Autoregressive Models

Statistically, time series are special among kinds of data because of their serial dependence. That is, the value of X_t is likely to depend on the value of X_{t-1}. The simplest form of dependence is linear dependence, as in the *autoregressive model*

$$X_t = \phi X_{t-1} + \epsilon_t.$$

This says that X_t has a regression on X_{t-1}, and otherwise is determined by noise. For consistency with later notation let us write the noise variables as W_t:

$$X_t = \phi X_{t-1} + W_t.$$

A natural generalization,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + W_t,$$

is called an *autoregressive model of order p*, written $AR(p)$. The W_t variables are usually assumed to be i.i.d. $N(0, \sigma^2)$.

Autoregressive Fitting via Lagged Regression

Let us consider an $AR(p)$ model for the core body temperature residuals. Because the model has the form of an ordinary linear regression model, we may apply it to data $x = (x_1, \ldots, x_n)$ using ordinary least squares regression. To do so we have to define *lagged* variables. In the simplest case, with $p = 1$, we begin by defining a pair of variables $x1, y$, each of length $n - 1$. We first take $y_{n-1} = x_n$, $y_{n-2} = x_{n-1}$, $\ldots, y_1 = x_2$. We then define the lag-1 version of y, which we call $x1$, given by $x1_1 = x_1$, $x1_2 = x_2$, $\ldots, x1_{n-1} = x_{n-1}$. We may fit the $AR(1)$ model

$$X_t = \phi X_{t-1} + W_t$$

by performing least-squares regression of y on $x1$.

More generally, to fit an $AR(p)$ model using ordinary least squares we begin by defining $y_{n-p} = x_n$, $y_{n-p-1} = x_{n-1}$, $\ldots, y_1 = x_{n-p+1}$ and then

define $x1$ to be the lag-1 version of y, $x2$ to be the lag-2 version of y, etc., until we reach xp. We then regress y on the variables $x1, x2, \ldots, xp$.

Let us apply this to the core body temperature data after removing the dominant 24-hour cycle. We take $p = 22$. The fitted coefficients are plotted in Figure 1. Here is the table of coefficients:

```
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
x1   1.042298   0.057213  18.218  < 2e-16 ***
x2  -0.220820   0.082367  -2.681  0.00774 **
x3  -0.124162   0.083325  -1.490  0.13723
x4   0.048982   0.083329   0.588  0.55709
x5  -0.127941   0.083324  -1.535  0.12569
x6   0.115148   0.083563   1.378  0.16921
x7   0.028210   0.083539   0.338  0.73583
x8   0.046588   0.083531   0.558  0.57743
x9  -0.111434   0.083441  -1.335  0.18271
x10  0.056760   0.083572   0.679  0.49754
x11 -0.014675   0.083232  -0.176  0.86017
x12  0.018977   0.083232   0.228  0.81980
x13  0.053313   0.083194   0.641  0.52211
x14 -0.039973   0.083114  -0.481  0.63090
x15  0.024029   0.083120   0.289  0.77271
x16 -0.098060   0.083109  -1.180  0.23895
x17  0.032837   0.082969   0.396  0.69254
x18  0.032120   0.082597   0.389  0.69764
x19 -0.008946   0.082618  -0.108  0.91384
x20 -0.007037   0.082626  -0.085  0.93219
x21 -0.020689   0.081560  -0.254  0.79992
x22  0.010073   0.056830   0.177  0.85943
```

It appears that only the first two lagged variables are likely to be helpful in predicting the response variable. More information is provided by the sample autocorrelation function (ACF) and partial autocorrelation function (PACF), also shown in Figure 1.

The sample ACF is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

where $\hat{\gamma}$ is the sample autocovariance function

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}),$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \ldots, n-1$. Under fairly general conditions, if x_t are i.i.d. with finite variance then

$$\sqrt{n}\hat{\rho}(h) \xrightarrow{\mathcal{D}} N(0,1).$$

In the middle plot of Figure 1, horizontal lines are drawn at $\pm 2/\sqrt{n}$. If the series were i.i.d., then roughly 95% of the sample autocorrelation coefficients should fall between theses lines.

