Time Series

Peter Bühlmann
ETH Zürich, Switzerland

August 2000

Many data occur in the form of *time series* where observations are recorded at different times. The presence of the notion of time often creates dependence among observations whose structure is of particular interest in time series analysis.

When viewed from a stochastic perspective, time series data x_t with $t = 1, 2, \ldots, n$ are one finite realization of sample size n from a stochastic process $\{X_t : \ t = 0, \pm 1, \pm 2, \ldots\}$: x_t and X_t are generally r-vectors whose components take values on the real line, or in some discrete space being categorical or having an order. The restriction to equi-spaced time recordings can be relaxed to $t_1 < t_2 < \ldots < t_n$ and the data generating process can be indexed by continuous time $\{X_t\}$, for example with $t \in [0, 1]$. Until the last section, we focus on the univariate case with $r = 1$.

Reduction to stationarity

For many statistical purposes, the data generating process $\{X_t\}$ should exhibit some redundancy: either $\{X_t\}$ is itself stationary or it can be reduced to a stationary component. For the latter, the following basic model is often used:

$$X_t = m_t + s_t + Z_t, t = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

with $\{m_t\}$, $\{s_t\}$ deterministic trend and seasonal effects and $\{Z_t\}$ a stationary stochastic process with mean zero. Trend and seasonal effects are either modeled directly, e.g. using the STL decomposition (Cleveland [9]); or with another approach as described below for a real dataset about ozone measurements. Or they are taken out by differencing

$$((1 - B)^d (1 - B^s) X)_t, \ d, s \in \{0, 1, 2, \ldots\}, \tag{2}$$

where $B(\cdot)$ denotes the backshift operator defined by $(Bx)_t = x_{t-1}$. For example, assume in (1) that $m_t = \alpha_0 + \alpha_1 t + \ldots + \alpha_d t^d$ is a polynomial trend of order d and $s_t = \beta_{t \bmod s}$ a seasonal component with period s, i.e. $s_t = s_{t \pm s} = s_{t \pm ks}$ for all t and all k. Then, $\{((1 - B^s)X)_t\}$ is free of seasonal variation and differencing again yields $\{((1 - B)^d (1 - B^s) X)_t\}$, which is also free of trend. Of course, the stationary component then also transforms to $((1 - B)^d (1 - B^s) Z)_t$ but still being a mean zero, stationary process. The differencing strategy is used in the popular SARIMA models (Box and Jenkins [3]; Brockwell and Davis [6]; Shumway and Stoffer [21]), see p. 5. Other decompositions can be thought of so that a stationary component $\{Z_t\}$ can be estimated from the data $x_1, \ldots, x_n$: modeling this stationary component is a prominent and maybe more difficult task than for the deterministic parts.

Example: ozone data

As an example for model based reduction to stationarity, consider the world's longest time series on total ozone measurements from Arosa, Switzerland, during the period 1926–1997. It is an important source to assess the ozone depletion in the northern mid-latitude hemisphere. The measurements are currently performed by the Swiss Meteorological Institute. The homogenized data set is available from http://www.lapeth.ethz.ch/doc/totozon.html. The raw monthly measurements $\{O_t\}$ exhibit large seasonal effects, which can be explained very well; see Figure 1.

Figure 1 about here

Assuming fixed monthly effects β_i $(i = 0, \ldots, 11)$ with $\sum_{i=0}^{11} \beta_i = 0$, we deseasonalize the series by preliminary smoothing with a running mean $X_t = \sum_{i=-6}^{6} c_i O_{t+i}$ with $c_i = 1/12$ $(i = -5, \ldots, 5)$ and $c_i = 1/24$ $(i = -6, 6)$. Figure 1 also displays the filtered data $(X_t)_{t=1}^{n}$ with $n = 814$ on the Dobson scale. We consider then the decomposition

$$X_t = m_t + Z_t, \;\; t = 1, \ldots, n = 814, \tag{3}$$

where $m_t = m(t/n)$, $m(\cdot)$ a smooth mean function from $[0, 1] \to \mathbb{R}$. Moreover, $\{Z_t\}$ is assumed stationary with $\mathbb{E}[Z_t] = 0$. The estimated trend $\hat{m}_t = \hat{m}(t/n)$, using a Gaussian *kernel estimator* with bandwidth $h = 0.05n^{-1/5} = 0.0131$ is also displayed in Figure 1. The remaining residual part $\hat{Z}_t = X_t - \hat{m}_t$, can now be further analyzed assuming (approximate) stationarity, see later on p. 10–11. The decomposition used here has the following empirical variances as summary statistics: 1219.4 Dobson^2 for raw data $\{O_t\}$, 91.7 Dobson^2 for deseasonalized data $\{X_t\}$, and 19.5 Dobson^2 for detrended data $\{\hat{Z}_t\}$. We will see that stationary modeling of $\hat{Z}_t$ on p. 10–11 with 1.0 Dobson^2 as remaining noise variance brings a substantial further reduction in variability. One main interest in this example is the study of a possibly varying mean trend: the statistical accuracy and variability of the curve estimator $\hat{m}_t$, displayed in Figure 1, is crucial for deciding whether the mean trend changes over time. This is discussed later on p. 11.

Sometimes, the reduction to a stationary component seems unrealistic or hard to do and it can be advantageous to use a *locally stationary* model (Dahlhaus [10]): the idea being that within a suitable window of time, the process is approximately stationary. (The window depends then on sample size and the local stationary model produces an array of random variables $X_t^{(n)}$ $(t = 1, \ldots, n;\; n = 1, 2, \ldots)$). Another concept to model conditionally non-stationary phenomena (e.g. processes with switching regimes) are given by *state space models*, see p. 7–8.

First and second moments of stationary time series

Suppose that $\{X_t\}$ is a univariate stationary time series. The first moment $\mathbb{E}[X_t]$ is often not of particular interest since it doesn't describe any dependence phenomena. It can be estimated by the mean $\overline{X}_n = n^{-1} \sum_{t=1}^{n} X_t$: the difference to the i.i.d. situation is its variance described below in (5) implying also different confidence regions based on $\overline{X}_n$.

Of more interest is the second moment structure. The autocovariance function is defined by

$$R(k) = \mathrm{Cov}(X_0, X_k), \;\; k = 0, \pm 1, \pm 2, \ldots,$$

which is symmetric around zero. The standardized, dimensionless version, taking values between -1 and 1, is the autocorrelation function

$$\rho(k) = R(k)/R(0).$$

A process satisfying $R(k) = \rho(k) = 0$ for all $k \neq 0$ is called *white noise*. This name is motivated from the spectral domain, see p. 13–16, which generally yields an equivalent description of the autocovariance structure but often with a more meaningful interpretation. In some cases, the interesting quantity is the covariance between X_0 and X_k after having removed the linear effect of the variables $X_1, \ldots, X_{k-1}$ in between. This leads to the partial autocorrelation

$$\alpha(k) = \mathrm{Corr}(X_k - \hat{X}_{k;1:k-1}, X_0 - \hat{X}_{0;1:k-1}), \ k = 1, 2, \ldots$$

and $\alpha(-k)$ defined to be equal to $\alpha(k)$. Here, $\hat{X}_{s;1:t-1}$ denotes the best linear predictor of X_s based on $X_1, \ldots, X_{t-1}$ (including a constant intercept term). We will see, that the partial autocorrelation function can be useful in identifying some types of models, see p. 5.

The estimate of the autocovariance is most often as follows,

$$\hat{R}(k) = n^{-1} \sum_{t=1}^{n-|k|} (X_t - \overline{X}_n)(X_{t+|k|} - \overline{X}_n), \ -n+1 \leq k \leq n-1. \tag{4}$$

From this we use the plug-in estimate for autocorrelations $\hat{\rho}(k) = \hat{R}(k)/\hat{R}(0)$. The sample version of the partial autocorrelation can be calculated with partial regression or with a Yule-Walker system of equations in terms of the sample autocorrelations as in (10), see Abraham and Ledolter [1].

