# Review (Lecture 1): Time series modelling and forecasting

1. Plot the time series.

   Look for trends, seasonal components, step changes, outliers.

2. Transform data so that residuals are **stationary**.

   (a) Remove trend and seasonal components.

   (b) Differencing.

   (c) Nonlinear transformations (log, $\sqrt{\cdot}$).

3. Fit model to residuals.

4. Forecast time series by forecasting residuals and inverting any transformations.

# Review: Time series modelling and forecasting

Stationary time series models: ARMA(p,q).

- $p = 0$: MA(q),
- $q = 0$: AR(p).

We have seen that any causal, invertible linear process has:
an MA($\infty$) representation (from causality), and
an AR($\infty$) representation (from invertibility).

Real data cannot be *exactly* modelled using a finite number of parameters.

We choose $p, q$ to give a simple but accurate model.

## Review: Time series modelling and forecasting

How do we use data to decide on $p, q$?

1. Use sample ACF/PACF to make preliminary choices of model order.
2. Estimate parameters for each of these choices.
3. Compare predictive accuracy/complexity of each (using, e.g., AIC).

NB: We need to compute parameter estimates for several different model orders.

Thus, recursive algorithms for parameter estimation are important.

We'll see that some of these are identical to the recursive algorithms for forecasting.

# Review: Time series modelling and forecasting

| **Model:** | **ACF:** | **PACF:** |
|---:|---:|---:|
| AR(p) | decays | zero for $h > p$ |
| MA(q) | zero for $h > q$ | decays |
| ARMA(p,q) | decays | decays |

# Parameter estimation

We want to estimate the parameters of an ARMA(p,q) model.
We will assume (for now) that:
1. The model order (p and q) is known, and
2. The data has zero mean.

If (2) is not a reasonable assumption, we can subtract the sample mean $\bar{y}$, fit a zero-mean ARMA model,

$$\phi(B)X_t = \theta(B)W_t,$$

to the mean-corrected time series $X_t = Y_t - \bar{y}$, and then use $X_t + \bar{y}$ as the model for $Y_t$.

# Parameter estimation: Maximum likelihood estimator

One approach:

Assume that $\{X_t\}$ is Gaussian, that is, $\phi(B)X_t = \theta(B)W_t$, where $W_t$ is i.i.d. Gaussian.

Choose $\phi_i, \theta_j$ to maximize the *likelihood*:

$$L(\phi, \theta, \sigma^2) = f(X_1, \ldots, X_n),$$

where $f$ is the joint (Gaussian) density for the given ARMA model.

(c.f. choosing the parameters that maximize the probability of the data.)

## Parameter estimation: Maximum likelihood estimator

**Advantages of MLE:**

Efficient (low variance estimates).

Often the Gaussian assumption is reasonable.

Even if $\{X_t\}$ is not Gaussian, the asymptotic distribution of the estimates $(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2)$ is the same as the Gaussian case.

**Disadvantages of MLE:**

Difficult optimization problem.

Need to choose a good starting point (often use other estimators for this).

# Preliminary parameter estimates

**Yule-Walker for AR(p):** Regress $X_t$ onto $X_{t-1}, \ldots, X_{t-p}$.

Durbin-Levinson algorithm with $\gamma$ replaced by $\hat{\gamma}$.

**Yule-Walker for ARMA(p,q):** Method of moments. Not efficient.

**Innovations algorithm for MA(q):** with $\gamma$ replaced by $\hat{\gamma}$.

**Hannan-Rissanen algorithm for ARMA(p,q):**

1. Estimate high-order AR.
2. Use to estimate (unobserved) noise $W_t$.
3. Regress $X_t$ onto $X_{t-1}, \ldots, X_{t-p}, \hat{W}_{t-1}, \ldots, \hat{W}_{t-q}$.
4. Regress again with improved estimates of $W_t$.

