Time Series Simulation

Rachael Phillips

In this project, we will be in a setting of multi-person multi-dimensional time series. We will use an online SL and can incorporate learners which pool across subjects and individualize subjects (via more aggressive modeling of the effect of the baseline covariates). The online cross-validated risk is individualized, but we will average across subjects because it's a more interpretable measure.

Simulation Overview

Simplistic settings which can capture the following:

- A situation where subjects are iid (baseline cov variation is null)
- Time series variation is very much a function of baseline covariates
- Time series variation is very much a function of baseline covariates that you don't measure (unexplained heterogeneity).
- Across time the process is stationary
- Time is not stationary (sudden jumps) so you cannot learn from the past how much it will change in the future.

Today I simulate from relatively simplistic ARIMA models.

Later I will model the binary or continuous MIMIC outcome data with GLARMA or ARIMA, respectively. If the GLARMA/ARIMA model seems to be suitable for modeling the MIMIC data, then I can create a simulated dataset with the fitted model.

Simple ARIMA Simulations

ARIMA Introduction

An auto-regressive integrated moving average model (ARIMA) is specified by three order parameters: (p, d, q).

p is the number of autoregressive terms The p is the auto-regressive (AR(p)) component and refers to the use of past values in the regression equation for the series. The auto-regressive parameter p specifies the number of lags used in the model. Intuitively, this would be similar to stating that it is likely to be warm tomorrow if it has been warm the past p days.

d is the number of nonseasonal differences The d represents the degree of differencing in the integrated (I(d)) component. Differencing a series involves subtracting its current and previous values d times. Often, differencing is used to stabilize the series when the stationarity assumption is not met. Intuitively, this would be similar to stating that it is likely to be same temperature tomorrow if the difference in temperature in the last d days has been very small.

q is the number of moving-averages terms A moving average (MA(q)) component represents the error of the model as a combination of previous error terms, where q defines the number of terms to include in the model.

Differencing, autoregressive, and moving average components make up a non-seasonal ARIMA model which can be written as a linear equation:

$$Y_t = c + \phi_1 y_{dt-1} + \phi_p y_{dt-p} + ... + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t$$

where y_d is Y differenced d times and c is a constant.

ARIMA models can be also specified through a seasonal structure. In this case, the model is specified by two sets of order parameters: (p, d, q) as described above and $(P, D, Q)_m$ parameters describing the seasonal component of m periods.

ARIMA methodology does have its limitations. These models directly rely on past values, and therefore work best on long and stable series. Also note that ARIMA simply approximates historical patterns and therefore does not aim to explain the structure of the underlying data mechanism.

Resources

- A Short Introduction to ARIMA
- Time Series: AR, MA, ARMA, ARIMA
- Hyndman and Athanasopoulos *Forecasting: Principles and Practice*

ARIMA Simulations with White Noise

White noise time series can be useful because the stochastic behavior of all time series can be explained in terms of the white noise model. We simulate Gaussian white noise, wherein $wt \sim_{iid} N(\mu, \sigma)$. μ is set to 0 or a baseline covariate value and σ is set to 1 or a baseline covariate value. We consider three scenarios:

1. μ is 0 and σ is 1, the time series is a function unexplained by baseline covariates.
2. μ depends on a baseline covariate and σ is 1, the time series is partially a function of baseline covariates.
3. μ and σ depend on a baseline covariate values, the time series variation is a function of baseline covariates.