Already for the mean, the inference part is much different from the i.i.d. set-up. The variance of $\overline{X}_n$ is

$$\mathrm{Var}(\overline{X}_n) = n^{-1} \sum_{k=-n+1}^{n-1} R(k)(1 - |k|/n). \tag{5}$$

If the autocovariances are absolutely summable, the expression behaves as $\mathrm{Var}(\overline{X}_n) \sim n^{-1} \sum_{k=-\infty}^{\infty} R(k)$. Under suitable additional conditions (e.g. mixing conditions describing the decay of dependence as a function of separating lags), the mean is asymptotically normally distributed

$$n^{1/2}(\overline{X}_n - \mathbb{E}[X_t]) \Rightarrow \mathcal{N}(0, \sigma^2_{as}), \ \sigma^2_{as} = \sum_{k=-\infty}^{\infty} R(k).$$

The asymptotic distribution of the estimated autocorrelations $\hat{\rho}(\cdot)$ at lags from 1 to p, centered around the true values $\rho(\cdot)$, is asymptotically normal with a complicated expression for the limiting variance. However, if the data generating process is white noise, i.e. $R(k) = 0$ for $k \neq 0$, the asymptotic distribution is

$$n^{1/2}(\hat{\rho}(1), \ldots, \hat{\rho}(p))^T \Rightarrow \mathcal{N}_p(0, I), \tag{6}$$

where I denotes the identity matrix. The result shows that the estimated autocorrelations at lags 1 to p are asymptotically i.i.d. standard normal variables. Thus, the region $\pm 1.96 n^{-1/2}$ contains with probability 0.95 the sample autocorrelations of a white noise process. When plotting the sample autocorrelations, this band is usually added as a reference for strength of correlation, see Figure 2. The analogous result as in (6) holds for the sample partial autocorrelation: if the underlying partial autocorrelations are all zero, the band $\pm 1.96 n^{-1/2}$ describes an acceptance region, which is plotted in Figure 2. Setting a confidence band for the auto- or partial autocorrelations is a more ambitious task. The block bootstrap (Künsch [16]) yields reasonable results for general stationary time series, see also p. 13.

Figure 2 about here

Example: ozone data (continued)
The sample autocorrelation and its partial version for the deseasonalized and detrended ozone series $\hat{Z}_t = X_t - \hat{m}_t$ are displayed in Figure 2. It suggests a substantial dependence structure being only non-significant for lags greater than 26.

Models for stationary time series

A major task in time series analysis is modeling the stationary component of a time series. Models can be classified as linear or nonlinear. We assume in the following that the stationary process has mean $\mathbb{E}[X_t] = 0$. This is formally achieved by replacing X_t with $X_t - \mu,\ \mu = \mathbb{E}[X_t]$ in all the model-specifications below. In practice, the mean $\overline{X}_n$ is often subtracted from the data to get a mean zero time series.

Linear models

Autoregressive models build a simple yet useful class of models: $\{X_t\}$ is called autoregressive of order p (AR(p)) if it satisfies the difference equation

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-1} + \varepsilon_t,\ t = 0, \pm 1, \pm 2, \dots\ , \tag{7}$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. disturbances with $\mathbb{E}[\varepsilon_t] = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$ and ε_t independent from all $X_s,\ s < t$. Stationarity of such models is ensured via a condition on the AR-polynomial in complex variables

$$\Phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j \neq 0 \text{ for all complex } z \text{ with } |z| \leq 1. \tag{8}$$

With $\Phi(\cdot)$ as in (8) and the backshift operator $B(\cdot)$ used in (2), AR-models are often compactly represented as

$$(\Phi(B)X)_t = \varepsilon_t,\ t = 0, \pm 1, \pm 2, \dots$$

Autoregressive models are obviously linear in terms of lagged observations. It can be useful to include also linear relations in terms of lagged, unobservable disturbances. This is incorporated with *autoregressive moving average models*: $\{X_t\}$ is called autoregressive moving average of orders p, q (ARMA(p, q)) if

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-1} + \sum_{k=1}^{q} \psi_k \varepsilon_{t-k} + \varepsilon_t,\ t = 0, \pm 1, \pm 2, \dots\ ,$$

with $\{\varepsilon_t\}$ as in (7). Stationarity is again ensured by condition (8). As a special case we can also consider the *moving average model* of order q (MA(q)) being an ARMA($0, q$). The inclusion of lagged disturbances can be viewed as a parsimonious parameterization for processes whose conditional distribution of X_t given past values depends on the infinite past $X_{t-1}, X_{t-2}, \dots$ For example, an MA(1) with $|\psi_1| < 1$ can be represented as an AR(∞),

$$X_t = \sum_{j=1}^{\infty} (-1)^{j+1} \psi_1^j X_{t-j} + \varepsilon_t.$$

In the class of ARMA models, inspection of the autocorrelation- and partial autocorrelation function gives some hints about reasonable models, see also p. 10–11 on model selection. For an MA(q) model, the autocorrelation $\rho(k) = 0$ for all $k > q$; for an AR(p) model, the partial autocorrelation $\alpha(k) = 0$ for all $k > p$. Moreover, $\rho(k)$ and $\alpha(k)$ for stationary ARMA(p,q) models decay exponentially as k increases. The ARMA models build a powerful class within linear processes

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j},\ t = 0, \pm 1, \pm 2, \ldots ,$$

being a linear combination of i.i.d. mean zero disturbances. Gaussian processes, i.e. all finite-dimensional distributions are multivariate normal, are special cases of linear processes. The converse is not true: there are non-Gaussian but linear processes. For example, an ARMA(p,q) with non-Gaussian innovations. The SARIMA(p,d,q;P,D,Q) model mentioned in connection with (2) is specified as

$$((1-B)^d (1-B^s)^D \Phi(B) \Phi_{season}(B^s) X)_t = (\Theta(B) \Theta_{season}(B^s) \varepsilon)_t,$$

where $\Phi(\cdot),\ \Phi_{season}(\cdot)$ are AR-generating polynomials of degrees p, P as in (8) and $\Theta(\cdot),\ \Theta_{season}(\cdot)$ are MA-polynomials of degrees q, Q, both of the form

$$\Theta(z) = 1 + \sum_{j=1}^{q} \theta_j z^j,\ z \text{ a complex number.} \tag{9}$$

Note that the SARIMA model goes beyond the differencing operation in (2): using the latter and modeling its output as ARMA(p,q) is a SARIMA(p,d,q;0,1,0) model. The SARIMA model tries additionally to capture non-stationarity in the second-order autocovariance structure. Note that $(\Phi(B)\Phi_{season}(B^s)X)_t = (\Theta(B)\Theta_{season}(B^s)\varepsilon)_t$ build an ARMA model of high order but parsimoniously parameterized with many coefficients equal to zero. For $d \neq 0$, the root condition in (8) fails which is a reason for non-stationarity of the SARIMA process (and analogously for $s, D \neq 0$). Such SARIMA models build a popular class for non-stationary (linear) processes.

Estimation of parameters in autoregressive models can be done with the *Yule-Walker method* based on estimating second moments,

$$\begin{aligned} &(\hat{\phi}_1, \ldots , \hat{\phi}_p)^T = \hat{\Gamma}^{-1}(\hat{\rho}(1), \ldots , \hat{\rho}(p))^T, \\ &\hat{\Gamma} \text{ a } p \times p \text{ matrix},\ (\hat{\Gamma})_{ij} = \hat{\rho}(|i-j|),\ 1 \leq i, j \leq p. \end{aligned} \tag{10}$$

For ARMA models, the *maximum likelihood estimator* (see p. 8–10) can be used. It has to be computed by some numerical optimization technique.

Markovian models

An important class of nonlinear stationary models have the Markov property, including the case with Markov order $p > 1$. That is, the distribution of X_t given the infinite past $X_{t-1}, X_{t-2}, \ldots$ depends only on the first p lagged values $X_{t-1}, \ldots , X_{t-p}$. For example, the generalization of the linear AR model is

$$X_t = f(X_{t-1}, \ldots , X_{t-p}) + \varepsilon_t, \tag{11}$$

with $\{\varepsilon_t\}$ as in (7). The real-valued function $f(\cdot)$, denoting the conditional mean of X_t given the past, may be of nonlinear parametric form or it can be also nonparametric. Denoting by x

the p-vector $(x_1, \ldots, x_p)^T$, the drift condition $\limsup_{x, \|x\| \to \infty} |f(x)|/\|x\| < 1$, ensuring that the function $f(\cdot)$ does not expand, implies stationarity of the process, provided the distribution for ε_t is suitably regular, e.g. Gaussian or Student t. A popular class of parametric specifications in (11) are *threshold autoregressive models*: a simple example is a SETAR$(2; p_1, p_2)$, where the conditional mean function is parameterized as

$$f(x_1, \ldots, x_p) = (\mu_1 + \sum_{j=1}^{p_1} \phi_{1,j} x_j) I_{[x_d \le \gamma]} + (\mu_2 + \sum_{j=1}^{p_2} \phi_{2,j} x_j) I_{[x_d > \gamma]}, \ \max(p_1, p_2, d) = p,$$

describing a piecewise linear model with one threshold. Other more sophisticated parameterizations to partition the predictor space with thresholds into regions with different linear dynamics are described in Tong [22]. For details: $\rightarrow$ *Threshold autoregressive models*. The model in (11), and also the AR model in (7) may be extended to include stochastically heteroscedastic disturbances,

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + g(X_{t-1}, \ldots, X_{t-r}) \varepsilon_t, \tag{12}$$

with $g(\cdot)$ a positive valued function and $\{\varepsilon_t\}$ as in (7) but now standardized (for identifiability reasons) to $\mathrm{Var}(\varepsilon_t) = 1$. A very popular parametric form of $g(\cdot)$, mainly used in econometric applications, is

$$g(x_1, \ldots, x_r) = \alpha_0 + \sum_{j=1}^{r} \alpha_j x_j^2, \ \alpha_j > 0 \text{ for all } j = 0, 1, \ldots, r,$$

from the ARCH (*autoregressive conditional heteroscedastic*) model.