# Yule-Walker estimation

For a causal AR(p) model $\phi(B)X_t = W_t$, we have

$$\mathrm{E}\left(X_{t-i}\left(X_t - \sum_{j=1}^{p}\phi_j X_{t-j}\right)\right) = \mathrm{E}(X_{t-i}W_t) \quad \text{for } i = 0, \ldots, p$$

$$\Leftrightarrow \qquad \gamma(0) - \phi'\gamma_p = \sigma^2 \quad \text{and}$$

$$\gamma_p - \Gamma_p\phi = 0,$$

where $\phi = (\phi_1, \ldots, \phi_p)'$, and we've used the causal representation

$$X_t = W_t + \sum_{j=1}^{\infty}\psi_j W_{t-j}.$$

# Yule-Walker estimation

**Method of moments:** We choose parameters for which the moments are equal to the empirical moments.

In this case, we choose $\phi$ so that $\gamma = \hat{\gamma}$.

$$\text{Yule-Walker equations for } \hat{\phi}: \qquad \begin{cases} \hat{\Gamma}_p \hat{\phi} = \hat{\gamma}_p, \\ \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}' \hat{\gamma}_p. \end{cases}$$

These are the forecasting equations.
We can use the Durbin-Levinson algorithm.

# Yule-Walker estimation: Confidence intervals

If $\{X_t\}$ is an AR(p) process, and $n$ is large,

- $\sqrt{n}(\hat{\phi}_p - \phi_p)$ is approximately $N(0, \hat{\sigma}^2 \hat{\Gamma}_p^{-1})$,
- with probability $\approx 1 - \alpha$, $\phi_p$ is in the ellipsoid

$$\left\{ \phi \in \mathbb{R}^p : \left( \hat{\phi}_p - \phi \right)' \hat{\Gamma}_p \left( \hat{\phi}_p - \phi \right) \le \frac{\hat{\sigma}^2}{n} \chi^2_{1-\alpha}(p) \right\},$$

where $\chi^2_{1-\alpha}(p)$ is the $(1 - \alpha)$ quantile of the chi-squared with $p$ degrees of freedom.

- with probability $\approx 1 - \alpha$, $\phi_{pj}$ is in the interval

$$\hat{\phi}_{pj} \pm \Phi_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \left( \hat{\Gamma}_p^{-1} \right)_{jj}^{1/2},$$

where $\Phi_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal.

# Yule-Walker estimation: Confidence intervals

If $\{X_t\}$ is an AR(p) process,

$$\hat{\phi} \sim AN\left(\phi, \frac{\sigma^2}{n}\Gamma_p^{-1}\right), \qquad \hat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

$$\hat{\phi}_{hh} \sim AN\left(0, \frac{1}{n}\right) \quad \text{for } h > p.$$

Thus, we can use the sample PACF to test for AR order, and we can calculate approximate confidence intervals for the parameters $\phi$.

## Yule-Walker estimation

It is also possible to define analogous estimators for ARMA(p,q) models with $q > 0$:

$$\hat{\gamma}(j) - \phi_1 \hat{\gamma}(j-1) - \cdots - \phi_p \hat{\gamma}(j-p) = \sigma^2 \sum_{i=j}^{q} \theta_i \psi_{i-j},$$

where $\psi(B) = \theta(B)/\phi(B)$.

Because of the dependence on the $\psi_i$, these equations are nonlinear in $\phi_i, \theta_i$.

There might be no solution, or nonunique solutions.

Also, the *asymptotic efficiency* of this estimator is poor: it has unnecessarily high variance.

# Efficiency of estimators

Let $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(2)}$ be two estimators. Suppose that

$$\hat{\phi}^{(1)} \sim AN(\phi, \sigma_1^2), \qquad \hat{\phi}^{(2)} \sim AN(\phi, \sigma_2^2).$$

The asymptotic efficiency of $\hat{\phi}^{(1)}$ relative to $\hat{\phi}^{(2)}$ is

$$e\left(\phi, \hat{\phi}^{(1)}, \hat{\phi}^{(2)}\right) = \frac{\sigma_2^2}{\sigma_1^2}.$$

If $e\left(\phi, \hat{\phi}^{(1)}, \hat{\phi}^{(2)}\right) \leq 1$ for all $\phi$, we say that $\hat{\phi}^{(2)}$ is a *more efficient* estimator of $\phi$ than $\hat{\phi}^{(1)}$.

