5 Parameter estimation and model identification for ARMA models

Edward Ionides

1/23/2018

Contents

Objectives

1. Develop likelihood-based inference in the context of ARMA models.

2. Discuss maximum likelihood parameter estimation and alternative methods.

3. Investigate strategies for model selection, also known as model identification, in the context of ARMA models.

4. Work on practical computational approaches for implementing these methods.

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).

- 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.

1

- 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. This is described by Shumway and Stoffer (Section 3.6) but is nowadays mostly of historical interest.

- There are occasionally time series situations where massively long data or massively complex models mean that it is 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

$$
\begin{aligned}
\hat{\theta}^* &= \hat{\theta}(y_{1:N}^*) & (1)\\
&= \arg\max_\theta \mathcal{L}(\theta) & (2)\\
&= \arg\max_\theta \ell(\theta). & (3)
\end{aligned}
$$

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

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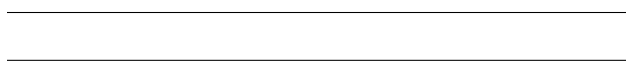
Standard errors for the MLE

- 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.

  - It also helps to remind you that it may be interesting and relevant to explain why the interval you present is an approximate confidence interval rather than an exact one.

- There are three main approaches to estimating the statistical uncertainty in an MLE.

1. The 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.

   - 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 fairly directly, but requires the most effort.

   - The simulation study takes time for you to develop and debug, time for you 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.

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

$$\hat{\theta}(Y_{1:N}) \approx N \left[\theta, [\hat{I}^*]^{-1} \right],$$

where θ is the true parameter value. This asserts that the MLE is asymptotically unbiased, with variance asymptotically attaining the Cramer-Rao lower bound. Thus, we say the MLE is **asymptotically efficient**.

- A corresponding approximate 95% confidence interval for θ_d is

$$\theta_d^* \pm 1.96 \big[\hat{I}^{*-1}\big]_{dd}^{1/2}.$$

- 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

- Let's 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

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

- This is known as a profile likelihood confidence interval. The cutoff 1.92 is derived using Wilks's theorem, which we will discuss in more detail when we develop likelihood ratio tests.

- Although 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 N is not so large—particularly when the log likelihood function is not close to quadratic near its maximum.

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, $\theta^* = \hat{\theta}(y_{1:N}^*)$.

- In particular, let's consider the problem of estimating uncertainty about θ_1. We want to assess the behavior of the maximum likelihood estimator, $\hat{\theta}(y_{1:N})$, and possibly the coverage of an associated confidence interval estimator, $[\hat{\theta}_{1,\mathrm{lo}}(y_{1:N}), \hat{\theta}_{1,\mathrm{hi}}(y_{1:N})]$. The confidence interval estimator could be constructed using either the Fisher information method or the profile likelihood approach.

- The following simulation study lets us address the following goals:

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

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

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

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,

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

and, if desired, the confidence interval estimator,

$$[\theta_{1,lo}^{[j]}, \theta_{1,hi}^{[j]}] = [\hat{\theta}_{1,lo}(X_{1:N}^{[j]}), \hat{\theta}_{1,hi}(X_{1:N}^{[j]})].$$

3. We can use these simulations to obtain solutions to our goals for uncertainty assessment:

(A) For large J, the coverage of the proposed confidence interval estimator is well approximated, for models in a neighborhood of $\hat{\theta}^*$, by the proportion of the intervals $[\theta_{1,lo}^{[j]}, \theta_{1,hi}^{[j]}]$ that include $\hat{\theta}_1^*$.

(B) The sample standard deviation of $\{\theta_1^{[j]}, j \in 1:J\}$ is a natural standard error to associate with $\hat{\theta}_1^*$.

(C) For large J, one can empirically calibrate a 95% confidence interval for θ_1 with exactly the claimed coverage in a neighborhood of $\hat{\theta}^*$. For example, using profile methods, one could replace the cutoff 1.92 by a constant α chosen such that 95% of the profile confidence intervals computed for the simulations cover $\hat{\theta}_1^*$.