Estimation in such Markovian models can be done with maximum likelihood for parametric specifications (see p. 8–10) and with *nonparametric methods* otherwise (see p. 10). Of course, semiparametric models with parametric and nonparametric components are possible as well. Due to the Markovian structure of finite order, say p, we can always view such models from a (stochastic) regression point of view with response variable $Y_t = X_t$ and explanatories $\mathbf{X}_t = (X_{t-1}, \ldots, X_{t-p})^T$. Estimation routines can then be borrowed from the regression framework.

Some data also take values in a finite space $\mathcal{X}$, equipped with a *scale*, or with an *order*, or it may be purely *categorical*. In all cases, the effort goes into modeling the transition probabilities $\mathbb{P}[X_t = x | X_{t-1}, \ldots, X_{t-p}]$, $x \in \mathcal{X}$. For scaled and ordinal valued time series, autoregressive-type *generalized linear models* (GLM) are useful; for categorical data, transition probabilities can be estimated with empirical frequencies. For the latter, however, the curse of dimensionality might be heavily present when p is moderate or large, saying that the number of observations needed to populate a space with pre-specified density increases exponentially as its dimension increases. Restricting to more parsimonious transition models is very essential.

Non-Markovian models with lagged disturbances

The moving average modeling idea can be extended to nonlinear functional relations with lagged disturbances. The following is a very general prototype including also a model component of such a form,

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) + g(X_{t-1}, \ldots, X_{t-r}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-s}) \varepsilon_t, \ t = 0, \pm 1, \pm 2, \ldots, \tag{13}$$

with $g(\cdot)$ and $\{\varepsilon_t\}$ as in (12). This may try to approximate the following decomposition of a general stationary time series (assuming first and second conditional moments exist),

$$X_t = \mu_t + \sigma_t \eta_t, \ t = 0, \pm 1, \pm 2, \ldots,$$

where $\mu_t = \mathbb{E}[X_t|X_{t-1}, X_{t-2}, \dots]$, $\sigma_t^2 = \mathrm{Var}(X_t|X_{t-1}, X_{t-2}, \dots)$ and η_t's are uncorrelated variables with $\mathbb{E}[\eta_t] = 0$ and $\mathrm{Var}(\eta_t) = 1$ for all t. The functions $f(\cdot)$ and $g(\cdot)$ in (13) model the conditional mean and variance, and the i.i.d. innovations in (13) can be viewed as an approximative stochastic component for the uncorrelated, but generally not i.i.d. distributed, variables η_t. If lagged disturbances are functionally related to X_t such as in (13), the models for μ_t and σ_t specified by the functions $f(\cdot)$ and $g(\cdot)$ are depending on the infinite past $X_{t-1}, X_{t-2}, \dots$, explaining the non-Markovian character of the process.

There is of course a vast number of possibilities for modeling the functions $f(\cdot)$ and $g(\cdot)$ in (11), (12) and (13). Similar to nonlinear and nonparametric regression problems, the modeling may incorporate parametric mechanistic understanding of the problem under study, or nonparametric techniques, or a combination of both.

Long memory processes

We refer to a stationary long memory process if its autocovariances are not absolutely summable. One very popular long memory model is a fractional ARIMA(p,d,q) process with $d \in (-1/2, 1/2)$,

$$((1-B)^d X)_t = Y_t, \ \{Y_t\} \text{ a stationary ARMA(p,q)}.$$

Here $B(\cdot)$ denotes the backshift operator as in (2) but with fractional d: $(1-B)^d$ is then represented with a binomial expansion. For more details about long memory processes and their use in statistical modeling see Beran [2].

State space models

State space models build a general, flexible class involving a latent, non-observable process. Stationarity of such a model is in many applications not an important issue. The model has the general form

$$\begin{aligned}
&\{X_t\} \text{ is a } \textit{Markov chain} \text{ (state process)} \\
&\text{all } Y_t\text{'s are conditionally independent given } \{X_s : s = 0, \pm 1, \pm 2, \dots\}, \\
&Y_t \text{ given } X_t \text{ is conditionally independent from } \{X_s, s \neq t\}.
\end{aligned} \tag{14}$$

If X_t is discrete, we refer to (14) as *Hidden Markov model*. An example would be a switching two-regime AR model

$$\begin{aligned}
&\{X_t\} \text{ binary Markov chain}, \\
&Y_t = \alpha_{X_t} + \beta_{X_t} Y_{t-1} + \varepsilon_t, \ t = 0, \pm 1, \pm 2, \dots
\end{aligned}$$

with i.i.d. disturbances ε_t, independent from all X_s and from Y_s with $s < t$. Note that this is of the form (14) by using the two-dimensional observation variable $\mathbf{Y}_t = (Y_t, Y_{t-1})^T$.

The linear Gaussian state space model is

$$\begin{aligned}
X_t &= G_t X_{t-1} + V_t, \\
Y_t &= H_t X_t + \varepsilon_t,
\end{aligned} \tag{15}$$

where G_t, H_t are deterministic matrices and $\{V_t\}, \{\varepsilon_t\}$ two independent i.i.d. Gaussian sequences $V_t \sim \mathcal{N}(0, \Sigma_V)$, $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$. Besides latent structure processes, this includes for example the ARMA model. Estimation of the best state variables $X_1, \dots X_n$ based on observed data $Y_1, \dots, Y_n$, namely the mode or expectation of the conditional distribution of $X_1, \dots, X_n$ given observed $Y_1, \dots, Y_n$, can be computed recursively with the famous *Kalman filter*. The technique also allows to estimate unknown parameters in (15). For more details: $\rightarrow$ *State space models*.

Regression with stationary errors

Another structural time series model, a very specific state space model, is as in (1): the process involves exogenous variables causing a trend $\{m_t\}$ overlaid by stationary noise $\{Z_t\}$,

$$X_t = m_t + Z_t,\ \ m_t = f(\mathbf{u}_t),$$

with $\mathbf{u}_t = (u_{t,1}, \dots , u_{t,d})^T$ the vector of exogenous variables. Instead of separating the estimation of m_t and Z_t, we can look at the problem simultaneously.

For example, $m_t = \beta^T \mathbf{u}_t$ a linear regression function with parameter vector β and the noise $Z_t = \phi Z_{t-1} + \varepsilon_t$ modeled as AR(1). Efficient estimation can then be implemented by using *nonlinear least squares* in the transformed model

$$X_t = \phi X_{t-1} + \sum_{i=1}^{p} \beta_i (u_{t,i} - \phi u_{t-1,i}) + \varepsilon_t,$$

see Abraham and Ledolter [1]. Inference can then be based on stochastic regression with i.i.d. noise variables. Another general approach is given by *two stage least squares*, see Hannan [13], Harvey [15]; the fundamentals are due to Grenander and Rosenblatt.

Estimation, model selection and inference

Estimation in parametric models

Estimation of a parameter θ can be generally done by maximum likelihood. If the time series is Markovian, say of order p, the log-likelihood (conditioned on the first p variables) is

$$\ell(\theta) = \sum_{t=p+1}^{n} \log f(X_t | X_{t-1}, \dots , X_{t-p}; \theta) \tag{16}$$

where $f(\cdot | X_{t-1}, \dots X_{t-p})$ denotes the conditional density of X_t given $X_{t-1}, \dots X_{t-p}$. This in exact analogy to the regression setting with response variable $Y_t = X_t$ and explanatories $\mathbf{X}_t = X_{t-1}, \dots , X_{t-p}$. Particularly, the computation of the maximum likelihood estimator $\hat{\theta}$ being the minimizer of $-\ell(\theta)$ may be implemented with computing routines from regression modeling; or using a numerical minimization algorithm such as **nlmin** from *S-Plus* or **nlm** from *R* (freely available from http://www.stat.auckland.ac.nz/r/r.html).

Example: river flow

We illustrate the fitting of Markovian models on data about daily flow of the Fisher river (near Dallas, Texas) from January 1, 1988 – December 31, 1991. The raw flow data $\{F_t : t = 1, \dots , 1461\}$, in cubic meters per second, and the transformed series $X_t = \log(F_t) - \log(F_{t-1})$ are displayed in Figure 3. The transformed series seems from a pure data-analytic view easier to analyze, since the extremes are much less pronounced. Also, using a first order Taylor expansion,

$$X_t = \log(F_t/F_{t-1}) \approx \frac{F_t - F_{t-1}}{F_{t-1}},$$

which describes the relative gain within a time unit (here one day). This transformation is often useful in economic applications, where X_t is usually referred as the return of the underlying quantity.

