that $\tilde{X}_t = \pi(B)X_t$ is approximately white noise.
  2. Examine $r_k(\tilde{X}, \tilde{Y})$ where $\tilde{Y}_t = \pi(B)Y_t$:
     $\rho_k(\tilde{X}, \tilde{Y}) \propto \beta_{-k}$ (tutorial!) and we expect that
     $|r_k(\tilde{X}, \tilde{Y})| < 1.96\sqrt{1/n}$ for 95% of all lags with $\beta_k = 0$.

Prewhitening: finding filter $\pi(B)$

- If X_t is an invertible ARIMA process, we can write:

$$\begin{aligned} X_t &= e_t + \sum_{j=1}^{\infty} \pi_j X_{t-j}, \\ e_t &= X_t - \sum_{j=1}^{\infty} \pi_j X_{t-j} = (1 - \sum_{j=1}^{\infty} \pi_j B^j) X_t = \pi(B) X_t. \end{aligned}$$

- Ah, then $\tilde{X}_t = \pi(B) X_t$ is white noise!
- For a real data set, $\pi(B)$ is unknown, so how to find the filter to prewhiten the X_t's?
  - The filter follows from the model choice for X_t!

The unemployment project

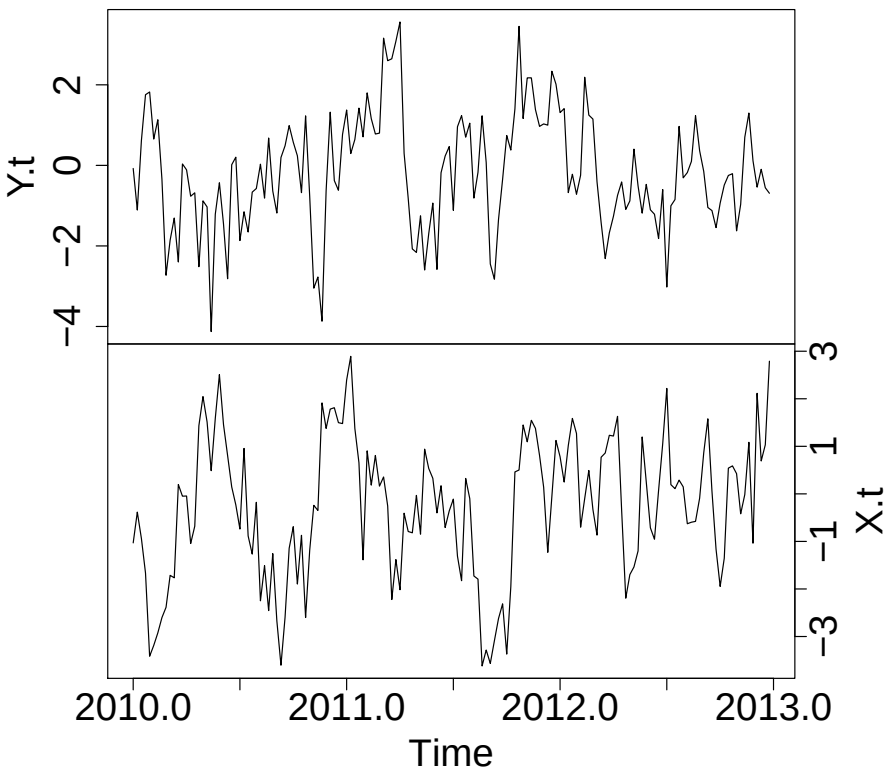
- Below is the listing of significant CCFs for Ireland.
- Note: Exact method is not given but I assume that whitening was used.

