Review: least squares linear prediction

Consider a **linear predictor** of X_{n+h} given $X_n = x_n$:

$$f(x_n) = \alpha_0 + \alpha_1 x_n.$$

For a stationary time series $\{X_t\}$, the best linear predictor is $f^*(x_n) = (1 - \rho(h))\mu + \rho(h)x_n$:

$$\begin{aligned} \mathrm{E}\left(X_{n+h} - (\alpha_0 + \alpha_1 X_n)\right)^2 &\geq \mathrm{E}\left(X_{n+h} - f^*(X_n)\right)^2 \\ &= \sigma^2(1 - \rho(h)^2). \end{aligned}$$

Linear prediction

Given $X_1, X_2, \ldots, X_n$, the best linear predictor

$$X_{n+m}^n = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

of X_{n+m} satisfies the **prediction equations**

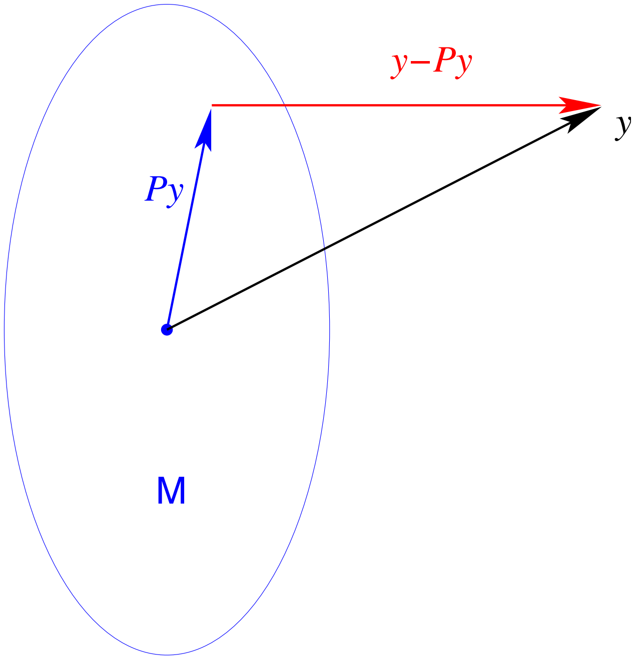
$$\begin{aligned} \mathrm{E}\left(X_{n+m} - X_{n+m}^n\right) &= 0 \\ \mathrm{E}\left[\left(X_{n+m} - X_{n+m}^n\right) X_i\right] &= 0 \qquad \text{for } i = 1, \ldots, n. \end{aligned}$$

This is a special case of the *projection theorem*.

Projection Theorem

If $\mathcal{H}$ is a Hilbert space,
$\mathcal{M}$ is a closed linear subspace of $\mathcal{H}$,
and $y \in \mathcal{H}$,
then there is a point $Py \in \mathcal{M}$
(the **projection of y on $\mathcal{M}$**)
satisfying

1. $\|Py - y\| \leq \|w - y\|$ for $w \in \mathcal{M}$,
2. $\|Py - y\| < \|w - y\|$ for $w \in \mathcal{M}, w \neq y$
3. $\langle y - Py, w\rangle = 0$ for $w \in \mathcal{M}$.



Projection theorem: Linear prediction

Let X_{n+m}^n denote the best linear predictor:

$$\|X_{n+m}^n - X_{n+m}\|^2 \le \|Z - X_{n+m}\|^2 \quad \text{for all } Z \in \mathcal{M}.$$

The projection theorem implies the orthogonality

$$\begin{aligned} & \langle X_{n+m}^n - X_{n+m}, Z\rangle = 0 \quad \text{for all } Z \in \mathcal{M} \\ \Leftrightarrow \quad & \langle X_{n+m}^n - X_{n+m}, Z\rangle = 0 \quad \text{for all } Z \in \{1, X_1, \ldots, X_n\} \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} \mathrm{E}\left(X_{n+m}^n - X_{n+m}\right) &= 0 \\ \mathrm{E}\left[\left(X_{n+m}^n - X_{n+m}\right) X_i\right] &= 0 \end{aligned}$$

That is, the *prediction errors ($X_{n+m}^n - X_{n+m}$) are uncorrelated with the prediction variables ($1, X_1, \ldots, X_n$).*

Linear prediction

Error $(X_{n+m}^n - X_{n+m})$ is uncorrelated with the prediction variable 1:

$$\begin{aligned} & & \mathrm{E}\left(X_{n+m}^n - X_{n+m}\right) &= 0 \\ &\Leftrightarrow & \mathrm{E}\left(\alpha_0 + \sum_i \alpha_i X_i - X_{n+m}\right) &= 0 \\ &\Leftrightarrow & \mu\left(1 - \sum_i \alpha_i\right) &= \alpha_0. \end{aligned}$$

So $\quad X_{n+m}^n = \alpha_0 + \sum_i \alpha_i X_i \quad \Leftrightarrow \quad X_{n+m}^n - \mu = \sum_i \alpha_i \left(X_i - \mu\right).$

Thus, for forecasting, we can assume $\mu = 0$. So we'll ignore α_0.

One-step-ahead linear prediction

Write $X_{n+1}^n = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \cdots + \phi_{nn} X_1$

Prediction equations: $\mathrm{E}\left((X_{n+1}^n - X_{n+1}) X_i\right) = 0, \text{ for } i = 1, \ldots, n$

$$\Leftrightarrow \quad \sum_{j=1}^{n} \phi_{nj} \mathrm{E}\left(X_{n+1-j} X_i\right) = \mathrm{E}(X_{n+1} X_i)$$

$$\Leftrightarrow \quad \sum_{j=1}^{n} \phi_{nj} \gamma(i-j) = \gamma(i)$$

$$\Leftrightarrow \quad \Gamma_n \phi_n = \gamma_n,$$

One-step-ahead linear prediction

Prediction equations: $\Gamma_n \phi_n = \gamma_n.$

$$\Gamma_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & & \gamma(n-2) \\ \vdots & & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix},$$

$$\phi_n = (\phi_{n1}, \phi_{n2}, \ldots, \phi_{nn})', \quad \gamma_n = (\gamma(1), \gamma(2), \ldots, \gamma(n))'.$$

Mean squared error of one-step-ahead linear prediction

$$
\begin{aligned}
P_{n+1}^{n} &= \mathrm{E}\left(X_{n+1} - X_{n+1}^{n}\right)^2 \\
&= \mathrm{E}\left(\left(X_{n+1} - X_{n+1}^{n}\right)\left(X_{n+1} - X_{n+1}^{n}\right)\right) \\
&= \mathrm{E}\left(X_{n+1}\left(X_{n+1} - X_{n+1}^{n}\right)\right) \\
&= \gamma(0) - \mathrm{E}\left(\phi_n' X X_{n+1}\right) \\
&= \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n,
\end{aligned}
$$

where $X = (X_n, X_{n-1}, \ldots, X_1)'$.

Backcasting: Predicting m steps in the past

Given $X_1, \ldots, X_n$, we wish to predict X_{1-m} for $m > 0$.
That is, we choose $Z \in \mathcal{M} = \bar{\text{sp}}\,\{X_1, \ldots, X_n\}$ to minimize $\|Z - X_{1-m}\|^2$.

The prediction equations are

$$\begin{aligned} & \langle X_{1-m}^n - X_{1-m}, Z \rangle = 0 \quad \text{for all } Z \in \mathcal{M} \\ \Leftrightarrow \qquad & \mathrm{E}\left(\left(X_{1-m}^n - X_{1-m}\right) X_i\right) = 0 \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

One-step backcasting

Write the least squares prediction of X_0 given $X_1, \ldots, X_n$ as

$$X_0^n = \phi_{n1} X_1 + \phi_{n2} X_2 + \cdots + \phi_{nn} X_n = \phi_n' X,$$

where the predictor vector is reversed: now $X = (X_1, \ldots, X_n)'$.
The prediction equations are

$$\begin{aligned} & \mathrm{E}\left(\left(X_0^n - X_0 \right) X_i \right) = 0 \quad \text{for } i = 1, \ldots, n \\ \Leftrightarrow \quad & \mathrm{E}\left(\left(\sum_{j=1}^{n} \phi_{nj} X_j - X_0 \right) X_i \right) = 0 \\ \Leftrightarrow \quad & \sum_{j=1}^{n} \phi_{nj} \gamma(j - i) = \gamma(i) \\ \Leftrightarrow \quad & \Gamma_n \phi_n = \gamma_n . \end{aligned}$$

One-step backcasting

The prediction equations are

$$\Gamma_n \phi_n = \gamma_n,$$

which is exactly the same as for forecasting, but with the indices of the predictor vector reversed: $X = (X_1, \ldots, X_n)'$ versus $X = (X_n, \ldots, X_1)'$.

