Lecture 6
Autoregressive Processes in Time

Dennis Sun
Stanford University
Stats 253

July 6, 2015

1. Review and Preview
2. Autoregressive Processes
3. Estimating Parameters of an AR process
4. Model-Based Approach and Simplifications for AR Processes

The "Hack" Approach

Model: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \mathrm{E}[\boldsymbol{\epsilon}|X] = \mathbf{0},\ \mathrm{Var}[\boldsymbol{\epsilon}|X] = \Sigma.$

- Obtain preliminary estimate $\hat{\boldsymbol{\beta}}^{OLS}$ of $\boldsymbol{\beta}$.
- Calculate residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}^{OLS}$.
- Assume a form of the covariance function and estimate it using $\hat{\boldsymbol{\epsilon}}$.
- Use estimated covariance function to obtain $\hat{\Sigma}$ and calculate the $\hat{\boldsymbol{\beta}}^{GLS}$ estimator.
- (Iterate the process if necessary.)

Today

Model: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \mathrm{E}[\boldsymbol{\epsilon}|X] = \mathbf{0},\ \mathrm{Var}[\boldsymbol{\epsilon}|X] = \Sigma.$

- Obtain preliminary estimate $\hat{\boldsymbol{\beta}}^{OLS}$ of $\boldsymbol{\beta}$.
- Calculate residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}^{OLS}$.
- Assume an autoregressive process for the errors and estimate it using $\hat{\boldsymbol{\epsilon}}$.
- Use estimated covariance function to obtain $\hat{\Sigma}$ and calculate the $\hat{\boldsymbol{\beta}}^{GLS}$ estimator.
- (Iterate the process if necessary.)

Explicit vs. Implicit Covariance Modeling

- Before, we modeled the covariance **explicitly**. That is, we specified $\text{Cov}(\epsilon_i, \epsilon_j)$ for every i and j.
- Today, we will look at modeling the covariance **implicitly**.
- For example, if we had a time series, we could let

$$\epsilon_t = \phi \epsilon_{t-1} + \delta_t,$$

  where the δ's are uncorrelated with each other and with past values of ϵ. Since each ϵ_t depends on the previous one, they will be correlated.
- This is an example of an **autoregressive process of order 1**, or AR(1) process.

Autoregressive Processes

$\{\epsilon_t\}$ is said to be an **AR(p) process** if

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_p \epsilon_{t-p} + \delta_t,$$

where $\mathrm{E}[\boldsymbol{\delta}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\delta}] = \tau^2 I$. Furthermore, δ_t is uncorrelated with past observations $\epsilon_{t-1}, \epsilon_{t-2}, ...$.

AR processes are usually assumed to be stationary. That is, $\mathrm{E}[\epsilon_t] = 0$ and

$$\mathrm{Cov}[\epsilon_t, \epsilon_{t+h}] = \Sigma(h).$$

Covariance Function of an AR process

Let's work out the covariance function of an AR(2) process.

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \delta_t.$$

Idea: Multiply both sides of equation by ϵ_{t-h} and take expectations. This gives us the **Yule-Walker equations**.

$$\mathrm{E}[\epsilon_t \epsilon_{t-h}] = \phi_1 \mathrm{E}[\epsilon_{t-1}\epsilon_{t-h}] + \phi_2 \mathrm{E}[\epsilon_{t-2}\epsilon_{t-h}] + \mathrm{E}[\delta_t \epsilon_{t-h}]$$

For $h \geq 1$: $\Sigma(h) = \phi_1 \Sigma(h-1) + \phi_2 \Sigma(h-2)$.

For $h = 0$: $\Sigma(0) = \phi_1 \Sigma(1) + \phi_2 \Sigma(2) + \tau^2$.

Correlation Function of an AR process

You can solve for the covariance function from those equations. But it's messy, and all we want to know is the general dependence as a function of the lag h.

Let's find the **correlation function** $\rho(h) = \Sigma(h)/\Sigma(0)$ instead. By definition, $\rho(0) = 1$.