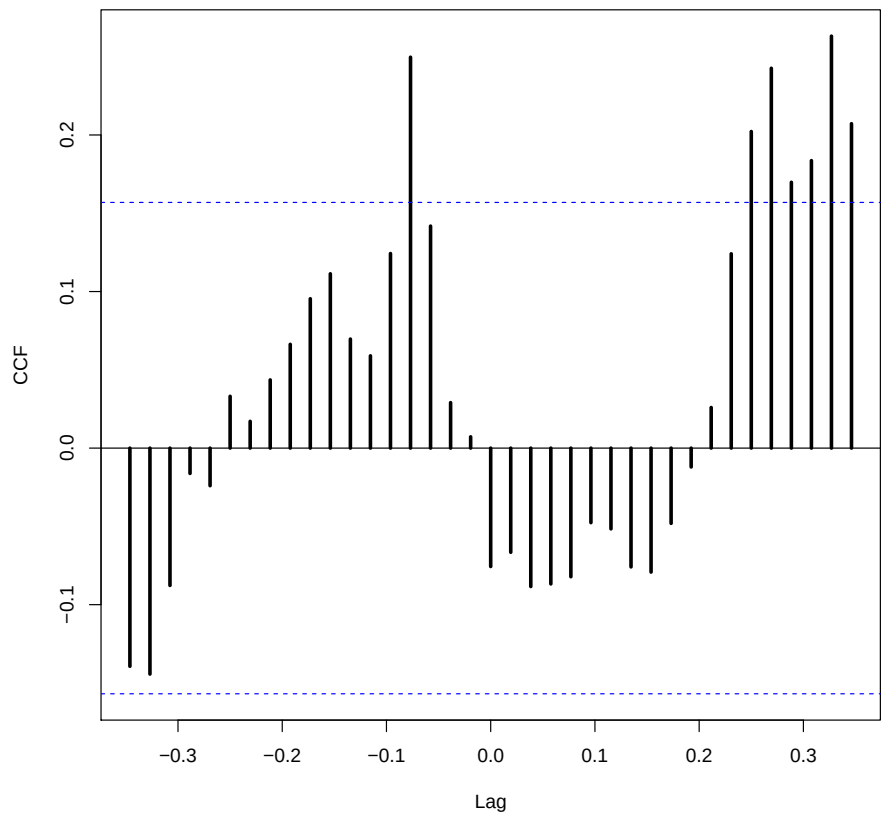
Table 1: Country Correlation, CCF and Significance Level

Country	Correlation Description	CCF	Significance Level
IRELAND	Anxious Mood increases 5 months before a spike in unemployment	.387	90%
IRELAND	Confused Mood increases 3 months before a spike in unemployment	.675	95%
IRELAND	Confident Mood decreases 2 months before a spike in unemployment	-.407	90%
IRELAND	Talk about changing transportation methods for the worse increases as unemployment increases	.380	90%
IRELAND	Talk about travel cancelations increases 3 months after an unemployment spike	.450	95%
IRELAND	Talk about changing housing situations for the worse increases 8 months after unemployment increases	.328	90%

A worked example

Suppose that the data set for the unemployment project was as specified below, where Y_t are (standardized) changes in unemployment and X_t are standardized employment-related measures of a "depressed mood".
Let's find out if this measure is related to unemployment!



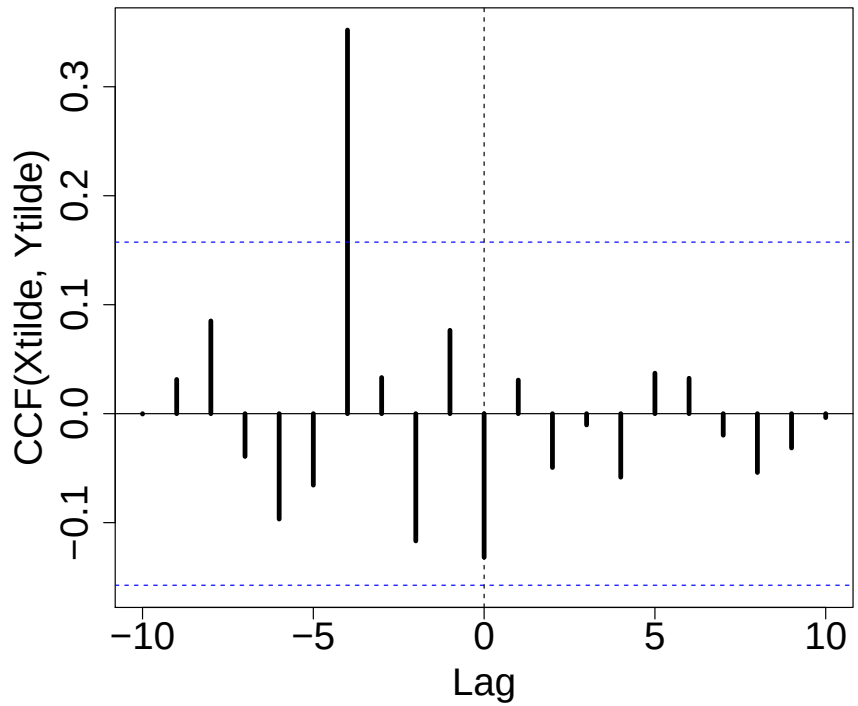


Approach: Model X & calculate $r_k(\tilde{X}, \tilde{Y})$

```
> auto.arima(X.t, ic = "bic")
Series: X.t
ARIMA(1,0,0)

m1=arima(X.t,order=c(1,0,0))

# use built-in function
prewhiten(x=as.vector(X.t),
  y=as.vector(Y.t),
  x.model=m1)
```