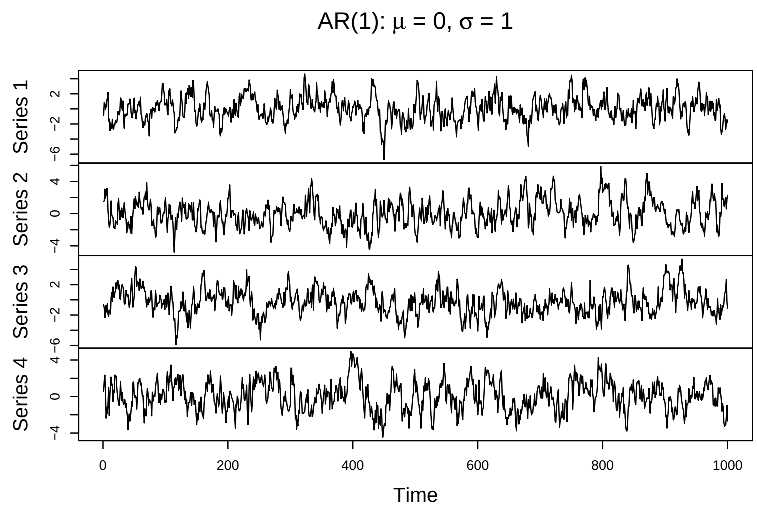
For each of the three scenarios, we simulate $N = 500$ subjects each with $n = 1000$ observations from the following models:

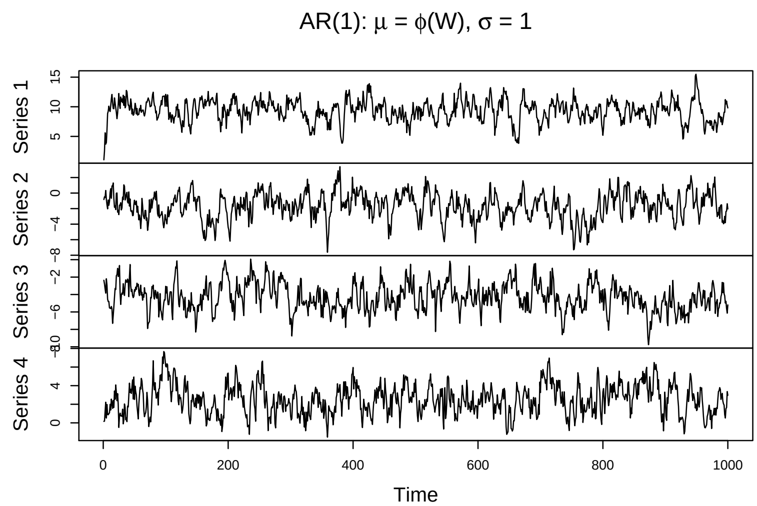
1. An autoregressive model of order 1 ($p = 1$), where each value of y equals the previous value times 0.8, plus the white noise.
2. A moving average of order 1 ($q = 1$), where each value of y equals the latest bit of white noise, plus 0.8 times the previous value of white noise.
3. An autoregressive moving average model of order (1, 1), combining the two above.
4. An ARIMA(1, 1, 1) model that is the cumulative sum of the ARMA(1, 1), so the first difference of the time series is stationary.

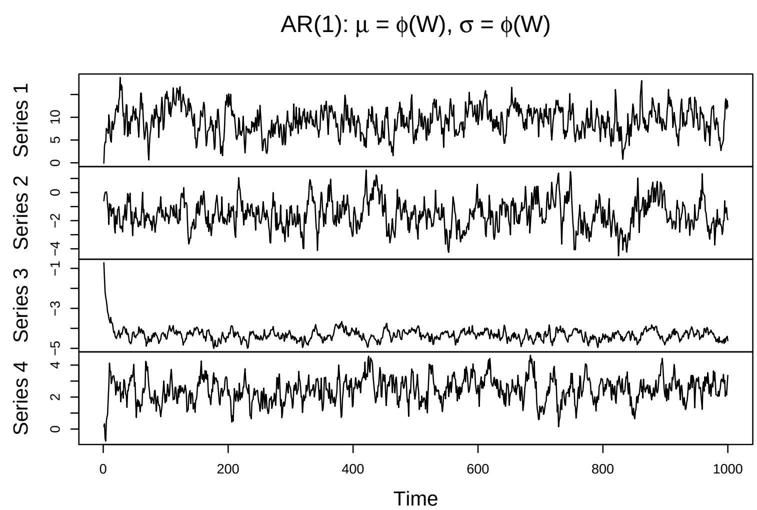
```
# mu is 0 and sigma is 1
wt_ts_0 <- sim_wt_ts()

# mu is W3 and sigma is 1
W1 <- runif(500, min = -1, max = 1)
W2 <- rbinom(500, prob = plogis(W1), size = 1)
W3 <- W1 + W2
wt_ts_W <- sim_wt_ts(mu = W3)

# mu is W3 and sigma is W4
W4 <- W1 + 1
wt_ts_WW <- sim_wt_ts(mu = W3, sigma = W4)
```

Now we can visualize the various models.

