Penalty terms for estimation of ARMA models: A Bayesian inspiration

ITISE Granada 2018

Helgi Tómasson
University of Iceland
helgito@hi.is

September 18-21, 2018

Plan of talk

- A brief review of parameterization of ARMA time-series models
- The role of the a prior distribution in the Bayesian estimation
- Review of partial fractions and residue calculus
- Implementation of smoothness priors in ARMA models
- Exact calculation of the distance between spectral shapes
- Implementation in R
- Conclusion and discussion

On the ARMA model

- A noise observation of a linear differential equation:

$$y'' + ay' + by = 0, \quad \text{non-stochastic}$$
$$y'' + ay' + by = \text{ a stochastic concept.}$$

- A classical discrete time version:

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \varepsilon_t \quad \text{white noise}$$

- Or a continuous time version:

$$Y'(t) + \alpha Y(t) = \sigma(dW(t) + \beta d^{(2)} W(t)), \quad dW(t) \quad \text{white-noise.}$$

- A representation of a continous-time ARMA(p,q), CARMA(p,q) process in terms of the differential operator D is:

$$
\begin{aligned}
&Y^{(p)}(t) + \alpha_1 Y^{(p-1)}(t) + \cdots + \alpha_p Y(t) = \\
&\sigma d(W(t) + \beta_1 W^1(t) + \cdots + \beta_q W^{(q))}(t)), \\
&\text{or } \boldsymbol{\alpha}(D)\, Y(t) = \sigma \boldsymbol{\beta}(D)\, dW(t), \\
&\boldsymbol{\alpha}(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p, \\
&\boldsymbol{\beta}(z) = 1 + \beta_1 z + \cdots + \beta_q z^q.
\end{aligned}
$$

- The spectral density of $Y(t)$ is a rational function:

$$
f(\omega) = \frac{\sigma^2}{2\pi} \frac{\boldsymbol{\beta}(i\omega)\boldsymbol{\beta}(-i\omega)}{\boldsymbol{\alpha}(i\omega)\boldsymbol{\alpha}(-i\omega)}.
$$

A **key feature** in this paper. Similar formulas apply for the ususal discrete-time ARMA models. Then the polynomials are in $exp(-i\omega)$.

- For stationarity we need the realpart of the roots of the polynomal $\alpha(z)$ to be negative. Similar to the discrete-time case where roots of a certain polynomial need to be on the correct side of the unit circle.
- In continuous-time we also need $p > q$.

The role of the prior in Bayesian estimation

- Bayesian inference about a parameter θ is based on the posterior-distribution which is proportional to the likelihood-function times the prior distribution.

$$\pi(\theta|\textit{data}) \propto \underbrace{\text{likelihood}(\text{data}|\theta)}_{\text{model}} \quad \underbrace{\pi(\theta)}_{\text{prior}}$$

- Then a Bayesian estimator can be calculated, e.g. by calculating the expected value, or the mode of posterior etc.

An example, the normal mean

- A possible approach for Bayesian inference on μ in $N(\mu, \sigma^2)$, σ known is:

$$X|\mu, \sigma \sim N(\mu, \sigma^2),$$
$$\mu|\sigma \sim N(\mu_0, \sigma_0^2), \quad \sigma_0 = \tau\sigma.$$

Given data, $x_1, \ldots, x_n$, and reparameterizing, $v = 1/\sigma$, $v_0 = k_0 v$, the log-posterior is (σ known):

$$log(p(\mu, v|x_1, \ldots, x_n, \mu_0, k_0)) =$$
$$\text{constant} + \underbrace{\frac{n}{2}\log(v) - v\sum_{i=1}^{n}(x_i - \mu)^2/2}_{A} + \underbrace{\frac{1}{2}\log(v) - k_0 v(\mu - \mu_0)^2/2}_{B}.$$

- The number k_0 expresses the certainty in the prior. If k_0 is set to zero and the log-posterior (as a function of μ) is maximized the result is the ML estimator and a nonzero k_0 biases the ML-estimate towards the prior (μ_0).
- Maximization of the log-posterior can be interpreted as a penalized maximum-likelihood.
- I.e. a deviance from μ_0 is penalized.
- AIC, BIC, R^2-adjusted are examples of penalizing terms.
- The added term penalizes for a more complicated (less reasonable) model.