The sample PACF at lag h is the normalized lag-h regression coefficient found from an $AR(h)$ model, normalized by dividing by the sample variance $\hat{\gamma}(0)$. In other words, the lag-h partial autocorrelation coefficient measures the lag-h correlation after adjusting for the effects of lags 1 through $h-1$, adjusting as in multiple linear regression. It is a special case of the partial correlation coefficient in multiple linear regression. The PACF in Figure 1 has nonzero lag-1 and lag-2 coefficients, but the remaining coefficients are not distinctly different from zero relative to statistical uncertainty.

Suppose X_t is a mean-zero stationary Gaussian series. Then the theoretical PACF is given by $\phi_{11} = Cor(X_t, X_{t+1})$ and for $h \geq 2$,

$$\phi_{hh} = Cor(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}).$$

More generally, for any mean-zero stationary series let $X_t^{h-1} = \sum_{j=1}^{h-1} \beta_j X_{t-j}$ where the coefficients $\beta_1, \ldots, \beta_{h-1}$ minimize $E((X_t - \sum_{j=1}^{h-1} \alpha_j X_{t-j})^2)$ over the α_js. Then, for $h \geq 2$,

$$\phi_{hh} = Cor(X_t - X_t^{h-1}, X_{t+h} - X_{t+h}^{h-1}).$$

Using an $AR(2)$ fit to the residuals added to the fitted 24-hour cycle produces the overall fit to the temperature data shown in Figure 2.

In general, autoregressive models may be fit by maximum likelihood. Let us write down the likelihood function for the $AR(1)$ model, assuming X_t is Gaussian with mean zero and $|\phi| < 1$. We have $X_1 \sim N(0, \sigma_1^2)$ where $\sigma_1^2 = \sigma_W^2/(1-\phi^2)$. We also have $X_t | X_{t-1} = x_{t-1} \sim N(\phi x_{t-1}, \sigma_W^2)$

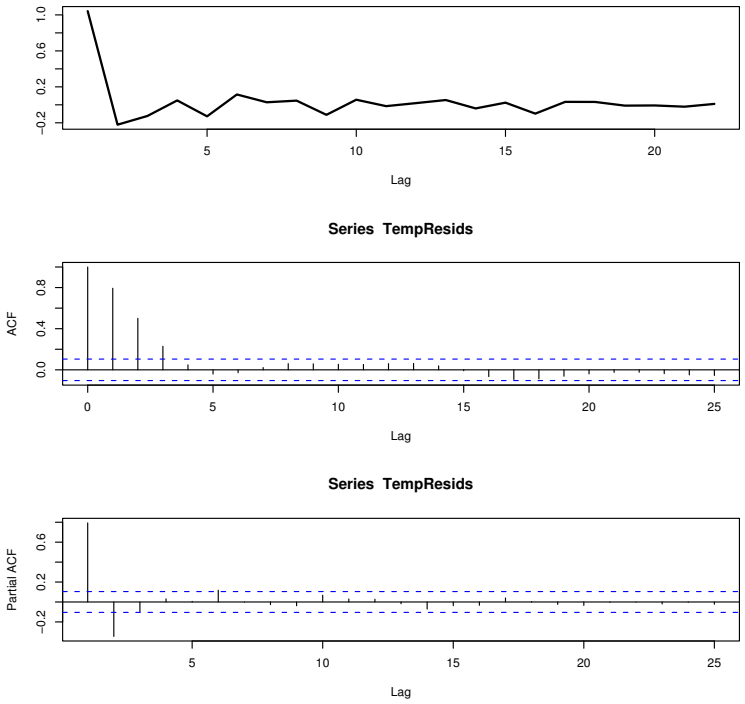

Figure 1: *Autoregressive model of order $p = 22$ for core body temperature residuals.* TOP: *Coefficients $\hat{\phi}_i$ as a function of lag i.* MIDDLE: *The sample autocorrelation function.* BOTTOM: *The sample partial autocorrelation function.*

for $t = 2, \ldots, n$. The joint pdf is

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|X_1 = x_1) \cdots f_{X_n|X_{n-1}}(x_n|X_{n-1} = x_{n-1}) \\ &= \frac{1}{\sigma_1} f_Z(\frac{x_1}{\sigma_1}) \prod_{t=2}^{n} \frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) \end{aligned}$$

where $f_Z(z)$ is the $N(0,1)$ pdf. The factors in the product are

$$\frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) = \frac{1}{\sqrt{2\pi}\sigma_W} \exp(-\frac{(x_t - \phi x_{t-1})^2}{2\sigma_W^2}).$$

