Administrative Details for Week 3

These lecture notes/slides, Hamilton, and the Hayashi texts are the required reading. I have listed several other readings that you should consult if you become particularly interested in a topic.

Schedule:

Lectures: *Old Plan:* Monday-Friday from 1:30PM – 3:00PM and from 3:30PM-5:00PM

*Updated Plan:*

I will post two recorded 50-minutes lectures on Sunday-Thursday evening. These will be the same Zoom lectures that you would have seen if we did them in person, but with the advantage that you can watch them when you want, rewind and review etc.

We will meet live on Zoom each day from 4:00PM – 5:00PM to go over the lectures. You can ask questions, and so forth. I will assume that you have watched the day's lectures **before** the 4:00-5:00PM Q&A meetings.

Exercises: Tuesday, Wednesday, Thursday from 7:30PM-8:30PM. We will discuss the format for these during our first live meeting on Monday.

I will send Zoom meeting numbers before each Q&A/exercise session.
Overview of week 3 topics:

1. Time series basics
2. Optimal prediction and Wold’s theorem
3. Likelihood function for time series models with examples
4. Recursive prediction and Kalman filter
5. Linear models with serially correlated data (GLS, HAC, HAR)
6. Dynamic causal effects and SVARs
7. Prediction using Big and Small Data
+ Additional topics as time allows

There will be a take-home exercise following weeks 3 and 4
Logarithm of U.S. Real GDP relative to value in 1947:Q1
Let $y_t = \ln(\text{GDP}_t)$ (quarterly time series)

Growth rate = $\Delta y_t = y_t - y_{t-1}$

Growth rate in Percentage points at an Annual rate: $400 \times \Delta y_t$

AR(1) model for $\Delta y_t$

$\Delta y_t = \beta_0 + \beta_1 \Delta y_{t-1} + \varepsilon_t$

$\varepsilon_t \sim \text{"white noise" } (0, \sigma^2)$
. generate ln_gdp = ln(GDPC1);
. generate d_ln_gdp = D.ln_gdp;
. reg d_ln_gdp L.d_ln_gdp, r;

|                | Coef.  | Std. Err. |     t  |     P>|t|  |      [95% Conf. Interval]|
|----------------|--------|-----------|--------|-------|--------------------------|
| d_ln_gdp       |        |           |        |       |                           |
| L1.            | .3599771| .0652932  | 5.51   | 0.000 | .2314552                 | .488499     |
| _cons          | .0049987| .0007872  | 6.35   | 0.000 | .0034492                 | .0065482    |

Linear regression                               Number of obs     =        285
F(1, 283)                                         =      30.40
Prob > F                                          =     0.0000
R-squared                                         =     0.1302
Root MSE                                          =     .00874
Other series

Generate $\Delta y_t = .0049987 + .3599771 \Delta y_{t-1} + .00874 e_t$

with $\Delta y_0 = 0$ and $e_t \sim iidN(0,1)$

then $y_t = \sum_{i=1}^{t} \Delta y_i + y_0$

(and set $y_0 = 0$)
Some Basic Time Series Concepts

Note: for many of you, much of this is a review, so we will go through this material quickly. The notes are more detailed than the ‘classroom’ discussion.

1. Stochastic Processes Jargon (Review from Hayashi Chapter 2 + elsewhere)

In linear difference equations, the input sequence \( \{ \varepsilon_t \}_{t=1}^{\infty} \) (and possibly, the initial conditions) are a sequence of random variables. The difference equation then generates another sequence of random variables \( \{ Y_t \}_{t=1}^{\infty} \). This is an example of a Stochastic Process. Some useful definitions:

**Stochastic Process**: The probability law governing \( \{ Y_t \}_{t=1}^{\infty} \)

**Realization**: One “draw” from the process, *i.e.*, \( \{ Y_t \}_{t=1}^{\infty} \)

**Stationarity (Strict Stationarity)**: The process is strictly stationary if the probability distribution of \( (Y_t, Y_{t+1}, \ldots, Y_{t+k}) \) is identical to the probability distribution of \( (Y_{\tau}, Y_{\tau+1}, \ldots, Y_{\tau+k}) \) for all \( t, \tau \) and \( k \). (Thus, all joint distributions are time invariant.)
**Autocovariances**. The autocovariances are $\lambda_{t,k} = \text{cov}(Y_t, Y_{t+k})$

**Autocorrelations**. The autocorrelations are $\rho_{t,k} = \text{cor}(Y_t, Y_{t+k})$

**Covariance Stationarity**: The process is covariance stationary if

$$\mu_t = E(Y_t) = \mu \quad \text{for all } t \text{ and } \lambda_{t,k} = \lambda_k \quad \text{for all } t \text{ and } k.$$  

Thus, the means and autocovariances do not depend on time. When $Y_t$ is a scalar covariance stationarity implies that $\lambda_k = \lambda_{-k}$ and when $Y_t$ is a vector $\lambda_k = \lambda_{-k}$.

(Exercise: Show this.)
**White noise:** A process is called white noise if it is covariance stationary, \( \mu = 0 \) and \( \lambda_k = 0 \) for \( |k| > 0 \).

**Martingale Process:** \( Y_t \) follows a martingale process if

\[
E\{Y_t | \Omega_{t-1}\} = Y_{t-1}
\]

where \( \Omega_{t-1} \subseteq \Omega_t \) is the time \( t \) information set. Often \( \Omega_t = \{Y_t\}_{t=0}^\infty \).

**Martingale Difference Process:** \( Y_t \) follows a martingale difference process if

\[
E\{Y_t | \Omega_{t-1}\} = 0
\]

where \( \Omega_{t-1} \subseteq \Omega_t \) is the time \( t \) information set.
Random Walk: $Y_t$ is a random walk if $Y_t - Y_{t-1}$ is white noise.

(Questions: Using these definitions: If $Y_t$ is a martingale, is $Y_t$ a random walk? If $Y_t$ is a random walk, is $Y_t$ a martingale?)

Notation: Throughout these notes, I will use $Y_{ij}$ to denote the vector $Y_{ij} = (Y_i, Y_{i+1}, \ldots, Y_j)'$. 
2. Autoregressive Processes

An example of a stochastic process is

\[ Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t \]

where \( \epsilon_t \sim iid(0, \sigma^2) \),

and the vector \( Y_{p+1:0} = (Y_{-p+1}, Y_{-p+2}, \ldots, Y_0)' \) has mean \( \mu_0 \), covariance matrix \( \Sigma_0 \) and is independent of \( \{\epsilon_t\}_{t=1}^{\infty} \).

This process is called an Autoregressive Process of Order \( p \), abbreviated as AR(\( p \)). (Note: In what follows I will ignore the constant term, \( \phi_0 \), for notational convenience.)
In the AR(1) model: \( Y_t = \phi Y_{t-1} + \varepsilon_t \), it is easy to verify that

\[
Y_t = \phi Y_0 + \sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i}
\]

so that

\[
\mu_t = \phi^t \mu_0
\]

\[
\lambda_{t,0} = \phi^{2t} \sigma_o^2 + \frac{1 - \phi^{2t}}{1 - \phi^2} \sigma^2 \
\text{for } |\phi| \neq 1
\]

and

\[
\lambda_{t,0} = \sigma_o^2 + t \sigma^2 \text{ for } |\phi| = 1.
\]
Suppose that

\[(i) |\phi| < 1\]

and

\[(ii) \mu_0 = 0, \sigma_0^2 = \frac{\sigma^2}{1 - \phi^2}\]

Then:

\[\lambda_{t,k} = \frac{\phi^k}{1 - \phi^2} \sigma^2 = \lambda_k \text{ and } \mu_t = 0.\]

The moments do not depend on \(t\), so (i) and (ii) are the necessary and sufficient conditions for covariance stationarity. Often only (i) is mentioned because, given (i), \(Y_0\) has only a transient effect on the process. Notice that \(\rho_k = \phi^k\) for a covariance stationary AR(1) process.
Vector autoregression with one lag, the VAR(1):