For example, for an AR(p) process, the moment estimator and the maximum likelihood estimator are as efficient as each other.

For an MA(q) process, the moment estimator is less efficient than the innovations estimator, which is less efficient than the MLE.

# Yule Walker estimation: Example

AR(1): $$\gamma(0) = \frac{\sigma^2}{1-\phi_1^2}$$

$$\hat{\phi}_1 \sim AN\left(\phi_1, \frac{\sigma^2}{n}\Gamma_1^{-1}\right) = AN\left(\phi_1, \frac{1-\phi_1^2}{n}\right).$$

AR(2): $$\begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} \sim AN\left(\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \frac{\sigma^2}{n}\Gamma_2^{-1}\right)$$

and $$\frac{\sigma^2}{n}\Gamma_2^{-1} = \frac{1}{n}\begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}.$$

# Yule Walker estimation: Example

Suppose $\{X_t\}$ is an AR(1) process and the sample size $n$ is large.

If we estimate $\phi$, we have

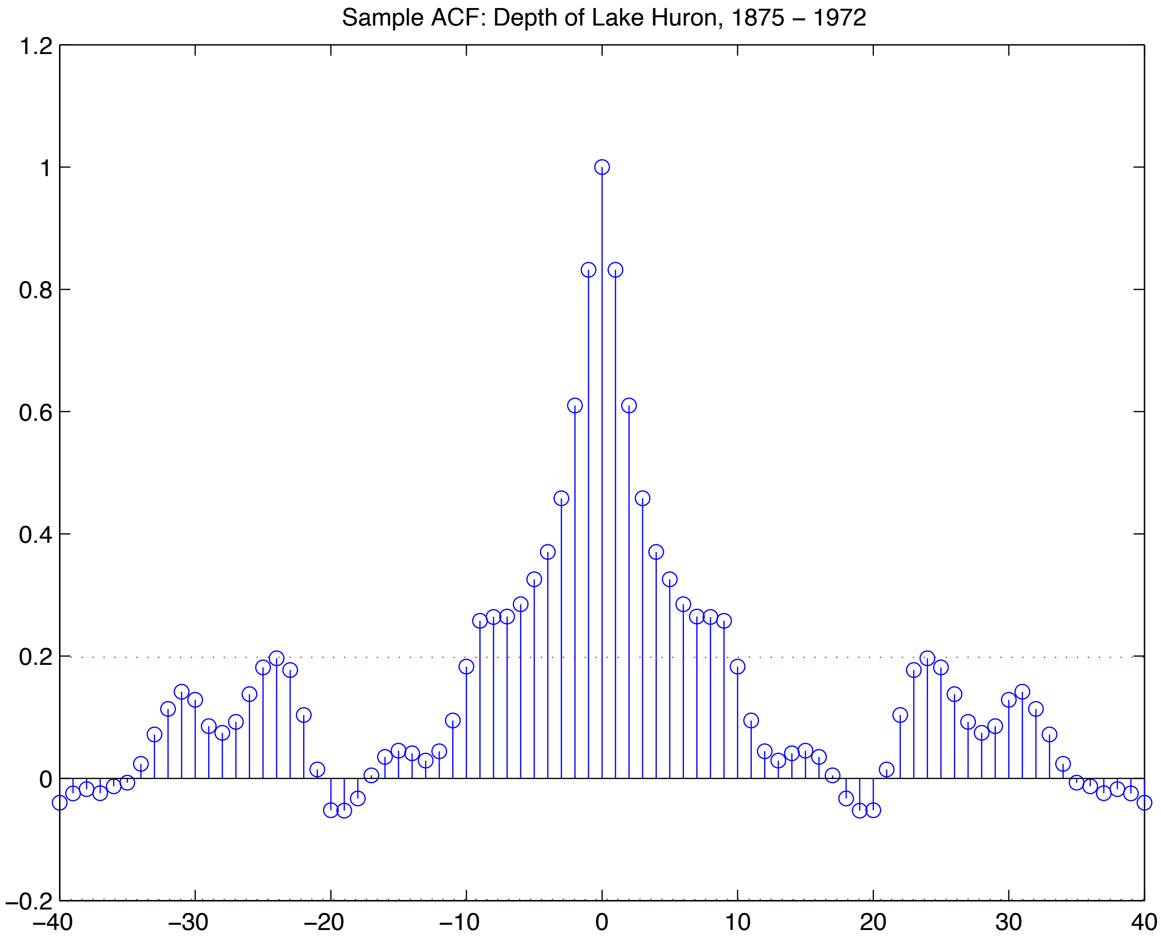
$$\text{Var}(\hat{\phi}_1) \approx \frac{1-\phi_1^2}{n}.$$

