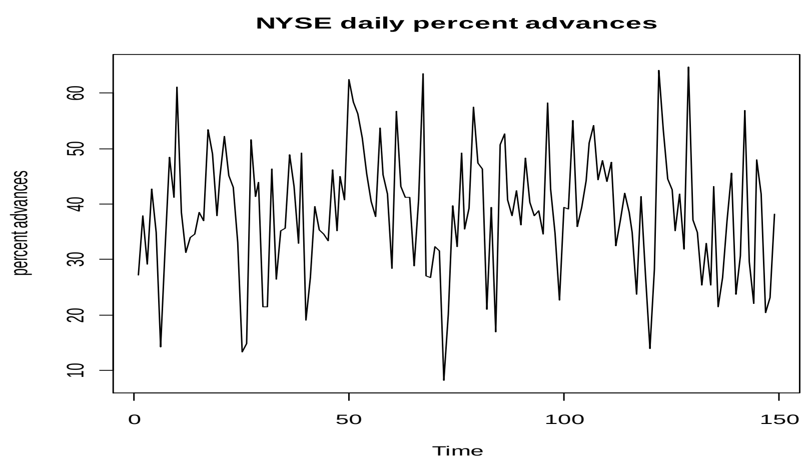
Statistics 435/535/711 23 March 2022
P. Shaman

Fitting ARMA Models in JMP: Percentage Advances on the NYSE

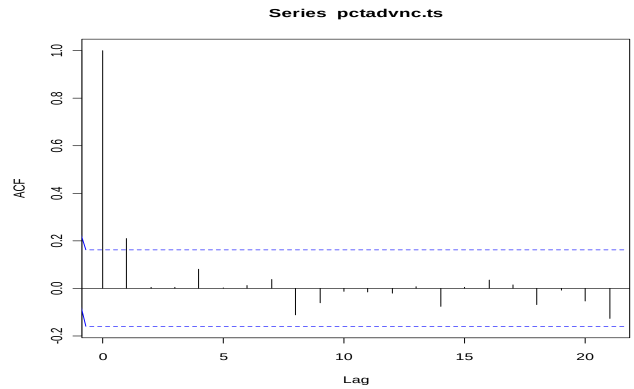
We consider the data set nyseadv.txt, 149 daily observations giving the percentage advances of issues traded on the New York Stock Exchange.

```
> nyseadv<-read.csv("F:/Stat71122Spring/nyseadv.txt")
> attach(nyseadv)
> head(nyseadv)
  pctadvnc
1    27.22
2    37.85
3    29.20
4    42.77
5    34.81
6    14.14

> pctadvnc.ts<-ts(pctadvnc)
> plot(pctadvnc.ts,ylab="percent advances",main="NYSE daily percent
advances")
```

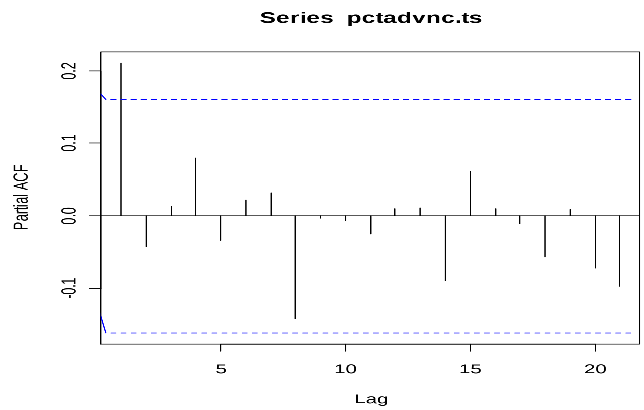


The acf and pacf plots follow.

```
> acf(pctadvnc.ts)
```




```
> pacf(pctadvnc.ts)
```



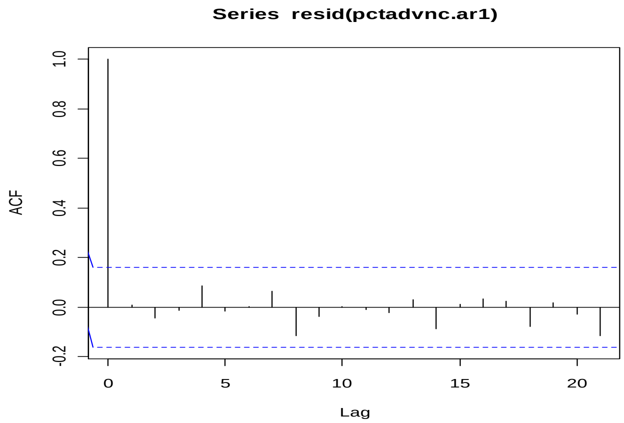
The acf and pacf plots suggest either an AR(1) or an MA(1) fit. These, as well as some overfitting models, are given below. The diagnostics (residual autocorrelations and partial autocorrelations) for both the AR(1) and MA(1) models suggest they are adequate fits.

```
> library("lmtest")
```