Suppose now that $Y_t$ is an $n \times 1$ vector and $\Phi$ is an $n \times n$ matrix of $\varepsilon_t$ is an $n \times 1$ vector white noise process with covariance matrix $\Sigma$. So

$$Y_t = \Phi Y_{t-1} + \varepsilon_t.$$ 

We can mimic the same steps:

$$Y_t = \Phi' Y_0 + \sum_{i=0}^{t-1} \Phi' \varepsilon_{t-i}.$$ 

Thus, letting $\mu_{Y,t} = E(Y_t)$, we have $\mu_{Y,0} = \Phi' \mu_{Y,0}$,

and letting $\Lambda_t = \text{var}(Y_t)$, we have

$$\Lambda_t = \Phi' \text{var}(Y_0) \Phi' + \sum_{i=0}^{t-1} \Phi' \Sigma \Phi'. $$

Evidently, an important restriction necessary for covariance stationarity is that $\Phi^t \to 0$ as $t$ grows large. This condition is equivalent to the condition that all of eigenvalues of $\Phi$ are less than one in modulus (absolute value).
Here's a quick calculation to understand this restriction on the eigenvalues of $\Phi$:

Assume for the moment that the eigenvalues of $\Phi$ are distinct. In this case we can write

$$\Phi = P \Gamma P^{-1}$$

where the columns of $P$ are eigenvectors of $\Phi$, and $\Gamma$ is a diagonal matrix with the eigenvalues of $\Phi$ on the diagonal (see Hamilton p. 730). Thus

$$\Phi^2 = P \Gamma P^{-1} \Gamma P^{-1} = P \Gamma^2 P^{-1}$$

and $\Phi^i = P \Gamma^i P^{-1}$.

Since $\Gamma$ is diagonal, $\lim_{t \to \infty} \Phi^t = 0$ requires $|\gamma_i| < 1$ for all $i$, where $\gamma_i$ denotes the $i^{th}$ eigenvalue of $\Phi$. When the eigenvalues of $\Phi$ are not distinct, a similar argument can be applied to the Jordan decomposition of $\Phi$ (see Hamilton, page 731.) This restriction, along with a restriction on the mean and variance of $Y_0$ imply that the process is covariance stationary. You can see the details worked out in Hamilton.
To compute the variance (say $\Lambda_0$) in the stationary model, note that $Y_t = \Phi Y_{t-1} + \varepsilon_t$ implies that $\Lambda_0 = \Phi \Lambda_0 \Phi' + \Sigma$.

Recall that $\text{vec}(ABC) = (C^\top \otimes A)\text{vec}(B)$ for conformable matrices $A$, $B$, and $C$.

Apply this to the equation for $\Lambda_0$ and solving yields:

$$\text{vec}(\Lambda_0) = [I - (\Phi \otimes \Phi)]^{-1} \text{vec}(\Sigma).$$
To study AR($p$) models, it is convenient to write the AR($p$) as a VAR(1). This is sometime called the *companion form* of the model. Let $Y_t$ be a scalar that follows an AR($p$) process. Then:

$$\begin{bmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or

$$Z_t = \Phi Z_{t-1} + \epsilon_t$$

with

$$Z_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}$$

From the VAR(1) results, we know that stationarity requires the eigenvalues of the "companion matrix" $\Phi$ to be less than one in absolute value.

VAR($p$) models can be manipulated similarly.
3. The Lag Operator

A useful notational device is the *lag operator*, denoted by $L$ (Note: some people use $B$ to denote this operator). In general, $L$ is an operator that maps the sequence $\{y_t\}_{t=-\infty}^{\infty}$ into another sequence $\{x_t\}_{t=-\infty}^{\infty}$. Specifically, $L$ “lags” the sequence one period. Thus

$$L y_t = y_{t-1}, \quad L^2 y_t = L(L y_t) = L y_{t-1} = y_{t-2}, \quad \text{and} \quad L^p y_t = y_{t-p}.$$ 

Two properties of $L$:

1. If $b$ denotes a constant, then
   $$bL y_t = by_{t-1} = L(by)$$

2. $L(y_t + x_t) = (y_{t-1} + x_{t-1}) = Ly_t + Lx_t$

(1) and (2) imply that $L$ is a linear operator.
We can use this operator to write the AR($p$) model as:

\[ Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t \]

or

\[ Y_t = \phi_1 L Y_t + \phi_2 L^2 Y_t + \ldots + \phi_p L^p Y_t + \epsilon_t \]

or

\[ (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) Y_t = \epsilon_t \]

or

\[ \phi(L) Y_t = \epsilon_t \]

with \( \phi(L) = (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) \).

Notice that the operator \( \phi(L) \) is a \( p \)'th order polynomial in the lag-operator \( L \). This polynomial is called the autoregressive polynomial.
As an exercise you should show that zeros or roots of the autoregressive polynomial are the reciprocals of the eigenvalues of the companion matrix $\Phi$. (The zeros of the autoregressive polynomial are the values of $z$ that make $\phi(z) = 0$.) Thus, often the conditions for covariance stationarity are equivalently stated as:

(1) The eigenvalues of the AR companion matrix are less than 1 in modulus.

or

(2) The roots of the AR polynomial are greater than 1 in modulus.
The lag operator is useful because it can be manipulated in familiar algebraic ways. For example, suppose

\[(1 - \phi L)Y_t = \varepsilon_t\]

and we want to find the operator that maps \(\{\varepsilon_t\}\) into \(\{Y_t\}\), that is, we seek a \(c(L)\) such that

\[Y_t = c(L)\varepsilon_t.\]

Since \((1 - \phi L)Y_t = \varepsilon_t\), it is natural to write \(c(L) = (1 - \phi L)^{-1}\).

What does this mean? Recall

\[(1 - x)^{-1} = 1 + x + x^2 + x^3 + \ldots\]

when \(|x| < 1\). This suggests writing

\[(1 - \phi L)^{-1} = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \ldots\]

which yields the solution

\[Y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\]

which is the same solution obtained by infinite recursive substitution. Thus, inverting \((1 - \phi L)\) algebraically gives the “right answer.”
A word of warning however: The algebraic inverse of the polynomial \((1-\phi z)\) where \(z\) is a variable, is not unique. To see this, note

\[
(1-\phi z)\phi^{-1}z^{-1} = \phi^{-1}z^{-1} - 1
\]

so that

\[
(1-\phi) = \frac{\phi^{-1}z^{-1} - 1}{\phi^{-1}z^{-1}}
\]

and thus

\[
(1-\phi)^{-1} = \phi^{-1}z^{-1} \times \left[-(1-\phi^{-1}z^{-1})^{-1}\right]
\]

\[
= -\phi^{-1}z^{-1}(1+\phi^{-1}z^{-1}+\phi^{-2}z^{-2}+...)
\]

which suggesting writing

\[
(1-\phi L)^{-1} = -\phi^{-1}L^{-1}(1+\phi^{-1}L^{-1}+\phi^{-2}L^{-2}+...).
\]

This yields the solution

\[
Y_t = -\phi^{-1} \sum_{i=1}^{\infty} \phi^{-i} \epsilon_{t+i}
\]

This is the “forward” solution to the difference equation which could be deduced, for example, by forward recursive substitution.
How do we choose between the two solutions

\[ Y_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i} \]

and

\[ Y_t = -\phi^{-1} \sum_{i=1}^{\infty} \phi^{-i} \epsilon_{t+i} \]

One approach is to impose a side condition on the solution: bounded input sequences \{\epsilon_i\} must lead to bounded output sequences. This would rule out the forward solution if \(|\phi|<1\), since \(\phi^t\) will explode as \(t\) grows large. Analogously, it would rule out the backward solution if \(|\phi|>1\). This yields a solution rule: “Solve stable roots backwards and unstable roots forward.”
4. Moving Average Models

We say that $Y_t$ follows a moving average process of order $q$ if

$$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

where $\epsilon_t \sim iid(0, \sigma^2)$.