1. An autoregressive model of order 1 ($p = 1$), where each value of y equals the previous value times 0.8, plus the white noise.

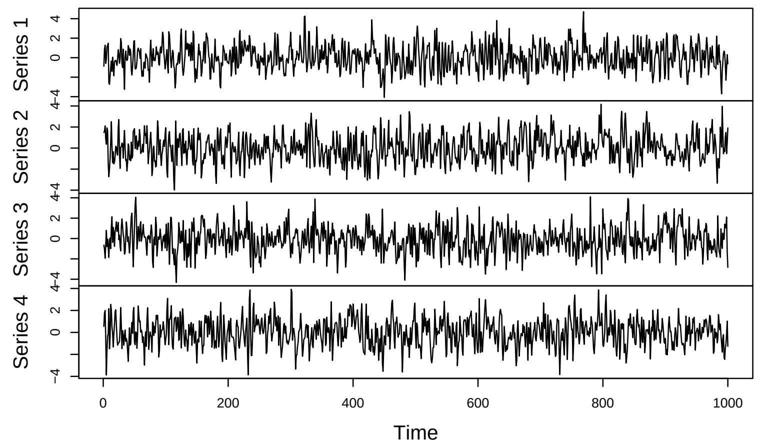




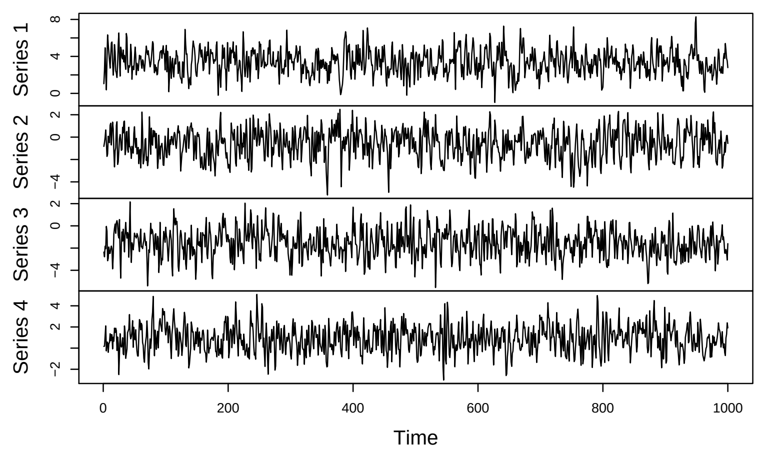


2. A moving average of order 1 ($q = 1$), where each value of y equals the latest bit of white noise, plus 0.8 times the previous value of white noise.

