Modeling and Analysis of Time Series Data

Chapter 5: Parameter estimation and model identification for ARMA models

Edward L. Ionides

Outline

1. Likelihood-based inference in the context of ARMA models
    - The maximum likelihood estimator
    - Fisher information
    - Profile likelihood confidence intervals
    - Bootstrap standard errors
2. Model selection for ARMA models
    - Likelihood ratio tests
    - Akaike's information criterion (AIC)
3. Fitting ARMA models in R
    - Examining the AR and MA roots
    - Assessing numerical correctness

Background on likelihood-based inference

- For any data $y^*_{1:N}$ and any probabilistic model $f_{Y_{1:N}}(y_{1:N}\,;\theta)$ we define the likelihood function to be

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y^*_{1:N}\,;\theta).$$

- It is often convenient to work with the logarithm to base e of the likelihood, which we write as

$$\ell(\theta) = \log \mathcal{L}(\theta).$$

- Using the likelihood function as a statistical tool is a very general technique, widely used since Fisher (1922) (Wikipedia: Likelihood_function).
- Time series analysis involves various situations where we can, with sufficient care, compute the likelihood function and take advantage of the general framework of likelihood-based inference.

- Computation of the likelihood function for ARMA models is not entirely straightforward.
- Computationally efficient algorithms exist, using a state space model representation of ARMA models that will be developed later in this course.
- For now, it is enough that software exists to evaluate and maximize the likelihood function for a Gaussian ARMA model. Our immediate task is to think about how to use that capability.

- Before evaluation of the ARMA likelihood became routine, it was popular to use a method of moments estimator called **Yule-Walker** estimation (Shumway and Stoffer, 2017, Section 3.5). This is nowadays mostly of historical interest.
- For massively long time series data and big ARMA models, it can be computationally infeasible to work with the likelihood function. However, we are going to focus on the common situation where we can (with due care) work with the likelihood.
- Likelihood-based inference (meaning statistical tools based on the likelihood function) provides tools for parameter estimation, standard errors, hypothesis tests and diagnosing model misspecification.
- Likelihood-based inference often (but not always) has favorable theoretical properties. Here, we are not especially concerned with the underlying theory of likelihood-based inference. On any practical problem, we can check the properties of a statistical procedure by simulation experiments.

The maximum likelihood estimator (MLE)

- A maximum likelihood estimator (MLE) is

$$\hat{\theta}(y_{1:N}) = \arg\max_{\theta} f_{Y_{1:N}}(y_{1:N}\,;\theta),$$

  where $\arg\max_{\theta} g(\theta)$ means a value of argument θ at which the maximum of the function g is attained, so $g\big(\arg\max_{\theta} g(\theta)\big) = \max_{\theta} g(\theta)$.
- If there are many values of θ giving the same maximum value of the likelihood, then an MLE still exists but is not unique.
- The maximum likelihood estimate (also known as the MLE) is

$$\begin{aligned} \hat{\theta} &= \hat{\theta}(y^*_{1:N}) \\ &= \arg\max_{\theta} \mathcal{L}(\theta) \\ &= \arg\max_{\theta} \ell(\theta). \end{aligned}$$

We have said "the" not "a" since most of the time we assume the MLE exists & is unique.

Question 5.1. Why are $\arg\max_\theta \mathcal{L}(\theta)$ and $\arg\max_\theta \ell(\theta)$ the same?

Because log is a strictly increasing function (monotonic).

- We can write $\hat{\theta}_{MLE}$ to denote the MLE if we are considering various alternative estimation methods. However, in this course, we will most often be using maximum likelihood estimation so we let $\hat{\theta}$ correspond to this approach.

Standard errors for the MLE

- As statisticians, it would be irresponsible to present an estimate without a measure of uncertainty!
- Usually, this means obtaining a confidence interval, or an approximate confidence interval.
- It is good to say **approximate** when you present something that is not exactly a confidence interval with the claimed coverage. For example, remind yourself of the definition of a 95% confidence interval.
- Saying "approximate" reminds you that there is some checking that could be done to assess how accurate the approximation is in your particular situation.