Consider the MA(1) model

$$Y_t = \epsilon_t - \theta \epsilon_{t-1}.$$ 

A direct calculation yields

$E(Y_t) = 0$

$Var(Y_t) = \sigma^2(1 + \theta^2)$

$Cov(Y_t Y_{t-1}) = -\sigma^2 \theta$, and

$Cov(Y_t Y_{t-k}) = 0$ for $|k| > 1$.

Similarly, for the MA($q$) model

$$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

$E(Y_t) = 0$

$Var(Y_t) = \sigma^2(1 + \sum_{j=1}^{q} \theta_j^2)$

$Cov(Y_t Y_{t-k}) = \sigma^2 \left( -\kappa + \sum_{j=1}^{q-k} \theta_j \theta_{k+j} \right)$ for $k \leq q$, and $Cov(Y_t Y_{t-k}) = 0$ for $|k| > q$.

These moments do not depend on $t$, so the MA($q$) process is covariance stationary.
Invertibility in the MA model:

Motivation: Consider forecasting in the MA(1) model $Y_t = \varepsilon_t - \theta \varepsilon_{t-1}$. Thus, the forecast of $Y_t$ constructed at time $t - 1$ would be $-\theta \varepsilon_{t-1}$. The problem is that $\varepsilon_{t-1}$ is not directly observed; it must be constructed from the lagged values of $Y_t$. How can this be done?

Note: $Y_t = \varepsilon_t - \theta \varepsilon_{t-1}$ implies that $\varepsilon_t = \theta \varepsilon_{t-1} + Y_t$, so that

$$
\varepsilon_t = \theta \varepsilon_0 + \sum_{i=0}^{t-1} \theta^i Y_{t-i}
$$

Thus, if the value of $\varepsilon_0$ was known, then the entire sequence of $\varepsilon$'s could be recovered from present and lagged $Y$'s. But, in general, the value $\varepsilon_0$ is unknown.

The process said to be invertible (meaning that the $\varepsilon$'s can be determined from lagged $Y$'s) if

$$
\hat{\varepsilon}_t = \sum_{i=0}^{t-1} \theta^i Y_{t-i}
$$

converges (say, in mean square) to $\varepsilon_t$:

$$
\hat{\varepsilon}_t = \sum_{i=0}^{t-1} \theta^i Y_{t-i} \to \varepsilon_t.
$$

That is, assuming that $\varepsilon_0 = 0$ has no lasting effect on the quality of the forecasts.
For the MA(1) model invertibility requires $|\theta|<1$. Equivalently, writing $Y_t = (1-\theta L)\varepsilon_t$, invertibility requires that the roots of $\theta(z) = (1-\theta z)$ are greater than 1 in modulus.

For the MA(q) model, we get an analogous result, namely that the process is invertible if the roots of $\theta(z) = (1-\theta_1 z - \theta_2 z^2 - \ldots - \theta_q z^q)$ are greater than 1 in modulus.
Another motivation for the invertibility restriction comes from considering the autocovariances of the MA(1) model

$$\lambda_0 = \sigma^2(1 + \theta^2)$$
$$\lambda_1 = -\sigma^2 \theta$$

Note that we could also describe these two autocovariances as

$$\tilde{\lambda}_0 = \sigma^2(1 + \tilde{\theta}^2)$$
$$\tilde{\lambda}_1 = -\tilde{\sigma}^2 \tilde{\theta}$$

with $\tilde{\theta} = \theta^{-1}$ and $\tilde{\sigma}^2 = \sigma^2(1 + \tilde{\theta}^2)^{-1}(1 + \theta^2)$. That is, the two MA(1) processes

$$Y_t = \varepsilon_t - \theta \varepsilon_{t-1}, \text{ with } \text{var}(\varepsilon_t) = \sigma^2$$

and

$$\tilde{Y}_t = \tilde{\varepsilon}_t - \tilde{\theta} \tilde{\varepsilon}_{t-1}, \text{ with } \text{var}(\tilde{\varepsilon}_t) = \tilde{\sigma}^2$$

have exactly the same autocovariances. Thus, given data on $Y_t$, we can’t tell the processes apart, at least using the first two moments of the data.
These two models are said to be *Observationally Equivalent* for their first two moments. Since $\tilde{\theta} = \theta^{-1}$ one way to (arbitrarily) choose between them is to impose the restriction that the MA parameter is $\leq 1$ in absolute value.

The issue of invertibility is important in dynamic linear structural models and we will discuss it more when we discuss structural VAR models.
5. Autocovariance Generating Functions

The autocovariance generating function for a covariance stationary process is given by

\[ \lambda(z) = \sum_{j=-\infty}^{\infty} \lambda_j z^j \]

so that the autocovariances are given by the coefficients on the argument \( z^j \). Its purpose (or one of the purposes) is the same as the moment generating function – namely it is a convenient way to “store” the autocovariances of a covariance stationary stochastic process.

For the MA process, the ACGF is particularly easy to construct. Suppose

\[ Y_t = \theta(L) \varepsilon_t \]

then

\[ \lambda(z) = \sigma^2 \theta(z) \theta(z^{-1}) \]

We will verify this formula for the MA(1) process – you should verify it for higher order MA processes.

For an MA(1) process \( \theta(z) = (1 - \theta z) \), so that

\[ \lambda(z) = \sigma^2 \theta(z) \theta(z^{-1}) = \sigma^2 [ -\theta z^{-1} + (1 + \theta^2) - \theta z ] \]

which implies autocovariances

\[ \lambda_{-1} = -\sigma^2 \theta \]
\[ \lambda_0 = \sigma^2 (1 + \theta^2) \]
\[ \lambda_1 = -\sigma^2 \theta \]

with all other autocovariances equal to 0. Thus, the formula yields the correct answer for an MA(1) process.
Now, return to the issue of invertibility. Consider the MA polynomial for the MA($q$) model

$$\theta(z) = (1 - \theta_1 z - \theta_2 z^2 - ... - \theta_q z^q)$$

Suppose that this polynomial has zeros at $z = \gamma_1^{-1}, \gamma_2^{-1}, ..., \gamma_q^{-1}$. In this case, we can factor the polynomial as

$$\theta(z) = (1 - \gamma_1 z)(1 - \gamma_2 z)...(1 - \gamma_q z)$$

so the ACGF is given by

$$\lambda(z) = \sigma^2 [(1 - \gamma_1 z)(1 - \gamma_2 z)...(1 - \gamma_q z)(1 - \gamma_1 z^{-1})(1 - \gamma_2 z^{-1})...(1 - \gamma_q z^{-1})].$$

But, since $(1 - \gamma z)(1 - \gamma z^{-1})$ is proportional to $(1 - \gamma^{-1} z)(1 - \gamma^{-1} z^{-1})$ (remember the discussion of invertibility in the MA(1) model), we can “flip” or invert the roots of the MA polynomial, change $\sigma^2$ (to adjust for the factor of proportionality), and obtain the same ACGF – hence a model with the same autocovariances.

Using the notation above, for the MA model $\theta(z) = (1 - \gamma z)$ and the ACGF is:

$$\lambda(z) = \sigma^2 \theta(z) \theta(z^{-1}) = \sigma^2 (1 - \gamma z)(1 - \gamma z^{-1})$$

which, from the invertibility results above is also equal to

$$\lambda(z) = \tilde{\sigma}^2 \left(1 - \frac{1}{\gamma_1} z \right) \left(1 - \frac{1}{\gamma_1} z^{-1} \right)$$

for an appropriate choice of $\tilde{\sigma}^2$.
In the MA(2) model, if

\[ \theta(z) = (1 - \gamma_1 z)(1 - \gamma_2 z) \]

then models with

\[ \theta_1(z) = (1 - \gamma_1^{-1} z)(1 - \gamma_2 z) \]
\[ \theta_2(z) = (1 - \gamma_1 z)(1 - \gamma_2^{-1} z) \]
\[ \theta_3(z) = (1 - \gamma_1^{-1} z)(1 - \gamma_2^{-1} z) \]

are proportional to one another and the factor of proportionality can be captured in the value of \( \sigma^2 \). So, these models are observationally equivalent.