If we fit a *larger* model, say an AR(2), to this AR(1) process,

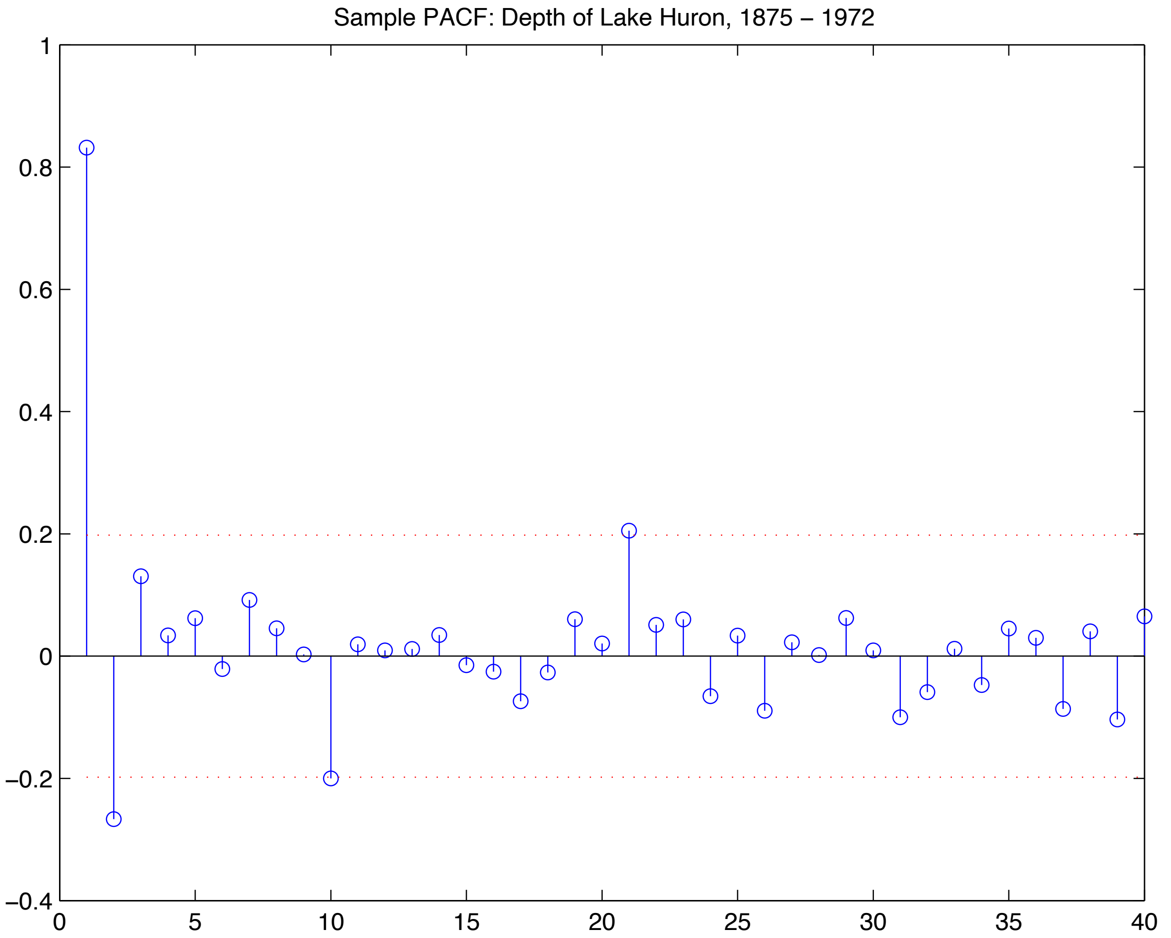
$$\text{Var}(\hat{\phi}_1) \approx \frac{1-\phi_2^2}{n} = \frac{1}{n} \quad > \quad \frac{1-\phi_1^2}{n}.$$