Vote: C.I. 95% means exactly 95% coverage 2 ✓

<u>or</u> at least 95% coverage 11 ×

wikipedia: definition means <u>exact</u>

common usage often omits "approximate".

Three ways to quantify statistical uncertainty in an MLE

1. Fisher information. This is computationally quick, but works well only when $\hat{\theta}(Y_{1:N})$ is well approximated by a normal distribution.
2. Profile likelihood estimation. This is a bit more computational effort, but generally is preferable to the Fisher information.
3. A simulation study, also known as a bootstrap.

Standard errors via the observed Fisher information

- We suppose that $\theta \in \mathbb{R}^D$ and so we can write $\theta = \theta_{1:D}$.
- The Hessian matrix of a function is the matrix of its second partial derivatives. We write the Hessian matrix of the log likelihood function as $\nabla^2 \ell(\theta)$, a $D \times D$ matrix whose (i, j) element is
$$\left[\nabla^2 \ell(\theta)\right]_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\theta).$$
- The observed Fisher information is
$$\hat{I} = -\nabla^2 \ell(\hat{\theta}).$$
- A standard asymptotic approximation to the distribution of the MLE for large N is
$$\hat{\theta}(Y_{1:N}) \underline{\approx} N\left[\theta, \hat{I}^{-1}\right],$$
where θ is the true parameter value. This asserts that the MLE is asymptotically unbiased, with variance asymptotically attaining the Cramer-Rao lower bound.

- Since the MLE attains the Cramer-Rao lower bound, under regularity conditions, we it is **asymptotically efficient**.
- We can interpret $\approx$ in the above normal approximation to mean "one could write a limit statement formally justifying this approximation in a suitable limit." Almost equivalently, $\approx$ can mean "this approximation is useful in the finite sample situation at hand."
- A corresponding approximate 95% confidence interval for θ_d is $\hat{\theta}_d \pm 1.96\big(\big[\hat{I}^{-1}\big]_{dd}\big)^{1/2}$. The R function `arima` computes standard errors for the MLE of an ARMA model in this way.
- We usually only have one time series, with some fixed N, and so we cannot in practice take $N \to \infty$. When our time series model is non-stationary it may not even be clear what it would mean to take $N \to \infty$. These asymptotic results should be viewed as nice mathematical reasons to consider computing an MLE, but not a substitute for checking how the MLE behaves for our model and data.

Confidence intervals via the profile likelihood

All but one of the class have not seen profile likelihood before.

- We consider the problem of obtaining a confidence interval for θ_d, the dth component of $\theta_{1:D}$.
- The **profile log likelihood function** of θ_d is defined to be

$$\ell_d^{\mathrm{profile}}(\theta_d) = \max_{\phi\in\mathbb{R}^D:\phi_d=\theta_d} \ell(\phi).$$

  In general, the profile likelihood of one parameter is constructed by maximizing the likelihood function over all other parameters.
- Check that $\max_{\theta_d} \ell_d^{\mathrm{profile}}(\theta_d) = \max_{\theta_{1:D}} \ell(\theta_{1:D})$. Maximizing the profile likelihood $\ell_d^{\mathrm{profile}}(\theta_d)$ gives the MLE, $\hat{\theta}_d$.
- An approximate 95% confidence interval for θ_d is given by

$$\big\{\theta_d : \ell(\hat{\theta}) - \ell_d^{\mathrm{profile}}(\theta_d) < 1.92\big\}.$$

$1.92 = \frac{1.96^2}{2}$

- This is known as a profile likelihood confidence interval.

Where does the 1.92 cutoff come from

- The cutoff 1.92 is derived using **Wilks's theorem**, which we will discuss in more detail when we develop likelihood ratio tests.
- Note that $1.92 = \frac{1.96^2}{2}$.
- The asymptotic justification of Wilks's theorem is the same limit that justifies the Fisher information standard errors.
- Profile likelihood confidence intervals tend to work better than Fisher information confidence intervals when the log likelihood function is not close to quadratic near its maximum. This is more common when N is not large.