Example: Forecasting AR(1)

AR(1) model: $X_t = \phi_1 X_{t-1} + W_t$

linear prediction of X_2: $X_2^1 = \phi_{11} X_1$

Prediction equation:

$$\begin{aligned} \gamma(0)\phi_{11} &= \gamma(1) \\ &= \mathrm{Cov}(X_0, X_1) \\ &= \phi_1 \gamma(0) \end{aligned}$$

$\Leftrightarrow$ $\phi_{11} = \phi_1.$

Example: Backcasting AR(1)

AR(1) model: $X_t = \phi_1 X_{t-1} + W_t$

linear prediction of X_0: $X_0^1 = \phi_{11} X_1$

Prediction equation:

$$\begin{aligned} \gamma(0)\phi_{11} &= \gamma(1) \\ &= \text{Cov}(X_0, X_1) \\ &= \phi_1 \gamma(0) \end{aligned}$$

$\Leftrightarrow$ $\phi_{11} = \phi_1.$

The prediction operator

For random variables $Y, Z_1, \ldots, Z_n$, define the **best linear prediction of Y given $Z = (Z_1, \ldots, Z_n)'$** as the operator $P(\cdot|Z)$ applied to Y:

$$
\begin{aligned}
 & & P(Y|Z) &= \mu_Y + \phi'(Z - \mu_Z) \\
 &\text{with} & \Gamma\phi &= \gamma, \\
 &\text{where} & \gamma &= \mathrm{Cov}(Y, Z) \\
 & & \Gamma &= \mathrm{Cov}(Z, Z).
\end{aligned}
$$

Properties of the prediction operator

1. $\mathrm{E}(Y - P(Y|Z)) = 0, \mathrm{E}((Y - P(Y|Z))Z) = 0.$

2. $\mathrm{E}((Y - P(Y|Z))^2) = \text{Var}(Y) - \phi'\gamma.$

3. $P(\alpha_1 Y_1 + \alpha_2 Y_2 + \alpha_0|Z) = \alpha_0 + \alpha_1 P(Y_1|Z) + \alpha_2 P(Y_2|Z).$

4. $P(Z_i|Z) = Z_i.$

5. $P(Y|Z) = \mathrm{E}Y$ if $\gamma = 0.$

Example: predicting m steps ahead

Write
$$\begin{aligned} X^n_{n+m} &= \phi^{(m)}_{n1} X_n + \phi^{(m)}_{n2} X_{n-1} + \cdots + \phi^{(m)}_{nn} X_1 \\ \Gamma_n \phi^{(m)}_n &= \gamma^{(m)}_n, \end{aligned}$$

with
$$\begin{aligned} \Gamma_n &= \mathrm{Cov}(X, X), \\ \gamma^{(m)}_n &= \mathrm{Cov}(X_{n+m}, X) \\ &= (\gamma(m), \gamma(m+1), \ldots, \gamma(m+n-1))'. \end{aligned}$$

Also, $\mathrm{E}((X_{n+m} - X^n_{n+m})^2) = \gamma(0) - \phi^{(m)\prime} \gamma^{(m)}_n.$

Partial autocovariance function

AR(1) model: $X_t = \phi_1 X_{t-1} + W_t$

$$\begin{aligned}
\gamma(1) &= \mathrm{Cov}(X_0, X_1) = \phi_1 \gamma(0) \\
\gamma(2) &= \mathrm{Cov}(X_0, X_2) \\
&= \mathrm{Cov}(X_0, \phi_1 X_1 + W_2) \\
&= \mathrm{Cov}(X_0, \phi_1^2 X_0 + \phi_1 W_1 + W_2) \\
&= \phi_1^2 \gamma(0).
\end{aligned}$$

Clearly, X_0 and X_2 are correlated through X_1.

In the PACF, we remove this dependence by considering the covariance of the *prediction errors* of X_2^1 and X_0^1.

Partial autocorrelation function

The Partial AutoCorrelation Function (PACF) of a stationary time series $\{X_t\}$ is

$$\phi_{11} = \mathrm{Corr}(X_1, X_0) = \rho(1)$$

$$\phi_{hh} = \mathrm{Corr}(X_h - X_h^{h-1}, X_0 - X_0^{h-1}) \quad \text{for } h = 2, 3, \ldots$$

This removes the linear effects of $X_1, \ldots, X_{h-1}$:

$$\ldots, X_{-1}, \underline{X_0}, \underbrace{X_1, X_2, \ldots, X_{h-1},}_{\text{partial out}} \underline{X_h}, X_{h+1}, \ldots$$

Partial autocorrelation function

The PACF ϕ_{hh} is also the last coefficient in the best linear prediction of X_{h+1} given $X_1, \ldots, X_h$:

$$\Gamma_h \phi_h = \gamma_h \qquad X_{h+1}^h = \phi_h' X$$

$$\phi_h = (\phi_{h1}, \phi_{h2}, \ldots, \phi_{hh}).$$

Example: Forecasting an AR(p)

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

$$\begin{aligned} X_{n+1}^n &= P(X_{n+1}|X_1, \ldots, X_n) \\ &= P\left(\sum_{i=1}^{p} \phi_i X_{n+1-i} + W_{n+1} | X_1, \ldots, X_n \right) \\ &= \sum_{i=1}^{p} \phi_i P\left(X_{n+1-i} | X_1, \ldots, X_n \right) \\ &= \sum_{i=1}^{p} \phi_i X_{n+1-i} \qquad \text{for } n \geq p. \end{aligned}$$

Example: PACF of an AR(p)

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

$$X_{n+1}^n = \sum_{i=1}^{p} \phi_i X_{n+1-i}.$$

$$\text{Thus, } \phi_{hh} = \begin{cases} \phi_h & \text{if } 1 \leq h \leq p \\ 0 & \text{otherwise.} \end{cases}$$

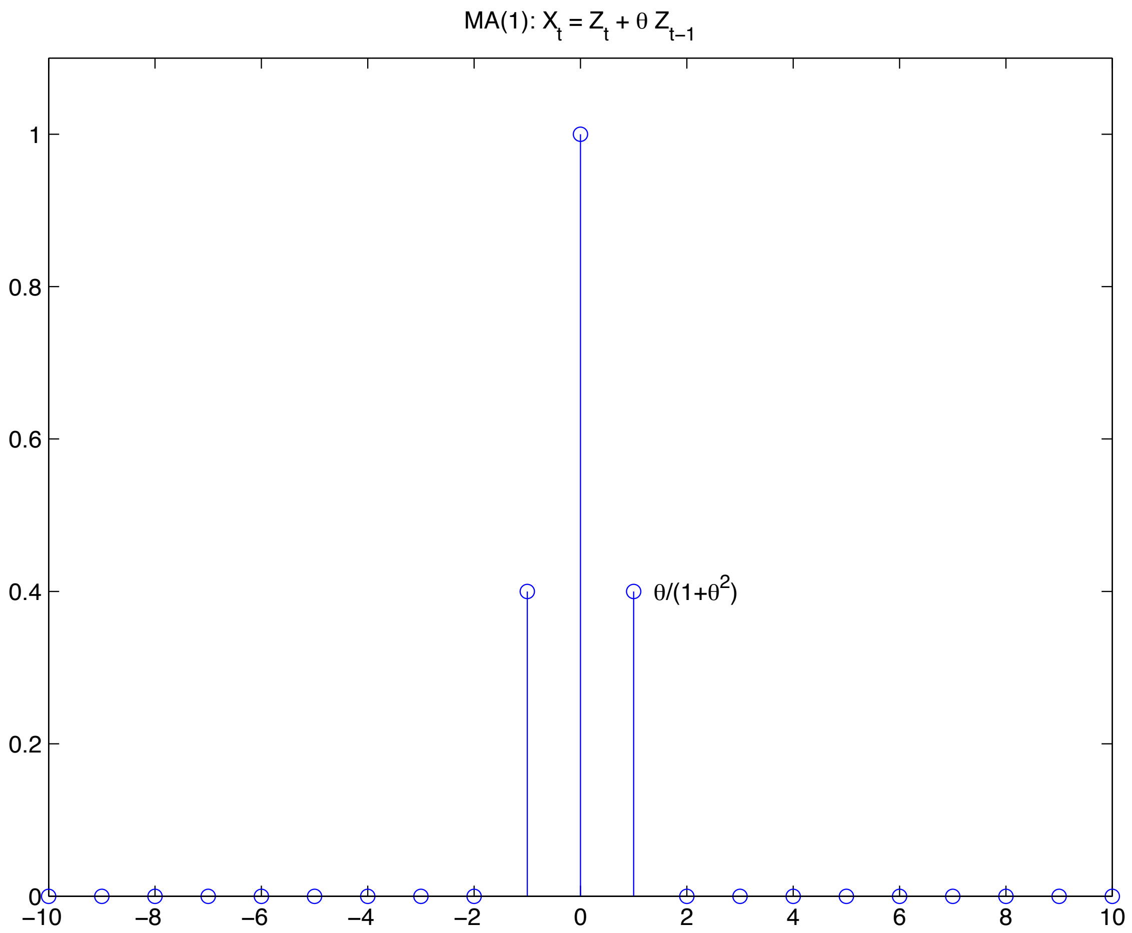
Example: PACF of an invertible MA(q)

For $X_t = \sum_{i=1}^{q} \theta_i W_{t-i} + W_t, \qquad X_t = -\sum_{i=1}^{\infty} \pi_i X_{t-i} + W_t,$