Thus, if we ignore x_1, maximizing the likelihood $L(\phi, \sigma_W)$ amounts to solving the ordinary least-squares problem in the regression of x_t on its lag-1 version. This maximization is called *conditional maximum likelihood* because we act as if the distribution of X_1 is given, i.e., it involves no unknown parameters. Because σ_1 is a function of ϕ and σ_W, when we include the factor due to X_1, which is $f_Z(x_1/\sigma_1)/\sigma_1$, the maximization problem changes and it is no longer solvable by least squares.

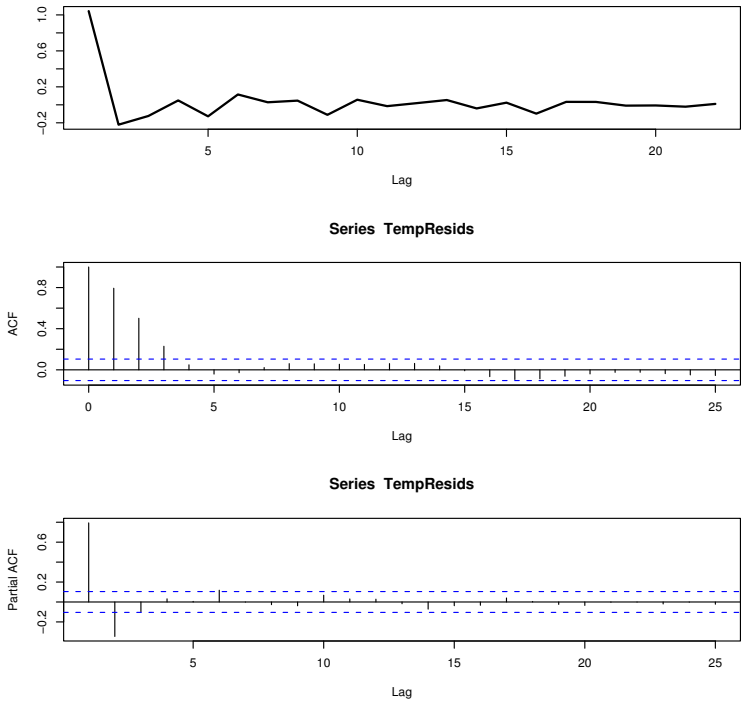

Figure 2: *Core body temperature data together with fit (red) based on the sum of an $AR(2)$ fit to residuals and the fitted 24-hour cycle.*

Thus, the MLE must be found by an iterative method, but it is likely to be very close to the conditional MLE. Similar considerations hold also for $AR(p)$ models: the likelihood is nonlinear in the autoregressive parameters, but if we condition on the first p values then ML estimation reduces to ordinary least squares lagged regression. Thus, lagged regression is essentially the same as ML estimation for large samples.

```
library(Hmisc) # to get Lag function
x=rep(0,330*22)
dim(x)=c(330,22)
y=TempResids[23:352]
for(i in 1:22){
  x[,i]=Lag(TempResids,i)[23:352]
}
ar.reg=lm(y~x-1)
summary(ar.reg)
par(mfrow=c(3,1))
plot(1:22,ar.reg$coef,type="l",xlab="Lag",ylab=" ",lwd=2)
```

```
acf(TempResids)
pacf(TempResids)
dev.print(device=postscript,file="temp.resids.ar.ps")

# compare lagged regression to ML estimation
temp.ar2=lm(x[,1]~x[,2:3]-1)
>x[, 2:3]1  x[, 2:3]2
   1.0765    -0.3531
temp.arima=arima(TempResids,order=c(2,0,0))
temp.arima
>Coefficients:
         ar1      ar2  intercept
      1.0858  -0.3564    -0.0023
s.e.  0.0498   0.0501     0.0221
```

Moving Average Models

Autoregressive models are a natural idea. In practice, however, the order p can be large, leading to over-parameterization. A reduction in the number of parameters is achieved by introducing an alternative form for linear processes. Let W_t be a Gaussian white noise process, i.e., the W_t variables are i.i.d. $N(0, \sigma^2)$. The process X_t defined by

$$X_t = W_t + \sum_{i=1}^{q} \theta_i W_{t-i}$$

is called a *moving average model of order q*, written $MA(q)$. More generally, a process of the form

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + W_t + \sum_{i=1}^{q} \theta_i W_{t-i}$$

with $\phi_p \neq 0$ and $\theta_q \neq 0$ is called an *auto-regressive moving average model of order (p,q)*, written $ARMA(p, q)$. (It is not always the case that W_t is assumed to be Gaussian.)

