Prediction

Jared Fisher

Lecture 7b

Announcements

Announcements

- Project Checkpoint 3 is due the Wednesday after Spring Break, March 31.

Recap

Where we are at: Signal

- We model time series Y_t as

$$Y_t = signal_t + noise_t$$

- The signal is [essentially] $E(Y_t)$, and we estimate it with $\hat{f}(t)$ or $\hat{Y}_t =$
  - trend plus seasonality $(m_t + s_t)$
  - a filter of past observations $\left(\sum_{j=1}^{\infty} a_j Y_{t-j}\right)$, like differencing or smoothing
  - or perhaps a combination of the two
- The noise terms in many other statistical models are assumed to be Gaussian noise. However, in this class they are not necessarily iid! In fact, often there is an autocorrelation structure.
- If the residuals $(Y_t - \hat{Y}_t)$ have means, variances or autocorrelation that changes with respect to t, then we have a quite-challenging modeling problem.
- So, we will pursue stationarity in our residuals, then address the autocorrelation structure.

Where we are at: Noise

- Our noise term is the stationary process X_t
- We look specfically at autocorrelation structures created by filters on white noise W_t:

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$$

- If stationary, then our main interest in the noise process is to model its ACF
- When choosing a theoretical model to apply to our observed data, we're essentially matching the sample ACF to the theoretical ACF as best as we can.

ARMA(p,q): a flexible way to model ACF's

- ARMA(p,q) models can fit many forms of the ACF, but this brings up two questions:

1. What we've seen are (essentially) the expected values of the autocorrelations. What is reasonable uncertainty around these values? see Bartlett's Formula
2. It seems to be much easier to distiguish different values of q (the MA order) than it does p (the AR order). In other words, some AR(1) ACFs look similar to AR(2) ACFs. How can we differentiate, especially considering p can be larger than 2?

- We will looked at #1 last class, and will start building the tool for #2 today.

Theorem: Bartlett's Formula

Under some general conditions on the white noise process $\{W_t\}$, if $\{X_t\}$ is an causal and invertible ARMA process $\phi(B)X_t = \theta(B)W_t$, then for any fixed lag k and n large enough the sample autocorrelations $(r_1, r_2, \ldots, r_k)$ are approximately multivariate normal distributed with mean $(\rho_X(1), \ldots, \rho_X(k))$ and covariance matrix W/n with (i,j)th entry equal to $W_{ij} =$

$$\begin{aligned}\sum_{m=1}^{\infty} [(\rho_X(m+i) + {} & \rho_X(m-i) - 2\rho_X(i)\rho_X(m)) \\ & * (\rho_X(m+j) + \rho_X(m-j) - 2\rho_X(j)\rho_X(m))]\end{aligned}$$

that is,

$$\sqrt{n}\left(\begin{pmatrix} r_1 \\ \vdots \\ r_k \end{pmatrix} - \begin{pmatrix} \rho_X(1) \\ \vdots \\ \rho_X(k) \end{pmatrix}\right) \to N(0, W) \qquad \text{as } n \to \infty.$$

Example: MA(1)

- Suppose $X_t = W_t + \theta W_{t-1}$.
- We have seen that $\rho_X(1) = \theta/(1+\theta^2)$ and $\rho_X(h) = 0$ for higher lags h.
- Bartlett's formula says that the variance of r_i is approximately W_{ii}/n where

$$W_{ii} = \sum_{m=1}^{\infty} \left(\rho(m+i) + \rho(m-i) - 2\rho(i)\rho(m)\right)^2 .$$

- What's the variance of r_1?

Example: MA(1)

- For $i = 1$ i.e., when we consider the first order sample autocorrelation, this formula gives
$$var(r_1) \approx W_{11}/n = (1 - 3\rho^2(1) + 4\rho^4(1))/n < 1/n$$
.
- This last inequality requires noting that for any θ we have $\rho^2(1) = \frac{\theta^2}{(1+\theta^2)^2} \leq 1/4$.
- In other words, r_1 for MA(1) is less variable than r_1 for white noise.

Example: MA(1)

- For higher values of i, the formula gives

$$W_{ii} = \sum_{m=1}^{\infty} \rho^2(m-i) = 1 + 2\rho^2(1) > 1.$$

- In other words, r_k for $k \geq 2$ are more variable for MA(1) than for white noise.
- Thus we can expect to see more r_k's sticking out the horizontal blue lines for MA(1).

Empirical Strategy

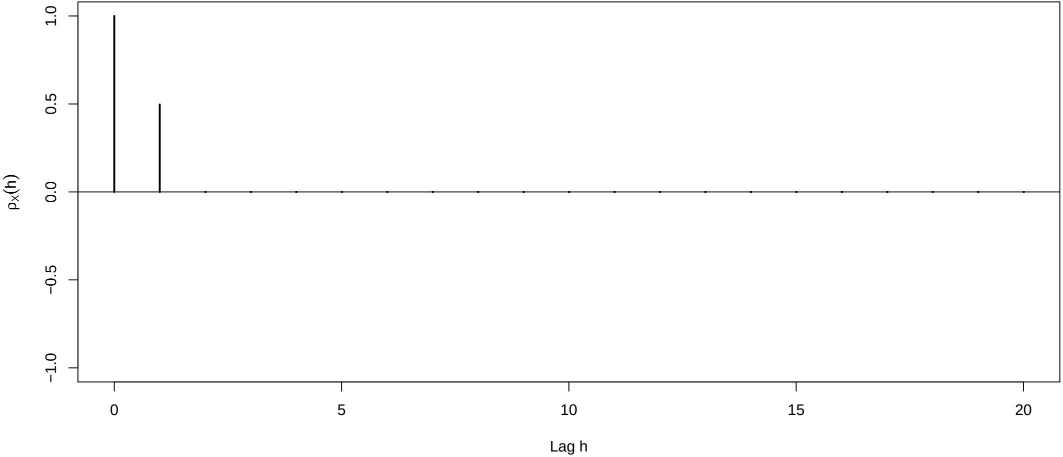
- A general strategy to find out whether ARMA(p,q) is a good model for data is as follows:

1. Plot the ACF correlogram $(r_1, \ldots, r_k)$.
2. Compare this with the theoretical ACF $\rho_X(h)$ of the ARMA(p,q) model.
3. Keep in mind the variability of the r_k's given by Bartlett's formula

- BUT, how can we get initial guesses at p and q in the first place? There are clues in the ACF plots.