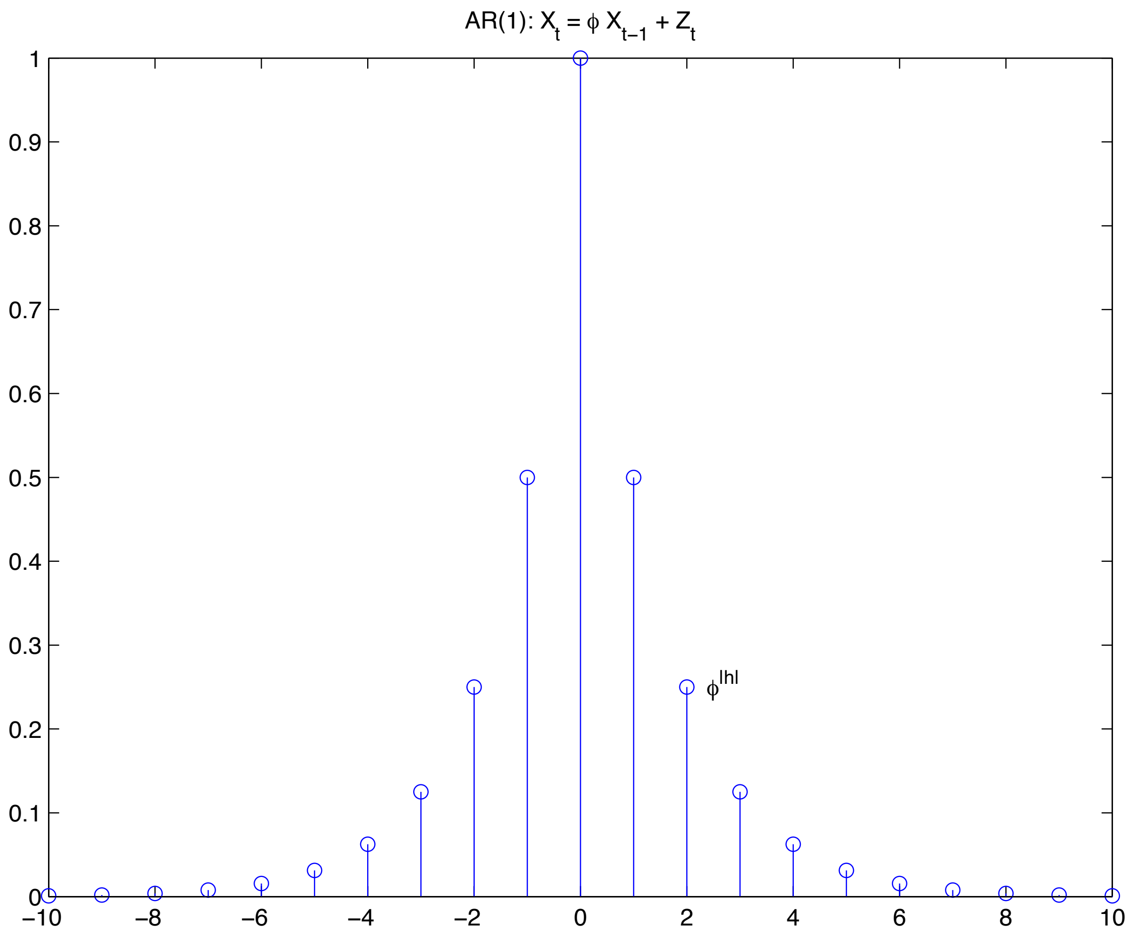
$$\begin{aligned} X_{n+1}^{n} &= P(X_{n+1} | X_1, \ldots, X_n) \\ &= P\left(\sum_{i=1}^{\infty} \pi_i X_{n+1-i} + W_t | X_1, \ldots, X_n \right) \\ &= \sum_{i=1}^{\infty} \pi_i P\left(X_{n+1-i} | X_1, \ldots, X_n \right) \\ &= \sum_{i=1}^{n} \pi_i X_{n+1-i} + \sum_{i=n+1}^{\infty} \pi_i P\left(X_{n+1-i} | X_1, \ldots, X_n \right). \end{aligned}$$

In general, $\phi_{hh} \neq 0$.