The Backshift Operator

Some important properties of ARMA models are most easily obtained using a formalism involving the *backshift operator* B defined by

$$BX_t = X_{t-1}$$

and then, using $B^2X_t = B(BX_t)$, etc., we have

$$B^kX_t = X_{t-k}.$$

This means that the $AR(p)$ model may be written in the form

$$\phi(B)X_t = W_t \tag{1}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$.

Now we examine the $AR(1)$ process. We have

$$X_t = \phi X_{t-1} + W_t = \phi(\phi X_{t-1} + w_{t-1}) + W_{t-2} = \phi^2 X_{t-2} + \phi W_{t-1} + W_t$$

and, iterating this process k times we have

$$X_t = \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j}. \tag{2}$$

Assuming the process is stationary with $|\phi| < 1$ we may continue iterating and pass to the limit to get

$$X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}. \tag{3}$$

To make sense of the infinite sum a notion of convergence is required. Here, it is not hard to show that the random sequence $X_t - \sum_{j=0}^{k-1} \phi^j W_{t-j}$ converges to 0 in probability. The representation (3) is important both conceptually and technically.

Now, consider what happens when $|\phi| > 1$. The first term in (2) ϕ^k will grow indefinitely in magnitude, so that the infinite sum will no longer converge. However, instead we may write $X_{t+1} = \phi X_t + W_{t+1}$ so that

$$X_t = \phi^{-1}X_{t+1} - \phi^{-1}W_{t+1} = \phi^{-1}(\phi^{-1}X_{t+2} - \phi^{-1}W_{t-2}) - \phi^{-1}W_{t+1}$$

and, again, continuing we get

$$X_t = \phi^k X_{t+k} - \sum_{j=1}^{k-1} \phi^{-j} W_{t+j}. \tag{4}$$

We have $|\phi^{-1}| < 1$ so, analogously to (3) we have

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} W_{t+j}. \tag{5}$$

However, the variables on the right-hand side will not have been observed at time t, so this form of the process is not likely to be helpful for predictive purposes. When a linear process depends only on the past it is usually called *causal*. Thus, the $AR(1)$ is causal when $|\phi| < 1$ but not when $|\phi| > 1$.

The forgoing manipulations may be performed using the backshift operator. We may define

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$$

and write (3) as

$$X_t = \psi(B) W_t \tag{6}$$

with $\psi_j = \phi^j$. Now, let us work backward by combining (1) and (6) to get

$$\phi(B)\psi(B) W_t = W_t$$

which implies $\phi(B)\psi(B) = 1$. Expanding, we have

$$(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_j B^j + \cdots) = 1$$

and because

$$\begin{aligned} & (1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_j B^j + \cdots) \\ = \;& 1 + (\psi_1 - \phi) B + (\psi_2 - \psi_1 \phi) B^2 + \cdots + (\psi_j - \psi_{j-1}\phi) B^j + \cdots \end{aligned}$$

this implies $\psi_1 = \phi$ and then $\psi_2 = \phi^2$ and so on, giving $\psi_j = \phi^j$, as before.

This algebra involving the backshift operator suggests a close analogy with polynomials, and this turns out to be useful. Generally, the *AR and MA polynomials* are

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

for $\phi_p \neq 0$, and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

for $\phi_q \neq 0$. Here z is a complex numer. An $ARMA(p, q)$ process $\{X_t, t \in \mathcal{Z}\}$, may be written in the form $\phi(B) X_t = \theta(B) W_t$. It is said to be *causal* if it can be written in the form

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} = \psi(B) W_t,$$

with $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. It is causal if and only if $\phi(z) \neq 0$ when $|z| \leq 1$. The coefficients of its linear process representation are found by solving

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$$

with $|z| \leq 1$.