A Simulation study, also called bootstrap

- If done carefully and well, this can be the best approach.
- A confidence interval is a claim about reproducibility. You claim, so far as your model is correct, that on 95% of realizations from the model, a 95% confidence interval you have constructed will cover the true value of the parameter.
- A simulation study can check this claim directly.
- The simulation study takes time to develop and debug, time to explain, and time for the reader to understand and check what you have done. We usually carry out simulation studies to check our main conclusions only.

Bootstrap methods for constructing standard errors and confidence intervals

- Suppose we want to know the statistical behavior of the estimator $\hat{\theta}(y_{1:N})$ for models in a neighborhood of the MLE.
- In particular, let's consider the problem of estimating uncertainty about θ_1, the first component of the vector θ.
- We use simulation to assess the behavior of the maximum likelihood estimator, $\hat{\theta}_1(y_{1:N})$, and possibly the coverage of an associated confidence interval estimator, $\big[\hat{\theta}_{1,\mathrm{l}o}(y_{1:N}),\hat{\theta}_{1,\mathrm{h}i}(y_{1:N})\big]$.
- The confidence interval estimator could be constructed using either the Fisher information method or the profile likelihood approach.

- We can design a simulation study to address the following goals:

(A) Evaluate the coverage of a proposed confidence interval estimator, $[\hat{\theta}_{1,\mathrm{l}o}, \hat{\theta}_{1,\mathrm{h}i}]$,

(B) Construct a standard error for $\hat{\theta}_1$,

(C) Construct a confidence interval for θ_1 with exact local coverage.

An approximate C.I. may have different coverage at different values of θ. A simulation study can guarantee exact coverage at the value of θ where the simulation study was carried out. Such a method likely has close to exact coverage in a local neighborhood of θ.

A simulation study

1. Generate J independent Monte Carlo simulations,

$$Y_{1:N}^{[j]} \sim f_{Y_{1:N}}(y_{1:N}\,;\hat{\theta}) \text{ for } j \in 1:J.$$

2. For each simulation, evaluate the maximum likelihood estimator,

$$\hat{\theta}^{[j]} = \hat{\theta}\big(Y_{1:N}^{[j]}\big) \text{ for } j \in 1:J,$$

and, if desired, the confidence interval estimator,

$$\big[\hat{\theta}_{1,\mathrm{lo}}^{[j]}, \hat{\theta}_{1,\mathrm{h}i}^{[j]}\big] = \big[\hat{\theta}_{1,\mathrm{lo}}(Y_{1:N}^{[j]}), \hat{\theta}_{1,\mathrm{h}i}(Y_{1:N}^{[j]})\big].$$

3. For large J, the coverage of the proposed confidence interval is well approximated, for models in a neighborhood of $\hat{\theta}$, by the proportion of the intervals $\big[\hat{\theta}_{1,\mathrm{lo}}^{[j]}, \hat{\theta}_{1,\mathrm{h}i}^{[j]}\big]$ that include $\hat{\theta}_1$.

4. The sample standard deviation of $\{\hat{\theta}_1^{[j]}, j \in 1:J\}$ is a natural standard error to associate with $\hat{\theta}_1$.

Likelihood ratio tests for nested hypotheses

- The whole parameter space on which the model is defined is $\Theta \subset \mathbb{R}^D$.
- Suppose we have two **nested** hypotheses

$$\begin{aligned} H^{\langle 0 \rangle} &: \theta \in \Theta^{\langle 0 \rangle}, \\ H^{\langle 1 \rangle} &: \theta \in \Theta^{\langle 1 \rangle}, \end{aligned}$$

  defined via two nested parameter subspaces, $\Theta^{\langle 0 \rangle} \subset \Theta^{\langle 1 \rangle}$, with respective dimensions $D^{\langle 0 \rangle} < D^{\langle 1 \rangle} \le D$.

- We consider the log likelihood maximized over each of the hypotheses,

$$\begin{aligned} \ell^{\langle 0 \rangle} &= \sup_{\theta \in \Theta^{\langle 0 \rangle}} \ell(\theta), \\ \ell^{\langle 1 \rangle} &= \sup_{\theta \in \Theta^{\langle 1 \rangle}} \ell(\theta). \end{aligned}$$

sup & max are equivalent here.