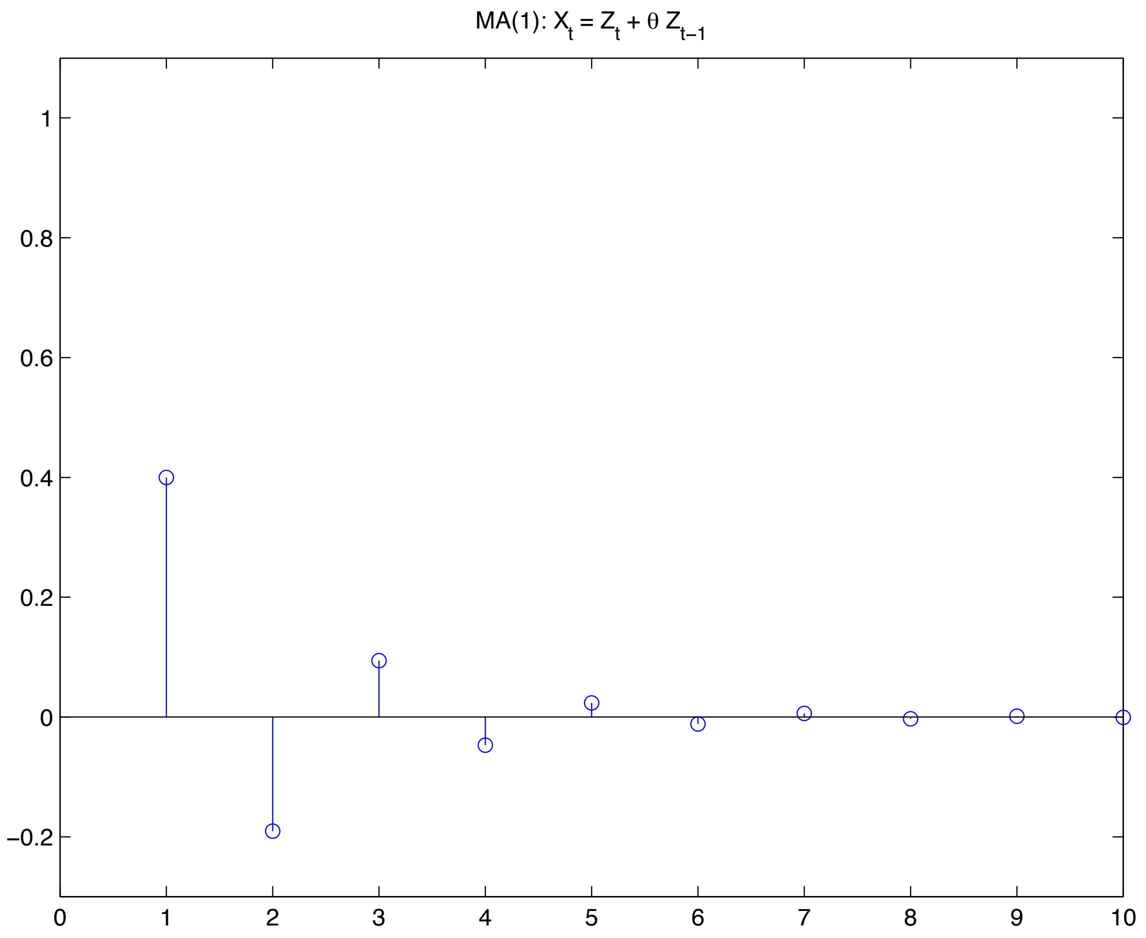
ACF of the MA(1) process



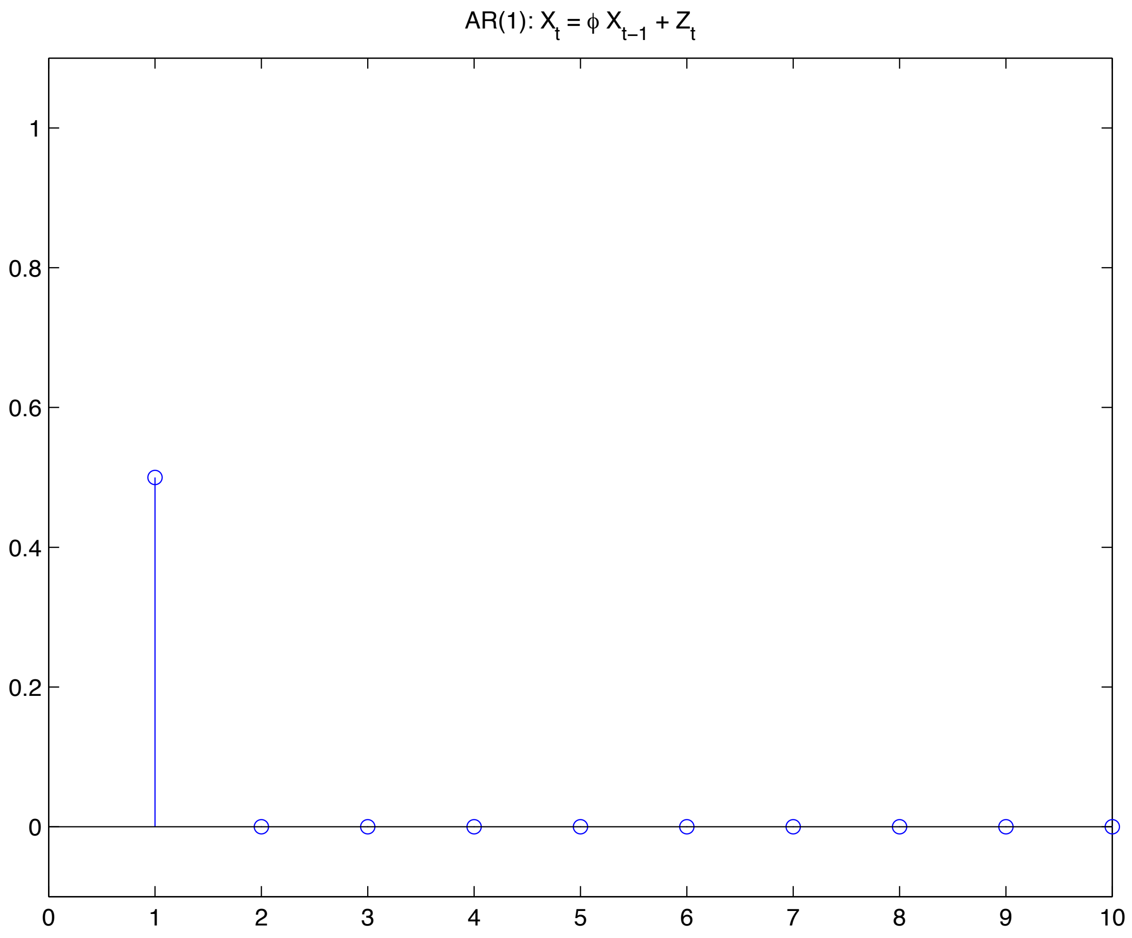
ACF of the AR(1) process



PACF of the MA(1) process



PACF of the AR(1) process



PACF and ACF

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

Sample PACF

For a realization $x_1, \ldots, x_n$ of a time series,
the **sample PACF** is defined by

$$\begin{aligned} \hat{\phi}_{00} &= 1 \\ \hat{\phi}_{hh} &= \text{ last component of } \hat{\phi}_h, \end{aligned}$$

where $\hat{\phi}_h = \hat{\Gamma}_h^{-1} \hat{\gamma}_h$.

The importance of P_{n+1}^n: Prediction intervals

$$X_{n+1}^n = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \cdots + \phi_{nn} X_1$$

$$\Gamma_n \phi_n = \gamma_n, \qquad P_{n+1}^n = \mathrm{E}\left(X_{n+1} - X_{n+1}^n\right)^2 = \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n.$$

After seeing $X_1, \ldots, X_n$, we forecast X_{n+1}^n. The expected squared error of our forecast is P_{n+1}^n. We can construct a prediction interval:

$$X_{n+1}^n \pm c_{\alpha/2} \sqrt{P_{n+1}^n}.$$

For a Gaussian process, the prediction error has distribution $\mathcal{N}(0, P_{n+1}^n)$, so $c_{0.05/2} = 1.96$ gives a 95% prediction interval. For any process with finite second moments, we can apply Chebyshev's inequality:

$$\Pr\left(|X - \mathrm{E}X| \geq t\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{t^2}.$$

Computing linear prediction coefficients

$$X_{n+1}^{n} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \cdots + \phi_{nn}X_1$$

$$\Gamma_n\phi_n = \gamma_n,$$

$$P_{n+1}^{n} = \mathrm{E}\left(X_{n+1} - X_{n+1}^{n}\right)^2 = \gamma(0) - \gamma_n'\Gamma_n^{-1}\gamma_n.$$

How can we compute these quantities recursively?
i.e., given the coefficients ϕ_{n-1} of X_n^{n-1}, how can we compute the coefficients ϕ_n of X_{n+1}^{n}, without solving another linear system $\Gamma_n\phi_n = \gamma_n$?

Durbin-Levinson

$$\phi_0 = 0, \qquad \phi_{00} = 0;$$

$$\phi_1 = \phi_{11}, \qquad \phi_{11} = \frac{\gamma(1)}{\gamma(0)};$$

$$\phi_n = \begin{pmatrix} \phi_{n-1} - \phi_{nn}\tilde{\phi}_{n-1} \\ \phi_{nn} \end{pmatrix}, \quad \phi_{nn} = \frac{\gamma(n) - \phi'_{n-1}\tilde{\gamma}_{n-1}}{\gamma(0) - \phi'_{n-1}\gamma_{n-1}}.$$

$$\phi_n = (\phi_{n1}, \ldots, \phi_{nn})' \qquad \tilde{\phi}_n = (\phi_{nn}, \ldots, \phi_{n1})',$$

$$\gamma_n = (\gamma(1), \ldots, \gamma(n))' \qquad \tilde{\gamma}_n = (\gamma(n), \ldots, \gamma(1))'.$$

Durbin-Levinson: Example

$$\begin{aligned} \phi_0 &= 0, & \phi_{00} &= 0; \\ \phi_1 &= \phi_{11}, & \phi_{11} &= \frac{\gamma(1)}{\gamma(0)}; \\ \phi_n &= \begin{pmatrix} \phi_{n-1} - \phi_{nn}\tilde{\phi}_{n-1} \\ \phi_{nn} \end{pmatrix}, & \phi_{nn} &= \frac{\gamma(n) - \phi'_{n-1}\tilde{\gamma}_{n-1}}{\gamma(0) - \phi'_{n-1}\gamma_{n-1}}. \end{aligned}$$

This algorithm computes $\phi_1, \phi_2, \phi_3, \ldots$, where

$$X_2^1 = X_1\phi_1, \quad X_3^2 = (X_2, X_1)\phi_2, \quad X_4^3 = (X_3, X_2, X_1)\phi_3, \ldots$$

Durbin-Levinson: Example

$$\phi_1 = \phi_{11}, \qquad \phi_{11} = \frac{\gamma(1)}{\gamma(0)};$$

$$\phi_n = \begin{pmatrix} \phi_{n-1} - \phi_{nn}\tilde{\phi}_{n-1} \\ \phi_{nn} \end{pmatrix}, \quad \phi_{nn} = \frac{\gamma(n) - \phi'_{n-1}\tilde{\gamma}_{n-1}}{\gamma(0) - \phi'_{n-1}\gamma_{n-1}}.$$

$$\phi_1 = \gamma(1)/\gamma(0),$$

$$\phi_2 = \begin{pmatrix} \phi_1 - \phi_{22}\phi_{11} \\ \phi_{22} \end{pmatrix} = \begin{pmatrix} \frac{\gamma(1)}{\gamma(0)}\left(1 - \frac{\gamma(2)-\gamma(1)}{\gamma(0)-\gamma(1)}\right) \\ \frac{\gamma(2)-\gamma(1)}{\gamma(0)-\gamma(1)} \end{pmatrix}, \text{ etc.}$$

The innovations representation

Instead of writing the best linear predictor as

$$X_{n+1}^{n} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \cdots + \phi_{nn} X_1,$$

we can write

$$X_{n+1}^{n} = \theta_{n1} \underbrace{\left(X_n - X_n^{n-1}\right)}_{\text{innovation}} + \theta_{n2} \left(X_{n-1} - X_{n-1}^{n-2}\right) + \cdots + \theta_{nn} \left(X_1 - X_1^{0}\right).$$

This is still linear in $X_1, \ldots, X_n$.

The innovations are uncorrelated:
$\mathrm{Cov}(X_j - X_j^{j-1}, X_i - X_i^{i-1}) = 0$ for $i \neq j$.

Comparing representations: $U_n = X_n - X_n^{n-1}$ versus X_n

$$\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\phi_{11} & 1 & & 0 \\ \vdots & & \ddots & \\ -\phi_{n-1,n-1} & -\phi_{n-1,n-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

$$\begin{pmatrix} X_1^0 \\ X_2^1 \\ \vdots \\ X_n^{n-1} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \theta_{11} & 0 & & 0 \\ \vdots & & \ddots & \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \cdots & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$$

Innovations Algorithm

$$X_1^0 = 0, \qquad X_{n+1}^n = \sum_{i=1}^{n} \theta_{ni} \left(X_{n+1-i} - X_{n+1-i}^{n-i} \right).$$

NB: error in text.

$$\theta_{n,n-i} = \frac{1}{P_{i+1}^i} \left(\gamma(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j} \theta_{n,n-j} P_{j+1}^j \right).$$

$$P_1^0 = \gamma(0) \qquad P_{n+1}^n = \gamma(0) - \sum_{i=0}^{n-1} \theta_{n,n-i}^2 P_{i+1}^i.$$

Innovations Algorithm: Example

$$\theta_{n,n-i} = \frac{1}{P_{i+1}^{i}} \left(\gamma(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j}\theta_{n,n-j}P_{j+1}^{j} \right).$$

$$P_1^0 = \gamma(0) \qquad P_{n+1}^{n} = \gamma(0) - \sum_{i=0}^{n-1} \theta_{n,n-i}^2 P_{i+1}^{i}.$$

$$\theta_{1,1} = \gamma(1)/P_1^0, \qquad P_2^1 = \gamma(0) - \theta_{1,1}^2 P_1^0$$

$$\theta_{2,2} = \gamma(2)/P_1^0, \quad \theta_{2,1} = \left(\gamma(1) - \theta_{1,1}\theta_{2,2}P_1^0\right)/P_2^1,$$

$$P_3^2 = \gamma(0) - \left(\theta_{2,2}^2 P_1^0 + \theta_{2,1}^2 P_2^1\right)$$

$$\theta_{3,3}, \quad \theta_{3,2}, \quad \theta_{3,1}, \quad P_4^3, \ldots$$

Predicting h steps ahead using innovations

The innovations representation for the one-step-ahead forecast is

$$P(X_{n+1}|X_1,\ldots,X_n)=\sum_{i=1}^{n}\theta_{ni}\left(X_{n+1-i}-X_{n+1-i}^{n-i}\right),$$

What is the innovations representation for $P(X_{n+h}|X_1,\ldots,X_n)$?