Thus, for the MA(1) model there are \( 2^1 = 2 \) observationally equivalent models.

For the MA(2) model there are \( 2^2 = 4 \) observationally equivalent models.

and for the MA(q) model there are \( 2^q \) observationally equivalent models.
6. Autoregressive-Moving Average (ARMA) Models

Autoregressive-Moving Average models combine the simple AR and MA models. The ARMA$(p,q)$ model is

\[ Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \]

or

\[ \phi(L)Y_t = \theta(L)\varepsilon_t \]

with

\[ \phi(L) = (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) \]

and

\[ \theta(L) = (1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_q L^q) \]

Conditions for covariance stationarity and invertibility are just as in the constituent AR and MA models: the roots of $\phi(z)$ and $\theta(z)$ are greater than 1 in modulus.
The ACGF for the ARMA can be derived as follows. Since
\[ \phi(L)Y_t = \theta(L)\varepsilon_t \]
then
\[ Y_t = c(L)\varepsilon_t \]
with \( c(L) = \phi(L)^{-1}\theta(L) \)
which is a well-defined (mean square convergent) polynomial in positive powers of \( L \) (i.e., backward looking) if the roots of \( \phi(z) \) are greater than 1 in absolute value.

Thus \( Y_t \) has the MA representation \( Y_t = c(L)\varepsilon_t \) so that
\[
\lambda(z) = \sigma^2 c(z)c(z^{-1})
\]
\[
= \sigma^2 \phi(z)^{-1}\theta(z)\phi(z^{-1}) \theta(z^{-1}) = \sigma^2 \frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})}
\]
7. Integrated Autoregressive-Moving Average (ARIMA) Models

Suppose $X_t$ is covariance stationary, and $Y_t$ is constructed as $(1-L)Y_t = X_t$, so that

$$Y_t = Y_0 + \sum_{i=1}^{\infty} X_i.$$ 

In this case, $Y_t$ is said to be a ‘integrated’ version of $X_t$.

More generally, if

$$(1-L)^d Y_t = X_t,$$ 

then $Y_t$ is said to be integrated of order $d$ or $I(d)$. Typically, $d$ is an integer.

If $X_t \sim \text{ARMA}(p,q)$, then $Y_t$ follows an ARIMA($p,d,q$) model. (The extra “I” in ARIMA is for “integrated”).
8. Optimal Prediction and Wold’s Theorem

Consider two sets of random variables, $Y$ and $X$, where $Y$ is a scalar and $X$ is a $k \times 1$ vector.

Suppose we want to predict $Y$ using $X$, i.e., we want to use $\hat{Y} = h(X)$ as a predictor of $Y$. How should the function $h(\cdot)$ be chosen? The answer depends on the forecaster’s loss function. Suppose we solve the problem for squared loss, i.e. we chose $h(X)$ to minimize the mean-square-prediction error: $E[(Y - \hat{Y})^2]$.

In this case the optimal predictor, $h(X)$, is

$$h(X) = E(Y \mid X).$$
A. Linear Predictors

Now, suppose that we restrict the predictor to be linear. That is, we restrict the predictor to be of the form

$$\hat{Y} = \alpha + X'\beta$$

where $\alpha$ is a scalar and $\beta$ is a $k \times 1$ vector. What values of $\alpha$ and $\beta$ minimize the prediction squared error?

Some useful notation. Let

$$\mu_y = E(Y)$$
$$\mu_x = E(X)$$
$$\Sigma_{xx} = var(X)$$
$$\Sigma_{yy} = var(Y)$$
$$\Sigma_{xy} = Cov(X,Y) = E[(X - \mu_x)(Y - \mu_y)]$$

Then we write the optimal linear prediction problem is

$$\min_{a,b} E[(Y - a - X' b)^2]$$

The first order necessary conditions are

$$E(Y - \alpha - X' \beta) = 0$$
$$E[X(Y - \alpha - X' \beta)] = 0$$

Which yield the solutions:

$$\beta = \Sigma_{xx}^{-1} \Sigma_{xy}$$

and

$$\alpha = \mu_y - \mu_x \beta$$
3 useful facts about the linear predictor:

(1) The prediction error $Y - \hat{Y}$ has a mean of 0. (This follows from the 1st first-order condition.)

(2) The prediction error is uncorrelated with $X$. (This follows from the 2nd first-order condition.)

(3) $\text{Var}(Y - \hat{Y}) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ This formula should look familiar – it is the variance of $Y \mid X$ when $Y$ and $X$ are jointly normally distributed. (Why must this be true?)
B. Wold’s Theorem

This theorem says that if $Y_t$ is generated by a covariance stationary process then it can be represented as an MA process.

**Wold’s Theorem:** Suppose $Y_t$ is generated by a “linearly indeterministic” covariance stationary process. Then $Y_t$ can be represented as

$$Y_t = \varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + ...$$

where

$E(\varepsilon_t) = 0$

$var(\varepsilon_t) = \sigma^2$

$cov(\varepsilon_t, \varepsilon_{t-k}) = 0$ for $k \neq 0$

and

$$\sum_{j=1}^{\infty} c_j^2 < \infty$$

Let me sketch the proof and this will make it clear what the term “linearly indeterministic” means.
To prove the theorem we consider the optimal linear predictor of $Y_t$ given $Y_{t-1}, Y_{t-2},\ldots$. Call this predictor $\hat{Y}_t$, which I write as

$$\hat{Y}_t = b_1 Y_{t-1} + b_2 Y_{t-2} + \ldots$$

The coefficients $b_1, b_2, \ldots$ are functions of the autocovariances of the process.

Since the autocovariances are time invariant, the values of $b_i$ are time invariant.

Let

$$\varepsilon_t = Y_t - \hat{Y}_t = Y_t - b_1 Y_{t-1} - b_2 Y_{t-2} + \ldots$$

so that

$$Y_t = \varepsilon_t + \hat{Y}_t = \varepsilon_t + b_1 Y_{t-1} + b_2 Y_{t-2} + \ldots$$

Recall that since $\varepsilon_t$ is the linear prediction error, it is uncorrelated with $\{Y_t\}_{t=\infty}^{t=1}$.

Since the values of $b_i$ are time invariant, we can also write

$$Y_{t-1} = \varepsilon_{t-1} + b_1 Y_{t-2} + b_2 Y_{t-3} + \ldots$$

Substituting this into the equation for $Y_t$ yields

$$Y_t = \varepsilon_t + b_1 \varepsilon_{t-1} + (b_2 + b_1^2) Y_{t-2} + (b_3 + b_1 b_2) Y_{t-3} + \ldots$$

Continuing,

$$Y_{t-2} = \varepsilon_{t-2} + b_1 Y_{t-3} + b_2 Y_{t-4} + \ldots$$

and this can substituted into the equation for $Y_t$, and the process can be further continued to yield an equation of the form
\[ Y_t = \varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + \ldots \]

A Few Notes:

(1) Since \( \varepsilon_t \) is a linear function of \((Y_t, Y_{t-1}, \ldots)\), and \( Y_t \) is covariance stationary with a mean of 0 then \( \varepsilon_t \) has a mean 0.

(2) Since \( \varepsilon_t \) is a time invariant linear function of \((Y_t, Y_{t-1}, \ldots)\), and \( Y_t \) is covariance stationary then \( \varepsilon_t \) has a variance that is time invariant.