We have lost efficiency.

# Yule Walker estimation: Example



# Yule Walker estimation: Example



# Maximum likelihood estimation

Suppose that $X_1, X_2, \ldots, X_n$ is drawn from a zero mean Gaussian ARMA(p,q) process. The likelihood of parameters $\phi \in \mathbb{R}^p$, $\theta \in \mathbb{R}^q$, $\sigma_w^2 \in \mathbb{R}_+$ is defined as the density of $X = (X_1, X_2, \ldots, X_n)'$ under the Gaussian model with those parameters:

$$L(\phi, \theta, \sigma_w^2) = \frac{1}{(2\pi)^{n/2} \left|\Gamma_n\right|^{1/2}} \exp\left( -\frac{1}{2} X' \Gamma_n^{-1} X \right),$$

where $|A|$ denotes the determinant of a matrix $A$, and $\Gamma_n$ is the variance/covariance matrix of $X$ with the given parameter values.

The maximum likelihood estimator (MLE) of $\phi, \theta, \sigma_w^2$ maximizes this quantity.

# Maximum likelihood estimation

We can simplify the likelihood by expressing it in terms of the *innovations*.

Since the innovations are linear in previous and current values, we can write

$$\underbrace{\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}}_{X} = C \underbrace{\begin{pmatrix} X_1 - X_1^0 \\ \vdots \\ X_n - X_n^{n-1} \end{pmatrix}}_{U}$$

where $C$ is a lower triangular matrix with ones on the diagonal.
Take the variance of both sides to see that

$$\Gamma_n = CDC' \qquad \text{where } D = \mathrm{diag}(P_1^0, \ldots, P_n^{n-1}).$$

# Maximum likelihood estimation

Thus, $|\Gamma_n| = |C|^2 P_1^0 \cdots P_n^{n-1} = P_1^0 \cdots P_n^{n-1}$ and

$$X'\Gamma_n^{-1}X = U'C'\Gamma_n^{-1}CU = U'C'C^{-T}D^{-1}C^{-1}CU = U'D^{-1}U.$$

So we can rewrite the likelihood as

$$\begin{aligned} L(\phi,\theta,\sigma_w^2) &= \frac{1}{\left((2\pi)^n P_1^0 \cdots P_n^{n-1}\right)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - X_i^{i-1})^2/P_i^{i-1}\right) \\ &= \frac{1}{\left((2\pi\sigma_w^2)^n r_1^0 \cdots r_n^{n-1}\right)^{1/2}} \exp\left(-\frac{S(\phi,\theta)}{2\sigma_w^2}\right), \end{aligned}$$

where $r_i^{i-1} = P_i^{i-1}/\sigma_w^2$ and

$$S(\phi,\theta) = \sum_{i=1}^n \frac{\left(X_i - X_i^{i-1}\right)^2}{r_i^{i-1}}.$$