Figure 3 about here

We consider two models for the relative gains. The first is an AR(9) model with Gaussian innovations ε_t (see p. 4): the order 9 is chosen with the AIC model selection criterion (see p. 10). The mean $\mathbb{E}[X_t]$ appears to be non-significantly different from zero and we exclude it from the model. The residuals from the AR(9) model are examined in Figure 4 (top row): they appear to be uncorrelated but still dependent due to the significant autocorrelations of their squares. The other model is a combination of AR and ARCH, of the form as in (12),

$$\begin{aligned} X_t &= \sum_{j=1}^{9} \phi_j X_{t-j} + \sigma_t \varepsilon_t,\ \varepsilon_t \sim \frac{\nu-2}{\nu} t_\nu, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2,\ \alpha_0, \alpha_1 \geq 0. \end{aligned} \tag{17}$$

The scaling with the factor $(\nu-2)/\nu$ ensures that ε_t has variance one; $\nu = \infty$ denotes standard normal innovations. The residuals from this AR(9)-ARCH(1) model in (17) are also displayed in Figure 4 (bottom row) indicating graphically that independence is much more likely.

Figure 4 about here

The negative log-likelihoods (the lower the better), as a goodness of fit measure, are as follows:

$$\begin{aligned} -\ell &= -712.1 \text{ for AR(9) with } \mathcal{N}(0,\sigma^2) \text{ innovations} \\ -\ell &= -1100.7 \text{ for AR(9)-ARCH(1) with } \mathcal{N}(0,1) \text{ innovations} \\ -\ell &= -1498.1 \text{ for AR(9)-ARCH(1) with scaled } t_\nu \text{ innovations, } \hat{\nu} = 2.42. \end{aligned}$$

The three models are nested: e.g., $\alpha_1 = 0$, $\nu = \infty$ in (17) yield the AR(9) with Gaussian innovations. There is a clear indication also from the values of the negative log-likelihoods that the largest model fits the data best (compare with model selection on p. 10–11). The parameter estimates in the model (17) are as follows,

$$\begin{aligned} &(\hat{\phi}_1, \ldots, \hat{\phi}_9)^T = (0.671, -0.156, 0.001, -0.038, -0.005, 0.003, -0.023, 0.024, -0.035)^T, \\ &(\hat{\alpha}_0, \hat{\alpha}_1)^T = (0.009, 1.063)^T, \\ &\hat{\nu} = 2.426. \end{aligned}$$

(The fact that $\alpha_1 > 1$ is not indicating non-stationarity). Uncertainty measures for the parameters can be given from asymptotic expressions: in this example, this would still be analytically possible. But since this kind of AR-ARCH model is not implemented in standard statistical software, an easy yet powerful way is given by the parametric *bootstrap*, see p. 12.

In non-Markovian settings, the log-likelihood (conditioned on the first variable) is

$$\ell(\theta) = \sum_{t=2}^{n} \log f(X_t | X_{t-1}, \ldots, X_1; \theta). \tag{18}$$

Typically, the conditional density $f(\cdot | X_{t-1}, \ldots, X_1; \theta)$ is defined recursively. For example,

$$\begin{aligned} &f(x_t | x_{t-1}, \ldots x_1; \theta) = \sigma_t^{-1} \varphi(x_t/\sigma_t), \\ &\sigma_t^2 = \theta_1 + \theta_2 x_{t-1}^2 + \theta_3 \sigma_{t-1}^2,\ \theta_i \geq 0\ (i = 1, 2, 3) \end{aligned}$$

from a GARCH(1,1) model with Gaussian innovations yielding non-Markovian stochastic heteroscedasticity.

Minimizing $-\ell(\theta)$ in (18) and in (16) can become difficult if θ is high dimensional, in complete analogy to regression.

Nonparametric estimation

In Markovian models as in (11) and (12), nonparametric techniques from regression with independent observations can be used. The curse of dimensionality can be heavily present with only a moderate number of lagged explanatory variables. The strategies and procedures aiming to deal intelligently with the problem can be borrowed from regression, e.g. *additive models* or *projection pursuit*.

The variance function $g(\cdot)$ in (12) can be estimated with essentially two nonparametric strategies. In a first step, without taking heteroscedasticity into account, get a nonparametric estimates $\hat{f}(\cdot)$. Then, construct residuals $R_t = X_t - \hat{f}(X_{t-1}, \ldots, X_{t-p})$. Now, recover $g(\cdot)$ from

$$\log(R_t^2) \approx \beta + 2\log(g(X_{t-1}, \ldots, X_{t-r}) + \log(\varepsilon_t^2) - \beta,\ \beta = \mathbb{E}[\log(\varepsilon_t^2)],$$

with β known or estimated via method of moments (using that in the original model, $\mathbb{E}[\varepsilon_t^2] = 1$). The log-transform yields here an (approximately) homoscedastic regression problem for the transformed variance function $g(\cdot)$. The procedure can be iterated by using weighted regression with weights $1/\hat{g}(X_{t-1}, \ldots, X_{t-r})$ for estimation of $f(\cdot)$.

Nonparametric estimates, such as smoothers, for models as in (13) are not sufficiently developed yet due to unobservability of disturbances.

Model selection

The general information criteria such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) or Minimum Description Length (MDL) can be used for many time series model classes, see the review paper by Hansen and Yu [14]. They are all of the form of a penalized log-likelihood,

$$-2\ell(\hat{\theta}) + \alpha \cdot df, \tag{19}$$

where $\ell(\hat{\theta})$ is the log-likelihood of a fitted model of interest, df denotes the degrees of freedom in this model (e.g. the number of free parameters in a parametric model) and α is a penalty factor being 2 for AIC and $\log(n)$ for BIC. The best estimated model is the minimizer of the criterion in (19). Other selection techniques may be used as well.

Example: ozone data (continued)
For the deseasonalized ozone series $\{X_t\}$ in (3), we first subtract an estimated trend $\{\hat{m}_t\}$ as on p. 2. For the residual process $\{\hat{Z}_t\} = \{X_t - \hat{m}_t\}$ we use an autoregressive model whose order is chosen by the AIC or BIC criterion. Figure 5 shows the information criteria: BIC employs a higher penalty for larger models, and complex models are more clearly rejected (due to the sharper increase in BIC as a function of dimensionality). The estimated models are AR(34) and AR(14) with AIC and BIC, respectively.

Figure 5 about here

The residuals from the fitted AR models appear to be uncorrelated, but still slightly dependent. See Figure 6.

Figure 6 about here

Bootstrap and subsampling techniques can be used for model selection as well: but for all of them, the appropriate version for time series should be used, i.e., block bootstrap (Künsch [16]) or block subsampling (Politis et al. [18]). See p. 13.

Variability and distribution of estimates

We often want to know variability and even the distribution of point estimates to construct tests or confidence intervals. The problem of variance and distribution estimation of a statistical estimator is generally much more difficult with time series than independent data. Therefore, the content of this section is on a more advanced technical level.

Already the simple arithmetic mean has relatively complicated variance, see (5). The asymptotic variance is an infinite-dimensional object involving all autocovariances $R(\cdot)$. This cannot be estimated with the parametric convergence rate $1/\sqrt{n}$, the slower rate is due to dependence.

Since the Central Limit Theorem is still working for a large class of dependent processes, the limiting distributions of many estimators are still normal (namely if the estimator is linearizable as a mean of influence functions, see Hampel et al. [12]). The distinction to the independent setting is again the asymptotic variance which is often very different. For example, in case of linear regression with stationary errors, the autocovariances of the errors as well as the structure of the predictors (the regression spectrum) appear in the asymptotic variance of ordinary least squares, see Hannan [13]. Although generally the problem to estimate the asymptotic variance or distribution of estimates in time series is difficult, there are some noticeable exceptions. Some nonparametric estimates of the mean function $f(\cdot)$ in the Markovian models (11) and (12) have the same asymptotic distribution as in the independent regression set-up. For example, with $p = 1$ using a kernel estimator $\hat{f}_n(\cdot)$ with kernel $K(\cdot)$,

$$\sqrt{nh}(\hat{f}_n(x) - f(x)) \Rightarrow \mathcal{N}(B_{as}(x), \sigma^2_{as}(x)) \ (n \to \infty),$$

where $h \to 0$ is a bandwidth, $B_{as}(x) = \lim(nh)^{1/2}h^2 f''(x) \int u^2 K(u)du/2$ and $\sigma^2_{as} = p(x)\,\mathrm{Var}(X_1|X_0 = x) \int K^2(u)du$ with $p(\cdot)$ the stationary density of X_1. The reason why this holds is sometimes called 'whitening by windowing effect', which asymptotically decorrelates dependence. The situation is different for nonparametric kernel estimators for a trend with stationary errors, as used for the deseasonalized ozone series on p. 2 in **Example: ozone data**. There, under suitable conditions,

$$\sqrt{nh}(\hat{m}_n(x) - m(x)) \Rightarrow \mathcal{N}(B_{as}(x), \sigma^2_{as}(x)) \ (n \to \infty),$$

with $B_{as}(\cdot)$ as above, but $\sigma^2_{as} = \sum_{k=-\infty}^{\infty} R_Z(k) \int K^2(u)du$ involving all autocovariances $R_Z(\cdot)$ of the noise process $\{Z_t\}$ which can be estimated from residuals. This can be used to construct pointwise confidence intervals for the mean function $m(\cdot)$. Better approximations can be constructed with the sieve bootstrap, see p. 13. An initial trend estimate $\tilde{m}_n(\cdot)$ is used for computing residuals $\hat{Z}_t = X_t - \tilde{m}_n(t/n)$ and the stationary sieve bootstrap is then applied to $\{\hat{Z}_t\}$. This yields bootstrapped disturbances $\{Z^*_t\}$ and the bootstrapped observations are constructed as $X^*_t = \tilde{m}(t/n) + Z^*_t$. These can then be used for bootstrapping the estimator $\hat{m}_n(\cdot)$ and hence setting confidence bands for $m(\cdot)$, see Bühlmann [8].