The role of parameters in ARMA models

- The parameters in the polynomials $\boldsymbol{\alpha}(z)$ and $\boldsymbol{\beta}(z)$ are auto-correlation parameters, and the parameter σ is a scale parameter.
- If normality is assumed and the polynomials $\boldsymbol{\alpha}(z)$ and $\boldsymbol{\beta}(z)$ were known inference about σ is similar to inference about σ in a normal model. E.g. a posterior like:

$$gamma(a + n/2, b + \mathbf{y}' M(\boldsymbol{\alpha}, \boldsymbol{\beta})^{-1} \mathbf{y}/2).$$

- It is very difficult to have a good intuition about the auto-correlation function.
- The interpretation of the spectral density is easier and therefore perhaps more natural to express a prior on the parameters in the polynomials $\boldsymbol{\alpha}(z)$ and $\boldsymbol{\beta}(z)$ based on properties of the spectral function.
- One might e.g. state a prior on the smoothness of the spectral function or its closeness to a particular spectral function.

More than the number of parameters

- The AIC, BIC and the R^2-adjusted all penalize by using a function of the number of estimated parameters. The number of parameters is not always the natural way of grading complexity. In regression it seems reasonable that the model:

  $$y = a + bx + e,$$

  is simpler than:

  $$y = \sin(\cos(ax))^a \exp(-bx)/x^b + e.$$

- The ARMA(1,0) model:

  $$dY + Y = dW, \quad ,$$

  is actually the same as:

  $$Y^{(4)} + 4Y^{(3)} + 6Y^{(2)} + 4Y^{(1)} + Y = d(W + 3W^{(1)} + 3W^{(2)} + W^{(3)}).$$

  That is the ARMA(1,0) is a special case of (many) ARMA(4,3) models. Estimation of six additional parameters might result in a spectral function with an unreasonable shape. However, it might be of interest to estimate a model which is more complicated than an AR(1). One might, however, want restrict the freedom of the additional parameters.

- In time-series analysis, just as in non-parametric regression a smoothness restriction may be enforced on the fitted values. That is the sharp spikes and turns are penalized. In economics a well known procedure of this type is the Hodrick-Prescott filter.
  In stationary time-series analysis a natural form of a priori information might consist of a specification of the spectral function or some features of the spectral function. In analogy with the Hodrick-Prescott filter one can introduce a term that penalizes for sharp spikes and turns, e.g., a term proportional to:

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega.$$

- One might also want that the estimated spectrum is close to a particular spectral function.

How to measure distance between functions?

- Here the I only discuss the Kullback-Leibler distance measure.

$$KLD(f, f^*) = \int_{-\infty}^{\infty} \log(\frac{f(\omega)}{f^*(\omega)}) f(\omega) \mathrm{d}\omega.$$

- Here I use f^* because I do not work directly with the distance between two spectral curves, but only with the proportionality between two spectral curves:

$$f(\omega) = \frac{\beta(i\omega)\beta(-i\omega)}{\alpha(i\omega)\alpha(-i\omega)},$$
$$f^*(\omega) = \sigma^* \frac{\beta_0(i\omega)\beta_0(-i\omega)}{\alpha_0(i\omega)\alpha_0(-i\omega)},$$

where σ^* is chosen such that,

$$\int_{-\infty}^{\infty} f(\omega)\mathrm{d}\omega = \int_{-\infty}^{\infty} f^*(\omega)\mathrm{d}\omega.$$

- The fact that the spectral function of an ARMA model is rational allows for an exact calculation of integrals like:

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega.$$

- By use of partial fractions the function:

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{\beta(i\omega)\beta(-i\omega)}{\alpha(i\omega)\alpha(-i\omega)} = \frac{\sigma^2}{2\pi} \frac{\prod_{j=1}^{q}(1+\mu_j^2\omega^2)}{\prod_{j=1}^{p}(\omega^2+\lambda_j^2)},$$

can be written as:

$$f(\omega) = \frac{\sigma^2}{2\pi}\left(\frac{a_1}{\omega - i\lambda_1} + \cdots \frac{a_p}{w - i\lambda_p} + \frac{b_1}{\omega + i\lambda_1} + \cdots \frac{b_p}{\omega + i\lambda_p}\right),$$

where λ_j are the roots of the AR polynomial, $\alpha(z)$. Another way is:

$$f(\omega) = \frac{\sigma^2}{2\pi}\left(\frac{c_1}{\omega^2 + \lambda_1^2} + \cdots + \frac{c_p}{\omega^2 + \lambda_p^2}\right).$$