- Conclusion? The CCF for the prewhitened series suggests that high job-related depression in month t is related to high unemployment in month $t + 4$.
- Can we fit a model $Y_t = \beta_0 + \beta_1 X_{t-4} + Z_t$ to try to predict Y_t, or obtain the relation between Y_t and X_{t-4}?

Modeling Y_t using X_{t-m}

- A model of the form

$$Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t$$

  is called a transfer-function model/distributed-lag model/dynamic regression model.
  - These models may include the covariate at several lags but we will discuss only the example with just one lagged covariate.
- After identifying which X_{t-m} to include, how to specify Z_t in the model $Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t$?
- In order to explore the model specification for Z_t, the following approach is used:
  (A) Regress Y_t on X_{t-m} (assume temporarily that $Y_t = \beta_0 + \beta_1 X_{t-m} + e_t$) and obtain residuals $\hat{Z}_t = Y_t - \hat{Y}_t$.
  (B) Explore $\hat{Z}_t$ to specify a candidate model for Z_t
  (C) Fit the complete model $Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t$, where Z_t is specified by the candidate model and check model diagnostics.

Continuing the unemployment "data example"

- The CCF for the prewhitened series suggested that high anxiety in month t is related to high unemployment in month $t+4$, thus we want to fit the model

$$Y_t = \beta_0 + \beta_1 X_{t-4} + Z_t.$$

- Step A: Regress Y_t on X_{t-4} and obtain residuals:

```
# we can only use pairs X(t), Y(t+4)
Y5.t <- Y.t[5:n]
X1.t <- X.t[1:(n-4)]
mod <- lm(Y5.t ~ X1.t)
> summary(mod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.11742    0.11664  -1.007   0.3157
X1.t         0.25803    0.07872   3.278   0.0013 **
```

Continuing the unemployment "data example"

- Model $Y_t = \beta_0 + \beta_1 X_{t-4} + Z_t$.
- Step A: Regress Y_t on X_{t-4} and obtain residuals.
- Step B: Analyze the residuals to find a candidate model for Z_t:

```
Zhat.t <- resid(mod)
> auto.arima(Zhat.t, ic = "aicc")
Series: Zhat.t
ARIMA(1,0,0) with zero mean
```

Continuing the unemployment "data example"

- Model $Y_t = \beta_0 + \beta_1 X_{t-4} + Z_t$.
- Step A: Regress Y_t on X_{t-4} and obtain residuals.
- Step B: Analyze the residuals to find a candidate model for Z_t.
- Step C: Fit the complete model:

```
mod =arima(Y5.t,order=c(1,0,0), xreg = X1.t)
> mod
ARIMA(1,0,0) with non-zero mean
         ar1  intercept    xreg
      0.6769    -0.0399  0.4102
s.e.  0.0609     0.2598  0.0883
```

  - and check diagnostics (see R-script, they're fine).