Question: Local coverage as an approximation to actual coverage for a confidence interval

- A true 95% confidence interval covers θ with probability 0.95 whatever the value of θ.

- The local coverage probability at a value $\theta = \tilde{\theta}$ is the chance that the confidence interval covers $\tilde{\theta}$ when the true parameter value is $\tilde{\theta}$. Typically, we compute local coverage at $\theta = \hat{\theta}^*$.

- Local coverage can be evaluated or calibrated via simulation; the actual (global) coverage is usually hard to work with.

- What properties of the model and data make local coverage a good substitute for global coverage? How would you check whether or not these properties hold?

Likelihood-based model selection and model diagnostics

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

$$
\begin{align}
H^{\langle 0 \rangle} \quad &: \quad \theta \in \Theta^{\langle 0 \rangle}, \tag{4}\\
H^{\langle 1 \rangle} \quad &= \quad \theta \in \Theta^{\langle 1 \rangle}, \tag{5}
\end{align}
$$

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

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

$$
\begin{aligned}
\ell^{\langle 0 \rangle} &= \sup_{\theta \in \Theta^{\langle 0 \rangle}} \ell(\theta), && (6) \\
\ell^{\langle 1 \rangle} &= \sup_{\theta \in \Theta^{\langle 1 \rangle}} \ell(\theta). && (7)
\end{aligned}
$$

- 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.

Exercise: 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 withing 1.92 log units of the maximum.

- It is a exercise to write out more details (to your own satisfaction) on how to use the Wilks approximation, together with the duality between hypothesis tests and confidence intervals, to derive a profile likelihood confidence interval.

Akaike's information criterion (AIC)

- Likelihood ratio tests provide an approach to model selection for nested hypotheses, but what do we do 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.

- Viewed as a hypothesis test, AIC may have weak statistical properties. It can be 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.

Exercise: Comparing AIC with likelihood ratio tests

- 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.

- We can view this model selection approach as a formal statistical test.

- 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).

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

Implementing 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.

- An additional current concern is the effects of 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 depth of Lake Huron.

Reading in the data

```
system("head huron_depth.csv")
```

- A bit of work has to be done manipulating the `Date` variable.
  - Moving between date formats is a necessary skill for time series analysis!
  - A standard representation of time is `POSIXct`, which is a signed real number representing the number of seconds since the beginning of 1970.
  - The raw data have a character string representing date. We convert this into the standard format using `strptime`. Than we can extract whatever we need. See `?DateTimeClasses` for more on manipulating date and time formats in R.

```
dat <- read.table(file="huron_depth.csv",sep=",",header=TRUE)
dat$Date <- strptime(dat$Date,"%m/%d/%Y")
dat$year <- as.numeric(format(dat$Date, format="%Y"))
dat$month <- as.numeric(format(dat$Date, format="%m"))
head(dat)
```

```
##          Date Average year month
## 1 1860-01-01 177.285 1860     1
## 2 1860-02-01 177.339 1860     2
## 3 1860-03-01 177.349 1860     3
## 4 1860-04-01 177.388 1860     4
## 5 1860-05-01 177.425 1860     5
## 6 1860-06-01 177.461 1860     6
```