Residue calculus

- The residue calculus of complex analysis offers a useful tool for calculating integrals of rational functions. The residue theorem states that

$$\int h(x)dx = 2\pi i \sum Res(h(z)), \quad \text{over a certain path},$$

where the sum is evaluated over the residues of the function h (Kreyszig, 1999).

-

$$f''(\omega) = \frac{\sigma^2}{2\pi}\left(\frac{2a_1}{(\omega - i\lambda_1)^3} + \cdots \frac{2a_p}{(w - i\lambda_p)^3} + \frac{2b_1}{(\omega + i\lambda_1)^3} + \cdots \frac{2b_p}{(\omega + i\lambda_p)^3}\right),$$

$f''(\omega)^2$ will contain p terms of the type $a_j/(\omega - i\lambda_j)^6$ and p terms $b_k/(\omega + i\lambda_k)^6$ and $p(p-1)$ terms, $k \neq j$, of the type $a_k a_j/((\omega - i\lambda_k)^3)(\omega - i\lambda_j)^3)$ and similarly $p(p-1)$ terms, $k \neq j$, $b_k b_j/((\omega + i\lambda_j)(\omega + i\lambda_j))$. The residues in the upper half-plane of these terms sums to zero. The integral will be the sum of the $2p^2$ terms of the type

$$a_k b_j/((\omega - i\lambda_k)^3(\omega + i\lambda_j)^3).$$

The residues of these terms are of the form:

$$3 \cdot 4 a_k b_j / (-(i\lambda_k + i\lambda_j)^5),$$

and the integral therefore,

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega = 2\pi i \cdot 2 \sum_{k=1}^{p} \sum_{j=1}^{p} 3 \cdot 4 \frac{a_k b_j}{-(i\lambda_k + i\lambda_j)^5}.$$

Similarly one can use residue calculus to create of measure of steep hills in the spectrum,

$$\int_{-\infty}^{\infty} (f'(\omega))^2 d\omega,$$

or weighing $(f'(\omega))^2$ or $(f''(\omega))^2$ with a rational function.

- I have checked this numerically. Dual roots will give somewhat more complicated formulas.

More partial fractions

The partial fraction trick can also be applied to calculate the Kullback-Leibner (KL) metric, as a measure of the distance between two functions, f and f_0 (e.g. a prior).

$$D(f_1; f_0) = \int f_1(\omega) \log(f_1(\omega)) d\omega - \int f_1(\omega) \log(f_0(\omega)) d\omega.$$

Using the second partial fraction formulation of the spectral density the termsthat need to be integrated will be of the form:

$$-\frac{c_{1,k}}{(\omega^2 + \lambda_{1,k}^2)} \log(w^2 + \lambda_{1,j}^2), \text{ and } \frac{c_{1,j}}{(\omega^2 + \lambda_{1,k}^2)} \log(1 + \mu_{1,j}^2 w^2).$$

$$\int_0^\infty \log(1+\mu^2 x^2)\frac{dx}{x^2+\lambda^2} = \frac{\pi}{\sqrt{\lambda^2}}\log(\sqrt{\lambda^2\mu^2}+1),$$

$$\int_0^\infty \log(p^2+x^2)/(q^2+r^2x^2) = \frac{\pi}{pr}log((q+pr/r)).$$

Here the square-root is take such that the real part of the square-root is positive (Gradshteyn & Ryzhik, 2007, eq 1, page 560). By use of partial fractions the KL distance can be written as:

$$\int \sum_{k=1}^{p_1}\frac{c_{1,k}}{\omega^2+\lambda_{1,k}^2}(\sum_{j=1}^{q_1}\log(1+\mu_{1,j}^2\omega^2))d\omega - \int \sum_{k=1}^{p_1}\frac{c_{1,k}}{\omega^2+\lambda_{1,k}^2}(\sum_{j=1}^{p_1}\log(\omega^2+\lambda_{1,j}^2))d\omega -$$