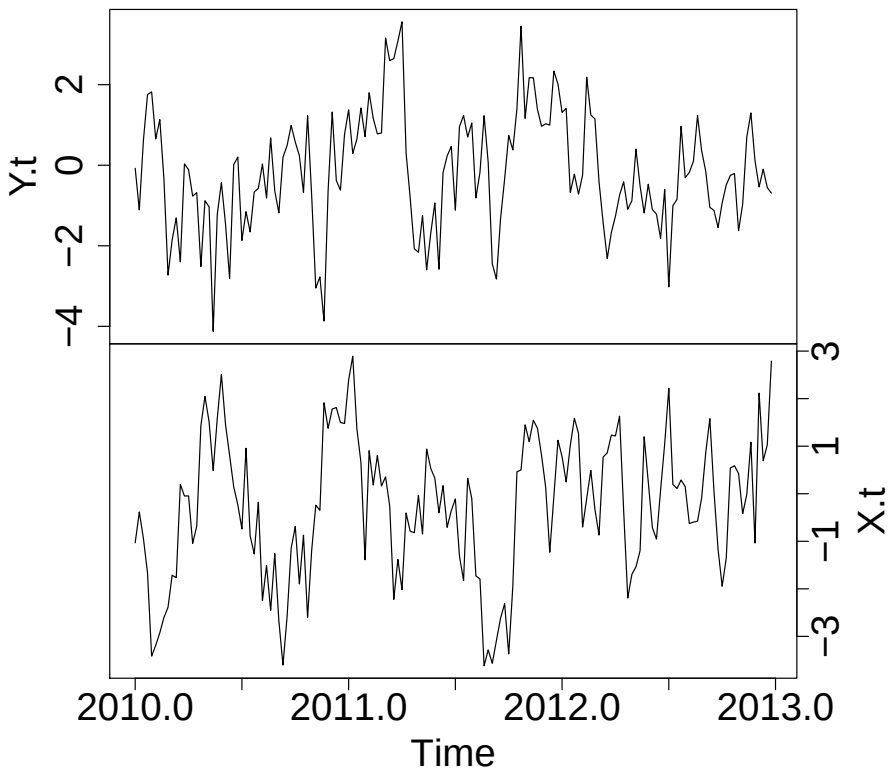
- Conclusion?

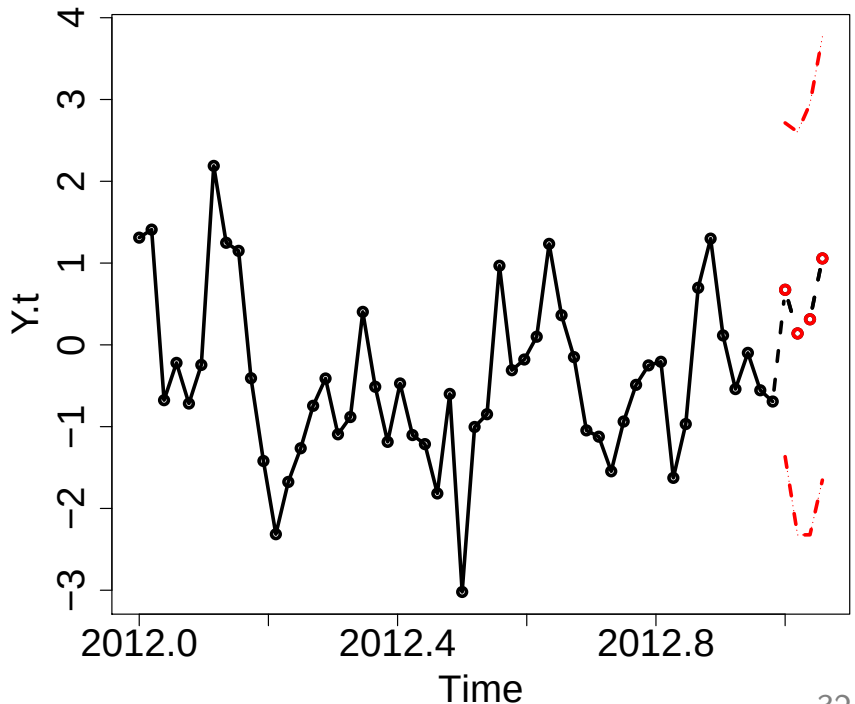
A worked example

Suppose that the data set for the unemployment project was as specified below, where Y_t are (standardized) changes in unemployment and X_t are standardized employment-related measures of a "depressed mood".
We find that

$$Y_t = \beta_0 + \beta_1 X_{t-4} + Z_t,$$

where Z_t follows an AR(1) model.





Summary

- Motivation:
  - Suppose we have a time series of interest $Y_1, Y_2, \ldots, Y_t$ (e.g. changes in the unemployment rate), here denoted by Y, and we want to explore whether/how another time series $X_1, X_2, \ldots, X_t$, here denoted by X (e.g. depression measured in employment-related social media output), relates to Y_t.
- We discussed:
  - How to summarize the correlation between X and Y using the (sample) cross-correlation function (CCF),
  - and that the sample CCF can show spurious correlation if

$$Y_t = \beta_0 + \beta_1 X_{t-m} + Z_t,$$

    if X and Z are both autocorrelated time series.
  - How to prewhiten X, and use the same procedure for Y, to obtain a new sample CCF which is informative of the relation between Y and X.
  - How to fit a dynamic regression model to model Y using X while accounting for autocorrelation in Y.
- The UN Global Pulse project can now hire you for their analyses!