- For now, let's 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.

```
dat <- subset(dat,month==1)
huron_depth <- dat$Average
year <- dat$year
plot(huron_depth~year,type="l")
```
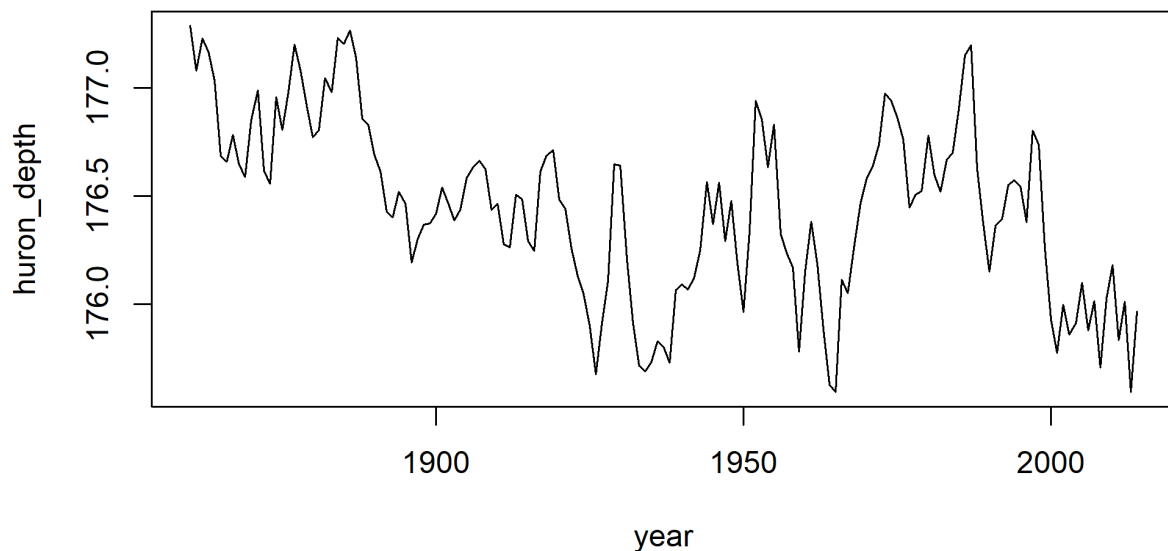
Fitting an ARMA model

- 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 150 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

$$
\begin{aligned}
\mu &= \mathbb{E}[Y_n] & (8) \\
\phi(x) &= 1 - \phi_1 x - \cdots - \phi_p x^p, & (9) \\
\psi(x) &= 1 + \psi_1 x + \cdots + \psi_q x^q, & (10) \\
\epsilon_n &\sim \text{iid } N[0, \sigma^2]. & (11)
\end{aligned}
$$

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

- In the code below, note the use of `kable` for formatting HTML tables when using Rmarkdown. The `"<b>"` and `""</b>"` tags in `dimnames` make the rownames boldface in HTML. By default, only column names are boldface in standard HTML.

```
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("<b> AR",0:P, "</b>", sep=""),paste("MA",0:Q,sep=""))
  table
}
huron_aic_table <- aic_table(huron_depth,4,5)
require(knitr)
kable(huron_aic_table,digits=2)
```

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	166.75	46.60	7.28	-14.97	-18.64	-26.09
AR1	-38.00	-37.41	-35.46	-33.82	-34.13	-32.20
AR2	-37.33	-38.43	-36.90	-34.93	-34.35	-33.08
AR3	-35.52	-35.17	-32.71	-31.38	-33.21	-32.98
AR4	-33.94	-34.91	-34.43	-37.48	-31.31	-30.90

Question: What do we learn by interpreting the results in the above table of AIC values?

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

- Let's fit the ARMA(2,1) model recommended by consideration of AIC.

```
huron_arma21 <- arima(huron_depth,order=c(2,0,1))
huron_arma21
```

```
##
## Call:
## arima(x = huron_depth, order = c(2, 0, 1))
##
## Coefficients:
##           ar1     ar2     ma1  intercept
##       -0.0525  0.7910  1.0000   176.4603
## s.e.   0.0522  0.0526  0.0242     0.1210
##
## sigma^2 estimated as 0.04188:  log likelihood = 24.21,  aic = -38.43
```

- We can examine the roots of the AR polynomial,

```
AR_roots <- polyroot(c(1,-coef(huron_arma21)[c("ar1","ar2")]))
AR_roots
```

```
## [1]  1.158083-0i -1.091668+0i
```

- These are just outside the unit circle, suggesting we have a stationary causal fitted ARMA.

- However, the MA root is -1, showing that the fitted model is at the threshold of non-invertibility.