- A useful approximation asserts that, under the hypothesis $H^{\langle 0\rangle}$,
$$\ell^{\langle 1\rangle} - \ell^{\langle 0\rangle} \approx (1/2)\chi^2_{D^{\langle 1\rangle}-D^{\langle 0\rangle}},$$
where χ^2_d is a chi-squared random variable on d degrees of freedom and $\approx$ means "is approximately distributed as."
- We will call this the **Wilks approximation**.
- The Wilks approximation can be used to construct a hypothesis test of the null hypothesis $H^{\langle 0\rangle}$ against the alternative $H^{\langle 1\rangle}$.
- This is called a **likelihood ratio test** since a difference of log likelihoods corresponds to a ratio of likelihoods.
- When the data are iid, $N \to \infty$, and the hypotheses satisfy suitable regularity conditions, this approximation can be derived mathematically and is known as **Wilks's theorem**.
- The chi-squared approximation to the likelihood ratio statistic may be useful, and can be assessed empirically by a simulation study, even in situations that do not formally satisfy any known theorem.

Using a likelihood ratio test to construct profile likelihood confidence intervals

- Recall the duality between hypothesis tests and confidence intervals:

  The estimated parameter θ^* does not lead us to reject a null hypothesis of $\theta=\theta^{\langle 0\rangle}$ at the 5% level

  $$\Updownarrow$$

  $\theta^{\langle 0\rangle}$ is in a 95% confidence interval for θ.

- We can check what the 95% cutoff is for a chi-squared distribution with one degree of freedom,

```
qchisq(0.95,df=1)
[1] 3.841459
```

- We can now see how the Wilks approximation suggests a confidence interval constructed from parameter values having a profile likelihood within 1.92 log units of the maximum.

Akaike's information criterion (AIC)

- Likelihood ratio tests provide an approach to model selection for nested hypotheses, but how about when models are not nested?
- A more general approach is to compare likelihoods of different models by penalizing the likelihood of each model by a measure of its complexity.
- Akaike's information criterion **AIC** is given by

  $$AIC = -2 \times \ell(\theta^*) + 2D$$

  "Minus twice the maximized log likelihood plus twice the number of parameters."
- We are invited to select the model with the lowest AIC score.
- AIC was derived as an approach to minimizing prediction error. Increasing the number of parameters leads to additional **overfitting** which can decrease predictive skill of the fitted model.

A caution for using AIC

- Viewed as a hypothesis test, AIC may have weak statistical properties.
- It is a mistake to interpret AIC by making a claim that the favored model has been shown to provides a superior explanation of the data.
- However, viewed as a way to select a model with reasonable predictive skill from a range of possibilities, it is often useful.

Comparing AIC with likelihood ratio tests

Wilks : $\ell^{\langle 1\rangle} - \ell^{\langle 0\rangle} \approx \frac{1}{2}\chi^2_{D^{\langle 1\rangle}-D^{\langle 0\rangle}}$

Question 5.2. Suppose we are in a situation in which we wish to choose between two nested hypotheses, with dimensions $D^{\langle 0\rangle} < D^{\langle 1\rangle}$. Suppose the Wilks approximation is valid. Consider the strategy of selecting the model with the lowest AIC value, and view this model selection approach as a formal statistical test.

(A) Find an expression for the size of this AIC test (i.e, the probability of rejecting the null hypothesis, $H^{\langle 0\rangle}$, when this null hypothesis is true).

(B) Evaluate this expression for $D^{\langle 1\rangle} - D^{\langle 0\rangle} = 1$.