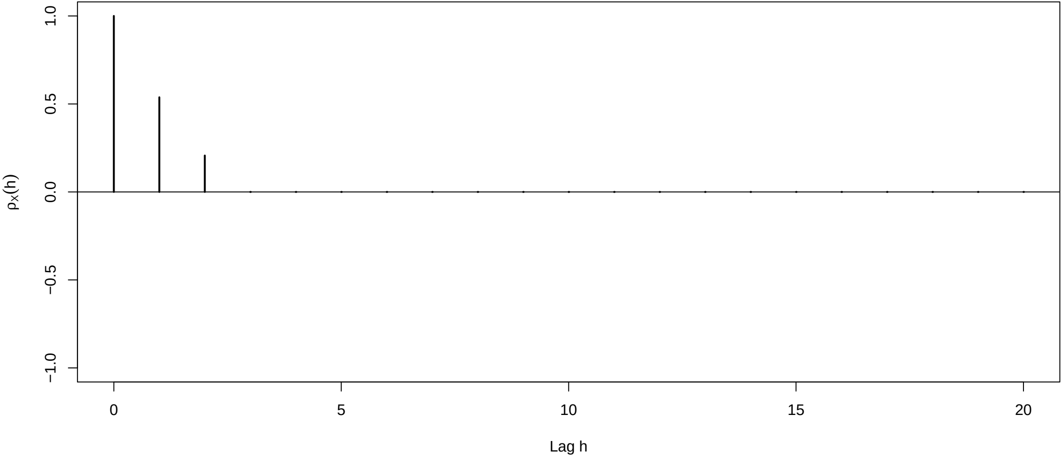
MA(1), $\theta = 0.9$

```
plot.theoretical.acf(ma=0.9)
```



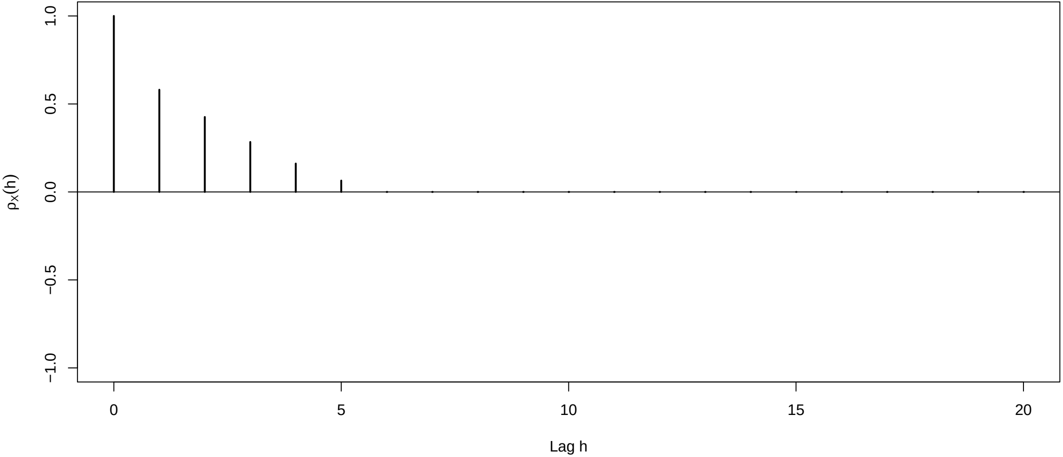
MA(2), $\theta_1 = 0.6$, $\theta_2 = 0.3$

```
plot.theoretical.acf(ma=c(0.6,0.3))
```



MA plays nice: MA(5), $\theta_1 = 0.5$, $\theta_2 = 0.4$, $\theta_3 = 0.3$, $\theta_4 = 0.2$, $\theta_5 = 0.1$

```r
plot.theoretical.acf(ma=c(0.5,0.4,0.3,0.2,0.1))
```

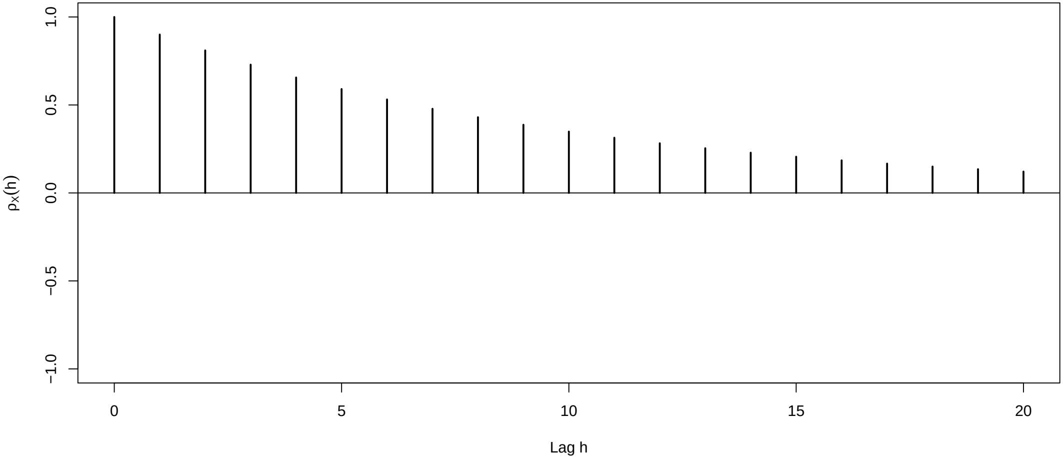


MA vs. AR

The ACFs of MA(q) have no significant autocorrelations after lag q (the “cutoff”). But p in AR(p) models is harder to distinguish.