$$\int \sum_{k=1}^{p_1}\frac{c_{1,k}}{\omega^2+\lambda_{1,k}^2}(\sum_{j=1}^{q_0}\log(1+\mu_{0,j}^2\omega^2))d\omega + \int \sum_{k=1}^{p_1}\frac{c_{1,k}}{\omega^2+\lambda_{1,k}^2}(\sum_{j=1}^{p_0}\log(\omega^2+\lambda_{0,j}^2))d\omega.$$

This integral consists of $p_1 \times q_1 + p_1^2 + p_1 \times q_0 + p_1 \times p_0$ terms and each of them can be calculated by the use of the above formula,

$$\pi(\sum_{k=1}^{p_1}\sum_{j=1}^{q_1}\frac{c_{1,k}\log(1+\lambda_{1,k}\mu_{1,j})}{\lambda_{1,k}} - \sum_{k=1}^{p_1}\sum_{j=1}^{p_1}\frac{c_{1,k}\log(1+\lambda_{1,j}\lambda_{1,k})}{\lambda_{1,j}} +$$

$$\sum_{k=1}^{p_1}\sum_{j=1}^{q_0}\frac{c_{1,k}\log(1+\lambda_{1,k}\mu_{0,j})}{\lambda_{0,j}} - \sum_{k=1}^{p_1}\sum_{j=1}^{p_0}\frac{c_{1,k}\log(1+\lambda_{1,k}\lambda_{0,j})}{\lambda_{0,j}}).$$

Implementation in R

- I have implemented the partial fractions and some of the above in R a R package `ctarmaRcpp`.
- $$1/(6+11x+6x^2+x^3) = \frac{1}{2(x+3)} - \frac{1}{x+2} + \frac{1}{2(x+1)},$$

  here the roots are -1,-2,-3, and the function `partfrac1` gives the coefficients in the partial fraction (all roots distinct).

```
partfrac1(c(6,11,6,1),1,c(-1,-2,-3),1)
[1]  0.5 -1.0  0.5
```

  The partial fraction enables the calculation of the Kullback-Leibler distance between two spectral shapes.

A data set on the Earth's temperature for the past 800.000 years is used as an illustration on an unevenly sampled time series. The `ctarmaRcpp` package bundles data and model into a R object. Similar to (Tómasson, 2015).The maximized log-likelihood of a continuous-time ARMA(2,1) is contained in `m2e`. The log-likelihood of `m2e` is calculate by:

```
> ctarma.loglik(m2e)
[1] -5701.584
```

An ARMA(4,3) gives log-likelihood of -5664.627, and an ARMA(6,5) a log-likelihood of -5660.819. The coefficients of the estimated ARMA(2,1), are

```
[1] 1792.32808   13.39429
> m2e$bhat
[1] 1.00000000 0.02315723
> m2e$sigma
[1] 1331.322
```

Similarly the estimated coefficients of the ARMA(4,3) are:

```
> m4e$ahat
[1] 1497.15420 3410.91710 2328.64602   28.11924
> m4e$bhat
[1] 1.0000000 1.2087125 0.3772288 0.0128648
> m4e$sigma
[1] 2239.939
```

The Kullback-Leibler distance is calculated with the function `kullbackDist` (here the implementation is between spectral shapes).

```
> kullbackDist(m4e$ahat,m4e$bhat,m4e$sigma,m2e$ahat,m2e$bhat)
[1] 1.172553
```

and for the ARMA(6,5)

```
> kullbackDist(m6e$ahat,m6e$bhat,m6e$sigma,m2e$ahat,m2e$bhat)
[1] 3.706201
```