- Is this non-invertibility a problem? Let's investigate a little, using profile and bootstrap methods. The claimed standard error on the MA1 coefficient, from the Fisher information approach used by `arima` is small.

- First, we can see if the approximate confidence interval constructed using profile likelihood is in agreement with the approximate confidence interval constructed using the observed Fisher information.

- To do this, we need to maximize the ARMA likelihood while fixing the MA1 coefficient at a range of values. This can be done using `arima` as follows. Note that the `fixed` argument expects a vector of length $p + q + 1$ corresponding to a concatenated vector $(\phi_{1:p}, \psi_{1:q}, \mu)$. Somehow, the Gaussian white noise variance, σ^2, is not included in this representation. Parameters with `NA` entries in `fixed` are estimated.

```r
K <- 500
ma1 <- seq(from=0.2,to=1.1,length=K)
profile_loglik <- rep(NA,K)
for(k in 1:K){
   profile_loglik[k] <- logLik(arima(huron_depth,order=c(2,0,1),
      fixed=c(NA,NA,ma1[k],NA)))
}
plot(profile_loglik~ma1,ty="l")
```
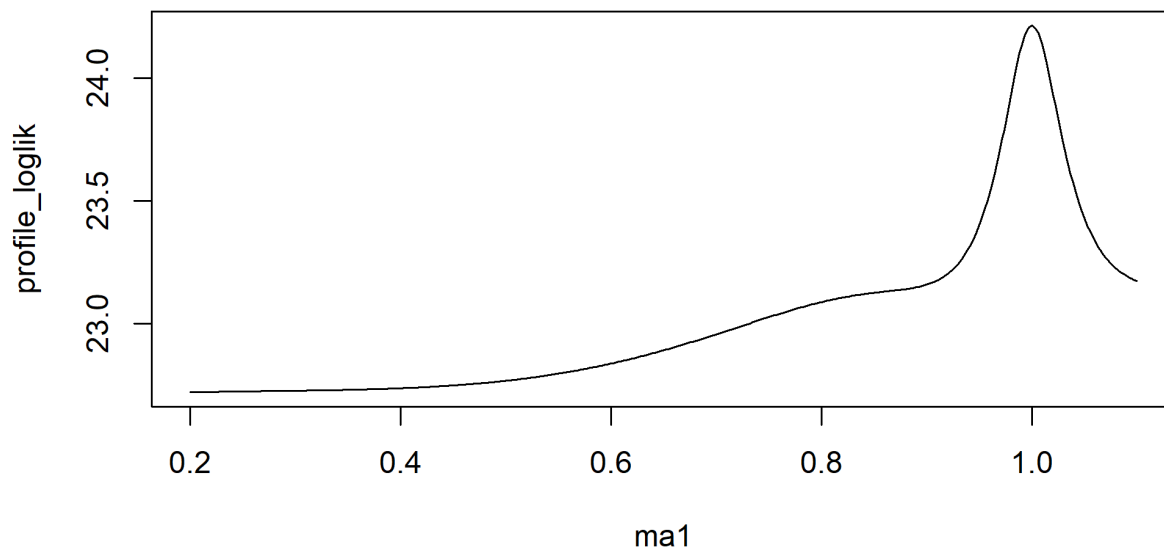

Question: Interpret the profile likelihood plot for ψ_1.

- What do you conclude about the Fisher information confidence interval proposed by `arima`?

- When do you think the Fisher information confidence interval may be reliable?

- Is this profile likelihood plot, and its statistical interpretation, reliable? How do you support your opinion on this?