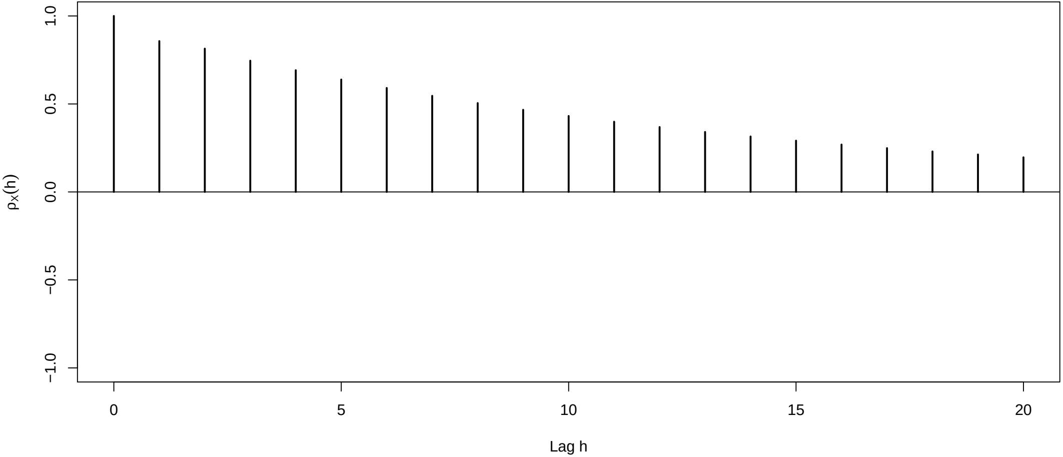
AR(1), $\phi = .9$

```r
plot.theoretical.acf(ar=.9)
```



AR(2), $\phi_1 = 0.6$, $\phi_2 = 0.3$

```
plot.theoretical.acf(ar=c(.6,.3))
```



Empirical Strategy

- For example, when the sample autocorrelations after lag q drop off and lie between the band for MA(q) given by Bartlett's formula, the MA(q) model might be appropriate.
- For AR(p) models the ACF does not drop to zero for large lags. Thus, it is more difficult to choose the order of an appropriate AR process for data by looking at the sample ACF.
- If we have time today, we will introduce the *partial autocorrelation function (PACF)*, which will be helpful for AR(p) models: the PACF of an AR(p) model is zero for lags strictly larger than p.
- But in order to define the PACF, we first must define "prediction".

Prediction

Prediction

- One of our major goals is to predict a future observation X_{n+1} from a given time series data set $X_1, \ldots, X_n$.
- We will assume that the stationary time series $\{X_t\}$ has mean zero.
- Otherwise, when $E(X_t) = \mu \neq 0$ we can just consider $\{X_t - \mu\}$ instead.
- When μ is unknown, estimate with the sample mean $\hat{\mu} = \overline{X} = \sum_{i=1}^{n} X_i / n$.
- First, we study the general problem of predicting the outcome of a random variable Y based on some other zero mean random variables $X_1, \ldots, X_n$. Y is a generic variable until the next section "Prediction of a Stationary Process"

Theorem - Best Prediction

Let $Y, X_1, \ldots, X_n$ be random variables. Then for the best **mean squared error** prediction $f^\star(X_1, \ldots, X_n)$ of Y, that is

$$E\left[(Y - f^\star(X_1, \ldots, X_n))^2\right] = \min_f E\left[(Y - f(X_1, \ldots, X_n))^2\right],$$

it holds that

$$f^*(X_1, \ldots, X_n) := E\left(Y | X_1, \ldots, X_n\right).$$

Best Prediction

- (The proof of this theorem may be a future exercise...)
- Problem with the best prediction: in general, we'd need to know the entire joint distribution of $Y, X_1, \ldots, X_n$ in order to compute it.

Best Linear Prediction?

- On the other hand, it is much easier to compute the best **linear** prediction of Y in terms of $X_1, \ldots, X_n$.
- Assume that X_i and Y all have finite second moments
- Let Δ denote the covariance matrix of $X = (X_1, \ldots, X_n)$
- Assume Δ is an invertible matrix (this just excludes the situation that a linear combination of the X_i's has variance zero), that is

$$\Delta_{ij} = cov(X_i, X_j) \quad \text{and} \quad \zeta_i = cov(Y, X_i).$$

Thoerem: Best Linear Prediction (“BLP”)

Let $Y, X_1, \ldots, X_n$ be zero mean random variables with finite second moments. Then for the best mean squared error linear prediction $a_1 X_1 + \ldots + a_n X_n$ of Y, that is

$$E\left[(Y - (a_1^\star X_1 + \ldots + a_n^\star X_n))^2\right] = \min_a E\left[(Y - (a_1 X_1 + \ldots + a_n X_n))^2\right],$$

it holds that

$$(a_1^\star, \ldots, a_n^\star)^\top = \Delta^{-1}\zeta.$$

Proof

We have that

$$\begin{aligned} F(\mathbf{a}) :=& E\left[(Y - a_1X_1 - \cdots - a_nX_n)^2\right] \\ =& E\left[\left(Y - \mathbf{a}^TX\right)^2\right] \\ =& EY^2 - 2E((\mathbf{a}^TX)Y) + E(\mathbf{a}^TXX^T\mathbf{a}) \\ =& EY^2 - 2\mathbf{a}^T\zeta + \mathbf{a}^T\Delta\mathbf{a}. \end{aligned}$$

Differentiate with respect to **a** and set equal to zero to get

$$\begin{aligned} -2\zeta + 2\Delta\mathbf{a} = 0 \\ \Rightarrow \mathbf{a} = \Delta^{-1}\zeta \end{aligned}$$

Therefore the best linear predictor of Y in terms of $X_1, \ldots, X_n$ equals $\zeta^T\Delta^{-1}X$.

Theorem Setup

There exists a useful equivalent characterization of the best linear predictor...