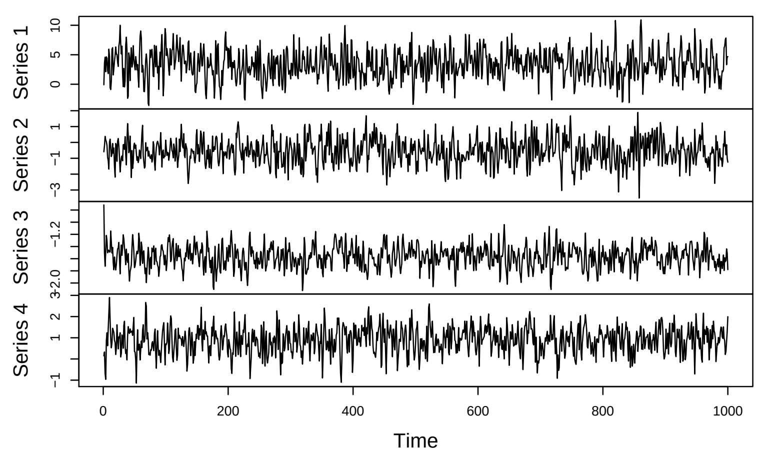
MA(1): μ = 0, σ = 1



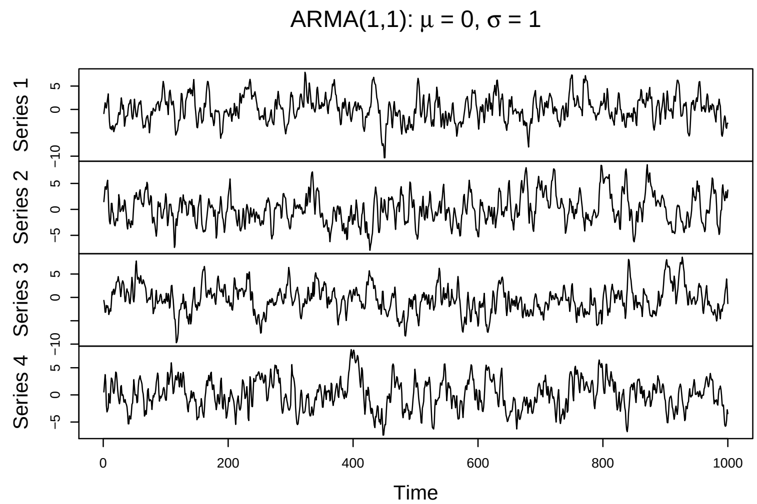
MA(1): μ = ϕ(W), σ = 1

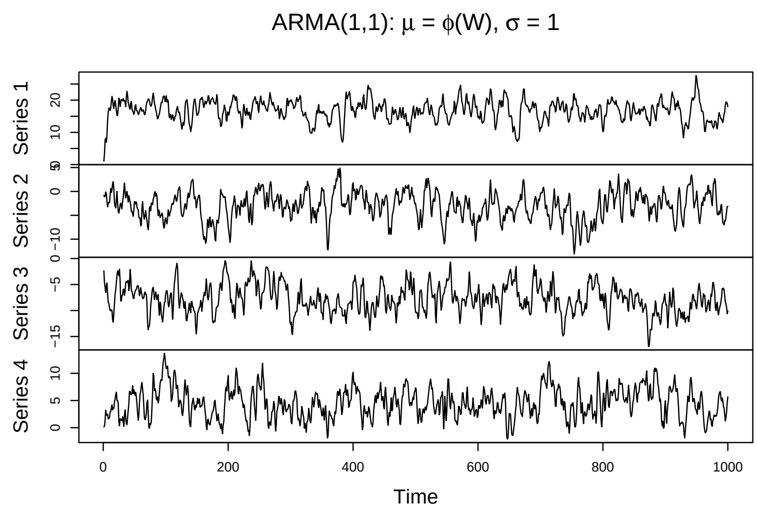


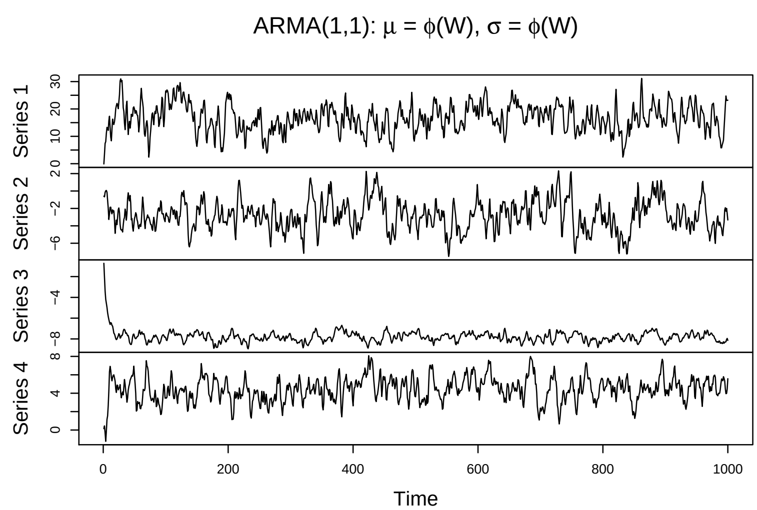
MA(1): μ = ϕ(W), σ = ϕ(W)