- Let's do a simulation study

```r
set.seed(57892330)
J <- 1000
params <- coef(huron_arma21)
ar <- params[grep("^ar",names(params))]
ma <- params[grep("^ma",names(params))]
intercept <- params["intercept"]
sigma <- sqrt(huron_arma21$sigma2)
theta <- matrix(NA,nrow=J,ncol=length(params),dimnames=list(NULL,names(params)))
for(j in 1:J){
   Yj <- arima.sim(
      list(ar=ar,ma=ma),
      n=length(huron_depth),
      sd=sigma
   )+intercept
   theta[j,] <- coef(arima(Yj,order=c(2,0,1),method="ML"))
}
hist(theta[,"ma1"],freq=FALSE)
```

Histogram of theta[, "ma1"]
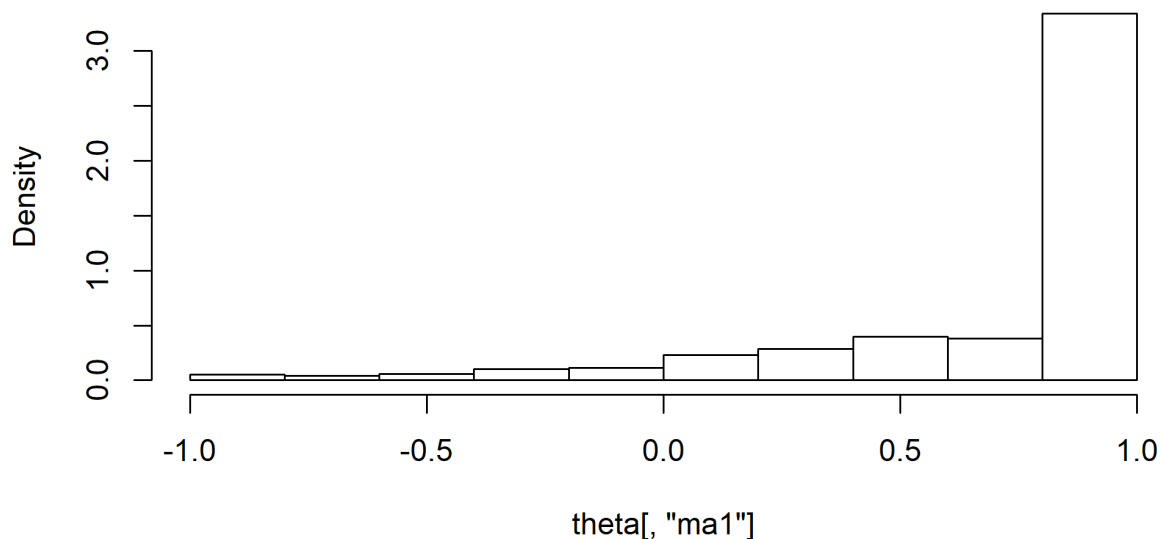
- This seems consistent with the profile likelihood plot.

- A density plot shows this similarity even more clearly.

```r
plot(density(theta[,"ma1"],bw=0.05))
```

density.default(x = theta[, "ma1"], bw = 0.05)

N = 1000   Bandwidth = 0.05
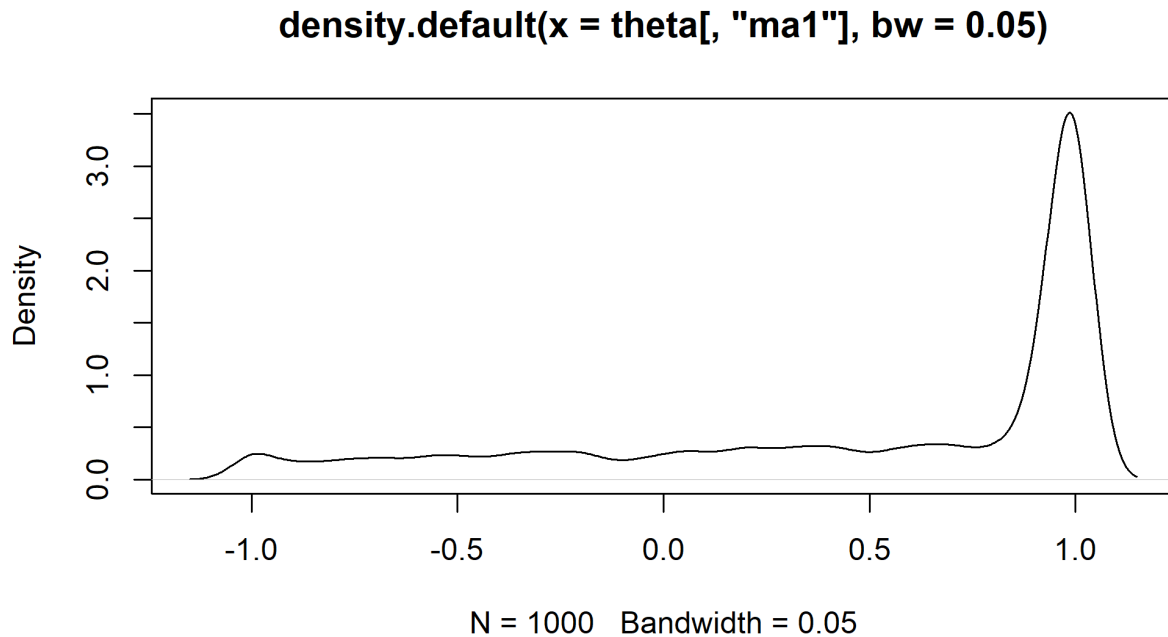
- Here, I'm showing the raw plot for instructional purposes. For a report, one should improve the default axis labels and title.