For $h \geq 1$, we have:

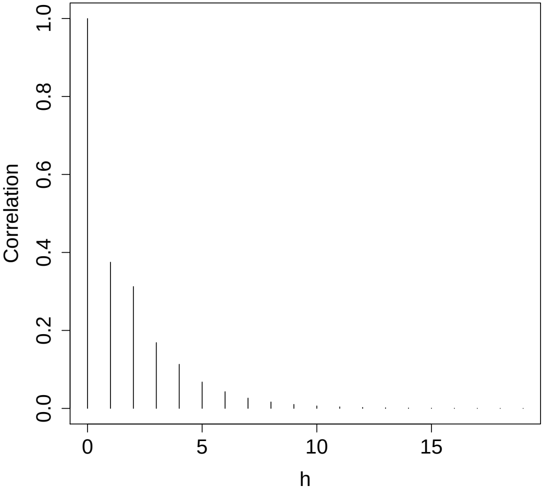
$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2).$$

This implies that $\rho(1) = \phi_1 \rho(0) + \phi_2 \rho(1)$, so $\rho(1) = \frac{\phi_1}{1-\phi_2}$.

Now that we have the initial conditions $\rho(0)$ and $\rho(1)$, we can calculate $\rho(2), \rho(3), ...$

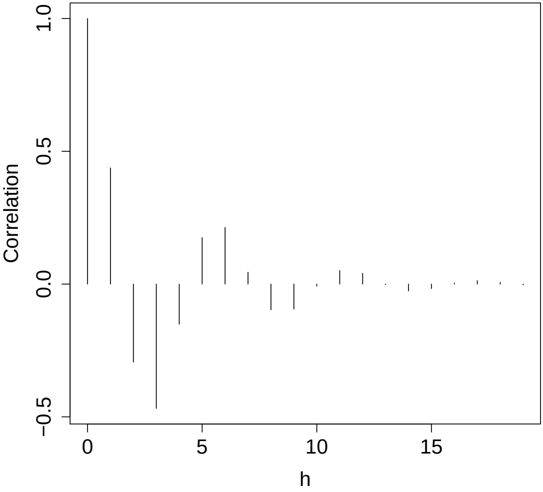
The Induced Correlation Function

$$\epsilon_t = .3\epsilon_{t-1} + .2\epsilon_{t-2} + \delta_t$$



The Induced Correlation Function

$$\epsilon_t = .7\epsilon_{t-1} - .6\epsilon_{t-2} + \delta_t$$



Forecasting an AR process

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \delta_t$$

Suppose we observe $\epsilon_1, ..., \epsilon_n$. How do we predict $\epsilon_{n+1}, \epsilon_{n+2}, ...$?

Intuitively:

$$\hat{\epsilon}_{n+1} = \phi_1 \epsilon_n + \phi_2 \epsilon_{n-1}$$

$$\hat{\epsilon}_{n+2} = \phi_1 \hat{\epsilon}_{n+1} + \phi_2 \epsilon_n$$

$$\hat{\epsilon}_{n+3} = \phi_1 \hat{\epsilon}_{n+2} + \phi_2 \hat{\epsilon}_{n+1}$$

and so forth...

This is called the **chain rule of forecasting**.

Formally: Want MMSE predictor $\hat{\epsilon}_t(\epsilon_1, ..., \epsilon_n)$:

$$\hat{\epsilon}_t = \operatorname*{argmin}_f \mathrm{E}(\epsilon_t - f(\epsilon_1, ..., \epsilon_n))^2 = \mathrm{E}[\epsilon_t | \epsilon_1, ..., \epsilon_n].$$

Applying this formula, we obtain:

$$\begin{aligned}
\hat{\epsilon}_{n+1} &= \mathrm{E}[\epsilon_{n+1} | \epsilon_1, ..., \epsilon_n] = \phi_1 \epsilon_n + \phi_2 \epsilon_{n-1} \\
\hat{\epsilon}_{n+2} &= \mathrm{E}[\epsilon_{n+2} | \epsilon_1, ..., \epsilon_n] = \phi_1 \mathrm{E}[\epsilon_{n+1} | \epsilon_1, ..., \epsilon_n] + \phi_2 \epsilon_n \\
&= \phi_1 \hat{\epsilon}_{n+1} + \phi_2 \epsilon_n
\end{aligned}$$

The Setup

Now suppose that we have observations $\epsilon_1, ..., \epsilon_n$ (or at least $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$) and wish to fit an AR($p$) model to the data.

How do we estimate $\phi_1, ..., \phi_p$?

Autoregression!