Here is an important property: for a causal $AR(p)$ model, the PACF is zero for lags greater than p. See Figure 1.

Let us consider the ACF and PACF for a causal $AR(1)$ model, for which $|\phi| < 1$. By Equation (3) we have

$$\begin{aligned} \gamma(h) &= Cov(\sum_{j=0}^{\infty} \phi^j W_{t-j}, \sum_{j=0}^{\infty} \phi^j W_{t+h-j}) \\ &= \sigma_W^2 \phi^h \sum_{j=0}^{\infty} (\phi^2)^j \\ &= \frac{\sigma_W^2 \phi^h}{1-\phi^2}. \end{aligned}$$

Note that the variance is $V(X_t) = \gamma(0) = \sigma_W^2/(1-\phi^2)$. Because the theoretical ACF is $\rho(h) = \gamma(h)/\gamma(0)$, the ACF for the causal $AR(1)$ is

$$\rho(h) = \phi^h.$$

This says that the autocorrelation for an $AR(1)$ vanishes exponentially fast as a function of lag. Now, suppose that X_t is a Gaussian causal $AR(1)$ model, so that for $h \geq 2$,

$$\phi_{hh} = Cor(X_t, X_{t+h} | X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \ldots, X_{t+h-1} = x_{t+h-1}).$$

Plugging in $X_{t+h} = \phi X_{t+h-1} + W_{t+h}$ we have

$$\begin{aligned} & Cov(X_t, X_{t+h} | X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \ldots, X_{t+h-1} = x_{t+h-1}) \\ = & \quad Cov(X_t, \phi x_{t+h-1} + W_{t+h}) \\ = & \quad Cov(X_t, W_{t+h}) = 0. \end{aligned}$$

Therefore, for all $h \geq 2$,

$$\phi_{hh} = 0.$$

This result also holds in the non-Gaussian case, though the calculations are somewhat more involved.

For an $MA(q)$ process the PACF does not vanish for high lags, but the ACF does. It is not hard to calculate the ACF. Suppose X_t is an $MA(q)$ process. Then

$$\gamma(h) = Cov(X_t, X_{t+h}) = E\left((\sum_{j=0}^{q}\theta_j W_{t-j})(\sum_{k=0}^{q}\theta_k W_{t+h-k})\right)$$

which is zero when $h > q$, as claimed, while for $h \leq q$ we have

$$\gamma(h) = \sigma_W^2 \sum_{j=0}^{q-h}\theta_j\theta_{j+h}.$$

Dividing by $\gamma(0) = \sigma_W^2 \sum_{j=0}^{q-h}\theta_j^2$ we have the ACF

$$\rho(h) = \frac{\sum_{j=0}^{q-h}\theta_j\theta_{j+h}}{1+\sum_{j=1}^{q-h}\theta_j^2}$$

for $h \leq q$ and

$$\rho(h) = 0$$

for $h > q$.

The practical implications are these: if we see an ACF that dampens rapidly in magnitude (magnitude is important as the coefficients may oscillate) and the PACF vanishes after some lag p, then an $AR(p)$ is likely to fit well. If, instead, the ACF vanishes after some lag q then an $MA(q)$ is likely to fit well. As shown in Box, Jenkins, and Reinsel (2008; Chapter 3), an $ARMA(p,q)$ process will have an ACF that dampens after the first $q-p$ lags and a PACF that dampens after the first $p-q$ lags.

It can also be helpful to use model selection criteria.

Model Selection via AIC and BIC

In linear regression of Y on explanatory variables $x_1, \ldots, x_p$, one may use a t-statistic to assess the contribution of x_p beyond that of $x_1, \ldots, x_{p-1}$. More generally, an F-statistic may be used to assess the contribution of $x_{k+1}, \ldots, x_p$ beyond that of $x_1, \ldots, x_k$. The latter may be considered a comparison of the fit from the model based on $x_1, \ldots, x_k$ to that from the model based on $x_1, \ldots, x_p$. (The latter fit is necessarily better; the

issue is whether the improvement is best described as arising solely from noise.) This is a comparison of *nested* models, because the first model is a special case of the second when the regression coefficients of the last $p-k$ variables are set equal to zero. On the other hand, when non-nested models of different dimensionality are to be compared, the F-statistic no longer applies. To deal with comparison of non-nested models, the criteria AIC and BIC are often applied. Furthermore, they can be helpful supplements even for nested models. AIC is the *Akaike information criterion* and BIC is the *Bayesian information criterion*. AIC was introduced by Akaike in 1974; BIC by Schwarz in 1978.