- Note that `arima` transforms the model to invertibility. Thus, the estimated value of θ_1 can only fall in the interval $(-1, 1)$ but can be arbitrarily close to -1 or 1.

```r
range(theta[,"ma1"])
```

```
## [1] -1  1
```

+ Estimated densities outside $[-1,1]$ are artifacts of the density estimation procedure.

+ How would you refine this procedure to get a density estimate respecting the range of the parameter estimati

- To understand what is going on better, it is helpful to do another simulation study for which we fit ARMA(2,1) when the true model is AR(1).

- When doing simulation studies, it is helpful to use multicore computing, which most of us have on our machines nowadays.

- A basic approach to multicore statistical computing is to tell R you want it to look for available processors, using the `doParallel` package.

```r
require(doParallel)
registerDoParallel()
```

- Then, we can use `foreach` to carry out a parallel `for` loop where jobs are sent to different processors.

```r
J <- 1000
huron_ar1 <- arima(huron_depth,order=c(1,0,0))
params <- coef(huron_ar1)
ar <- params[grep("^ar",names(params))]
```

13

```
intercept <- params["intercept"]
sigma <- sqrt(huron_ar1$sigma2)
t1 <- system.time(
  huron_sim <- foreach(j=1:J) %dopar% {
      Yj <- arima.sim(list(ar=ar),n=length(huron_depth),sd=sigma)+intercept
      try(coef(arima(Yj,order=c(2,0,1),method="ML")))
  }
)
#method = 'ML' not previously in code, added by me due to errors
```