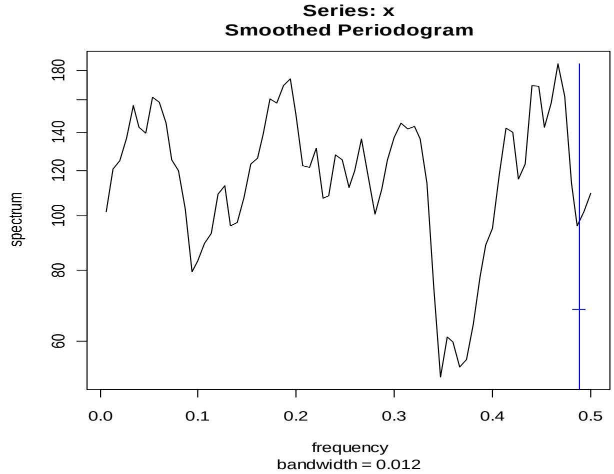
We start with an AR(1) fit.

```
> pctadvnc.ar1<-arima(pctadvnc.ts,order=c(1,0,0))
> coeftest(pctadvnc.ar1)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1        0.210642   0.080039  2.6317  0.008495 **
intercept 38.579566   1.142103 33.7794 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.ar1

sigma^2 estimated as 121.5:  log likelihood = -569.05,  aic = 1144.11
```

Here are the residual autocorrelations:



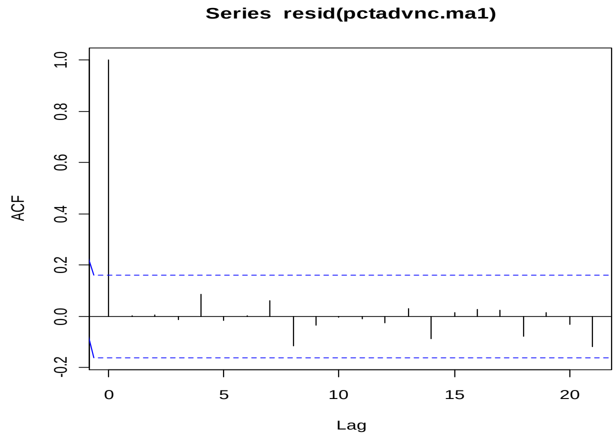
And the residual spectral density:

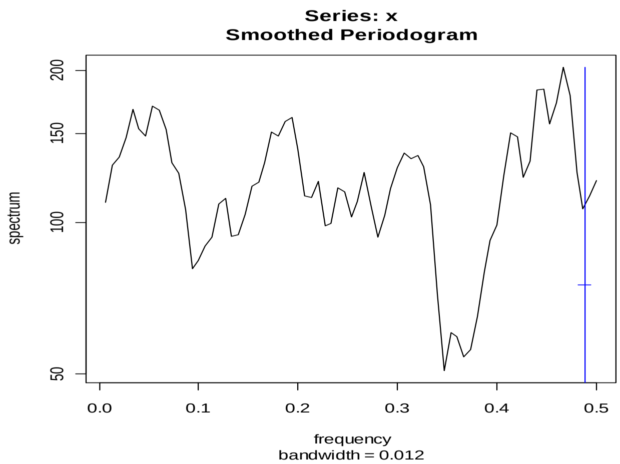


Now, an MA(1) fit.

```
> pctadvnc.ma1<-arima(pctadvnc.ts,order=c(0,0,1))
> coeftest(pctadvnc.ma1)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ma1        0.216791   0.078352  2.7669  0.005659 **
intercept 38.586830   1.096694 35.1847 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.ma1

sigma^2 estimated as 121.3:  log likelihood = -568.93,  aic = 1143.87
```

Here are the residual diagnostics:





The AR(1) and MA(1) residual diagnostics are virtually identical.