# Maximum likelihood estimation

The log likelihood of $\phi, \theta, \sigma_w^2$ is

$$\begin{aligned} l(\phi, \theta, \sigma_w^2) &= \log(L(\phi, \theta, \sigma_w^2)) \\ &= -\frac{n}{2}\log(2\pi\sigma_w^2) - \frac{1}{2}\sum_{i=1}^{n}\log r_i^{i-1} - \frac{S(\phi, \theta)}{2\sigma_w^2}. \end{aligned}$$

Differentiating with respect to $\sigma_w^2$ shows that the MLE $(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2)$ satisfies

$$\frac{n}{2\hat{\sigma}_w^2} = \frac{S(\hat{\phi}, \hat{\theta})}{2\hat{\sigma}_w^4} \qquad \Leftrightarrow \qquad \hat{\sigma}_w^2 = \frac{S(\hat{\phi}, \hat{\theta})}{n},$$

and $\hat{\phi}, \hat{\theta}$ minimize $\quad \log\left(\frac{S(\hat{\phi}, \hat{\theta})}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}\log r_i^{i-1}.$

# Maximum likelihood estimation

Minimization is done numerically (e.g., Newton-Raphson).

Computational simplifications:

- *Unconditional least squares*. Drop the $\log r_i^{i-1}$ terms.
- *Conditional least squares*. Also approximate the computation of $x_i^{i-1}$ by dropping initial terms in $S$. e.g., for AR(2), all but the first two terms in $S$ depend linearly on $\phi_1, \phi_2$, so we have a least squares problem.

The differences diminish as sample size increases. For example, $P_t^{t-1} \to \sigma_w^2$ so $r_t^{t-1} \to 1$, and thus $n^{-1} \sum_i \log r_i^{i-1} \to 0$.

# Maximum likelihood estimation: Confidence intervals

For an ARMA(p,q) process, the MLE and un/conditional least squares estimators satisfy

$$\begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} - \begin{pmatrix} \phi \\ \theta \end{pmatrix} \sim AN\left(0, \frac{\sigma_w^2}{n} \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta}, \end{pmatrix}^{-1}\right),$$

where $\begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta}, \end{pmatrix} = \text{Cov}((X,Y),(X,Y)),$

$$X = (X_1, \ldots, X_p)' \qquad \phi(B)X_t = W_t,$$

$$Y = (Y_1, \ldots, Y_p)' \qquad \theta(B)Y_t = W_t.$$

# Integrated ARMA Models: ARIMA(p,d,q)

For $p, d, q \geq 0$, we say that a time series $\{X_t\}$ is an **ARIMA (p,d,q) process** if $Y_t = \nabla^d X_t = (1 - B)^d X_t$ is ARMA(p,q). We can write

$$\phi(B)(1 - B)^d X_t = \theta(B) W_t.$$

Recall the random walk: $X_t = X_{t-1} + W_t$.
$X_t$ is not stationary, but $Y_t = (1 - B) X_t = W_t$ is a stationary process.
In this case, it is white, so $\{X_t\}$ is an ARIMA(0,1,0).

Also, if $X_t$ contains a trend component plus a stationary process, its first difference is stationary.

# ARIMA models example

Suppose $\{X_t\}$ is an ARIMA(0,1,1): $X_t = X_{t-1} + W_t - \theta_1 W_{t-1}$. If $|\theta_1| < 1$, we can show