(3) Since \( \varepsilon_{t-i} \) is a linear function of \((Y_{t-i}, Y_{t-i-1}, \ldots)\) and \( \varepsilon_t \) is uncorrelated with \((Y_{t-i}, Y_{t-i-1}, \ldots)\) for \( i \geq 1 \), then \( \varepsilon_t \) is uncorrelated with \( \varepsilon_{t-i} \) for \( i \geq 1 \). Covariance stationarity then implies that \( \varepsilon_t \) is uncorrelated with \( \varepsilon_{t+i} \) for \( i \geq 1 \).

(4) Since the \( \varepsilon \)'s are uncorrelated, \( \text{var}(Y_t) = \sigma^2(1 + \sum_{i=2}^{\infty} c_i^2) \), which is finite by the assumption of covariance stationarity.

Jargon: These 1-step-ahead forecast errors \( \{\varepsilon_t\} \) are sometimes called *Wold shocks* or *fundamental shocks*.

The assumption that \( Y_t \) is “linearly indeterministic” means that no part of \( Y_t \) can be perfectly predicted from the infinite past of the process. This means that there no pieces left in the predicting equation, except lagged values of \( \varepsilon_t \) after continued recursive substitution.
A cheat in the ‘proof’:

Formally, the results about optimal linear predictors in the last section need to generalized to handle the problem here. In the last section we considered the prediction problem using a $k \times 1$ vector $X$. In this section $X = (Y_{t-1}, Y_{t-2}, \ldots)$ and so $k$ is infinite.

The results can be generalized using by considering sequences of regressions and letting $k \to \infty$. These sequences can be studied as projections in Hilbert space.

The interested student should consult, for example,

T.W. Anderson’s book *Time Series Analysis*

or

Brockwell and Davis *Time Series: Theory and Methods* (2nd edition, section 5.7)

for a careful discussion of this and a complete proof of Wold’s theorem.
9. The Likelihood Function for Time Series Models with Some Examples

Let \( Y_{1:T} = (Y_1, Y_2, ..., Y_T) \) and denote the joint density of \( Y_{1:T} \) by \( f(Y_{1:T}) \). Recall that the joint pdf factors as \( f(a, b) = f(a | b)f(b) \) for arbitrary random variables \( a \) and \( b \). Thus

\[
f(Y_{1:T}) = f(Y_T | Y_{1:T-1}) f(Y_{1:T-1})
\]

\[
= f(Y_T | Y_{1:T-1}) f(Y_{T-1} | Y_{1:T-2}) f(Y_{1:T-2})
\]

\[
= \prod_{t=2}^{T} f(Y_t | Y_{1:t-1}) f(Y_t)
\]

This shows how the density of the \( T \times 1 \) vector \( Y_{1:T} \) (equivalently, the likelihood) can constructed recursively from \( f(Y_t | Y_{1:T-1}) \) together with a marginal density for the initial observations.
Example 1: Gaussian Model. Suppose $Y_t|Y_{t-1} \sim N(\mu_{t-1}, \sigma^2_{t-1})$, where $\mu_{t-1}$ and $\sigma^2_{t-1}$ are functions of $Y_{1:t-1}$. Then

$$f(Y_t) = \left(\frac{1}{\sqrt{2\pi}}\right)^{-T-1} \left[\prod_{t=2}^T \frac{1}{\sqrt{\sigma^2_{t-1}}}\right] \exp\left(-\frac{1}{2} \sum_{t=2}^T \left(\frac{Y_t - \mu_{t-1}}{\sigma_{t-1}}\right)^2\right) \times f(Y_t)$$

Note: $f(Y_t|Y_{t-1})$ can be interpreted as the density of the one-step-ahead forecast error in $Y_t$ constructed from its past values.

Hence, the factorization

$$f(Y_T) = \prod_{t=2}^T f(Y_t|Y_{t-1}) f(Y_1)$$

is often called the prediction error decomposition of the likelihood.
Example 2: Gaussian AR(1) Model. \( Y_t = \phi Y_{t-1} + \varepsilon_t \), where \( \varepsilon_t \sim \text{Niid}(0, \sigma^2) \).

Since \( \varepsilon_t \) is \( \text{Niid}(0, \sigma^2) \), \( Y_t | Y_{1:t-1} \sim N(\mu_{t-1}, \sigma^2) \), with \( \mu_{t-1} = \phi Y_{t-1} \),

(From Example 1), the likelihood is

\[
 f(Y_{1:T}) = \left( \frac{1}{\sqrt{2\pi \sigma}} \right)^{T-1} \exp \left( -\frac{1}{2} \sum_{t=2}^{T} \frac{(Y_t - \phi Y_{t-1})^2}{\sigma} \right) \times f(Y_1)
\]

where in the stationary model, \( Y_t \sim N \left( 0, \frac{\sigma^2}{1 - \phi^2} \right) \).

Note: ignoring the term \( f(Y_1) \), the MLE is the least squares estimator from the regression of \( Y_t \) onto \( Y_{t-1} \). For comparison with the exact MLE see Hamilton and Davidson and MacKinnon (1993).
**Example 2: Gaussian AR(p) Model.**

\[ Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \epsilon_t, \text{ where } \epsilon_t \sim N \text{iid}(0, \sigma^2). \]

As in the AR(1) model, \( Y_t | Y_{1:t-1} \sim N(\mu_{t-1}, \sigma^2) \), but now \( \mu_{t-1} = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} \).

Thus

\[
f(Y_{1:p}) = \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^{p+1} \exp \left( -\frac{1}{2} \sum_{t=p+1}^{T} \frac{(Y_t - (\phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p}))^2}{\sigma} \right) \times f(Y_{1:p})
\]

Note: The exact likelihood now depends on the density of \( Y_{1:p} \).
Example 3: Gaussian MA(1) model:

\[ Y_t = (1-\theta L)\varepsilon_t, \text{ with } \varepsilon_t \sim \text{Niid}(0, \sigma^2). \]

Assume that the MA process is invertible, so that \( Y_t | Y_{t-1} \sim N(\mu_{t-1}, \sigma^2) \),
but now \( \mu_{t-1} = -\theta \varepsilon_{t-1} = -\theta(1-\theta L)^{-1} Y_{t-1}. \)

Questions:

1. What if the MA process is not invertible?

2. What about initial conditions?

3. The regression function is now a nonlinear function of \( \theta \).
   How do we find the “nonlinear” least squares estimator of \( \theta \)?
From the examples above, the key for constructing the likelihood is that $\mu_{t-1}$ and $\sigma_{t-1}^2$ can be constructed from $Y_{1:t-1}$ (or other data available at time $t-1$).

Here are two others:

**Example 4:** “Threshold AR” Model (TAR): Suppose that

$$Y_t = \phi_{t-1}Y_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim N iid(0, \sigma^2).$$

and the AR coefficient $\phi_{t-1}$ depends on $Y_{t-1}$ as

$$\phi_{t-1} = \begin{cases} 
\phi \text{ if } Y_{t-1} \geq \gamma \\
\phi + \delta \text{ if } Y_{t-1} < \gamma 
\end{cases}$$

where $\gamma$ is a “threshold.”
Example 5: “Autoregressive Heteroskedasticity” or ARCH.

Suppose that $Y_t$ follows an AR(1) model

but with $e_t | Y_{1:t-1} \sim (0, \sigma^2_{t-1})$, where $\sigma^2_{t-1} = \omega + \alpha e^2_{t-1}$.

Noting that

$$\varepsilon^2_{t-1} = (Y_{t-1} - \phi Y_{t-2})^2,$$

then

$$\sigma^2_{t-1} = \omega + \alpha (Y_{t-1} - \phi Y_{t-2})^2$$

and the this fits into the framework outlined above.