In AR(p) models, the parameter estimates $\hat{\phi}_1, \ldots \hat{\phi}_p$ from least squares (Gaussian MLE), from the Yule-Walker method or also from the robust GM-estimators behave asymptotically as in the (stochastic) regression set-up, provided that the model is correct, see p. 12–13. It implies that computing routines for least squares and robust regression yield reasonable values in autoregressive models for point estimates and variability thereof. This analogy translates to more general Markovian models and parameter estimates whose influence functions are uncorrelated (for influence functions, see Hampel et al. [12]), see p. 13.

Many time series models, whether Markovian or not, are not routinely implemented in statistical software. Deriving and calculating expressions for the variance and distribution estimates of fitted parameters can then become an involved and in some circumstances rather complicated task. A user-friendly way out of this problem is given by the model-based, (semi-) parametric bootstrap. The bootstrap samples are generated by simulating from the fitted model: it typically involves i.i.d. resampling from residuals (mimicking the i.i.d. innovations in the model) and then using them to construct a bootstrapped time series $X_1^*, \ldots, X_n^*$. Having such a pseudo-sample, the construction of variance estimates and confidence intervals is as in the independent set-up, see Efron and Tibshirani [11]. Such model-based bootstrap estimates are typically even better than their counterparts from asymptotic considerations: e.g., the bootstrap still picks up finite-sample skewness, in contrast to a normal approximation.

Example: river flow (continued)
We illustrate the procedure for the Fisher river data, introduced on p. 8–9. Based on the model in (17), the purely parametric bootstrap would be implemented as follows:
(1) Generate innovations $\varepsilon_1^*, \ldots, \varepsilon_{m+n}^*$ i.i.d. $(\hat{\nu}-2)/\hat{\nu} t_{\hat{\nu}}$.
(2) Use starting values $X_{-8}^*, \ldots, X_0^* = 0$ and compute recursively

$$X_t^* = \sum_{j=1}^{9} \hat{\phi}_j X_{t-j}^* + (\hat{\alpha}_0 + \hat{\alpha}_1 X_{t-1}^2)^{1/2} \varepsilon_t^*, \ t = 1, \ldots, m+n.$$

(3) Use the last stretch $X_{m+1}^*, \ldots, X_{m+n}^*$ as the bootstrap pseudo-sample. (This is often re-denoted by $X_1^*, \ldots, X_n^*$).
The typically large number m (e.g. $m = 2000$) ensures approximate stationarity of the bootstrap sample. The (semi-)parametric approach would replace step (1) by the following:
(1') Compute residuals

$$\hat{\varepsilon}_t = (X_t - \sum_{j=1}^{9} \hat{\phi}_j X_{t-j})/\hat{\sigma}_t, \ \hat{\sigma}_t = (\hat{\alpha}_0 + \hat{\alpha}_1 X_{t-1}^2)^{1/2}.$$

Center them to mean zero and in this example, additionally standardize to variance one. Then, resample $\varepsilon_1^*, \ldots, \varepsilon_{m+n}^*$ i.i.d. from the empirical c.d.f. of these centered, standardized residuals.
The BC_a corrected bootstrap confidence intervals (see Efron and Tibshirani [11]) for some of the parameters are then

$$\begin{aligned} &[0.611, 0.721] \text{ for } \phi_1, \\ &[0.700, 2.015] \text{ for } \alpha_1, \\ &[2.125, 2.829] \text{ for } \nu, \end{aligned}$$

There is particularly large uncertainty about the ARCH parameter α_1, but with a clear indication that α_1 is significant, implying stochastic heteroscedasticity.

The (model-based) asymptotic variance and distribution can be completely wrong under model-misspecification. To illustrate this subtle point, consider the following example. Given data from a mean zero process, the working model is a Gaussian AR(1). The MLE parameter estimate then behaves as

$$\hat{\phi}_n \approx n^{-1} \sum_{t=2}^{n} \mathrm{IF}(X_{t-1}, X_t; P),$$
$$\mathrm{IF}(X_{t-1}, X_t; P) = \frac{(X_t - \phi X_{t-1}) X_{t-1}}{R(0)}.$$

IF$(\cdot; P)$ denotes here the influence function of the estimator at the data generating distribution P, $\phi = R(1)/R(0)$ denotes the first autocorrelation of the process. Suppose that the true model is an AR(1), i.e. $X_t = \phi_0 X_{t-1} + \varepsilon_t$: then $\phi = \phi_0$ and the estimator behaves as above with uncorrelated influence function terms. This implies that

$$n^{1/2}(\hat{\phi}_n - \phi_0) \Rightarrow \mathcal{N}(0, \sigma_{as}^2),$$
$$\sigma_{as}^2 = \mathbb{E}|\,\mathrm{IF}(X_{t-1}, X_t; P)|^2 = \sigma_\varepsilon^2/R(0),$$

which is the same formula as in (stochastic) linear regression for independent data. However, if the data-generating process is not an AR(1), e.g. an AR(2), then the influence function terms are correlated and

$$n^{1/2}(\hat{\phi}_n - \phi) \Rightarrow \mathcal{N}(0, \sigma_{as}^2),\ \ \phi = R(1)/R(0)$$
$$\sigma_{as}^2 = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(\mathrm{IF}(X_0, X_1; P), \mathrm{IF}(X_k, X_{k+1}; p)).$$

In the independent (stochastic) linear regression framework, the asymptotic variance in the misspecified model would still be $\mathbb{E}|\,\mathrm{IF}(Y_t, X_t; P)|^2$ which is estimated by standard computing packages for regression and which is still a reasonable value for uncertainty. In contrast to time series with model misspecification, where the whole dependence structure in terms of all autocovariances of influence functions is involved. The standard error from a regression (independent observation) software can be very misleading.

A general tool tailored to overcome such difficulties with assessing variability and confidence regions without relying on a specific model, is given by versions of the bootstrap for time series: the *block bootstrap* (Künsch [16]) for general stationary time series or the *sieve bootstrap* (Bühlmann [7]) for linear time series. The block bootstrap relies on the idea to resample independently with replacement blocks of consecutive observations. The sieve bootstrap is based on an AR-model, typically of high order: it is from an implementational point of view as the (semi-)parametric autoregressive model based bootstrap (compare with another model based bootstrap on p. 12), with the property that some finite-sample model-misspecification will not destroy consistency and accuracy of the method, provided the data generating process is linear. Bootstrap methods are often restricted for estimating distributions which are asymptotically normal. The approach with *subsampling* (Politis et al. [18]) overcomes this difficulty: but unless sample size is very large, subsampling can be inefficient and misleading in finite sample problems.

Frequency analysis

Harmonic analysis for stochastic processes says that a stationary process $\{X_t\}$ can be written as a sum of cosinusoids/sinusoids. The analysis of stationary processes in terms of such harmonic oscillations is referred to frequency analysis, cf. Brillinger [4]. The so-called Cramér representation of a mean zero stationary process $\{X_t\}$ is

$$X_t = \int_0^{2\pi} \exp(i\lambda t) dZ(\lambda),$$

where $\{Z(\lambda)\}$ is an orthogonal increment process, see Brillinger [4]. This stochastic integral can be interpreted as follows:

$$X_t \approx \frac{2\pi}{m} \sum_{j=0}^{m-1} \{A_j \cos(\frac{2\pi j}{m} t) + B_j \sin(\frac{2\pi j}{m} t)\} \ \text{ for } m \text{ large},$$

where $\mathbb{E}[A_j] = \mathbb{E}[B_j] = 0$, A_j's uncorrelated, B_j's uncorrelated, and A_j uncorrelated with B_s for all j, s. A sensible measure for the contribution of the random coefficients for the cosinusoid/sinusoid at frequency $2\pi j/m$ is given by their variances $\mathbb{E}|A_j|^2 = \mathbb{E}|B_j|^2 = f(2\pi j/m)$, where $f(\cdot)$ is the spectral density

$$f(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} R(k) \exp(-i\lambda k), \ \lambda \in \mathbb{R}. \tag{20}$$

We have assumed here that the autocovariances satisfy $\sum_{k=-\infty}^{\infty} |R(k)| < \infty$. Particularly, when the process consists of uncorrelated variables, i.e. $R(k) = 0$ for $k \neq 0$, we see that $f(\lambda) \equiv R(0)/(2\pi)$ is a constant with respect to frequency λ: such processes are called white noise since the contribution in the above Cramér representation for cosinusoids/sinusoids is the same for all frequencies, in analogy to white light.