An overfit, an AR(2) model, is next.

```
> pctadvnc.ar2<-arima(pctadvnc.ts,order=c(2,0,0))
> coeftest(pctadvnc.ar2)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1        0.219707   0.081871  2.6836  0.007284 **
ar2       -0.042644   0.082134 -0.5192  0.603617
intercept 38.589249   1.095395 35.2286 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.ar2

sigma^2 estimated as 121.3:  log likelihood = -568.92,  aic = 1145.84
```

An ARMA(1,1) fit is another overfit.

```
> pctadvnc.arma11<-arima(pctadvnc.ts,order=c(1,0,1))
> coeftest(pctadvnc.arma11)

z test of coefficients:

           Estimate Std. Error z value Pr(>|z|)
ar1        0.043682   0.326227  0.1339   0.8935
ma1        0.175928   0.319221  0.5511   0.5816
intercept 38.585350   1.108121 34.8205   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.arma11

sigma^2 estimated as 121.3:  log likelihood = -568.92,  aic = 1145.85
```

We explore still two more overfits. First is an ARMA(2,1) model.

```
> pctadvnc.arma21<-arima(pctadvnc.ts,order=c(2,0,1))
> coeftest(pctadvnc.arma21)

z test of coefficients:

           Estimate Std. Error z value Pr(>|z|)
ar1        0.149363   0.853461  0.1750   0.8611
ar2       -0.028200   0.199419 -0.1414   0.8875
ma1        0.070581   0.850887  0.0829   0.9339
intercept 38.588179   1.098090 35.1412   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.arma21

sigma^2 estimated as 121.3:  log likelihood = -568.92,  aic = 1147.83
```

And finally, an ARMA(1,2) model.

```
> pctadvnc.arma12<-arima(pctadvnc.ts,order=c(1,0,2))
> coeftest(pctadvnc.arma12)

z test of coefficients:

          Estimate Std. Error z value Pr(>|z|)
ar1       -0.58140    0.53664 -1.0834   0.2786
ma1        0.80934    0.53631  1.5091   0.1313
ma2        0.16309    0.14138  1.1536   0.2487
intercept 38.58649    1.12302 34.3595   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pctadvnc.arma12

sigma^2 estimated as 121.1:  log likelihood = -568.82,  aic = 1147.63
```

Here is a table of the AIC values for all the fitted models:

Model	AIC
MA(1)	1143.87
AR(1)	1144.11
AR(2)	1145.84
ARMA(1,1)	1145.85
ARMA(1,2)	1147.63
ARMA(2,1)	1147.83

Both the MA(1) and AR(1) models are adequate fits to the data, and AIC slightly favors the former. According to AIC, the overfits contain unneeded additional parameters, and for this dataset the penalty exacted by AIC increases as the number of unneeded additional parameters increases.

Now we turn to interpretation of the various ARMA fits. The AR(1) model can be written

$$(1-0.21064B)(y_t-38.57957)=\varepsilon_t,$$

or

$$y_t-38.57957=0.21064(y_{t-1}-38.57957)+\varepsilon_t.$$

This states that the deviation of today's percentage of advances from the estimate of the long-run average, 38.58 percent, is predicted to be 0.21 times yesterday's deviation. The equation also states that today's deviation is estimated to be 0.21 times yesterday's deviation, plus an unpredictable random shock with standard deviation estimated to be 11.02 (take the square root of 121.5, the variance estimate). There is positive correlation between two successive daily deviations from the estimated long-run average, but it is weak.

The signal is weak (the autoregressive coefficient is close to 0). Today's value cannot be predicted well from the past—there is simply too much new noise (relative to the signal) entering the system each day.

The MA(1) fit is

$$y_t = 38.58683 + (1 + 0.21679B)\varepsilon_t,$$

which can be rewritten in autoregressive form as

$$(1 + 0.21679B)^{-1}(y_t - 38.58683) = \varepsilon_t,$$

or

$$(1 - 0.21679B + 0.04700B^2 - 0.01019B^3 + \cdots)(y_t - 38.58683) = \varepsilon_t,$$

close to the AR(1) fit.

Overfitting is often viewed as a method of confirming that the model we have fit is appropriate. But overfitting should be done properly to avoid trouble (see the discussion below). Both the AR(1) and MA(1) fits seem to be good, and we can consider all of the other fits shown above to be overfits.

The lag 2 coefficient of the AR(2) overfit is not significant. The AR(2) fit is

$$(1 - 0.21971B + 0.04264B^2 + \cdots)(y_t - 38.58925) = \varepsilon_t.$$

This is very close to the MA(1) fit.