Theorem: Characterization of the Best Linear Prediction (CBLP)

The best linear predictor $(a_1^\star, \ldots, a_n^\star)^\top$ in the BLP Theorem is uniquely characterized by the property that

$$cov(Y - a_1X_1 - \cdots - a_nX_n, X_i) = 0 \text{ for all } i = 1, \ldots, n.$$

(Proof of this theorem is omitted as a potential future exercise)

Example: BLP for $n = 1$

For $n = 1$ (when there is only one predictor X_1), we have $\zeta_1 = \text{cov}(Y, X_1)$ and $\Delta_{11} = var(X_1)$. Thus, the best predictor of Y in terms of X_1 is

$$\frac{cov(Y, X_1)}{var(X_1)} X_1.$$

(just like simple linear regression)

Prediction of a Stationary Process

Prediction of a Stationary Process

- Let $\{X_t\}$ be a stationary zero mean time series with ACVF $\gamma_X(h)$
- Apply Theorem BLP to obtain that the best linear predictor of X_{n+1} in terms of k previous observations $X_n, X_{n-1}, \dots, X_{n-k+1}$ is

$$\text{predict} \quad X_{n+1} \quad \text{by} \quad (X_n, X_{n-1}, \dots, X_{n-k+1})\Delta^{-1}\zeta$$

with

$$\begin{aligned} \Delta_{ij} &= cov(X_{n-i+1}, X_{n-j+1}) = \gamma_X(i-j) \quad \text{for } i, j = 1, \dots, k, \\ \zeta_i &= cov(X_{n+1}, X_{n-i+1}) = \gamma_X(i) \quad \text{for } i = 1, \dots, k. \end{aligned}$$

Toeplitz Form

- Note that for a stationary process, the matrix Δ has a very specific structure, namely it is of so called *Toeplitz* form:

$$\Delta = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \dots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \ddots & & \vdots \\ \gamma(2) & \gamma(1) & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \gamma(1) & \gamma(2) \\ \vdots & & \ddots & \gamma(1) & \gamma(0) & \gamma(1) \\ \gamma(k-1) & \dots & \dots & \gamma(2) & \gamma(1) & \gamma(0) \end{pmatrix}$$

- Linear systems of equation $\Delta a = \zeta$, where Δ is of Toeplitz form, can be solved very efficiently with iterative algorithms (which do not invert the matrix Δ explicitly). One example is the *Durbin-Levinson algorithm* (TSA4e Property 3.4)

Example of Prediction with AR(p)

- First: take a couple minutes on your own to try to calculate the BP and BLP for AR(p)
- Zero mean AR(p): $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t$.
- Causality $\implies$ $X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p} = W_n$ is uncorrelated with $X_{n-1}, X_{n-2}, \dots, X_1$.
- Theorem CBLP implies that when $n > p$ the best linear predictor of X_n in terms of $X_{n-1}, X_{n-2}, \dots, X_1$ equals

$$\phi_1 X_{n-1} + \phi_2 X_{n-2} + \cdots + \phi_p X_{n-p}.$$

- Note that for AR(p) models, where the white noise process $\{W_t\}$ is i.i.d. and $n > p$, the Best Linear Prediction coincides with the Best Prediction. To see this, just take the conditional expectation on both sides of the AR(p) equation.

Prediction in ARMA Models

- In general, for causal ARMA(p,q) models, consider their AR(∞) representation $X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + W_t$.
- Then, for n large enough, this is well approximated by the AR(n) model $X_t \approx -\sum_{j=1}^{n} \pi_j X_{t-j} + W_t$ (recall that $\sum_j |\pi_j| < \infty$ and thus $|\pi_j| \to 0$ as $j \to \infty$).
- Hence, the Best Linear Prediction (and for i.i.d. noise $\{W_t\}$ also the Best Prediction) of X_{n+1} in terms of $X_n, X_{n-1}, \ldots, X_1$ is well approximated by $-\sum_{j=1}^{n} \pi_j X_{n-j+1}$.

Last Thoughts on Prediction

- The Best Linear Prediction is, in general, worse than the Best Prediction (a constrained solution cannot exceed the optimal solution in optimality).
- However, it is much easier to compute because it only requires knowledge of the covariances between the variables while the best predictor requires knowledge of the entire joint distribution.
- In the special case when $Y, X_1, \dots, X_n$ are jointly Gaussian, one can show that the best prediction and best linear prediction coincide, (TSA4e Theorem B.3).

PACF : Partial Autocorrelation Function

Definition

Let $\{X_t\}$ be a mean zero stationary process. The **Partial Autocorrelation** at lag h, denoted by pacf(h) is defined as the coefficient of X_{t-h} in the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h}$.

h=1

- For AR(1) models, the PACF at lag h=1:

$$
\begin{aligned}
pacf(1) &= \Delta^{-1}\zeta \\
&= cov(X_{t-1}, X_{t-1})^{-1} cov(X_t, X_{t-1}) \\
&= \gamma(0)^{-1} cov(\phi X_{t-1} + W_t, X_{t-1}) \\
&= \gamma(0)^{-1} \left(\phi cov(X_{t-1}, X_{t-1}) + cov(W_t, X_{t-1})\right) \\
&= \gamma(0)^{-1} \left(\phi\gamma(0) + 0\right) \\
&= \phi\gamma(0)^{-1}\gamma(0) \\
&= \phi \\
&= \rho(1)
\end{aligned}
$$

- But $pacf(h)$ for $h > 1$ can be quite different from $\rho(h)$.

AR(p)

Recall that for an AR(p) model we have $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t$ and hence by Theorem CBLP we immediately get the following theorem:

For the partial autocorrelation function of a causal AR(p) model $\phi(B)X_t = W_t$ it holds that $pacf(p) = \phi_p$ and $pacf(h) = 0$ for $h > p$.

Best Linear Prediction

- From the definition, it is not quite clear why this is called a correlation. We make this more clear in the following.
- pacf(h) is the correlation between X_t and X_{t-h} “with the linear effect of everything ‘in the middle’ removed” (TSA4e)
- corr$(X_t - \hat{X}_t \ , X_{t-h} - \hat{X}_{t-h})$, where $\hat{X}_t$ and $\hat{X}_{t-h}$ are the best linear predictors using $X_{t-1}, \dots, X_{t-h+1}$.

Best Linear Prediction

- Let $a_1 X_{t-1} + \cdots + a_{h-1} X_{t-h+1}$ denote the best linear predictor of X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$.
- By stationarity, the two sequences

  $$X_t, X_{t-1}, \ldots, X_{t-h+1}$$

  and

  $$X_{t-h}, X_{t-h+1}, \ldots, X_{t-1}$$

  have the same covariance matrix.
- Therefore, the best linear prediction of X_{t-h} in terms of $X_{t-h+1}, \ldots, X_{t-1}$ equals $a_1 X_{t-h+1} + \cdots + a_{h-1} X_{t-1}$.