Note, $e_t$ is sometimes written as $e_t = \sigma_{t-1} e_t$, where $e_t \sim \text{Niid}(0, 1)$. 
Some models have the same form as the models presented above but with \( \mu_{t-1} \) and/or \( \sigma_{t-1} \) functions of “latent” variables that cannot be recovered from \( Y_{1:t-1} \). The likelihood for models such as this are more complicated. Here are two well-known examples:

**Example 6:** Hamilton’s “Markov Switching” model. Suppose that

\[
Y_t = \phi_{t-1} Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{Niid}(0, \sigma^2),
\]

as in the TAR model, but now with

\[
\phi_{t-1} = \begin{cases} 
\phi & \text{if } S_{t-1} = 0 \\
\phi + \delta & \text{if } S_{t-1} = 1 
\end{cases}
\]

where \( S_t \) follows a Markov process with

\[
P(S_t = 1 | S_{t-1} = 0) = p_{01} \quad \text{and} \quad P(S_t = 1 | S_{t-1} = 1) = p_{11}.
\]

Here it is not possible to perfectly determine the value of \( S_{t-1} \) from \( Y_{1:t-1} \), so the value of \( \phi_{t-1} \) can’t be determined from \( Y_{1:t-1} \).
Example 7: Stochastic volatility. Suppose that $Y_t$ follows an AR(1) model, with

$$\varepsilon_t | Y_{1:t-1} \sim (0, \sigma_{t-1}^2)$$

as in the ARCH model.

But now suppose that $\sigma_t$ evolves as

$$\ln(\sigma_t) = \ln(\sigma_{t-1}) + v_t, \text{ where } v_t \sim \text{Niid}(0, \omega^2)$$

and where $v$ and $e$ (introduced above) and independent.

These two examples include latent (unobserved variables), $S_t$ is the first and $\sigma_t$ in the second.

When the latent variables are Gaussian, and enter the model linearly, they can be handled using the Kalman filter, which is the subject of the next section.
10. Optimal (Recursive) Prediction: The Kalman Filter

Reference: Hamilton, Chapter 13

A. The Basic Linear Model

Measurement Equation

\[ y_t = A'x_t + H'\xi_t + w_t \]

with

\[ E(w_tw'_t) = R \]

Transition (or State) Equation

\[ \xi_t = F\xi_{t-1} + v_t \]

with

\[ E(v_tv'_t) = Q \]

Variables

\( y_t \) is a vector of observed variables

\( x_t \) is a vector or deterministic components (constants, trends, seasonal components, etc.)

\( \xi_t \) is vector of unobserved “state” variables.

\( w_t \) and \( v_t \) are unobserved, mutually uncorrelated and serially uncorrelated noise variables.

\( A, H, R, F, \) and \( Q \) are non-random “system” matrices that may depend on unknown parameters.
Simple example: $y_t \sim AR(p)$

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

can be represented with the following state space model

$$\xi_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}$$

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}$$

$$v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and $w_t = 0$, $A = 0$ and $H' = [1 \ 0 \ \ldots \ 0]$. 
B. Signal Extraction and the Kalman Filter

Let \( y_{1:t} = \{y_i\}_{i=1}^t \),

\( \xi_{t|k} = E(\xi_t | y_{1:k}) \), and

\( P_{t|k} = Var(\xi_t | y_{1:k}) \)

The Kalman filter is a recursive algorithm for constructing \( \xi_{t|k} \) and \( P_{t|k} \). That is, the Kalman Filter is a function that constructs \( \xi_{t|k} \) and \( P_{t|k} \) from \( (\xi_{t-1|t-1}, P_{t-1|t-1}, y_t, x_t) \).

To derive the filter, assume

\[
\begin{bmatrix}
  w_t \\
  v_t
\end{bmatrix}
\sim \text{Niid}\left(\begin{bmatrix}
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  R & 0 \\
  0 & Q
\end{bmatrix}\right).
\]

This will imply that the variables in the model follow a joint normal distribution so that the optimal formula are the linear minimum mean square error formula discussed in Week 1.

Recall the following fact from the multivariate normal distribution:

Suppose

\[
\begin{bmatrix}
  z_1 \\
  z_2
\end{bmatrix}
\sim N\left(\begin{bmatrix}
  \mu_1 \\
  \mu_2
\end{bmatrix}, \begin{bmatrix}
  \Sigma_{11} & \Sigma_{12} \\
  \Sigma_{21} & \Sigma_{22}
\end{bmatrix}\right),
\]

then

\[
E(z_1 | z_2) = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (z_2 - \mu_2)
\]

and

\[
Var(z_1 | z_2) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}
\]

The Kalman Filter is an application of these formulae. We suppose \( \xi_{t-1|t-1} \) and \( P_{t-1|t-1} \) are known. With all noise Gaussian, all joint, marginal, and conditionals are Gaussian, and in particular:

\[
\begin{bmatrix}
  \xi_t \\
  y_{t-1}
\end{bmatrix}
\sim \text{Normal} \quad y_{t-1} \text{ is normally distributed}
\]
We then use the multivariate normal formula with \( z_1 = \xi_t \) and \( z_2 = y_t \). The details are:

\[
y_t = A' x_t + H' \xi_t + w_t
\]
\[
\xi_t = F \xi_{t-1} + v_t
\]

Kalman Filter equations:

1. \( \xi_{t|t-1} = F \xi_{t-1|t-1} \) \( (\mu_1) \)

2. \( y_{t|t-1} = A' x_t + H' \xi_{t|t-1} \) \( (\mu_2) \)

3. \( P_{t|t-1} = FP_{t-1|t-1}F' + Q \) \( (\Sigma_{11}) \)

4. \( h_t = H' P_{t|t-1} H + R \) \( (\Sigma_{22}) \)

5. \( K_t = P_{t|t-1} H h_t^{-1} \) \( (\Sigma_{12} \Sigma_{22}^{-1}) \)

6. \( \eta_t = y_t - y_{t|t-1} \) \( (z_2 - \mu_2) \)
\[ (7) \quad \xi_{t/t} = \xi_{t/t-1} + K_t \eta_t \quad \left( \mu + \Sigma_{12}^{-1} \Sigma_{22} (z_t - \mu) \right) \]

\[ (8) \quad P_{t/t} = P_{t/t-1} - K_t H_{t-1} P_{t/t-1} \quad \left( \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right) \]
Beginning the recursion – two cases:

1. $\xi_t$ is covariance stationary: In this case $\xi_{t=0} = E(\xi_0) = 0$, $P_{0/0} = Var(\xi_0)$ so that

   $$P_{0/0} = FP_{0/0}F' + Q$$

   or

   $$vec(P_{0/0}) = \left[ I - (F \otimes F) \right]^{-1} vec(Q)$$

   where this uses the result that $vec(ABC) = (C' \otimes A)vec(B)$ for conformable matrices $A, B, C$.

2. $\xi_t$ is not covariance stationary: See the discussion in Hamilton.
A related recursion, called a “Smoother” can be used to construct $\xi_{t,T}$ and $P_{t,T}$ (again, see Hamilton)
C. Likelihood function

As above, the Gaussian density function/likelihood function for \( y_{1:T} \) can be factored as

\[
f(y_{1:T}) = f(y_T | y_{1:T-1}) f(y_{T-1} | y_{1:T-2}) \ldots = \prod_{t=1}^{T} f(y_t | y_{t-1}) , \text{ where } y_{1:0} = \{ \emptyset \}.
\]

The conditional densities are normal in this model:

\[
f(y_t | y_{1:t-1}) = \left( \frac{1}{\sqrt{2\pi}} \right)^n | h_t |^{-1/2} \exp \left( -\frac{1}{2} \eta_t ' h_t^{-1} \eta_t \right)
\]

where \( n \) is the number of elements in \( y_t \).

The likelihood is

\[
f(Y_{1:T}) = \left( \frac{1}{\sqrt{2\pi}} \right)^n \prod_{t=1}^{T} | h_t |^{-1/2} \exp \left( -\frac{1}{2} \sum_{t=1}^{T} \eta_t ' h_t^{-1} \eta_t \right)
\]

where \( \eta_t = y_t - y_{t:t-1} \).