Fact: If $h\geq 1$ and $1\leq i\leq n$, we have
$\mathrm{Cov}(X_{n+h}-P(X_{n+h}|X_1,\ldots,X_{n+h-1}),X_i)=0.$

Thus, $P(X_{n+h}-P(X_{n+h}|X_1,\ldots,X_{n+h-1})|X_1,\ldots,X_n)=0.$
That is, the best prediction of X_{n+h} is the
best prediction of the one-step-ahead forecast of X_{n+h}.

Predicting h steps ahead using innovations

$$
\begin{aligned}
&P(X_{n+h}|X_1,\ldots,X_n)\\
&= P\left(P(X_{n+h}|X_1,\ldots,X_{n+h-1})|X_1,\ldots,X_n\right)\\
&= P\left(\sum_{i=1}^{n+h-1} \theta_{n+h-1,i}\left(X_{n+h-i} - X_{n+h-i}^{n+h-i+1}\right)|X_1,\ldots,X_n\right)\\
&= \sum_{i=1}^{n+h-1} \theta_{n+h-1,i} P\left(\left(X_{n+h-i} - X_{n+h-i}^{n+h-i+1}\right)|X_1,\ldots,X_n\right)\\
&= \sum_{i=h}^{n+h-1} \theta_{n+h-1,i} P\left(\left(X_{n+h-i} - X_{n+h-i}^{n+h-i+1}\right)|X_1,\ldots,X_n\right)\\
&= \sum_{i=h}^{n+h-1} \theta_{n+h-1,i}\left(X_{n+h-i} - X_{n+h-i}^{n+h-i+1}\right)
\end{aligned}
$$

Predicting h steps ahead using innovations

$$P(X_{n+1}|X_1,\ldots,X_n) = \sum_{i=1}^{n} \theta_{ni} \left(X_{n+1-i} - X_{n+1-i}^{n-i}\right)$$

$$P(X_{n+h}|X_1,\ldots,X_n) = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j} \left(X_{n+h-j} - X_{n+h-j}^{n+h-j+1}\right)$$

$$= \sum_{i=1}^{n} \theta_{n+h-1,h-1+i} \left(X_{n+1-i} - X_{n+1-i}^{n-i}\right)$$

$(j = i + h - 1)$

Mean squared error of h-step-ahead forecasts

From orthogonality of the predictors and the error,
$\mathrm{E}\left((X_{n+h} - P(X_{n+h}|X_1,\dots,X_n))\,P(X_{n+h}|X_1,\dots,X_n)\right) = 0.$

That is, $\mathrm{E}\left(X_{n+h}P(X_{n+h}|X_1,\dots,X_n)\right) = \mathrm{E}\left(P(X_{n+h}|X_1,\dots,X_n)^2\right)$.

Hence, we can express the mean squared error as

$$\begin{aligned} P^n_{n+h} &= \mathrm{E}\left(X_{n+h} - P(X_{n+h}|X_1,\dots,X_n)\right)^2 \\ &= \gamma(0) + \mathrm{E}\left(P(X_{n+h}|X_1,\dots,X_n)\right)^2 \\ &\qquad - 2\mathrm{E}\left(X_{n+h}P(X_{n+h}|X_1,\dots,X_n)\right) \\ &= \gamma(0) - \mathrm{E}\left(P(X_{n+h}|X_1,\dots,X_n)\right)^2. \end{aligned}$$

Mean squared error of h-step-ahead forecasts

But the innovations are uncorrelated, so

$$
\begin{aligned}
P_{n+h}^{n} &= \gamma(0) - \mathrm{E}\left(P(X_{n+h}|X_1,\ldots,X_n)\right)^2 \\
&= \gamma(0) - \mathrm{E}\left(\sum_{j=h}^{n+h-1} \theta_{n+h-1,j}\left(X_{n+h-j} - X_{n+h-j}^{n+h-j-1}\right)\right)^2 \\
&= \gamma(0) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2\, \mathrm{E}\left(X_{n+h-j} - X_{n+h-j}^{n+h-j-1}\right)^2 \\
&= \gamma(0) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2\, P_{n+h-j}^{n+h-j-1}.
\end{aligned}
$$

Example: Innovations algorithm for forecasting an MA(1)

Suppose that we have an MA(1) process $\{X_t\}$ satisfying

$$X_t = W_t + \theta_1 W_{t-1}.$$

Given $X_1, X_2, \ldots, X_n$, we wish to compute the best linear forecast of X_{n+1}, using the innovations representation,

$$X_1^0 = 0, \qquad X_{n+1}^n = \sum_{i=1}^{n} \theta_{ni} \left(X_{n+1-i} - X_{n+1-i}^{n-i} \right).$$

Example: Innovations algorithm for forecasting an MA(1)

An aside: The linear predictions are in the form

$$X_{n+1}^n = \sum_{i=1}^n \theta_{ni} Z_{n+1-i}$$

for uncorrelated, zero mean random variables Z_i. In particular,

$$X_{n+1} = Z_{n+1} + \sum_{i=1}^n \theta_{ni} Z_{n+1-i},$$

where $Z_{n+1} = X_{n+1} - X_{n+1}^n$ (and all the Z_i are uncorrelated).
This is suggestive of an MA representation.

Why isn't it an MA?

Example: Innovations algorithm for forecasting an MA(1)

$$\theta_{n,n-i} = \frac{1}{P_{i+1}^{i}} \left(\gamma(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j}\theta_{n,n-j}P_{j+1}^{j} \right).$$

$$P_1^0 = \gamma(0) \qquad P_{n+1}^{n} = \gamma(0) - \sum_{i=0}^{n-1} \theta_{n,n-i}^2 P_{i+1}^{i}.$$

The algorithm computes $P_1^0 = \gamma(0)$, $\theta_{1,1}$ (in terms of $\gamma(1)$); P_2^1, $\theta_{2,2}$ (in terms of $\gamma(2)$), $\theta_{2,1}$; P_3^2, $\theta_{3,3}$ (in terms of $\gamma(3)$), etc.

Example: Innovations algorithm for forecasting an MA(1)

$$\boxed{\theta_{n,n-i} = \frac{1}{P_{i+1}^{i}} \left(\gamma(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j} \theta_{n,n-j} P_{j+1}^{j} \right).}$$

For an MA(1), $\gamma(0) = \sigma^2(1+\theta_1^2)$, $\gamma(1) = \theta_1 \sigma^2$.

Thus: $\theta_{1,1} = \gamma(1)/P_1^0$;

$\theta_{2,2} = 0$, $\theta_{2,1} = \gamma(1)/P_2^1$;

$\theta_{3,3} = \theta_{3,2} = 0$; $\theta_{3,1} = \gamma(1)/P_3^2$, etc.

Because $\gamma(n-i) \neq 0$ only for $i = n-1$, only $\theta_{n,1} \neq 0$.

Example: Innovations algorithm for forecasting an MA(1)

For the MA(1) process $\{X_t\}$ satisfying

$$X_t = W_t + \theta_1 W_{t-1},$$

the innovations representation of the best linear forecast is

$$X_1^0 = 0, \qquad X_{n+1}^n = \theta_{n1} \left(X_n - X_n^{n-1} \right).$$

More generally, for an MA(q) process, we have $\theta_{ni} = 0$ for $i > q$.

Example: Innovations algorithm for forecasting an MA(1)

For the MA(1) process $\{X_t\}$,

$$X_1^0 = 0, \qquad X_{n+1}^n = \theta_{n1}\left(X_n - X_n^{n-1}\right).$$

This is consistent with the observation that

$$X_{n+1} = Z_{n+1} + \sum_{i=1}^{n} \theta_{ni} Z_{n+1-i},$$

where the uncorrelated Z_i are defined by $Z_t = X_t - X_t^{t-1}$ for $t = 1, \ldots, n+1$.

Indeed, as n increases, $P_{n+1}^n \to \mathrm{Var}(W_t)$ (recall the recursion for P_{n+1}^n), and $\theta_{n1} = \gamma(1)/P_n^{n-1} \to \theta_1$.

Recall: Forecasting an AR(p)

For the AR(p) process $\{X_t\}$ satisfying

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

$$X_1^0 = 0, \qquad X_{n+1}^n = \sum_{i=1}^{p} \phi_i X_{n+1-i}$$

for $n \geq p$. Then

$$X_{n+1} = \sum_{i=1}^{p} \phi_i X_{n+1-i} + Z_{n+1},$$

where $Z_{n+1} = X_{n+1} - X_{n+1}^n$.