Let's write the model

$$\epsilon_t = \phi_1 \epsilon_{t-1} + ... + \phi_p \epsilon_{t-p} + \delta_t, \;\; t = 1, ..., n$$

in matrix form:

$$\underbrace{\begin{pmatrix} \epsilon_{p+1} \\ \epsilon_{p+2} \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} \epsilon_p & \epsilon_{p-1} & \cdots & \epsilon_1 \\ \epsilon_{p+1} & \epsilon_p & \cdots & \epsilon_2 \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n-1} & \epsilon_{n-2} & \cdots & \epsilon_{n-p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \delta_{p+1} \\ \delta_{p+2} \\ \vdots \\ \delta_n \end{pmatrix}}_{\boldsymbol{\epsilon}}$$

Solution: Run OLS of $\boldsymbol{\epsilon}$ on lagged copies of itself.

* Notice that we had to throw away the first p observations. This is not a problem when n is large.

Justification 1: Yule-Walker Equations (Method of Moments)

Remember that the Yule-Walker equations were:

$$\Sigma(h) = \phi_1 \Sigma(h-1) + \ldots + \phi_p \Sigma(h-p), \;\; h \geq 1.$$

In matrix form, they are:

$$\underbrace{\begin{pmatrix} \Sigma(1) \\ \Sigma(2) \\ \vdots \\ \Sigma(p) \end{pmatrix}}_{\mathrm{E}\left[\frac{1}{n-p} X^T \mathbf{y}\right]} = \underbrace{\begin{pmatrix} \Sigma(0) & \Sigma(1) & \cdots & \Sigma(p-1) \\ \Sigma(1) & \Sigma(0) & \cdots & \Sigma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(p-1) & \Sigma(p-2) & \cdots & \Sigma(0) \end{pmatrix}}_{\mathrm{E}\left[\frac{1}{n-p} X^T X\right]} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix},$$

where X and $\mathbf{y}$ are as defined on the last slide. Now replace expected values by their sample versions to solve for $\boldsymbol{\phi}$. But now we just have an OLS problem.

Justification 2: Maximum Likelihood

$$\epsilon_t = \phi_1 \epsilon_{t-1} + ... + \phi_p \epsilon_{t-p} + \delta_t, \;\; t = p+1, ..., n$$

If we further assume that δ_t are normally distributed, then the log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\phi}) &= -\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{t=p+1}^{n} \delta_t^2 \\ &= -\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{t=p+1}^{n} (\epsilon_t - \phi_1\epsilon_{t-1} - ... - \phi_p\epsilon_{t-p})^2, \end{aligned}$$

so the MLE can be obtained by regressing $\boldsymbol{\epsilon}$ on lagged versions of itself.

Review of Model-Based Approach

The "hack" estimates the trend and covariance in two separate stages. This is unsatisfying.

If we're willing to assume that the errors $\boldsymbol{\epsilon}$ are Gaussian, then we can write down the log-likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2} \log \det \Sigma_{\boldsymbol{\theta}} - \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^T \Sigma_{\boldsymbol{\theta}}^{-1} (\mathbf{y} - X\boldsymbol{\beta}) + \text{const.}$$

and optimize jointly over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

To do this, we first *partially* optimize over $\boldsymbol{\beta}$ for $\boldsymbol{\theta}$ fixed to obtain the **partial likelihood**:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2} \log \det \Sigma_{\boldsymbol{\theta}} - \frac{1}{2} \mathbf{y}^T \Sigma_{\boldsymbol{\theta}}^{-1} (I - X(X^T \Sigma_{\boldsymbol{\theta}}^{-1} X)^{-1} X^T \Sigma_{\boldsymbol{\theta}}^{-1}) \mathbf{y}.$$

Now optimize over $\boldsymbol{\theta}$.

Computational Challenges

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log\det\Sigma_{\boldsymbol{\theta}} - \frac{1}{2}\mathbf{y}^T\Sigma_{\boldsymbol{\theta}}^{-1}(I - X(X^T\Sigma_{\boldsymbol{\theta}}^{-1}X)^{-1}X^T\Sigma_{\boldsymbol{\theta}}^{-1})\mathbf{y}.$$

This likelihood is expensive to evaluate, let alone to optimize!

The most expensive operations:

- Evaluating $\log\det\Sigma_{\boldsymbol{\theta}}$: requires Cholesky decomposition of $\Sigma_{\boldsymbol{\theta}}$, $O(n^3)$ operations
- "Inverting" $\Sigma_{\boldsymbol{\theta}}$: requires $O(n^3)$ operations in general.

Simplifications for AR Processes

For an AR(p) process, the inverse covariance matrix is **banded**.