The spectral density is 2π-periodic and symmetric around zero: hence all the information is given by frequencies λ between zero and π which is the usual range for describing $f(\cdot)$. The interpretation of the spectral density at frequency λ in connection with the Cramér representation above can be made as a measure of contribution for the cosinusoids/sinusoids at that frequency for describing the stochastic process $\{X_t\}$. For example, peaks of $f(\cdot)$ at frequencies $\lambda_1, \lambda_2, \ldots$ indicate a quasi-periodic behavior (in the stochastic sense) of the stochastic process with periods $2\pi/\lambda_1, 2\pi/\lambda_2, \ldots$

The spectral density in (20) is the (finite) Fourier transform of the autocovariance function $\{R(k)\}$. The spectral density $f(\cdot)$ thus carries the equivalent information as the autocovariance function, a second moment structure. Due to the interpretation given above, it is often more useful to represent second moments in the frequency domain with the spectral density $f(\cdot)$. On the other hand, it becomes clear that the spectral density cannot describe more than second-order moment structure.

In linear systems, the spectral density follows a simple transformation rule. For a stationary stochastic process $\{X_t\}$ with spectral density $f_X(\cdot)$, we consider the linear transform (from the *time-invariant linear filter* $\{\psi_j\}$,

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$$

with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Then, $\{Y_t\}$ is again a stationary process with spectral density

$$f_Y(\lambda) = H(\lambda) f_X(\lambda),$$

where $H(\lambda) = |\sum_{j=-\infty}^{\infty} \psi_j \exp(-i\lambda j)|^2$ is the *transfer function* of the linear filter $\{\psi_j\}$ which describes amplification of the input spectral density $f_X(\cdot)$. See Brockwell and Davis [6]. From this, we easily obtain the spectral densities in AR and ARMA models which are both linear transform of disturbances,

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\Theta(\exp(-i\lambda))|^2}{|\Phi(\exp(-i\lambda))|^2}, \tag{21}$$

where $\Theta(\cdot)$ and $\Phi(\cdot)$ are the generating polynomials for the MA- and AR-part in the model, see (9) and (8).

Long memory processes exhibit a singularity at frequency zero because the autocovariances are not absolutely summable. The behavior around zero can be exploited to characterize and estimate the degree of long range dependence, e.g. the *self-similarity parameter*. See Beran [2].

Higher order spectra, the Fourier transforms of *higher order cumulants*, may be used for examining Gaussianity or linearity of a stationary process. For an overview see Brillinger [5].

Estimation of the spectral density

Estimation of the spectral density in (20) can be based on the Fourier transform of the data, the so-called *periodogram*

$$I_n(\lambda) = (2\pi)^{-1} \left| \sum_{t=1}^{n} (X_t - \overline{X}_n) \exp(-i\lambda t) \right|^2.$$

Applying the Fourier transform, the periodogram can be represented as

$$I_n(\lambda) = (2\pi)^{-1} \sum_{k=-n+1}^{n-1} \hat{R}(k) \exp(-i\lambda k) \ \ (\lambda \neq 0)$$

with $\hat{R}(\cdot)$ as in (4). This is an overparameterized plug-in estimate for the spectral density in (20). Since it involves n parameters, it is intuitively clear that it is not a consistent estimate for the spectral density. However, it has the properties that it is asymptotically unbiased for $f(\lambda)$ $(\lambda \neq 0)$ and the estimates at different Fourier frequencies $\lambda_j = 2\pi j/n$ are asymptotically uncorrelated (assuming suitable decay of the autocovariance function $R(k)$ as $|k| \to \infty$). This suggests that a consistent estimate can be constructed by smoothing the periodogram,

$$\hat{f}(\lambda) = \sum_{k=-m}^{m} w(k) I_n(\lambda - \lambda_k), \ \ \lambda_k = 2\pi k/n, \tag{22}$$

with weights $w(k) = w_n(k)$ being symmetric in k, summing up to one, $\sum_{k=-m}^{m} w_n^2(k) \to 0$ $(n \to \infty)$, and $m = m_n \to \infty$, $m_n/n \to 0$ $(n \to \infty)$. The simplest (not necessarily best) example is $w(k) \equiv 1/(2m+1)$. See for example Brockwell and Davis [6]. An asymptotically equivalent form is given by lag-window estimates,

$$\hat{f}_{LW}(\lambda) = (2\pi)^{-1} \sum_{k=-q}^{q} v(k/q) \hat{R}(k) \exp(-i\lambda k),$$

with $q = q_n \to \infty$, $q_n/n \to 0$ $(n \to \infty)$ and $v(\cdot)$ a symmetric window function with $v(0) = 1$ and $v(x) = 0$ for $|x| > 1$; see Priestley [19]. In both cases, the smoothing parameter is given by m_n or q_n respectively. The estimator for $\hat{f}(\cdot)$ (and $\hat{f}_{LW}$) shares the same asymptotic properties as other smoothing techniques: asymptotic normality holds and the optimal mean squared error smoothing parameter is such that the asymptotic bias and variance are both of the same order. See for example Brockwell and Davis [6]. For more details: $\rightarrow$ *Spectral analysis*.

Example: ozone data (continued)

We illustrate spectral density estimation for the deseasonalized and detrended ozone series $\{\hat{Z}_t\}$. The raw periodogram and a smoothed version as in (22) using a modified rectangular window with $m = 5$ are shown in Figure 7. In addition, an AR-spectral density estimate based on (21) with $p = 34, q = 0$ and plugging-in the estimated AR-parameters is given for comparison.

Figure 7 about here

For both spectral density estimates, the main frequency is at $\hat{\lambda}_1 = 0.178$ (nonparametric smoothing estimate) and $\hat{\lambda}_1 = 0.140$ (AR estimate), corresponding to a quasi-period of 35.2 or 44.9 months, respectively.

The same strategy allows also estimation of higher order spectra, see Brillinger [4],[5].

Multiresolution analysis with wavelets and nonstationary processes

The discrete wavelet transform is an orthonormal transform of data $X_1, \ldots, X_n$ to a set of wavelet basis functions. The time series can then be analyzed, on different levels of resolution, in terms of wavelets. Due to orthonormality of the transform, a nice analysis of variance decomposition follows. Wavelet analysis is ideally suited for describing time series with nonstationary or transient components. It provides a succinct way of quantifying the nonstationary characteristics of time series. A comprehensible short description of the wavelet method with illustration on sea level fluctuations is given by Percival and Mofjeld [17].

Multivariate time series

The analysis of multivariate time series $X_t = (X_{t,1}, \ldots, X_{t,r})^T$ $(t = 0, \pm 1, \pm 2, \ldots)$ can be done with similar concepts and tools as in the univariate case. Of course, many of them become more difficult to handle due to the additional notion of cross-dependence between components, e.g. the dependence between $X_{t,i}$ and $X_{s,j}$ for the same and for different s, t. Reinsel [20] gives a detailed description of aspects for multivariate models. Box and Jenkins [3] or Brockwell and Davis [6] devote some chapters to the multivariate case, and Brillinger [4] includes the frequency domain for multivariate observations.

References

[1] Abraham, B. and Ledolter, J. (1983). Statistical Methods for Forecasting. Wiley.

[2] Beran, J. (1994). Statistical Methods for Long Memory Processes. Chapman & Hall.

[3] Box, G.E.P. and Jenkins, G.M. (1976). Forecasting and Control. Revised Edition. Holden-Day.

[4] Brillinger, D.R. (1975). Time Series: Data Analysis and Theory. Holden-Day.

[5] Brillinger, D.R. (1991). Some history of the study of higher-order moments and spectra. Statistica Sinica **1**, 465–476.

[6] Brockwell, P.J. and Davis, R.A. (1991). Time Series. Theory and Methods (Second Edition). Springer.

[7] Bühlmann, P. (1997). Sieve bootstrap for time series. Bernoulli **3**, 123–148.

[8] Bühlmann, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. The Annals of Statistics **26**, 48–83.

[9] Cleveland, W.S. (1994). The Elements of Graphing Data (Revised Edition). Hobart Press.

[10] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. The Annals of Statistics **25**, 1–37.

[11] Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall.

[12] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley.