The Durbin-Levinson algorithm is convenient for AR(p) processes.

The innovations algorithm is convenient for MA(q) processes.

Linear prediction based on the infinite past

So far, we have considered linear predictors based on n observed values of the time series:

$$X_{n+m}^n = P(X_{n+m}|X_n, X_{n-1}, \ldots, X_1).$$

What if we have access to *all* previous values, $X_n, X_{n-1}, X_{n-2}, \ldots$?

Write

$$\begin{aligned} \tilde{X}_{n+m} &= P(X_{n+m}|X_n, X_{n-1}, \ldots) \\ &= \sum_{i=1}^{\infty} \alpha_i X_{n+1-i}. \end{aligned}$$

Linear prediction based on the infinite past

$$\tilde{X}_{n+m} = P(X_{n+m}|X_n, X_{n-1}, \ldots) = \sum_{i=1}^{\infty} \alpha_i X_{n+1-i}.$$

The orthogonality property of the optimal linear predictor implies

$$\mathrm{E}\left[(\tilde{X}_{n+m} - X_{n+m})X_{n+1-i}\right] = 0, \quad i = 1, 2, \ldots$$

Thus, if $\{X_t\}$ is a zero-mean stationary time series, we have

$$\sum_{j=1}^{\infty} \alpha_j \gamma(i-j) = \gamma(m-1+i), \quad i = 1, 2, \ldots$$

Linear prediction based on the infinite past

If $\{X_t\}$ is a causal, invertible, *linear* process, we can write

$$X_{n+m} = \sum_{j=1}^{\infty} \psi_j W_{n+m-j} + W_{n+m}, \quad W_{n+m} = \sum_{j=1}^{\infty} \pi_j X_{n+m-j} + X_{n+m}.$$

In this case,

$$\begin{aligned}\tilde{X}_{n+m} &= P(X_{n+m}|X_n, X_{n-1}, \ldots) \\ &= P(W_{n+m}|X_n, \ldots) - \sum_{j=1}^{\infty} \pi_j P(X_{n+m-j}|X_n, \ldots) \\ &= -\sum_{j=1}^{m-1} \pi_j P(X_{n+m-j}|X_n, \ldots) - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}.\end{aligned}$$

Linear prediction based on the infinite past

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j P(X_{n+m-j}|X_n,\ldots) - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}.$$

That is, $\tilde{X}_{n+1} = -\sum_{j=1}^{\infty} \pi_j X_{n+1-j},$

$$\tilde{X}_{n+2} = -\pi_1 \tilde{X}_{n+1} - \sum_{j=2}^{\infty} \pi_j X_{n+2-j},$$

$$\tilde{X}_{n+3} = -\pi_1 \tilde{X}_{n+2} - \pi_2 \tilde{X}_{n+1} - \sum_{j=3}^{\infty} \pi_j X_{n+3-j}.$$

The invertible (AR(∞)) representation gives the forecasts $\tilde{X}^n_{n+m}$.

Linear prediction based on the infinite past

To compute the mean squared error, we notice that

$$\begin{aligned}\tilde{X}_{n+m} = P(X_{n+m}|X_n, X_{n-1}, \ldots) &= \sum_{j=1}^{\infty} \psi_j P(W_{n+m-j}|X_n, X_{n-1}, \ldots) \\ &\quad + P(W_{n+m}|X_n, X_{n-1}, \ldots) \\ &= \sum_{j=m}^{\infty} \psi_j W_{n+m-j}.\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X_{n+m} - P(X_{n+m}|X_n, X_{n-1}, \ldots)\right)^2 &= \mathrm{E}\left(\sum_{j=0}^{m-1} \psi_j W_{n+m-j}\right)^2 \\ &= \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2.\end{aligned}$$

Linear prediction based on the infinite past

That is, the mean squared error of the forecast based on the infinite history is given by the initial terms of the causal (MA(∞)) representation:

$$\mathrm{E}\left(X_{n+m}-\tilde{X}_{n+m}\right)^2=\sigma_w^2\sum_{j=0}^{m-1}\psi_j^2.$$

In particular, for $m=1$, the mean squared error is σ_w^2.

The truncated forecast

For large n, truncating the infinite-past forecasts gives a good approximation:

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}$$

$$\tilde{X}^n_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}^n_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j X_{n+m-j}.$$

The approximation is exact for AR(p) when $n \geq p$, since $\pi_j = 0$ for $j > p$. In general, it is a good approximation if the π_j converge quickly to 0.

Example: Forecasting an ARMA(p,q) model

Consider an ARMA(p,q) model:

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

Suppose we have $X_1, X_2, \ldots, X_n$, and we wish to forecast X_{n+m}.

We could use the best linear prediction, X^n_{n+m}.

For an AR(p) model (that is, $q = 0$), we can write down the coefficients ϕ_n.

Otherwise, we must solve a linear system of size n.

If n is large, the truncated forecasts $\tilde{X}^n_{n+m}$ give a good approximation. To compute them, we could compute π_i and truncate.

There is also a recursive method, which takes time $O((n+m)(p+q))$...

Recursive truncated forecasts for an ARMA(p,q) model

$$\tilde{W}_t^n = 0 \quad \text{for } t \leq 0. \qquad \tilde{X}_t^n = \begin{cases} 0 & \text{for } t \leq 0, \\ X_t & \text{for } 1 \leq t \leq n. \end{cases}$$

$$\begin{aligned} \tilde{W}_t^n &= \tilde{X}_t^n - \phi_1 \tilde{X}_{t-1}^n - \cdots - \phi_p \tilde{X}_{t-p}^n \\ &\qquad - \theta_1 \tilde{W}_{t-1}^n - \cdots - \theta_q \tilde{W}_{t-q}^n \qquad \text{for } t = 1, \ldots, n. \end{aligned}$$

$$\tilde{W}_t^n = 0 \qquad \text{for } t > n.$$

$$\begin{aligned} \tilde{X}_t^n &= \phi_1 \tilde{X}_{t-1}^n + \cdots + \phi_p \tilde{X}_{t-p}^n + \theta_1 \tilde{W}_{t-1}^n + \cdots + \theta_q \tilde{W}_{t-q}^n \\ &\qquad\qquad \text{for } t = n+1, \ldots, n+m. \end{aligned}$$

Example: Forecasting an AR(2) model

Consider the following AR(2) model.

$$X_t + \frac{1}{1.21} X_{t-2} = W_t.$$

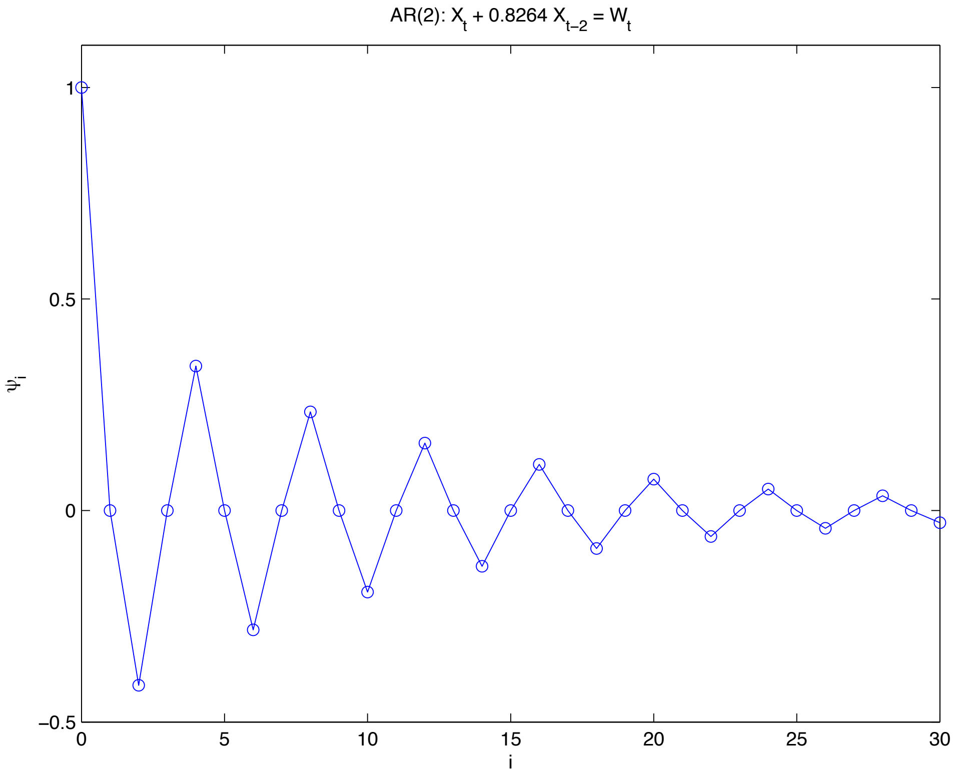
The zeros of the characteristic polynomial $z^2 + 1.21$ are at $\pm 1.1i$. We can solve the linear difference equations $\psi_0 = 1, \phi(B)\psi_t = 0$ to compute the MA(∞) representation:

$$\psi_t = \frac{1}{2} 1.1^{-t} \cos(\pi t/2).$$

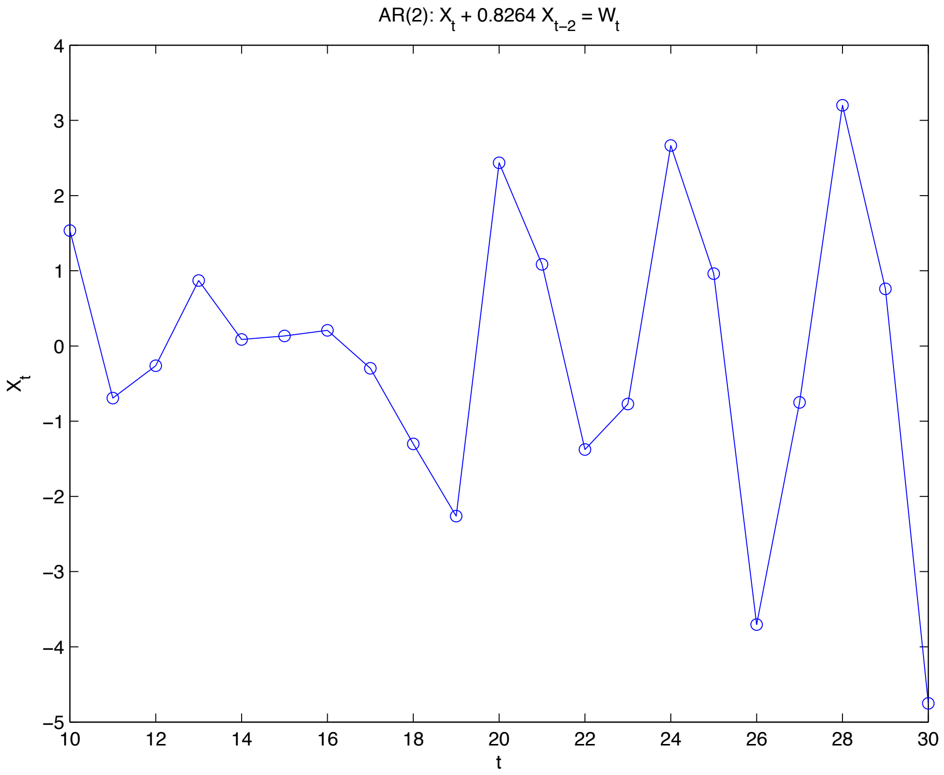
Thus, the m-step-ahead estimates have mean squared error

$$\mathrm{E}(X_{n+m} - \tilde{X}_{n+m})^2 = \sum_{j=0}^{m-1} \psi_j^2.$$

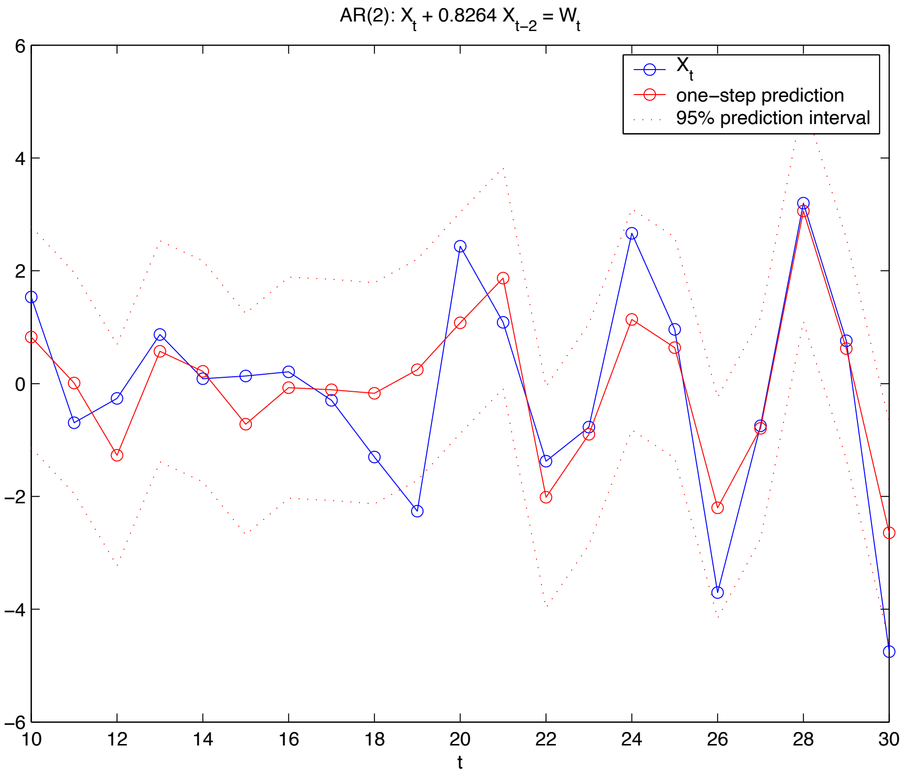
Example: Forecasting an AR(2) model



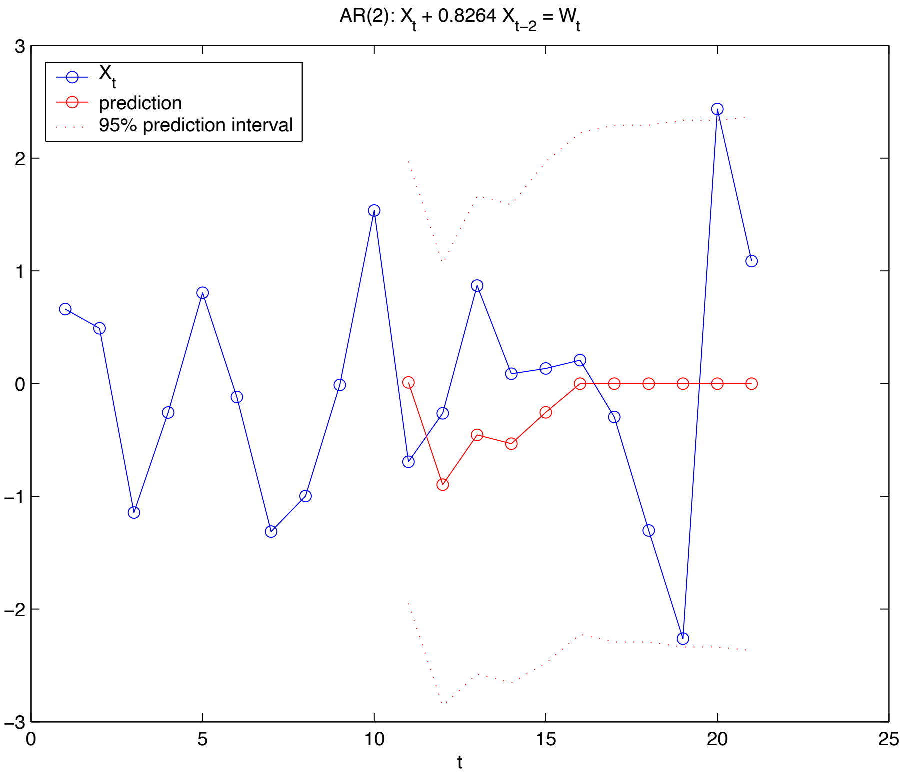
Example: Forecasting an AR(2) model



Example: Forecasting an AR(2) model



Example: Forecasting an AR(2) model



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(∞) representation (from causality), and
an AR(∞) 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 ϕ_i, θ_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 γ replaced by $\hat{\gamma}$.

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

Innovations algorithm for MA(q): with γ 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 ϕ 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 $\approx 1 - \alpha$, ϕ_p is in the ellipsoid

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

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

- with probability $\approx 1 - \alpha$, ϕ_{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 ϕ.

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 ψ_i, these equations are nonlinear in ϕ_i, θ_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 ϕ, we say that $\hat{\phi}^{(2)}$ is a *more efficient* estimator of ϕ 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 ϕ, we have