These model selection criteria begin with the observation that models with more parameters should provide better fits—in the case of nested models they necessarily do so when fit is judged by likelihood. Thus, it is natural to try to *penalize* models according to dimensionality. AIC and BIC include the maximized loglikelihood and then subtract a penality for dimensionality. By convention (to match the loglikelihood ratio statistic) the criterion includes a multiplier of -2. We then have

$$\text{criterion} = -2 \cdot \max \text{ loglikelihood} + \text{penalty}$$

and

$$\text{AIC penalty} = 2p$$

where p is the number of parameters in the model, while

$$\text{BIC penalty} = p \log n$$

where n is the sample size. (Many variants on these model selection criteria have also been proposed; they begin with the same idea, and have more or less the same general form.) Note that in this form smaller values of the criterion indicate better models.

The motivation for AIC is as follows. An appealing general measure of discrepancy of a pdf $f(x)$ from another pdf $g(x)$ is the Kullback-Leibler number

$$K(g,f) = E_g(\log \frac{g(X)}{f(X)}) = \int g(x) \log \frac{g(x)}{f(x)} dx.$$

Note that

$$K(g,f) = E_g(\log(g(X))) - E_g(\log(f(X))).$$

The Kullback-Liebner number is essentially unique among all discrepancies $D(g,f)$ that satisfy

(i) $D(g, f) = E_g(\varphi(g(X))) - E_g(\varphi(f(X)))$ for some differentiable function φ, and

(ii) $D(g, f)$ is minimized over f by $f = g$.

(See Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, Springer, Section 3.1.) Now, suppose we let $g(x)$ be the true pdf and we wish to obtain a model with pdf $f(x)$ that is a close as possible to $g(x)$ in the sense of minimzing $K(g, f)$. When we minimize over $f(x)$ we are maximizing $E_g(\log(f(X)))$. Consider the special case of trying to determine the value of a single scalar parameter θ, where the true value is θ_0, based on data x. Then we are trying to find the closest pdf $f(x|\theta)$ to $g(x) = f(x|\theta_0)$. It is not hard to show that the expectation $E_g(\log f(X|\theta))$ is maximized by $\theta = \theta_0$. Because θ_0 is unknown we might use the loglikelihood $\log f(x|\theta)$ as an estimate of $E_g(\log f(X|\theta)$, and thus might maximize to get the maximized loglikelihood $\log f(x|\hat{\theta})$. But this is, in general, a biased estimate of $E_g(\log f(X|\theta)$. Akaike proposed to subtract off an estimate of the bias, and then showed that the bias is, in general, approximately equal to the dimensionality of θ. (See Konishi and Kitagawa (2008) for full details.) Multiplying the maximized loglikelihood by -2 gives the form of AIC above.

BIC begins, instead, with the Bayesian formulation of choosing between models M_1 and M_2 based on posterior probability:

$$p(M_1|x) = \frac{f_1(x)p(M_1)}{f_1(x)p(M_1) + f_2(x)p(M_2)}$$

where $f_i(x)$ are the pdfs and $p(M_i)$ are prior probabilities under model M_i $(i = 1, 2)$. To eliminate the prior probabilities one may use the *Bayes factor*, which is the ratio of posterior odds to prior odds:

$$BF = \frac{p(M_1|x)}{p(M_2|x)} \div \frac{p(M_1)}{p(M_2)}.$$

Equivalently, one may consider the log odds $\log BF$. It may then be shown that asymptotic approximation of $\log BF$, as $n \to \infty$, leads to the form for BIC given above. See Kass, R.E. and Raftery, A. (1995) Bayes factors, *J. Amer. Statist. Assoc.*, 90: 773–795. More accurate approximations provide additional intuition, as reviewed by Kass and Raftery. From a more general perspective, BIC is consistent in the sense that, for sufficiently large samples, the probability of BIC choosing the correct model will get arbitrarily close to 1. In practice,

the most important fact is that BIC is conservative compared to AIC in the sense of imposing a larger penalty for dimensionality.