[13] Hannan, E.J. (1970). Multiple Time Series. Wiley.

[14] Hansen, M.H. and Yu, B. (2000). Model selection and the principle of minimum description length. To appear in J. of the American Statistical Association.

[15] Harvey, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

[16] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics **17**, 1217–1241.

[17] Percival, D.B. and Mofjeld, H.O. (1997). Analysis of subtidal coastal sea level fluctuations using wavelets. Journal of the American Statistical Association **92**, 868–880.

[18] Politis, D.N., Romano, J.P. and Wolf, M. (1999). Subsampling. Springer.

[19] Priestley, M.B. (1981). Spectral Analysis and Time Series. (Vol. 1): Univariate Series. Academic.

[20] Reinsel, G.C. (1997). Elements of Multivariate Time Series Analysis (Second Edition). Springer.

[21] Shumway, R.H. and Stoffer, D.S. (2000). Time Series Analysis and its Applications. Springer.

[22] Tong, H. (1990). Non-linear Time Series. A Dynamical System Approach. Oxford University Press.

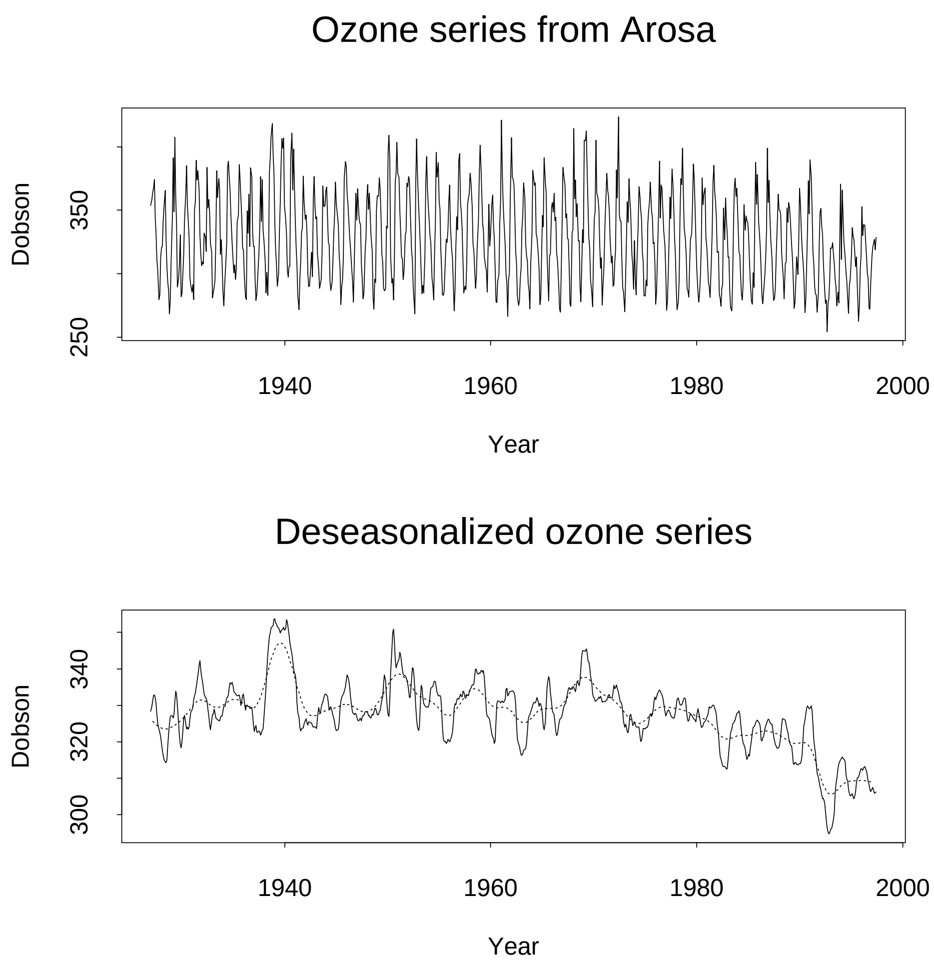

Figure 1: Top: Monthly ozone measurements. Bottom: Filtered ozone levels, removing seasonal effects (solid line) and trend estimate $\hat{m}_t$ (dotted line).

Figure 2: Left panel: sample autocorrelation function. Right panel: sample partial autocorrelation function. Both based on deseasonalized and detrended ozone time series.

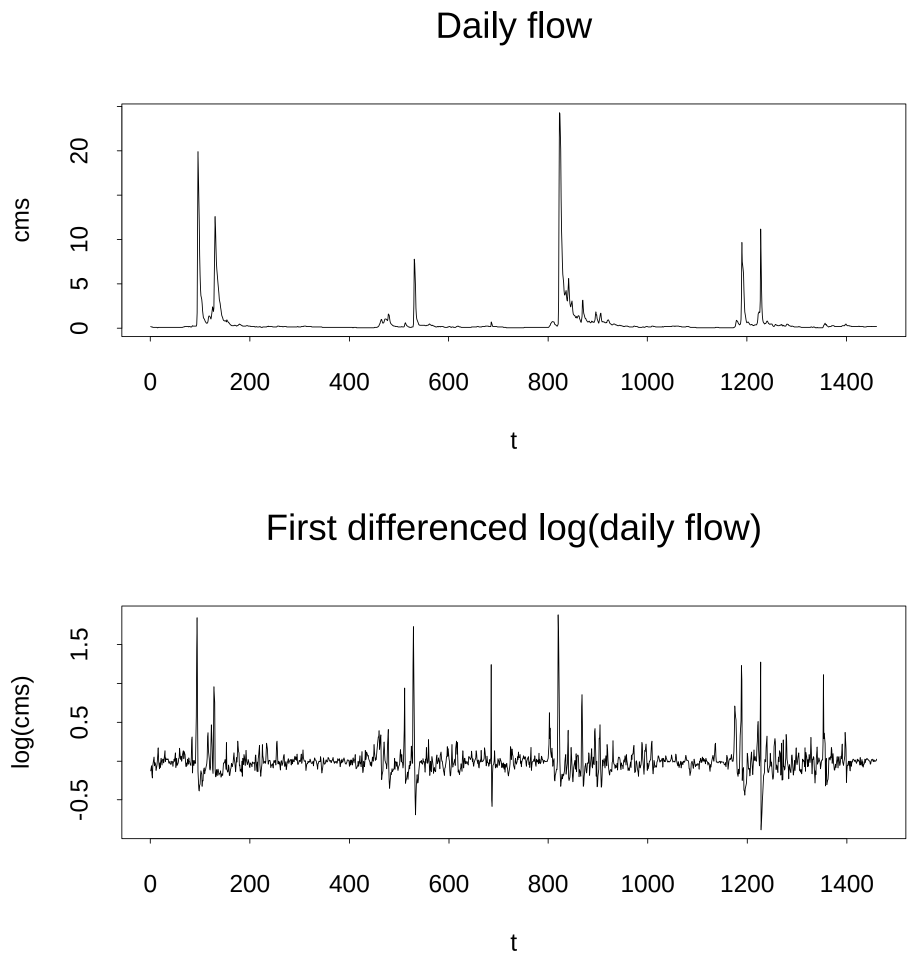

Figure 3: Daily river flow (top) and first differenced log-transformed series (bottom) of Fisher river from January 1, 1988 – December 31, 1991.