Temperature on Earth for 800 Kyears

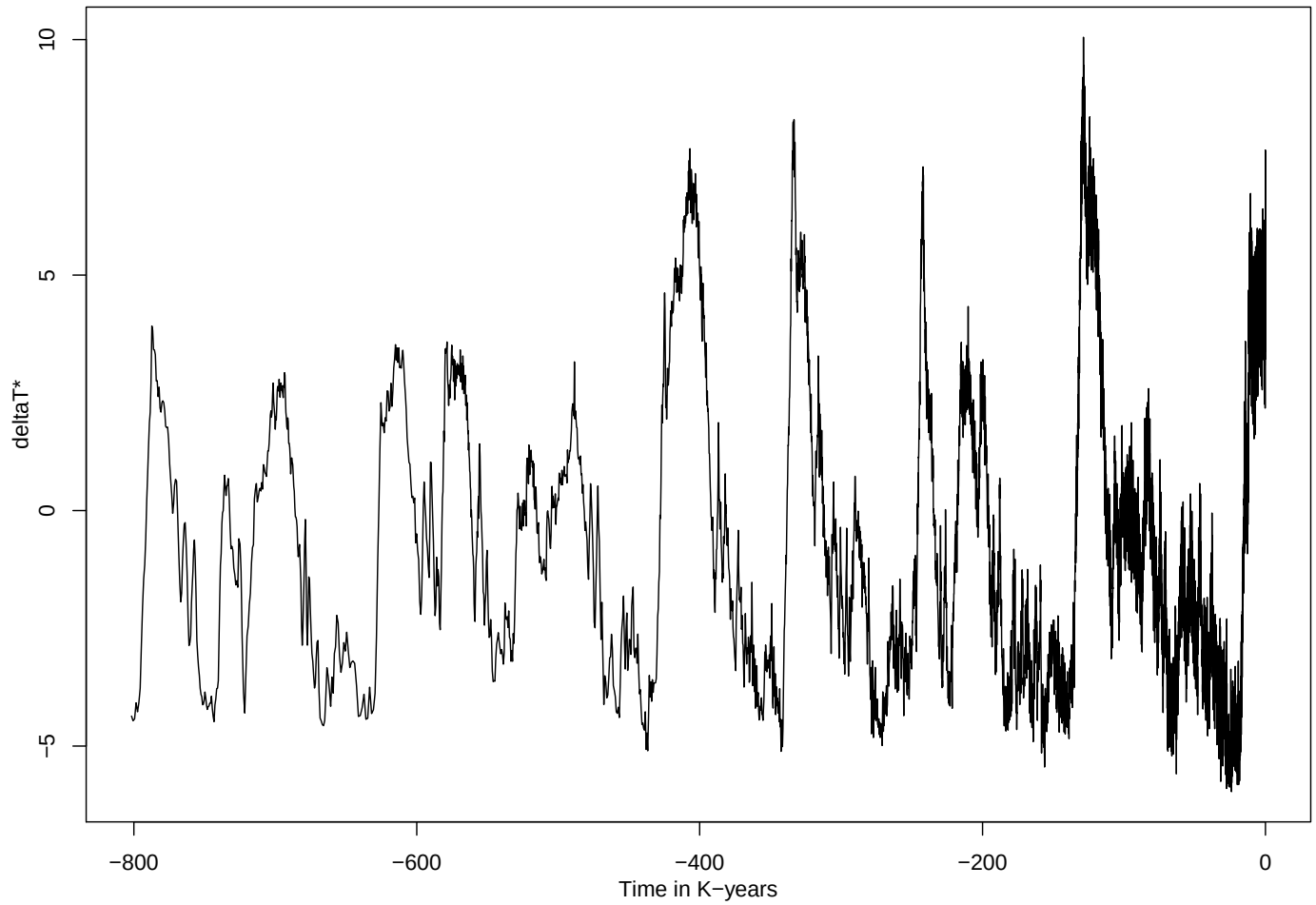

Figure: Temperature on Earth. About 5500 observations over 800.000 years.