In Other Words

- $$\begin{aligned} pacf(h) = &\text{corr}(X_t - a_1 X_{t-1} - \cdots - a_{h-1} X_{t-h+1}, \\ &X_{t-h} - a_1 X_{t-h+1} - \cdots - a_{h-1} X_{t-1}). \end{aligned}$$
- $pacf(h)$ is the correlation between the errors in the best linear predictions of X_t and X_{t-h} in terms of the intervening variables $X_{t-1}, \ldots, X_{t-h+1}$.
- That is, the correlation between X_t and X_{t-h} with the effect of the intervening variables $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$ removed.

h>p?

- $pacf(h)$ equals zero for lags $h > p$ for an AR(p) model
- Note that for $h > p$, the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$ equals $\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}$.
- In other words, $a_1 = \phi_1, \ldots, a_p = \phi_p$ and $a_i = 0$ for $i > p$.
- Therefore for $h > p$, we have by causality

$$\begin{aligned} pacf(h) &= \text{corr}(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}, \\ &\qquad X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}) \\ &= \text{corr}\left(W_t, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) \\ &= 0. \end{aligned}$$

So does it work???

- Recall $\rho(2) = \phi^2$ for AR(1).
- We'll briefly show on the board that pacf(2) $= 0$ for AR(1).

Estimating the PACF with data

- Estimate the entries in Δ and ζ by the respective sample autocorrelations
- To choose p for AR(p), a natural approach is to plot the sample pacf.
- As the true pacf for an AR(p) model is zero for lags larger than p, the sample pacf should be close to zero for lags larger than p.
- Similar to Bartlett's Theorem for the autocorrelation function, one can quantify the variability of the pacf function precisely, as the following theorem shows

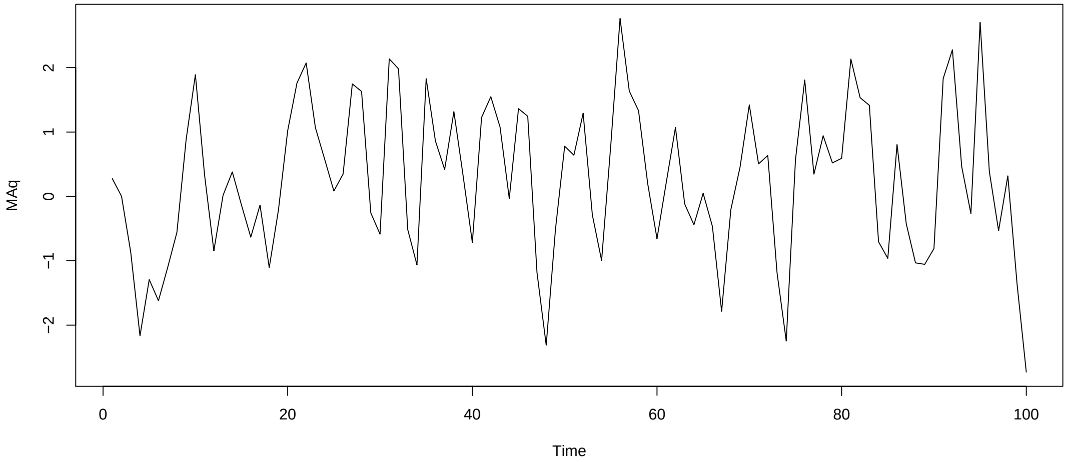
PACF Approximate Distribution

- Theorem: Let $\{X_t\}$ be a causal AR(p) process with i.i.d. noise $\{Z_t\}$. Let p_k denote the sample pacf at lag k defined above. Then for $k > p$ we have that the p_k's are approximately independent normally distributed with mean zero and variance $1/n$.
- Thus for $h > p$, the pacf() plot bands at $\pm 1.96 n^{-1/2}$ can be used for checking if an AR(p) model is appropriate.

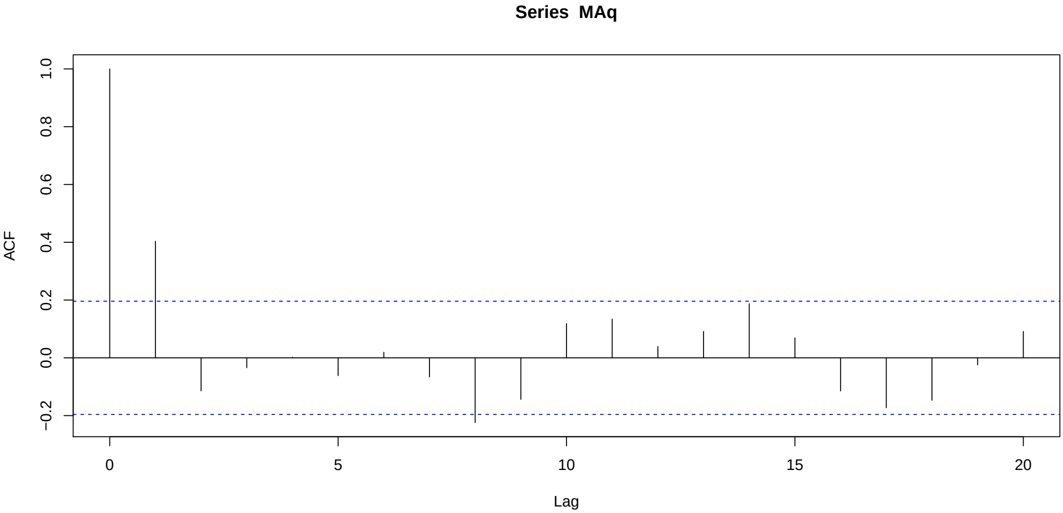
Returning to our example

MA(3)