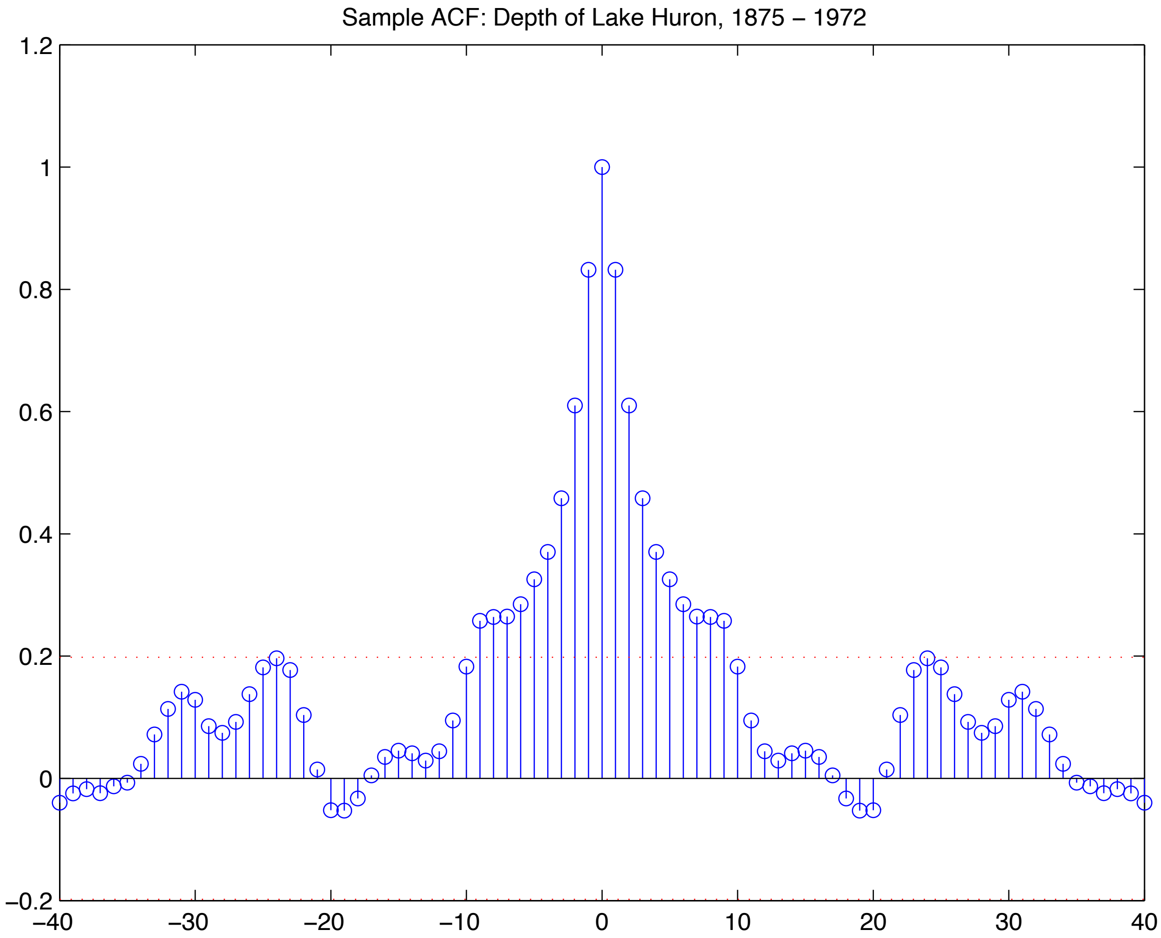
$$\mathrm{Var}(\hat{\phi}_1) \approx \frac{1-\phi_1^2}{n}.$$

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

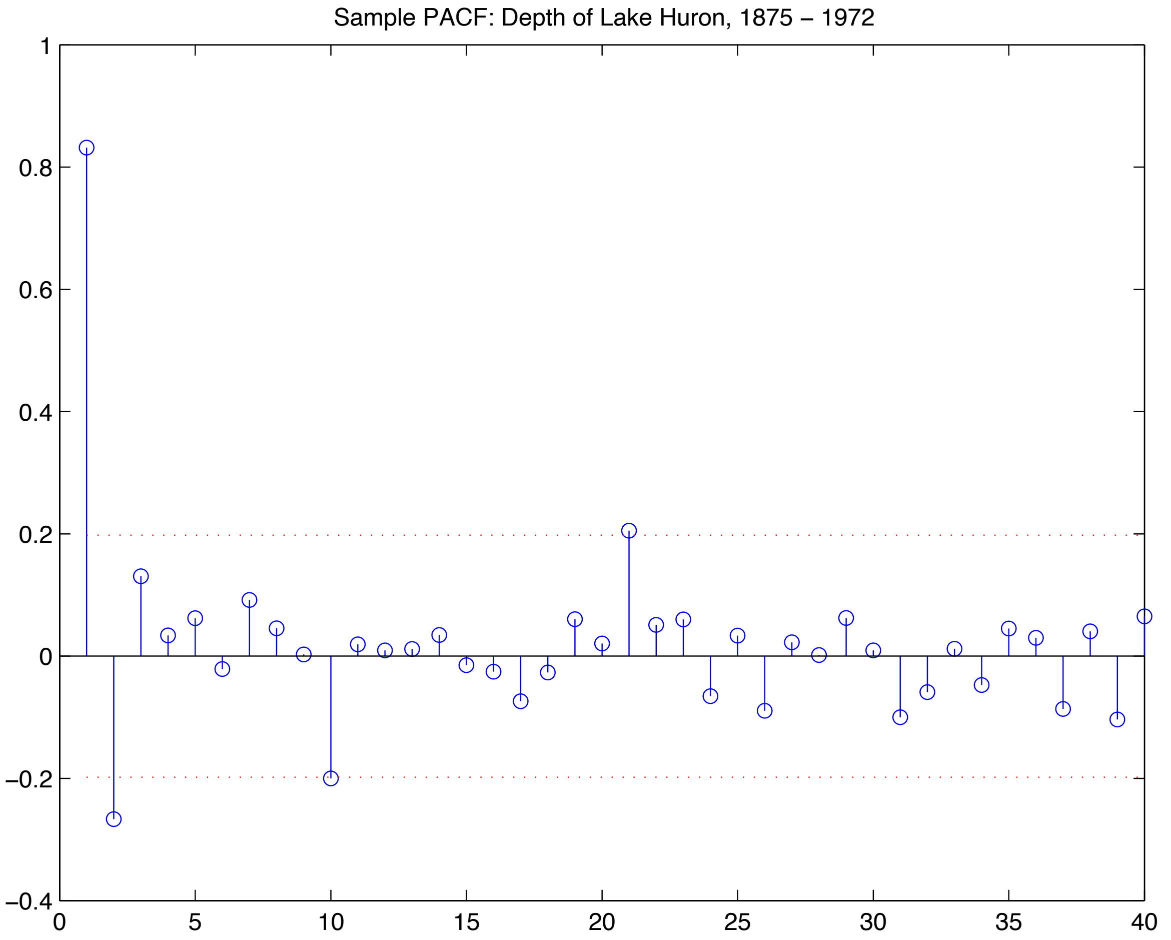
$$\mathrm{Var}(\hat{\phi}_1) \approx \frac{1-\phi_2^2}{n} = \frac{1}{n} \quad > \quad \frac{1-\phi_1^2}{n}.$$

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:

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

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

The maximum likelihood estimator (MLE) of ϕ, θ, σ_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 ϕ, θ, σ_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 σ_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 $\qquad \log\left(\frac{S(\hat{\phi},\hat{\theta})}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}\log r_i^{i-1}.$

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 ϕ_1, ϕ_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

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

$$\text{and so} \quad \tilde{X}_{n+1} = \sum_{j=1}^{\infty} (1 - \theta_1)\theta_1^{j-1} X_{n+1-j}$$

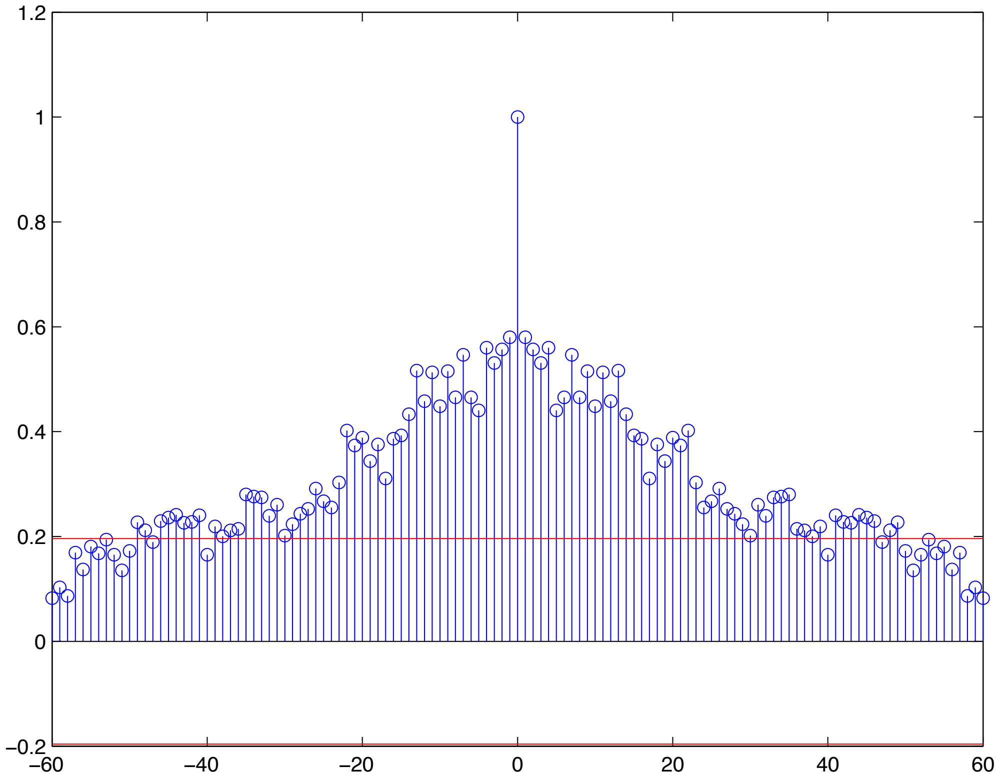
$$= (1 - \theta_1) X_n + \sum_{j=2}^{\infty} (1 - \theta_1)\theta_1^{j-1} X_{n+1-j}$$

$$= (1 - \theta_1) X_n + \theta_1 \tilde{X}_n.$$

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.