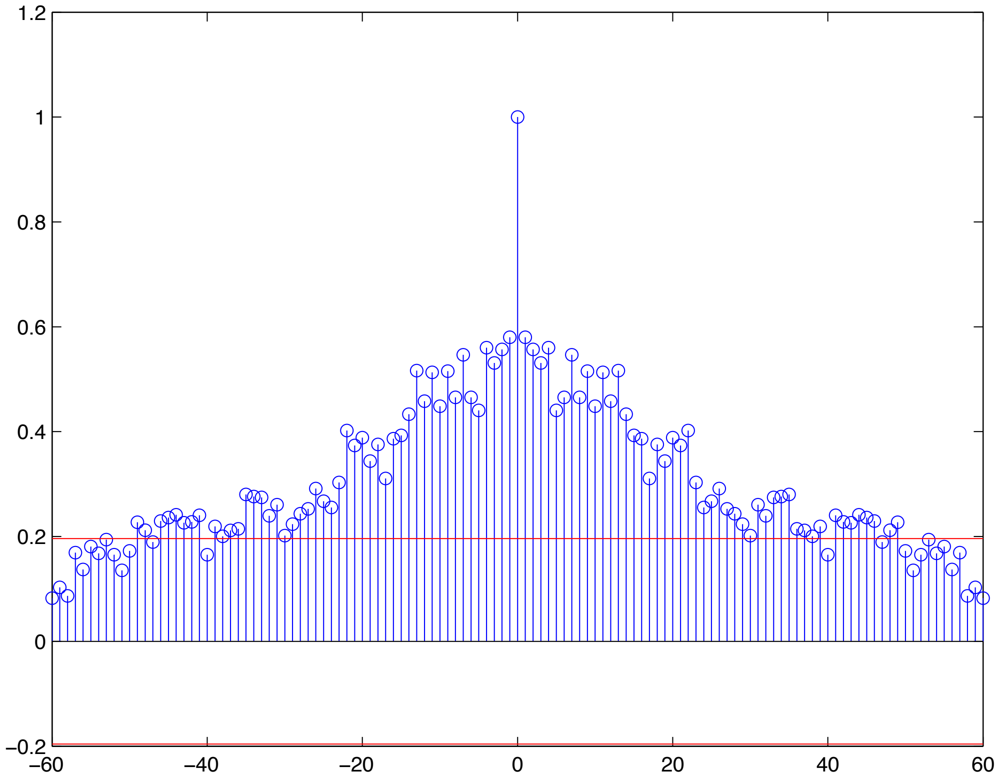
$$X_t = \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{t-j} + W_t,$$

$$\begin{aligned} \text{and so} \quad \tilde{X}_{n+1} &= \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{n+1-j} \\ &= (1-\theta_1)X_n + \sum_{j=2}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{n+1-j} \\ &= (1-\theta_1)X_n + \theta_1 \tilde{X}_n. \end{aligned}$$

Exponentially weighted moving average.

## Identifying preliminary values of $d$: Sample ACF

Trends lead to slowly decaying sample ACF:



# Identifying preliminary values of $d$, $p$, and $q$

For identifying preliminary values of $d$, a time plot can also help.

Too little differencing: not stationary.
Too much differencing: extra dependence introduced.

For identifying $p, q$, look at sample ACF, PACF of $(1-B)^d X_t$:

| **Model:** | **ACF:** | **PACF:** |
|---:|---:|---:|
| AR(p) | decays | zero for $h > p$ |
| MA(q) | zero for $h > q$ | decays |
| ARMA(p,q) | decays | decays |

# Diagnostics

How do we check that a model fits well?

The residuals (innovations, $x_t - x_t^{t-1}$) should be white.
Consider the *standardized innovations*,

$$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}_t^{t-1}}}.$$

This should behave like a mean-zero, unit variance, iid sequence.

- Check a time plot
- Turning point test
- Difference sign test
- Rank test
- Q-Q plot, histogram, to assess normality

# Model Selection

We have used the data $x$ to estimate parameters of several models. They all fit well (the innovations are white). We need to choose a single model to retain for forecasting. How do we do it?

If we had access to independent data $y$ from the same process, we could compare the likelihood on the new data, $L_y(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2)$.

We could obtain $y$ by leaving out some of the data from our model-building, and reserving it for model selection. This is called *cross-validation*. It suffers from the drawback that we are not using all of the data for parameter estimation.

# Model Selection: AIC

We can approximate the likelihood defined using independent data: asymptotically

$$-\ln L_y(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2) \approx -\ln L_x(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2) + \frac{(p+q+1)n}{n-p-q-2}.$$

AIC$_c$: corrected Akaike information criterion.