$\Delta AIC = AIC^{\langle 0\rangle} - AIC^{\langle 1\rangle}$

(A). Difference of AIC, $\Delta AIC = 2(\ell^{\langle 1\rangle} - \ell^{\langle 0\rangle}) - 2(D^{\langle 1\rangle} - D^{\langle 0\rangle})$

$$\frac{1}{2}\Delta AIC + D^{\langle 1\rangle} - D^{\langle 0\rangle} \approx \frac{1}{2}\chi^2_{D^{\langle 1\rangle}-D^{\langle 0\rangle}}$$

AIC rule: $\mathbb{P}_{H^{\langle 0\rangle}}[\Delta AIC > 0] = \mathbb{P}\left[\frac{1}{2}\chi^2_{D^{\langle 1\rangle}-D^{\langle 0\rangle}} > D^{\langle 1\rangle} - D^{\langle 0\rangle}\right]$

(B). If $D_{\langle 1\rangle} - D_{\langle 0\rangle} = 1$,

$$\mathbb{P}_{H^{\langle 0\rangle}}[\Delta AIC > 0] = \mathbb{P}[\chi^2_1 > 2] = 1 - \text{pchisq}(2, df=1)$$
$$= 0.157.$$

Likelihood-based inference for ARMA models in R

- The Great Lakes are an important resource for leisure, agriculture and industry in this region.
- A past concern has been whether human activities such as water diversion or channel dredging might be leading to a decline in lake levels.
- A current concern has been high levels leading to coastal erosion.
- Are lake levels affected by climate change?
- The physical mechanisms are not always obvious: for example, evaporation tends to be highest when the weather is cold but the lake is not ice-covered.
- We look at monthly time series data on the level of Lake Huron, which is essentially the same as Lake Michigan.

Reading in the data

Here is the head of the file huron_level.csv

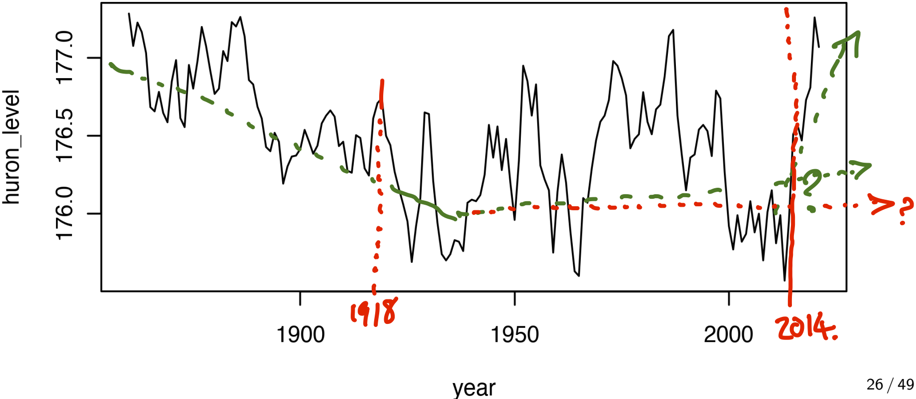
```
# Lake Michigan-Huron:, Monthly Average Water Levels (meters)
# 1860-1917 : Harbor Beach, MI
# https://www.glerl.noaa.gov/data/dashboard/data/levels/1860_1917/miHuron1860.csv
# 1918+ : Monthly Lake-Wide Average Water Level
# https://www.glerl.noaa.gov/data/dashboard/data/levels/1918_PRES/miHuron1918.csv
# Source:, NOAA/NOS; CHS. Downloaded: Jan 20, 2022
year,jan,feb,mar,apr,may,jun,jul,aug,sep,oct,nov,dec
1860,177.285,177.339,177.349,177.388,177.425,177.461,177.473,177.416,177.355,177.26
1861,177.077,177.105,177.224,177.254,177.382,177.431,177.47,177.544,177.449,177.413
```

```r
dat <- read.table(file="huron_level.csv",sep=",",header=TRUE)
head(dat[,1:7],2)
```