Notice that both \( h_t \) and \( \eta_t \) are computed by the Kalman filter.

(This is the same expression as in the last section with \( h_t = \sigma_{t-1}^2 \) and \( y_{t:t-1} = \mu_{t-1} \).)
D. (Some Additional Information): A General Formulation.

Models and objects of interest

General Model (Nonlinear, non-Gaussian state-space model)

\[
y_t = H(s_t, \varepsilon_t)
\]
\[
s_t = F(s_{t-1}, \eta_t)
\]
\[\varepsilon_t \text{ and } \eta_t \sim i.i.d.\]

Jargon: This is sometimes called a "hidden Markov model" because \( s_t \) is "hidden" by the measurement error \( \varepsilon_t \).

Example 1: Linear Gaussian Model

\[
y_t = Hs_t + \varepsilon_t
\]
\[
s_t = Fs_{t-1} + \eta_t
\]
\[
\begin{pmatrix}
\varepsilon_t \\
\eta_t
\end{pmatrix}
\overset{iid}{\sim} \mathcal{N}
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\Sigma_{\varepsilon} & 0 \\
0 & \Sigma_{\eta}
\end{pmatrix}
\right)
\]

Example 2: Hamilton Regime-Switching Model

\[
y_t = \mu(s_t) + \sigma(s_t) \varepsilon_t
\]
\[
s_t = 0 \text{ or } 1 \text{ with } P(s_t = i \mid s_{t-1} = j) = p_{ij}
\]

(using \( s_t = F(s_{t-1}, \eta_t) \) notation:

\[
s_t = 1(\varepsilon_t \leq p_{10} + (p_{11} - p_{10})s_{t-1}, \text{ where } \eta \sim \mathcal{U}[0,1])
\]
Example 3: Stochastic volatility model

\[ y_t = e^{\varepsilon_t} \]
\[ s_t = \mu + \phi(s_{t-1} - \mu) + \eta_t \]

with, say, \( \varepsilon_t \sim iid \ N(0,1) \) and

\[ e^{\varepsilon_t} = \sigma_t \]

the model for \( y \) is

\[ y_t | s_t \sim N(0, \sigma_t^2) \]
Model:

\[ y_t = H(s_t, \varepsilon_t) \]
\[ s_t = F(s_{t-1}, \eta_t) \]

\( \varepsilon_t \) and \( \eta_t \) ~ i.i.d.

Notation: \( y_{1:t} = (y_1, y_2, \ldots, y_t) \), \( s_{1:t} = (s_1, s_2, \ldots, s_t) \),

\( f(\cdot|\cdot) \) a generic density function.

Some things you might want to calculate

A. Prediction and Likelihood

(i) \( f(s_t | y_{1:t-1}) \)

(ii) \( f(y_t | y_{1:t-1}) \)

(iii) \( f(y_{1:T}) = \prod_{t=1}^{T} f(y_t | y_{1:t-1}) \) is the likelihood

B. Filtering: \( f(s_t | y_{1:t}) \)

C. Smoothing: \( f(s_t | y_{1:T}) \).
Some General Formulae

Model: \(y_t = H(s_t, \varepsilon_t), \ s_t = F(s_{t-1}, \eta_t), \ \varepsilon \text{ and } \eta \sim \text{iid}\)

A. Prediction of \(s_t\) and \(y_t\) given \(Y_{t-1}\).

(i)
\[
f(s_t | y_{1:t-1}) = \int f(s_t, s_{t-1} | y_{1:t-1}) ds_{t-1}
\]
\[
= \int f(s_t | s_{t-1}, y_{1:t-1}) f(s_{t-1} | y_{1:t-1}) ds_{t-1}
\]
\[
= \int f(s_t | s_{t-1}) f(s_{t-1} | y_{1:t-1}) ds_{t-1}
\]

(ii)
\[
f(y_t | y_{1:t-1}) = \int f(y_t, s_t | y_{1:t-1}) ds_t
\]
\[
= \int f(y_t | s_t, y_{1:t-1}) f(s_t | y_{1:t-1}) ds_t
\]
\[
= \int f(y_t | s_t) f(s_t | y_{1:t-1}) ds_t
\]

("t" component of likelihood)

---

B. Filtering

\[ f(s_t | y_{1:t}) = f(s_t | y_t, y_{1:t-1}) \]

\[
= \frac{f(y_t | s_t, y_{1:t-1}) f(s_t | y_{1:t-1})}{f(y_t | y_{1:t-1})}
\]

\[
= \frac{f(y_t | s_t) f(s_t | y_{1:t-1})}{f(y_t | y_{1:t-1})}
\]
C. Smoothing

\[ f(s_t \mid y_{1:T}) = \int f(s_t, s_{t+1} \mid y_{1:T}) ds_{t+1} \]

\[ = \int f(s_t \mid s_{t+1}, y_{1:T}) f(s_{t+1} \mid y_{1:T}) ds_{t+1} \]

\[ = \int f(s_t \mid s_{t+1}, y_{1:T}) f(s_{t+1} \mid y_{1:T}) ds_{t+1} \]

\[ = \int \left[ \frac{f(s_{t+1} \mid s_t, y_{1:T}) f(s_t \mid y_{1:T})}{f(s_{t+1} \mid y_{1:T})} \right] f(s_{t+1} \mid y_{1:T}) ds_{t+1} \]

\[ = f(s_t \mid y_{1:T}) \int f(s_{t+1} \mid s_t) \frac{f(s_{t+1} \mid y_{1:T})}{f(s_{t+1} \mid y_{1:T})} ds_{t+1} \]
Solving these integral equations depends on the structure of the problem.

Easy: Linear and normal. Analytic formulae available (e.g., Kalman filter)

Pretty Easy: Hamilton model

\[ y_t = \mu(s_t) + \sigma(s_t) \varepsilon_t \]

\[ s_t = 0 \text{ or } 1 \text{ with } P(s_t = i \mid s_{t-1} = j) = p_{ij} \]

(simple recursive formulae for likelihood and filter – exercise: work this out).

Harder: Stochastic volatility

\[ y_t = e^\mu \varepsilon_t \]

\[ s_t = \mu + \phi(s_{t-1} - \mu) + \eta_t \]

Numerical methods are used to evaluate the required integrals: Importance sampling, MCMC and Particle Filtering.
Some Examples
Standard Signal Extraction

$(y, x)$ are scalars

\[ y_t = x_t + \varepsilon_t \]

\[ x_t = \phi x_{t-1} + e_t \]

$x$: signal

$y$: measurement

$\varepsilon$: measurement error

An example with $\sigma_\varepsilon = 1$, $\sigma_e = 2$, $\phi = 0.9$ and $T = 200$ follows.
EX1.1: $y(t)$ (Grey) and $x(t)$ (Red)
EX1.2: $y(t)$ (Black), $x(t)$ (Red), $x(t/t)$ (Green)

(could also compute $x_{0|T}$)
Fig. 1. GDP and unemployment data. $GDP_E$ and $GDP_I$ are in growth rates and $U_t$ is in changes. All are measured in annualized percent.
\[ \begin{bmatrix} GDP_{t} \\ GDP_{t-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} GDP_{t-1} + \begin{bmatrix} \varepsilon_{Et} \\ \varepsilon_{It} \end{bmatrix} \]

\[ GDP_{t} = \alpha + \rho GDP_{t-1} + \varepsilon_{Gt} \]

\[ \text{var} \begin{bmatrix} \varepsilon_{g} \\ \varepsilon_{E} \\ \varepsilon_{I} \end{bmatrix} = \Sigma = \begin{bmatrix} \sigma_{GG} & 0 & 0 \\ 0 & \sigma_{EE} & \sigma_{EI} \\ 0 & \sigma_{EI} & \sigma_{II} \end{bmatrix} \]

(identification issues)
Results:

For the 2-equation model with $\Sigma$ block-diagonal, we have

$$GDP_t = \begin{bmatrix} 3.06 (1 - 0.62) + 0.62 & GDP_{t-1} + \epsilon_{Gt} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 5.17 & 0 & 0 \\ 0 & 3.86 & 1.43 \\ 0 & 1.43 & 2.70 \end{bmatrix}$$
Fig. 3. GDP sample paths, 1960Q1–2011Q4. In each panel we show the sample path of GDP\(_M\) (light color) together with posterior interquartile range with shading and we show one of the competitor series (dark color). For GDP\(_M\) we use our benchmark estimate from the 2-equation model with \(\gamma = 0.80\).
Figure 4: GDP Sample Paths, 2007Q1-2009Q4

Notes: In each panel we show the sample path of GDP in red together with a light-red posterior interquartile range, and we show one of the competitor series in black. For GDP we use our benchmark estimate from the 2-equation model with $\beta = 0.80$.