Notice that:

- More parameters incur a bigger penalty.
- Minimizing the criterion over all values of $p, q, \hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2$ corresponds to choosing the optimal $\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2$ for each $p, q$, and then comparing the penalized likelihoods.

There are also other criteria: BIC.

# Pure seasonal ARMA Models

For $P, Q \geq 0$ and $s > 0$, we say that a time series $\{X_t\}$ is an **ARMA(P,Q)$_s$ process** if $\Phi(B^s)X_t = \Theta(B^s)W_t$, where

$$\Phi(B^s) = 1 - \sum_{j=1}^{P} \Phi_j B^{js},$$

$$\Theta(B^s) = 1 + \sum_{j=1}^{Q} \Theta_j B^{js}.$$

It is **causal** iff the roots of $\Phi(z^s)$ are outside the unit circle.
It is **invertible** iff the roots of $\Theta(z^s)$ are outside the unit circle.

## Pure seasonal ARMA Models

Example: $P = 0, Q = 1, s = 12$. $X_t = W_t + \Theta_1 W_{t-12}$.

$$\begin{aligned}
\gamma(0) &= (1 + \Theta_1^2)\sigma_w^2, \\
\gamma(12) &= \Theta_1 \sigma_w^2, \\
\gamma(h) &= 0 \qquad \text{for } h = 1, 2, \ldots, 11, 13, 14, \ldots.
\end{aligned}$$

Example: $P = 1, Q = 0, s = 12$. $X_t = \Phi_1 X_{t-12} + W_t$.

$$\begin{aligned}
\gamma(0) &= \frac{\sigma_w^2}{1 - \Phi_1^2}, \\
\gamma(12i) &= \frac{\sigma_w^2 \Phi_1^i}{1 - \Phi_1^2}, \\
\gamma(h) &= 0 \qquad \text{for other } h.
\end{aligned}$$

## Pure seasonal ARMA Models

The ACF and PACF for a seasonal ARMA(P,Q)$_s$ are zero for $h \neq si$. For $h = si$, they are analogous to the patterns for ARMA(p,q):

| **Model:** | **ACF:** | **PACF:** |
|---:|---:|---:|
| AR(P)$_s$ | decays | zero for $i > P$ |
| MA(Q)$_s$ | zero for $i > Q$ | decays |
| ARMA(P,Q)$_s$ | decays | decays |

# Multiplicative seasonal ARMA Models

For $p, q, P, Q \geq 0$ and $s > 0$, we say that a time series $\{X_t\}$ is a **multiplicative seasonal ARMA model** (ARMA(p,q)×(P,Q)$_s$) if $\Phi(B^s)\phi(B)X_t = \Theta(B^s)\theta(B)W_t$.

If, in addition, $d, D > 0$, we define the **multiplicative seasonal ARIMA model** (ARIMA(p,d,q)×(P,D,Q)$_s$)

$$\Phi(B^s)\phi(B)\nabla_s^D \nabla^d X_t = \Theta(B^s)\theta(B)W_t,$$

where the *seasonal difference operator of order $D$* is defined by

$$\nabla_s^D X_t = (1 - B^s)^D X_t.$$

# Multiplicative seasonal ARMA Models

Notice that these can all be represented by polynomials

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d = \Xi(B), \qquad \Theta(B^s)\theta(B) = \Lambda(B).$$

But the difference operators imply that $\Xi(B)X_t = \Lambda(B)W_t$ does not define a stationary ARMA process (the AR polynomial has roots on the unit circle). And representing $\Phi(B^s)\phi(B)$ and $\Theta(B^s)\theta(B)$ as arbitrary polynomials is not as compact.

How do we choose $p, q, P, Q, d, D$?

First difference sufficiently to get to stationarity. Then find suitable orders for ARMA or seasonal ARMA models for the differenced time series. The ACF and PACF is again a useful tool here.