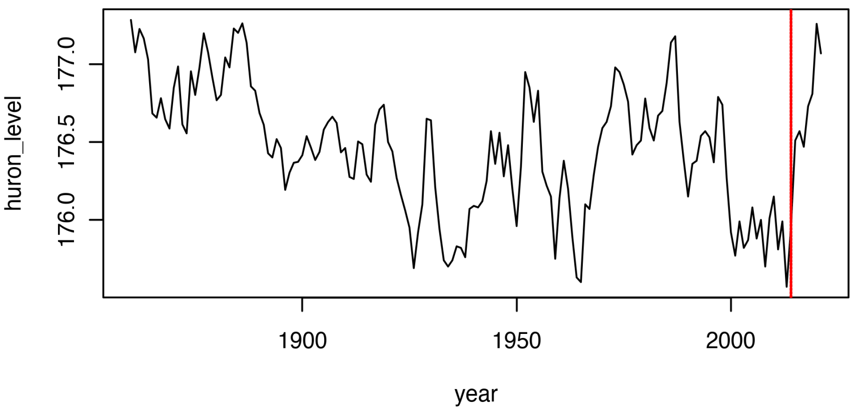
```
 year     jan     feb     mar     apr     may     jun
 1860 177.285 177.339 177.349 177.388 177.425 177.461
 1861 177.077 177.105 177.224 177.254 177.382 177.431
```

For now, we avoid monthly seasonal variation by considering an annual series of January depths. We will investigate seasonal variation later in the course, but sometimes it is best avoided.

```
huron_level <- dat$jan
year <- dat$year
plot(huron_level~year,type="l")
```



- Until the recent surge in water level, there was concern about a long-run decline in lake level due to dredging or water diversion or climate change.
- We put ourselves back in 2014 and temporarily ignore subsequent data



Fitting an ARMA model

- Later, we will consider hypotheses of trend. For now, let's start by fitting a stationary ARMA(p,q) model under the null hypothesis that there is no trend. This hypothesis, which asserts that nothing has substantially changed in this system over the last 160 years, is not entirely unreasonable from looking at the data.
- We seek to fit a stationary Gaussian ARMA(p,q) model with parameter vector $\theta = (\phi_{1:p}, \psi_{1:q}, \mu, \sigma^2)$ given by

$$\phi(B)(Y_n - \mu) = \psi(B)\epsilon_n,$$

where

$$\begin{array}{rcl} \mu &=& \mathbb{E}[Y_n] \\ \phi(x) &=& 1 - \phi_1 x - \cdots - \phi_p x^p, \\ \psi(x) &=& 1 + \psi_1 x + \cdots + \psi_q x^q, \\ \epsilon_n &\sim& \mathrm{iid}\, N[0, \sigma^2]. \end{array}$$

Choosing p and q

ARMA(p,q) is nested within ARMA(p',q') when $p \le p'$ and $q \le q'$

- We need to decide where to start in terms of values of p and q.
- We tabulate AIC values for a range of different choices of p and q.

```
aic_table <- function(data,P,Q){
  table <- matrix(NA,(P+1),(Q+1))
  for(p in 0:P) {
    for(q in 0:Q) {
       table[p+1,q+1] <- arima(data,order=c(p,0,q))$aic
    }
  }
  dimnames(table) <- list(paste("AR",0:P, sep=""),
    paste("MA",0:Q,sep=""))
  table
}
huron_aic_table <- aic_table(huron_level,4,5)
require(knitr)
kable(huron_aic_table,digits=2)
```

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	174.67	50.63	11.69	-11.90	-15.83	-20.21
AR1	-33.15	-33.52	-31.71	-29.73	-28.85	-27.06
AR2	-33.30	-31.85	-32.03	-30.03	-27.12	-26.89
AR3	-31.71	-30.08	-28.10	-28.19	-25.54	-24.97
AR4	-29.80	-30.08	-28.53	-30.15	-26.93	-24.94

Mathematically inconsistent: computational issue with maximization and/or evaluation.

Question 5.3. What do we learn by interpreting the results in the above table of AIC values? Taken at face value, the recommended model is ARMA(1,1), which has lowest AIC. We do not have to follow the recommendation of AIC, but we probably want to choose from models with competitive AIC. Simplicity & stability are not fully valued by AIC; our goal is not simply to predict.

Question 5.4. In what ways might we have to be careful not to over-interpret the results of this table?

Look for evidence that AIC is not correctly calculated.

Mathematically AIC for ARMA ... If $p \le p'$, $q \le q'$

$$AIC(ARMA(p',q')) < AIC(ARMA(p,q)) + 2[(p'-p)+(q'-q)]$$