- Some of these `arima` calls did not successfully produce parameter estimates. The `try` function lets the simulation proceed despite these errors. Let's see how many of them fail:

```
sum(sapply(huron_sim, function(x) inherits(x,"try-error")))
```
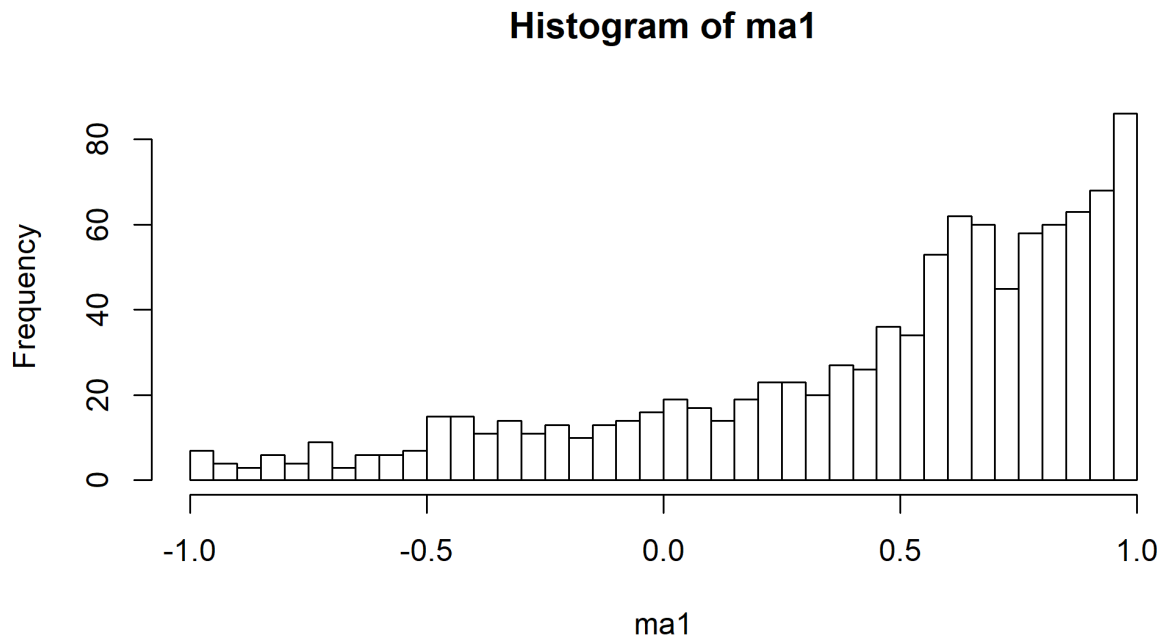
```
## [1] 0
```

- Now, for the remaining ones, we can look at the resulting estimates of the MA1 component:

```
ma1 <- unlist(lapply(huron_sim,function(x) if(!inherits(x,"try-error"))x["ma1"] else NULL ))
hist(ma1,breaks=50)
```

Histogram of ma1

- When the true model is AR1 and we fit ARMA(2,1), it seems that we often obtain a model with estimated MA1 coefficient on the boundary of invertibility.

- It is clear from this that we cannot reject an AR1 hypothesis, even though the Fisher information based analysis appears to give strong evidence that the data should be modeled with a nonzero MA1 coefficient.

- It may be sensible to avoid fitted models too close to the boundary of invertibility. This is a reason not to blindly accept whatever model AIC might suggest.

Question: what else could we look for to help diagnose, and understand, this kind of model fitting problem?

- Hint: pay some more attention to the roots of the fitted ARMA(2,1) model.

Assessing the numerical correctness of evaluation and maximization of the likelihood function

- We can probably suppose that `arima` has negligible numerical error in evaluating the likelihood.
    - Likelihood evaluation is a linear algebra computation which should be numerically stable away from singularities.
    - Possibly, numerical problems could arise for models very close to reducibility (canceling AR and MA roots).
- Numerical optimization is more problematic.
    - `arima` calls the general purpose optimization routine `optim`.
    - We know the likelihood surface can be multimodal and have nonlinear ridges; both these are consequences of the possibility of reducibility or near reducibility (AR and MA roots which almost cancel).
    - No optimization procedure is reliable for maximizing awkward, non-convex functions.
    - Evidence for imperfect maximization (assuming negligible likelihood evaluation error) can be found in the above AIC table, reproduced here:

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	166.75	46.60	7.28	-14.97	-18.64	-26.09
AR1	-38.00	-37.41	-35.46	-33.82	-34.13	-32.20
AR2	-37.33	-38.43	-36.90	-34.93	-34.35	-33.08
AR3	-35.52	-35.17	-32.71	-31.38	-33.21	-32.98
AR4	-33.94	-34.91	-34.43	-37.48	-31.31	-30.90

Question: How is this table inconsistent with perfect maximization?

- Here are two hints:
    - Recall that, for nested hypotheses $H^{\langle 0 \rangle} \subset H^{\langle 1 \rangle}$, the likelihood maximized over $H^{\langle 1 \rangle}$ cannot be less than the likelihood maximized over $H^{\langle 0 \rangle}$.
    - Recall also the definition of AIC, AIC = -2× maximized log likelihood + 2× number of parameters

---