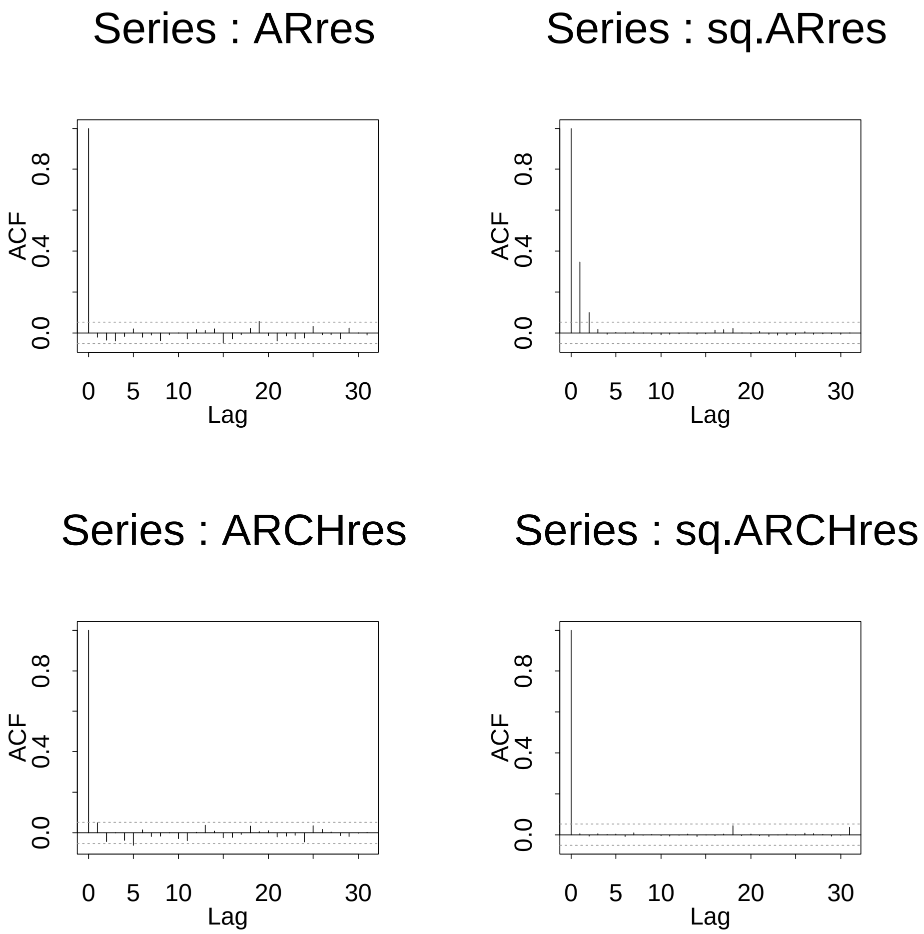

Figure 4: Sample autocorrelations of residuals (left panels) and squared residuals (right panels) for first differenced daily flow from Fisher river. Top row: residuals from AR(9) model with Gaussian innovations. Bottom row: residuals from AR(9)-ARCH(1) model with scaled t_ν innovations.

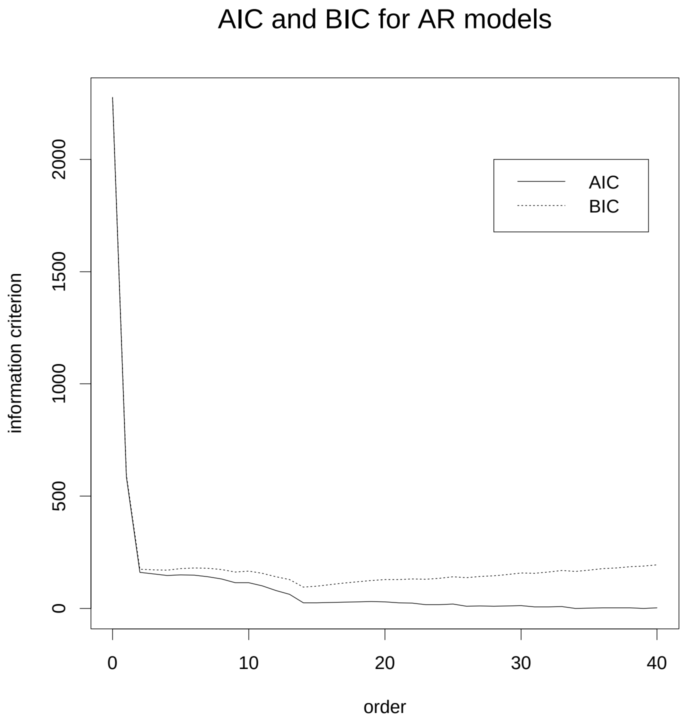

Figure 5: AIC and BIC with AR models for deseasonalized and detrended ozone series $\{\hat{Z}_t\}$. On the y-axis, the information criteria are given up to an additive constant. The estimated orders are 34 and 14 for AIC and BIC, respectively.

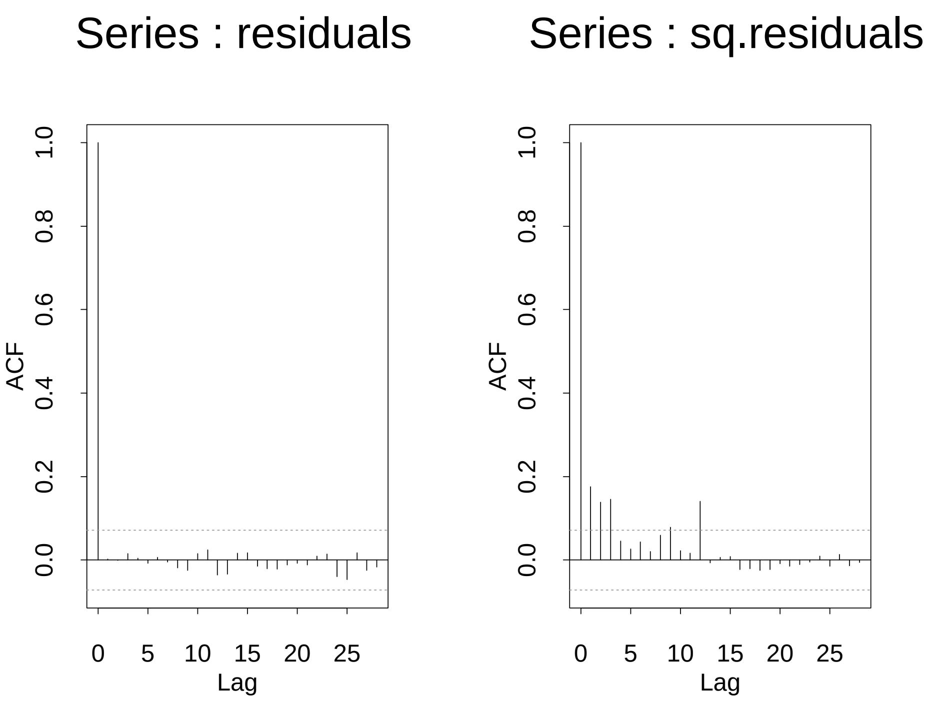

Figure 6: Autocorrelation function of residuals from AR(34) model for deseasonalized and detrended ozone series $\{\hat{Z}_t\}$. Left panel: residuals. Right panel: squared residuals.

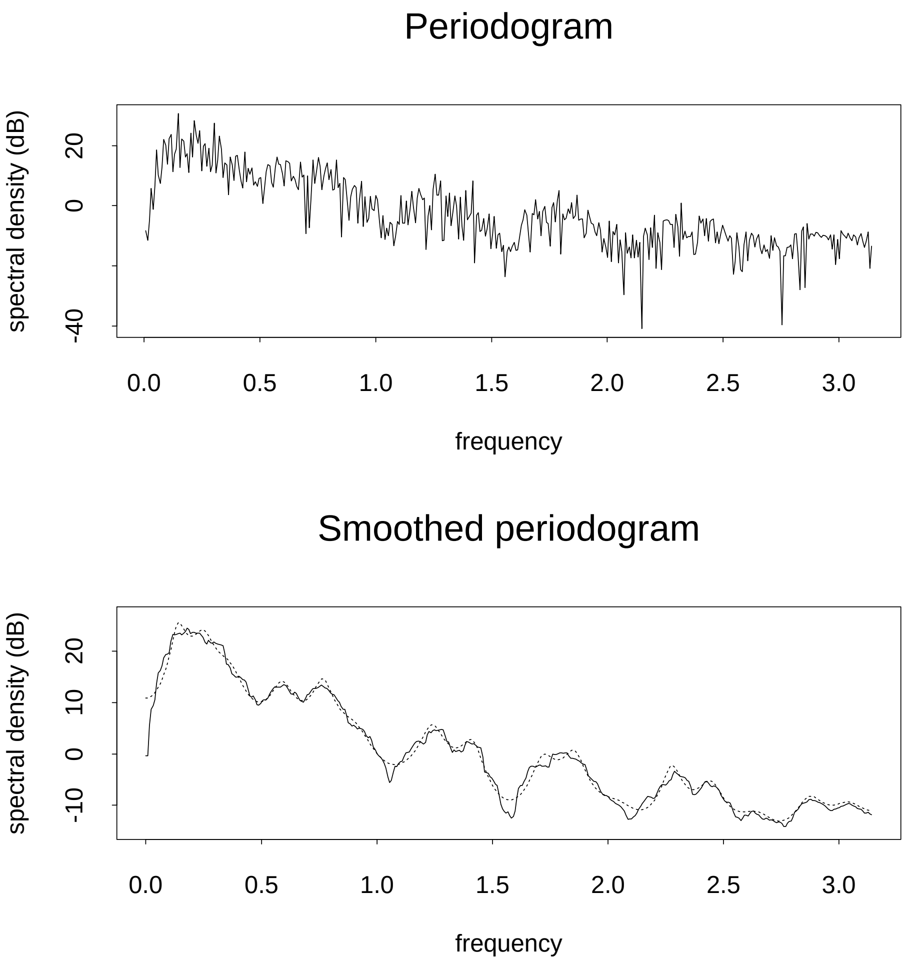

Figure 7: Top: Periodogram. Bottom: Smoothed periodogram with modified rectangular window, m=5, non-tapered (solid line). Overlaid is AR(34) AR-spectral estimate (dotted line). Y-axis on the dB-scale $10 * \log_{10}(f(\cdot))$.