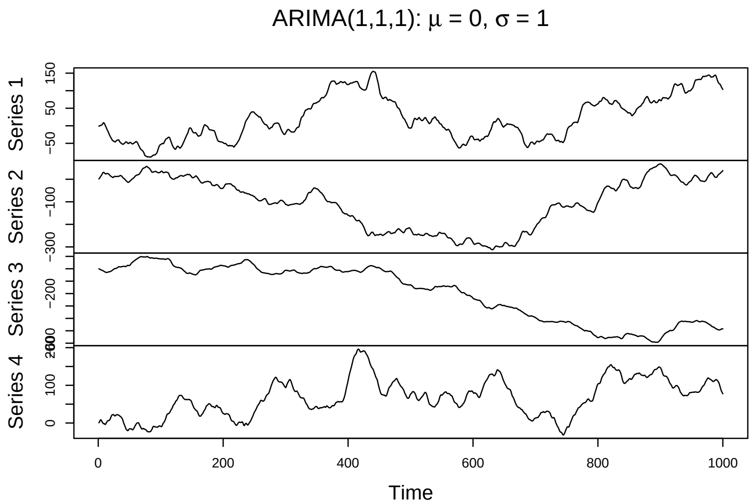
3. An autoregressive moving average model of order (1, 1), combining the two above.

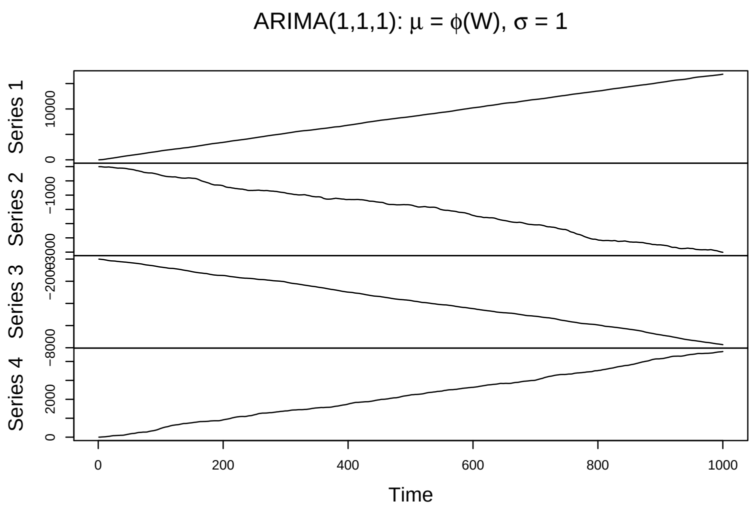


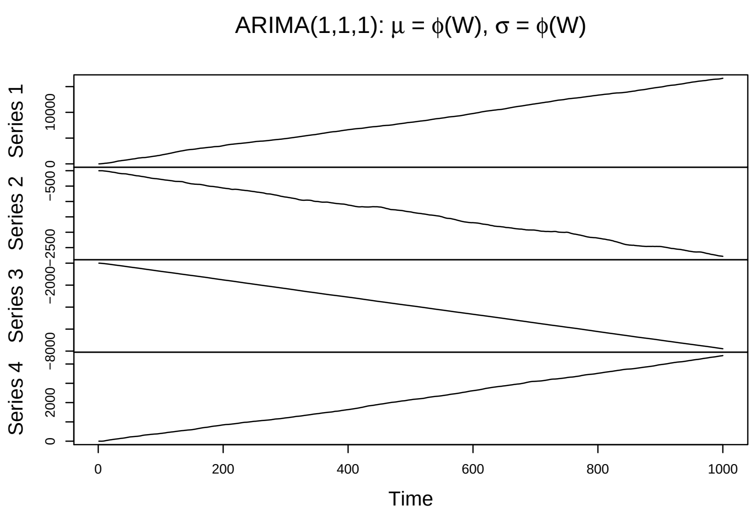




4. An ARIMA(1, 1, 1) model that is the cumulative sum of the ARMA(1, 1), so the first difference of the time series is stationary.







Simulations are based on those presented in free range statistics

MIMIC-based Simulation

For next time…