Consider the ARMA(1,1) fit. Both estimated ARMA coefficients are insignificant, because neither contributes significantly beyond the impact of the other. The fit is

$$(1 - 0.04368B)(y_t - 38.58535) = (1 + 0.17593B)\varepsilon_t.$$

which is, in autoregressive form,

$$(1 - 0.21961B + 0.03864B^2 + \cdots)(y_t - 38.58535) = \varepsilon_t,$$

essentially the same as the MA(1) and AR(2) fits.

And descriptions of the ARMA(2,1) and ARMA(1,2) overfits follow. The ARMA(2,1) fit is

$$\left(1-0.14936B+0.02820B^2\right)\left(y_t-38.58818\right)=\left(1+0.07058B\right)\varepsilon_t,$$

and after some calculation we obtain its autoregressive representation,

$$\left(1-0.21994B+0.04372B^2+\cdots\right)\left(y_t-38.58818\right)=\varepsilon_t.$$

The ARMA(1,2) model fit is

$$\left(1+0.58140B\right)\left(y_t-38.58649\right)=\left(1+0.80934B+0.16309B^2\right)\varepsilon_t,$$

which is, after some calculation,

$$\left(1-0.22794B+0.02139B^2+\cdots\right)\left(y_t-38.58649\right)=\varepsilon_t.$$

In summary, we can say the following:

- The data are well fit by the AR(1) and MA(1) models. This follows from the AIC criterion, examination of the residual diagnostics, and by looking at the lack of significance of coefficients in the overfits. We could also look at the residual spectra. I prefer the AR(1) model because of its clear interpretation.

- The AR(1) and MA(1) fits are close to each other. This occurs because the lag 1 autoregressive coefficient (lag 1 moving average coefficient) is relatively small in magnitude, here about 0.2. However, when the AR(1) coefficient increases in magnitude, the MA(1) model acquires a different structure from that of the AR(1) model.

- The models above arising from overfitting are seen to be essentially the same as the AR(1) and MA(1) fits. This is seen by writing all of the models in autoregressive form.

There are several different estimation schemes for ARMA parameters (including conditional least squares and maximum likelihood) and calculations of parameter estimates and standard errors performed by different statistical programs do not always agree. In addition, there is not a unique way to report AIC calculations. For AIC, all that matters is the ordering of the AIC values for the different models being considered.

Both the ARMA(1,2) and ARMA(2,1) fits above should not be used. The reason is that the data are properly fit by AR(1) and MA(1) models, and use of ARMA(1,2) and ARMA(2,1) involves adding an extra lag to both the autoregressive and moving average sides of the equation defining the model. Doing this often leads to some awkward results and can be counterproductive—the estimated coefficients may not make sense, and their standard error calculations may be unstable. For the present dataset, the ARMA(1,2) and

ARMA(2,1) fits given here have been shown to have essentially the same structures as the AR(1) and MA(1) fits.

Summary and additional remarks

1. Well-formulated ARMA fits can have multiple representations. For example, MA and ARMA fits can be written in autoregressive form, and AR and ARMA fits can be written in moving average form. Thus, it is often the case that a given data set can be fit equally well by more than one ARMA model. Different model fits are often useful, in that they provide alternative interpretations.

2. After one has fit a valid ARMA model, one may construct overfits to explore whether additional lagged terms are significant. *In doing so, avoid adding extra lags to both sides of the model equation simultaneously*. Unnecessary overfitting will often lead to an inappropriate model. When extra lags are added unnecessarily to both sides of the model equation, parameter estimates often are not interpretable, and calculation of parameter estimate standard errors can become unstable.

3. ARMA models require that the zeros of the autoregressive polynomial be properly positioned, either with magnitudes strictly greater than 1 (if the polynomial is written in ascending order) or strictly less than 1 (if the polynomial is written in descending order). In attempting to fit an ARMA model, sometimes there is failure, in that the iterations don't converge. When this occurs, change what model you are attempting to fit.

4. ARMA models are fit via maximum likelihood estimation, and several options are available to provide starting values for the procedure. It is common to find that different software packages produce slightly different parameter estimates and standard errors for the same model fit.

5. The AIC values produced by different programs may differ by an additive constant.

6. When an ARMA model is fit, it is useful to examine the zeros of both the AR and MA polynomials. If an AR polynomial zero and an MA polynomial are equal or almost equal, there is unnecessary overfitting, and the model should not be used.