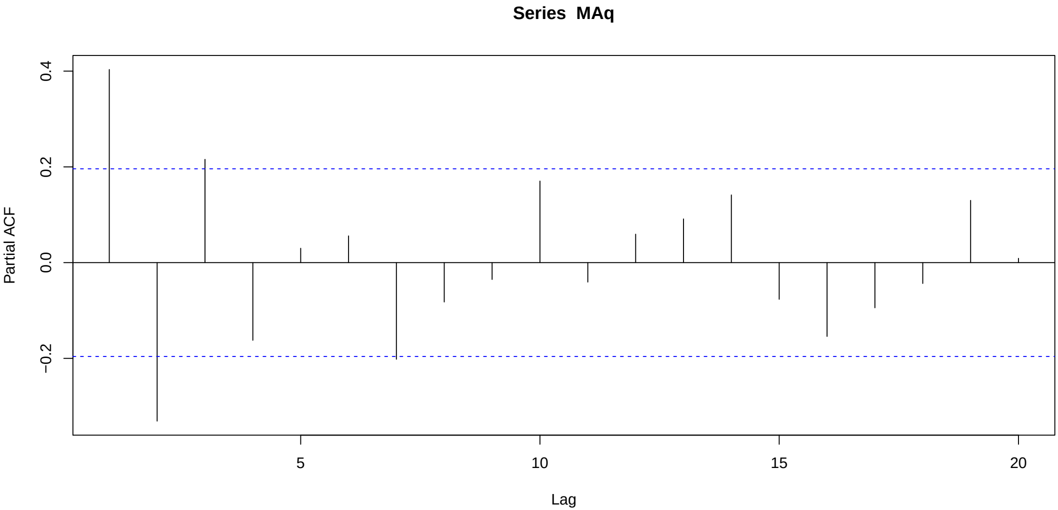
```
MAq = arima.sim(n=100,model=list(ma=c(.9,0,-.2)))
plot.ts(MAq)
```



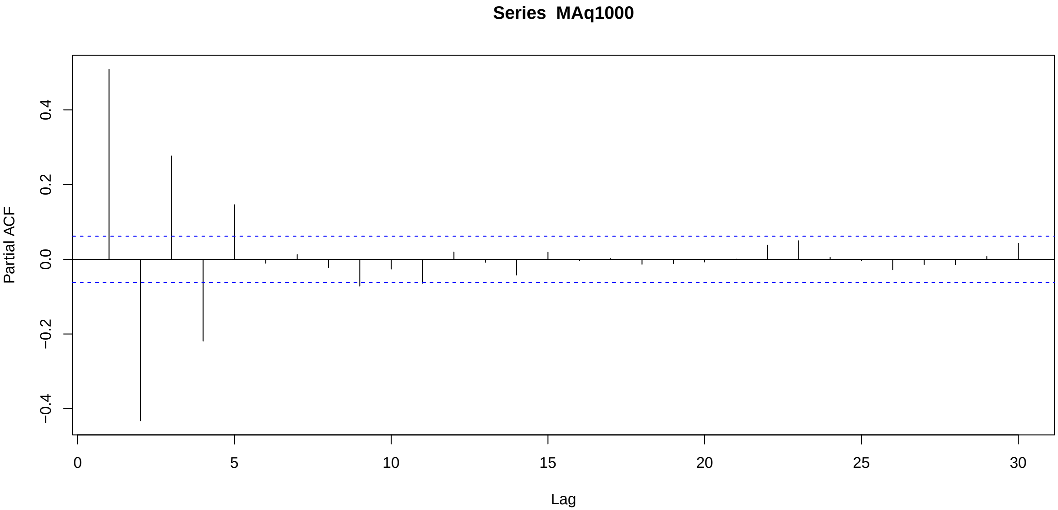
MA(3)



MA(3)

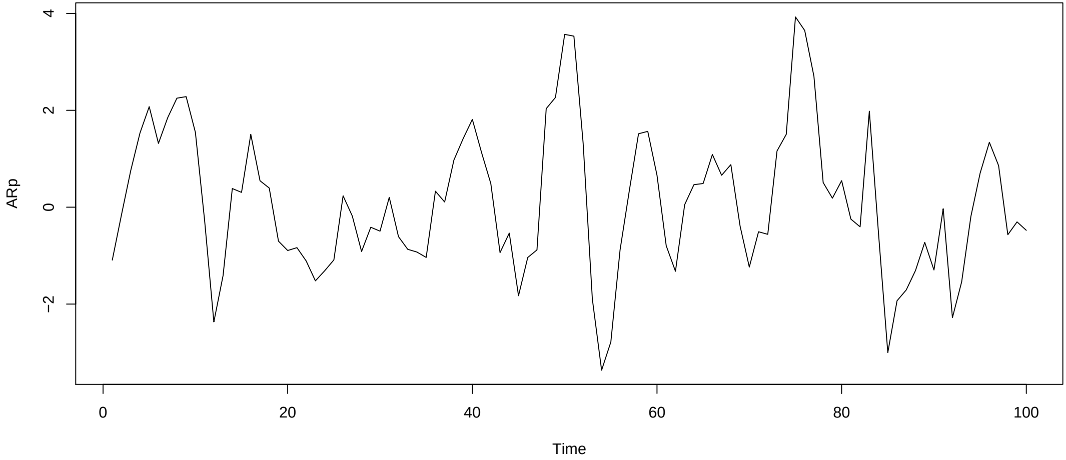


MA(3), n=1000

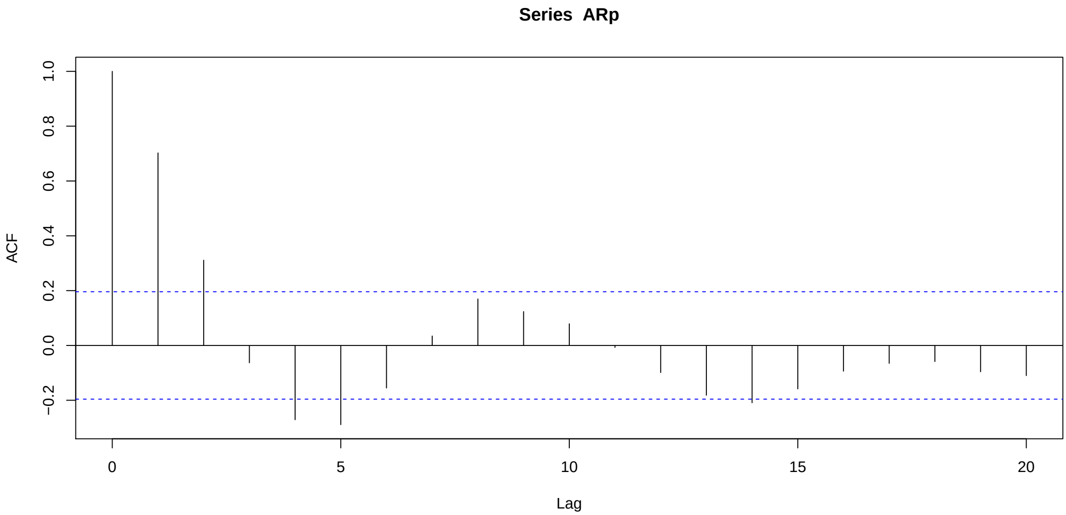


AR(p)