Log CARMA(20,19) spectrum

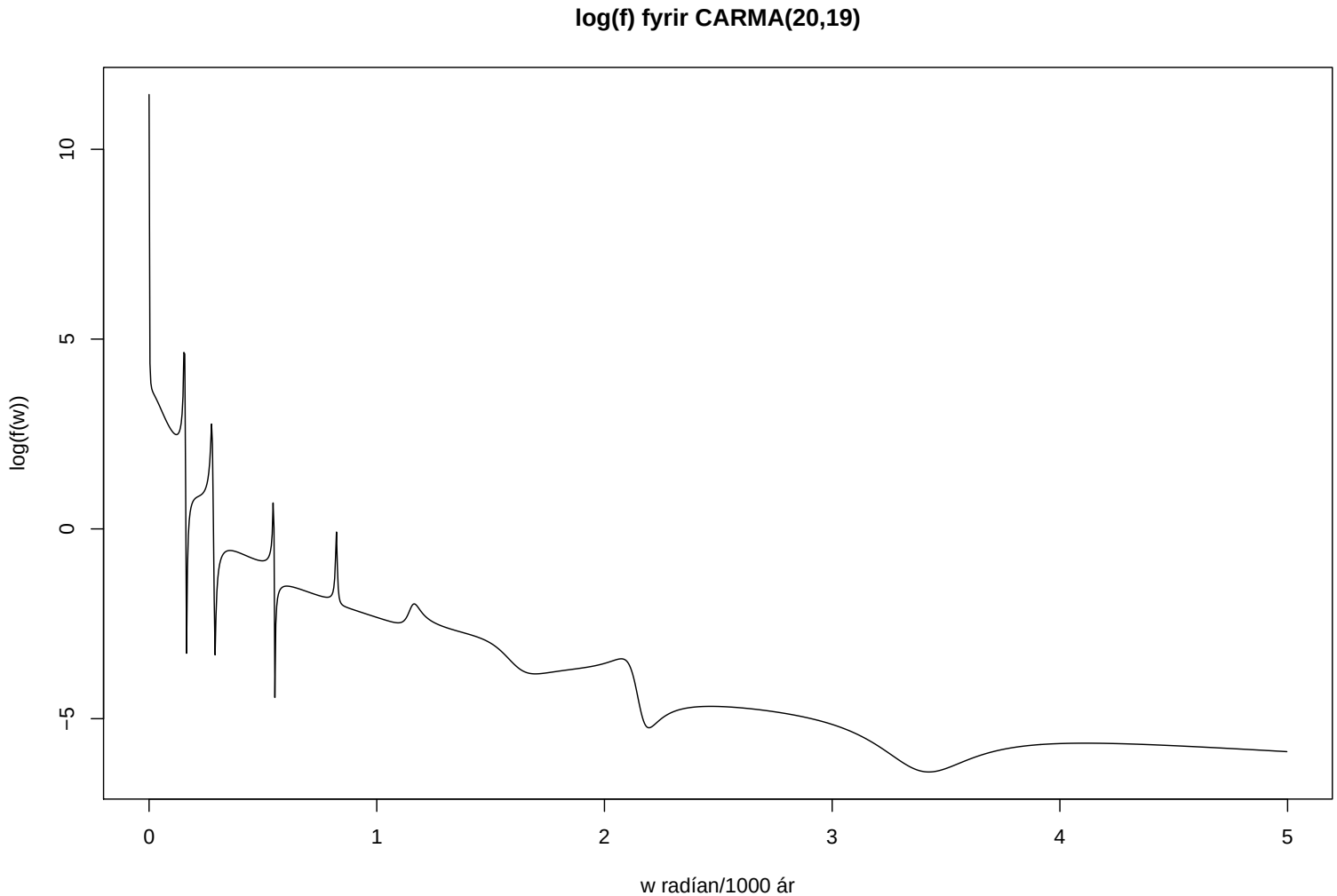

Figure: Log of ML-estimated CARMA(20,19) spectrum of Earth data.

Conclusion and discussion

- Technically it might be boring to try to find all the roots of a polynomial of degree 20. Perhaps it is better to perform numercial optimization directly in terms of the roots of the polynomial.
- The fact that the spectral function is rational can be exploited in more ways than shown here.
- The partial fraction trick along with the residue calculus can be used in calculation Bayesian, and semi-Bayesian estimators.

Gradshteyn, I. S. & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (Seventh ed.). Elsevier/Academic Press, Amsterdam. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

Kreyszig, E. (1999). *Advanced Engineering Mathematics* (8 ed.). John Wiley & Sons. Residue theorem, page 723-724.

Tómasson, H. (2015). Some computational aspects of gaussian CARMA modelling. *Statistics and Computing*, *25*(2), 375–387.