Spectral Densities for ARMA Processes

We previously noted the following. If $\{X_t; t \in \mathcal{Z}\}$ is a stationary process with spectral density $f_X(\omega)$, and $\{a_h; h \in \mathcal{Z}\}$ satisfies

$$\sum_{h=-\infty}^{\infty} |a_h| < \infty$$

and we write

$$A(\omega) = \sum_{h=-\infty}^{\infty} a_h e^{-2\pi i \omega h}$$

then the filtered process $\{Y_t; t \in \mathcal{Z}\}$ defined by

$$Y_t = \sum_{h=-\infty}^{\infty} a_h X_{t-h}$$

is stationary with spectral density

$$f_Y(\omega) = |A(\omega)|^2 f_X(\omega).$$

Note that when W_t is a white noise process with variance σ_W^2 its spectral density is $f_W(\omega) = \sigma_W^2$ for all ω. We may apply the filtering result to the $AR(p)$ model

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + W_t$$

by taking $a_0 = 1$, $a_j = \phi_j$ for $j = 1, \ldots, p$ and $a_h = 0$ otherwise to get the spectral density

$$f_X(\omega) = \frac{\sigma_W^2}{|\phi(e^{-2\pi i \omega})|^2}$$

where $\phi(z)$ is the AR polynomial. A similar argument applied to an $ARMA(p, q)$ process gives

$$f_X(\omega) = \frac{\sigma_W^2 |\theta(e^{2\pi i \omega})|^2}{|\phi(e^{-2\pi i \omega})|^2}$$

where $\theta(z)$ is the MA polynomial.

Here is some further analysis of the core body temperature data. First, the residuals were detrended. Then the AR regression was done a bit more carefully. Finally, an $AR(3)$ spectral density estimate was plotted.

Lagged regression using 22 lags:

```
     Estimate Std. Error t value Pr(>|t|)
x1   0.905660   0.057030  15.881  < 2e-16 ***
x2  -0.205181   0.076792  -2.672  0.00794 **
x3  -0.147316   0.077676  -1.897  0.05882 .
x4   0.005267   0.077790   0.068  0.94606
x5  -0.153942   0.077764  -1.980  0.04864 *
x6   0.073491   0.078220   0.940  0.34819
x7   0.005699   0.077944   0.073  0.94176
x8   0.014181   0.077937   0.182  0.85574
x9  -0.134053   0.077764  -1.724  0.08574 .
x10  0.032195   0.078051   0.412  0.68027
x11 -0.040730   0.077737  -0.524  0.60069
x12 -0.012715   0.077719  -0.164  0.87015
x13  0.026797   0.077706   0.345  0.73044
x14 -0.057330   0.077444  -0.740  0.45970
x15  0.003742   0.077577   0.048  0.96156
x16 -0.114838   0.077512  -1.482  0.13949
x17  0.003559   0.077632   0.046  0.96346
x18  0.011683   0.077102   0.152  0.87966
x19 -0.021889   0.077123  -0.284  0.77673
x20 -0.016421   0.077049  -0.213  0.83137
x21 -0.030719   0.076008  -0.404  0.68638
x22 -0.011373   0.057028  -0.199  0.84206
```

Lagged regression using 3 lags:

```
          Estimate Std. Error t value Pr(>|t|)
x[, 1:3]1  0.94064    0.05444  17.278  < 2e-16 ***
x[, 1:3]2 -0.19712    0.07445  -2.648  0.00850 **
x[, 1:3]3 -0.17676    0.05438  -3.250  0.00127 **
```

Lagged regression using 4 lags:

```
          Estimate Std. Error t value Pr(>|t|)
```

```
x[, 1:4]1  0.93463    0.05536  16.884  < 2e-16 ***
x[, 1:4]2 -0.20387    0.07533  -2.707  0.00716 **
x[, 1:4]3 -0.14470    0.07533  -1.921  0.05562 .
x[, 1:4]4 -0.03402    0.05525  -0.616  0.53852
```

This suggests $AR(3)$ might be appropriate. By default, the R function `ar` uses AIC to select among AR models. It selected $AR(3)$. Here are the AIC values, differenced from the value for $AR(3)$, for the first 9 lags:

```
         0          1          2          3          4          5          6
326.758200  59.099883   6.843735   0.000000   1.689885   2.743205   2.931410
         7          8          9
  4.080634   4.286393   3.330592
```