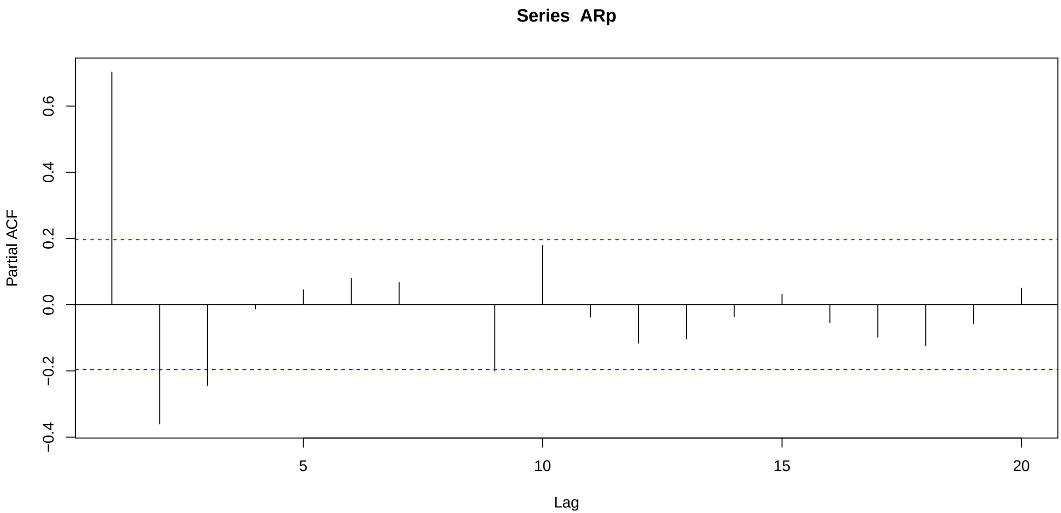
```
ARp = arima.sim(n=100,model=list(ar=c(.9,0,-.2)))
plot.ts(ARp)
```



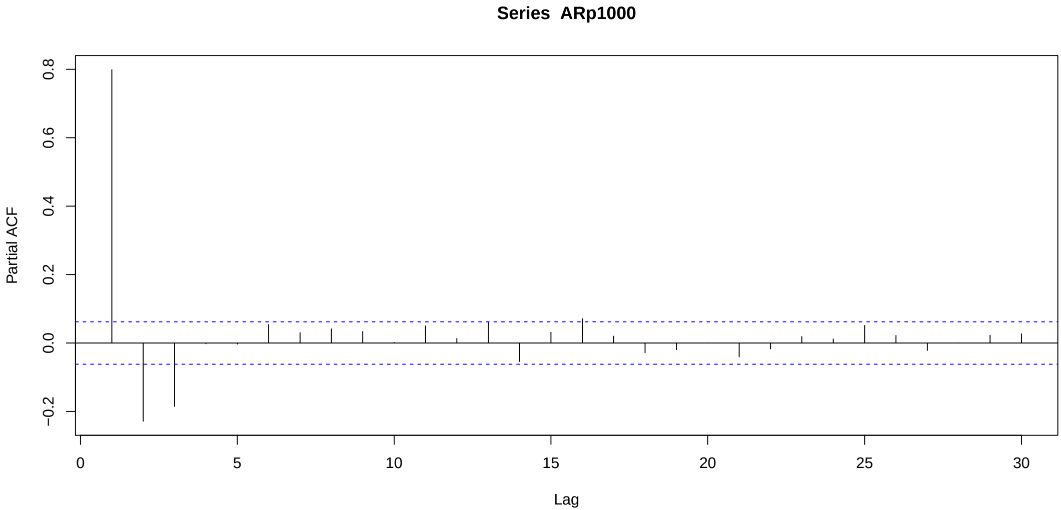
AR(p)



AR(p)