First consider Figure 5. Across measurement-error models $M$, GDP is robustly more serially correlated than both GDP and GDP, and it also has a smaller innovation variance. Hence most of our models achieve closely-matching unconditional variances, but they are composed of very different underlying ($\alpha$, $\beta$) values from those corresponding to GDP.

GDP has smaller shock volatility, but much more shock persistence – roughly double that of GDP ($\beta$ of roughly 0.60 for GDP vs. 0.30 for GDP).

Now consider Table 1. The various GDP series are all less volatile than each of GDP, GDP, and GDP, and a bit more skewed left. Most noticeably, the GDP series are much more strongly serially correlated than the GDP, GDP, and GDP series, and with smaller innovation variances. This translates into much higher predictive R^2's for GDP. Indeed GDP is twice as predictable as GDP or GDP, which in turn are twice as predictable as GDP.
**Example: Missing Data**

Scalar $y_t$ is generated by the ARMA(1,1) model

$$y_t = \phi y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$$

Observations available for $t = 1:45$, 50:90, 100:130, 140:170, 180:200

How can the likelihood be formed? How can the missing values be estimated?
State-space form of model:

\[
\begin{align*}
\xi_t &= \begin{bmatrix} y_t \\ \epsilon_t \end{bmatrix}, \\
F &= \begin{bmatrix} \phi & -\theta \\ 0 & 0 \end{bmatrix}, \\
\nu_t &= \begin{bmatrix} \epsilon_t \\ \epsilon_t \end{bmatrix}
\end{align*}
\]

\[
\text{var}(\nu_t) = Q = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}
\]

\[
y_t = H_t \xi_t + 0
\]

\[
H_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \text{ or } H_t = \begin{bmatrix} 0 & 0 \end{bmatrix}
\]

Example \( \sigma_\epsilon = 1, \phi = 0.9 \) and \( \theta = 0.5 \).
EX2.2: $y(t)$ -- with Missing observations
EX2.4: $y(t)$, $x(t)$ and $x(t/T)$
Example: **Nowcasting** (Good reference: Banbura, Giannoni, Modugno, and Reichlin (2013).)

- **Problem:** $y_t$ is a variable of interest (e.g., GDP growth rate in quarter $t$). It is available with a lag (say in $t+1$ or $t+2$). $X_t$ is a vector of variables that are measured *during* period $t$ (and perhaps earlier). How do you guess the value of $y_t$ given the $X$ data that has been revealed.

- **‘Solution’:** Suppose $X_{t_1}$ denotes the information known at time $t_1$. Then best guess of $y_t$ is $E(y_t | X_{t_1})$.
  - But how do you compute $E(y_t | X_{t_1})$?
  - How do you update the estimate as another element of $X_t$ is revealed?
Giannone, Reichlin, *et al* modeling approach:

\[
\begin{bmatrix}
y_t \\
X_{1t} \\
\vdots \\
X_{nt}
\end{bmatrix}
= 
\begin{bmatrix}
\lambda_y \\
\lambda_1 \\
\vdots \\
\lambda_n
\end{bmatrix}
F_t 
+ 
\begin{bmatrix}
e_{yt} \\
e_{1t} \\
\vdots \\
e_{nt}
\end{bmatrix}
\]

\[\Phi(L)F_t = \eta_t\]

- \(E(y_t|X_{t_1}) = \lambda_y \times E(F_t | X_{t_1})\)
- \(E(F_t | X_{t_1})\) computed by Kalman filter

(Lots of details left out)
### Data Flow (Mar 06, 2020)

<table>
<thead>
<tr>
<th>Model Update</th>
<th>Release Date</th>
<th>Data Series</th>
<th>Reference Period</th>
<th>Units</th>
<th>Actual</th>
<th>Forecast</th>
<th>Weight</th>
<th>Impact</th>
<th>Nowcast GDP Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 06</td>
<td>8:30AM Mar 06</td>
<td>Civilian unemployment rate</td>
<td>Feb</td>
<td>Ppt. chg.</td>
<td>-0.10</td>
<td>-0.05</td>
<td>-0.121</td>
<td>0.01</td>
<td>1.71</td>
</tr>
<tr>
<td></td>
<td>8:30AM Mar 06</td>
<td>Imports: Goods and services</td>
<td>Jan</td>
<td>MoM % chg.</td>
<td>-1.62</td>
<td>-0.62</td>
<td>0.064</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8:30AM Mar 06</td>
<td>Exports: Goods and services</td>
<td>Jan</td>
<td>MoM % chg.</td>
<td>-0.43</td>
<td>0.26</td>
<td>0.069</td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8:30AM Mar 06</td>
<td>All Employees: Total nonfarm</td>
<td>Feb</td>
<td>Level chg. (thousands)</td>
<td>273.00</td>
<td>171.81</td>
<td>0.326*</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 05</td>
<td>Total business inventories</td>
<td>Jan</td>
<td>MoM % chg.</td>
<td>-0.10</td>
<td>-0.05</td>
<td>-0.130</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 04</td>
<td>ISM non-mfg.: NMI composite index</td>
<td>Feb</td>
<td>Index</td>
<td>57.30</td>
<td>55.52</td>
<td>0.006</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8:10AM Mar 04</td>
<td>ADP nonfarm private payroll</td>
<td>Feb</td>
<td>Level chg. (thousands)</td>
<td>183.00</td>
<td>161.54</td>
<td>0.329*</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>employment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 02</td>
<td>ISM mfg.: Employment index</td>
<td>Feb</td>
<td>Index</td>
<td>46.90</td>
<td>52.88</td>
<td>0.027</td>
<td>-0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 02</td>
<td>Value of construction put in place</td>
<td>Jan</td>
<td>MoM % chg.</td>
<td>1.77</td>
<td>0.16</td>
<td>0.025</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 02</td>
<td>ISM mfg.: Prices index</td>
<td>Feb</td>
<td>Index</td>
<td>45.90</td>
<td>55.99</td>
<td>0.009</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10:00AM Mar 02</td>
<td>ISM mfg.: PMI composite index</td>
<td>Feb</td>
<td>Index</td>
<td>50.10</td>
<td>54.53</td>
<td>0.049</td>
<td>-0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data revisions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Foerster, Hornstein, Sarte, Watson (2020)

Figure 1: U.S. private-sector GDP growth rates 1950-2016
$n$ production sectors

\[ \left\{ z_{i,t} \right\}_{i=1}^{n} \]

→

\[
\begin{bmatrix}
\text{inter-related production functions} \\
\text{preferences} \\
\text{market structure}
\end{bmatrix}
\]

→

Final Goods
Model for $z_{i,t}$

$$\Delta z_{i,t} = \lambda_{\tau,i} \tau_{c,t} + \lambda_{\varepsilon,i} \varepsilon_{c,t} + \tau_{i,t} + \varepsilon_{i,t}$$

$\Delta \tau_{c,t}, \varepsilon_{c,t}, \{\Delta \tau_{i,t}, \varepsilon_{i,t}\}$ mutually independent white noises
Estimated $\tau$ components for Sectoral TFP growth rates