And here are the fitted coefficients via ML for the $AR(3)$:

```
arima(x = TempBARSdetrend, order = c(3, 0, 0))
Coefficients:
         ar1      ar2      ar3  intercept
      0.9496  -0.2204  -0.1572     0.0004
s.e.  0.0526   0.0721   0.0526     0.0131

sigma^2 estimated as 0.01112:  log likelihood = 291.85,  aic = -573.71
```

Finally, the $AR(3)$ spectral density estimate is shown in Figure 3. Note that it is very smooth. The Whittle smoothed periodogram is shown for comparison. The main distinction between these two smooth estimates is at the lower frequencies. In any case, there appears to be a peak near $\omega_j = .1$. To interpret this, we need units. The temperature was sampled every 20 minutes, and there were 352 observations. If $\omega_j = .1$, then the frequency is .1 per time unit (or 35.2 per 352 time units). To get frequency per day we multiply by 72 and get roughly 7. There appears to be a quasi-oscillation with a period of roughly 3.5 hours.

R Code

```
library(Hmisc) # to get Lag function
x=rep(0,330*22)
```

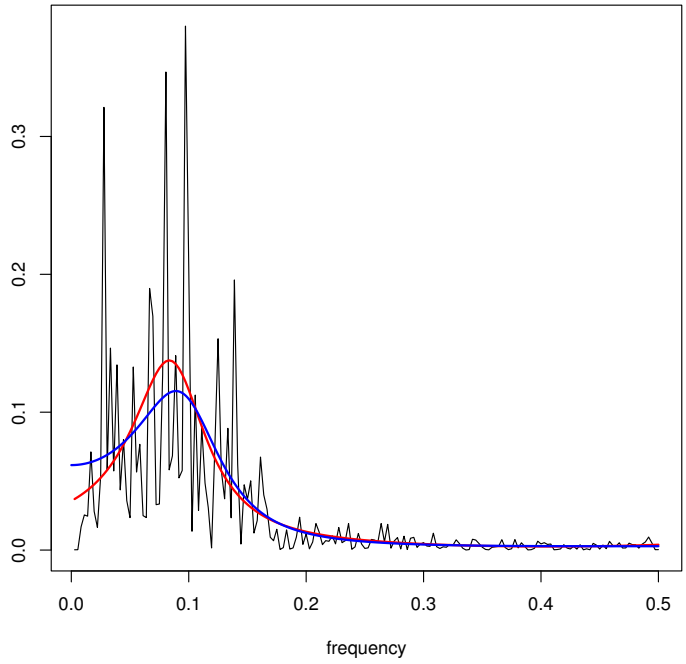

Figure 3: *Spectral density estimates for the BARS-detrended residuals from the core body temperature data, after removing the fitted 24-hour cycle. The tapered periodogram is in black; the Whittle smoothed version is in red; and the estimate from the AR(3) model is in blue.*

```
dim(x)=c(330,22)
y=TempBARSdetrend[23:352]
for(i in 1:22){
  x[,i]=Lag(TempBARSdetrend,i)[23:352]
}
ar.reg=lm(y~x-1)
summary(ar.reg)
ar.reg3=lm(y~x[,1:3]-1)
summary(ar.reg3)
ar.reg4=lm(y~x[,1:4]-1)
summary(ar.reg4)
temp.spec.ar=spec.ar(TempBARSdetrend,plot=FALSE)
temp.spec=spec.pgram(TempBARSdetrend,plot=FALSE)
temp.sm.wh=glm(temp.spec$spec~ns(temp.spec$freq,df=5),
                                family=Gamma)$fitted.values
plot(temp.spec$freq,temp.spec$spec,type="l",
```

```
                                xlab="frequency",ylab=" ")
lines(temp.spec$freq,temp.sm.wh,col="red",lwd=2)
lines(temp.spec.ar$freq,temp.spec.ar$spec,col="blue",lwd=2)
dev.print(device=postscript,file="temp.spec.ar.ps")

temp.ar.results=ar(TempBARSdetrend,order.max=9)
temp.ar.results$aic
arima(TempBARSdetrend,order=c(3,0,0))
```