For example, for an AR(2):

$$\Sigma_{\phi}^{-1} = \begin{pmatrix} c_0 & c_1 & c_2 & & & & & \\ c_1 & c_0 & c_1 & c_2 & & & & \\ c_2 & c_1 & c_0 & c_1 & c_2 & & & \\ & c_2 & c_1 & c_0 & c_1 & c_2 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & & & & \\ & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & c_2 & c_1 & c_0 \end{pmatrix}$$

Why does this happen? To answer this, we have to understand what the entries in the inverse covariance matrix represent....

Inverse Covariance Matrix

Let $\Sigma = \mathrm{Var}[\boldsymbol{\epsilon}]$. What is $(\Sigma^{-1})_{ij}$?

Let's look at $(\Sigma^{-1})_{12}$. (Otherwise, we could just simply reorder the rows and columns.) First, let's partition Σ:

$$\Sigma = \begin{pmatrix} \Sigma_{1:2,1:2} & \Sigma_{1:2,3:n} \\ \Sigma_{3:n,1:2} & \Sigma_{3:n,3:n} \end{pmatrix}.$$

Now calculate Σ^{-1}. We only want the upper left hand corner:

$$\begin{aligned}
(\Sigma^{-1})_{1:2,1:2} &= (\Sigma_{1:2,1:2} - \Sigma_{1:2,3:n}(\Sigma_{3:n,3:n})^{-1}\Sigma_{3:n,1:2})^{-1} \\
&= \left(\mathrm{Var}\left[\boldsymbol{\epsilon}_{1:2} \,\middle|\, \boldsymbol{\epsilon}_{3:n}\right]\right)^{-1} = \Sigma^{-1}_{1:2|3:n} \\
&= \frac{1}{\Sigma_{1|3:n}\Sigma_{2|3:n} - \Sigma_{1,2|3:n}} \begin{pmatrix} \Sigma_{2,2|3:n} & -\Sigma_{1,2|3:n} \\ -\Sigma_{1,2|3:n} & \Sigma_{1,1|3:n} \end{pmatrix}
\end{aligned}$$

So $(\Sigma^{-1})_{12}$ is roughly like $-\mathrm{Cov}[\epsilon_1, \epsilon_2 | \boldsymbol{\epsilon}_{3:n}]$ (times some normalizing constant).

Inverse Covariance Matrix

$(\Sigma^{-1})_{ij}$ measures $-\text{Cov}[\epsilon_i, \epsilon_j | \boldsymbol{\epsilon}_{-ij}]$. One is zero if and only if the other is.

So to argue that $(\Sigma^{-1})_{ij}$ for AR(2) is zero for all $|i-j|>2$, we can equiva lently look at $\text{Cov}[\epsilon_i, \epsilon_j | \boldsymbol{\epsilon}_{-ij}]$:

$$\begin{aligned}
\text{Cov}[\epsilon_t, \epsilon_{t+3} | \boldsymbol{\epsilon}_{-t,t+3}] &= \text{Cov}[\epsilon_t, \phi_1 \epsilon_{t+2} + \phi_2 \epsilon_{t+1} + \delta_{t+3} | \boldsymbol{\epsilon}_{-t,t+3}] \\
&= 0.
\end{aligned}$$

The same argument shows that $\text{Cov}[\epsilon_t, \epsilon_{t+h} | \boldsymbol{\epsilon}_{-t,t+h}] = 0$ for any $h > 2$. So $(\Sigma^{-1})_{ij} = 0$ for all $|i-j| > 2$.

How do Banded Inverse Covariances Help?

$$\ell(\boldsymbol{\phi}) = -\frac{1}{2}\log\det\Sigma_{\boldsymbol{\phi}} - \frac{1}{2}\mathbf{y}^T\Sigma_{\boldsymbol{\phi}}^{-1}(I - X(X^T\Sigma_{\boldsymbol{\phi}}^{-1}X)^{-1}X^T\Sigma_{\boldsymbol{\phi}}^{-1})\mathbf{y}.$$

- We can evaluate $\Sigma_{\boldsymbol{\phi}}^{-1}\mathbf{v}$ in $O(np)$ operations. (Since typically $p \ll n$, this means the second term can be evaluated in $O(n)$ operations.)
- We can row-reduce $\Sigma_{\boldsymbol{\phi}}$ to an upper triangular matrix in $O(np^2)$ operations. (Again, since $p \ll n$, this is just $O(n)$.) Then, the determinant is just the product of the values along the diagonal.

So we can evaluate the likelihood in $O(n)$ operations with AR processes, instead of $O(n^3)$ more generally.