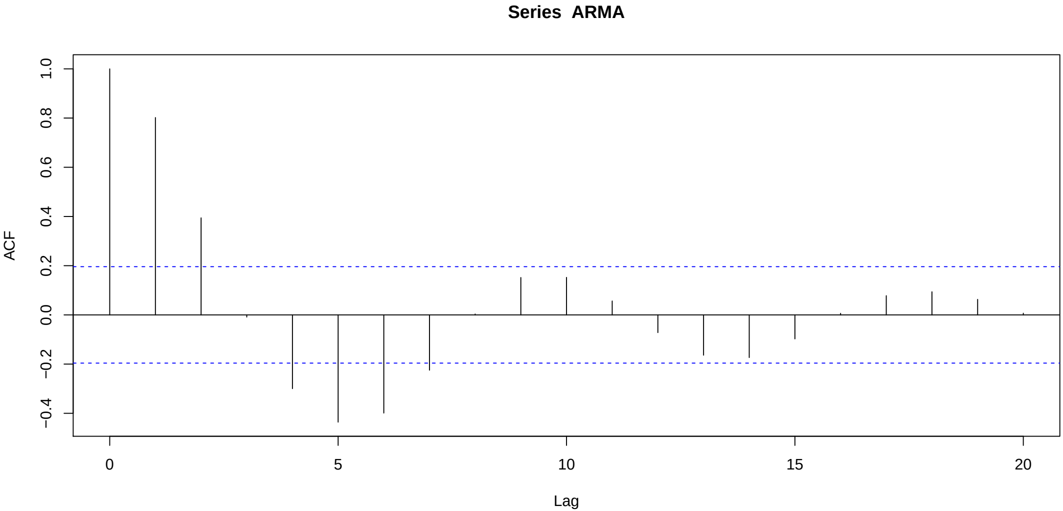


AR(p), n=1000

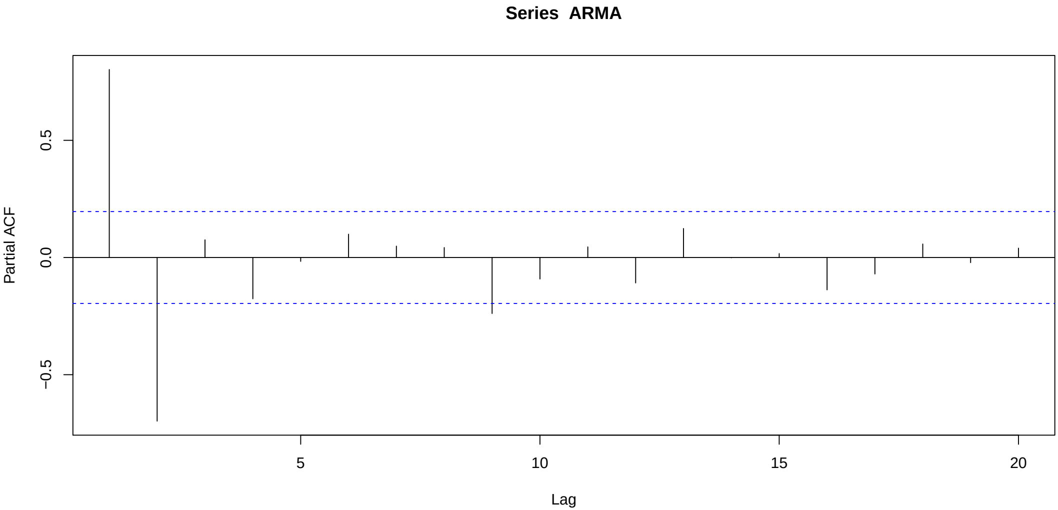

Series ARp1000
Partial ACF
−0.2
0.0
0.2
0.4
0.6
0.8
Lag
0
5
10
15
20
25
30

p=3 (!)

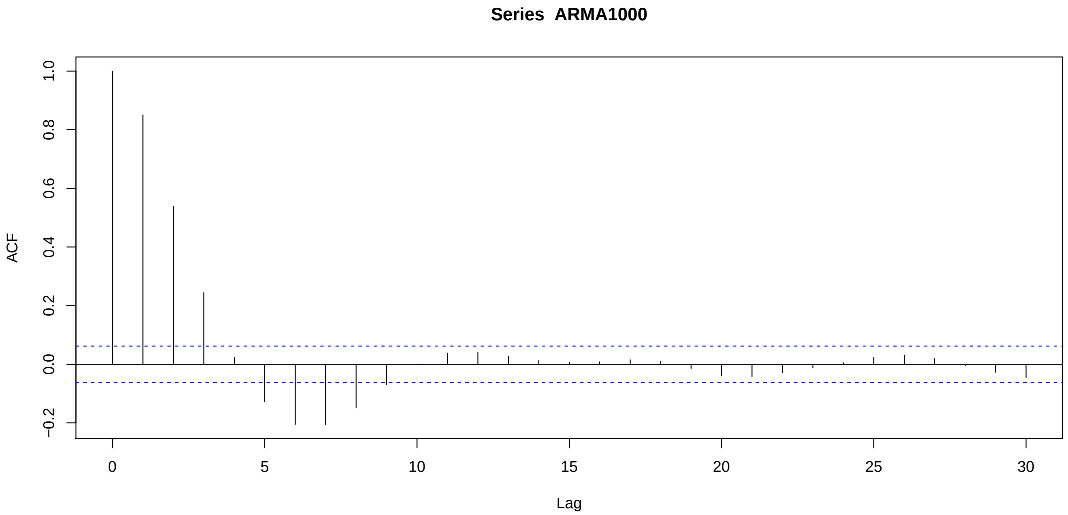
ARMA(p,q)



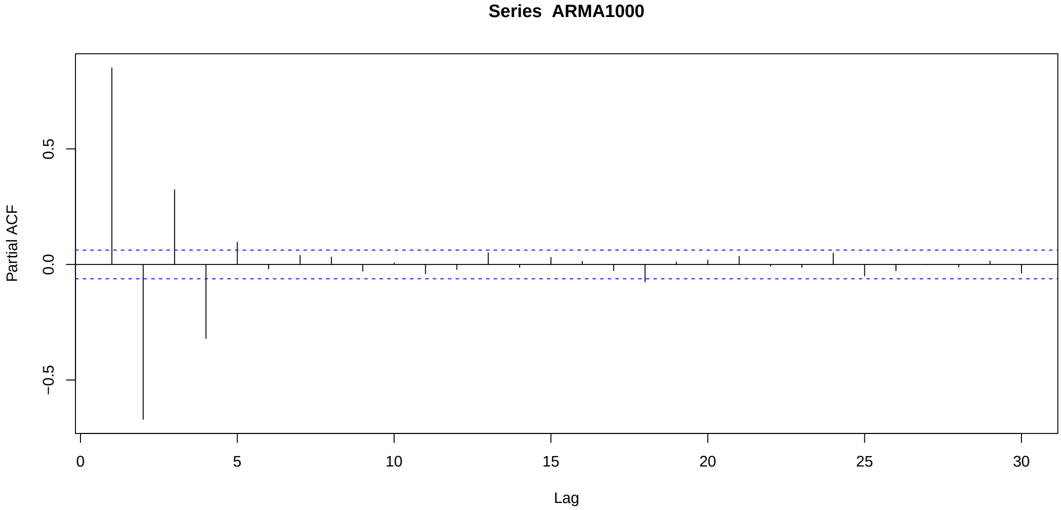
ARMA(p,q)



ARMA(p,q), n=1000



ARMA(p,q), n=1000



p=???, q=???

Conclusion

- For an $MA(q)$ model,
  - The autocorrelation function $\rho_X(h)$ equals zero for $h > q$.
  - Also for $h > q$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance w_{hh}/n where $w_{hh} := 1 + 2\rho^2(1) + \cdots + 2\rho^2(q)$ by Bartlett's Theorem.
- For an $AR(p)$ model,
  - The partial autocorrelation function $pacf(h)$ equals zero for $h > p$
  - For $h > p$, the sample partial autocorrelations are approximately normal with mean 0 and variance $1/n$.
- If the sample acf for a data set cuts off at some lag, we use an MA model. If the sample pacf cuts off at some lag, we use an AR model.
- What if neither of the above happens? In principle this is a model selection problem! To be continued.