Empirical Analysis II-1.
Winter 2019

(Harald Uhlig)

Simon Sangmin Oh

University of Chicago

Note:

These lecture notes are based on the lectures by Professor Harald Uhlig in 2018-19 Winter Quarter. The topics are techniques in time-series econometrics.

Contents

1 Preliminaries

1.1 Recap of Measure Theory

You want to measure sets, so you come up with μ – a measure – that assigns values to sets, $\mu(A)$. To do this, we must write down some rules for this measure.

▷ *What kind of sets?* A is a member of a set of measurable sets, denoted by $\mathcal{A}$. In fact, we want $\mathcal{A}$ to be a σ-algebra on $X \supseteq A$, i.e. with the following properties:

* Empty set: $\emptyset \in \mathcal{A}$
* Complements: $A \in \mathcal{A} \implies X \backslash A \in \mathcal{A}$
* Countable Unions: $A_i \in \mathcal{A} \implies \cup_{i=1}^{\infty} A_i \in \mathcal{A}$
* $\mu(X) = 1$

▷ *How do we define a measure?* A measure $\mu : A \to \mathbb{R}$ must satisfy:

* Empty set: $\mu(\emptyset) = 0$
* Positive: $\mu(A) \geq 0$
* Countable Additivity: A_i disjoint $\implies \mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$

Filtrations are defined by a sequence of σ-algebras. In most cases, we call $\mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \mathcal{A}_2$ a filtration where $\mathcal{A}_t$ denotes an information set in time t.

2 Bayesian Inference

2.1 Introduction

Sufficiency Principle + Conditionality Principle ⇒ Likelihood Principle ⇒ Stopping Rule Principle

2.1.1 Likelihood Principle

There is some unknown parameter $\theta \in \Theta$ with measure $\mu(d\theta)$ that we wish to estimate using observations $\boldsymbol{x}$ with measure $\nu(dx)$ that live in measure spaces. We talk about measures because we will be doing (a lot!) of integrations later. Given a certain true parameter θ, I know the probabilistic distribution $f(\boldsymbol{x}|\theta)$. Since we are interested in θ, define the likelihood function $L(\theta|\boldsymbol{x})$.

Definition 2.1. A function ("statistic") $T(x)$ is sufficient if the distribution of x conditional on $T(x)$ does not depend on θ.

For example, $T(x) = [\bar{x}, s^2]$ is a sufficient statistic for $\theta = [x, \sigma^2]$ where $x_i \sim N(\mu, \sigma^2)$.

Proposition 2.1. *(Sufficiency Principle)* Two observations x and y lead to the same value of a sufficient statistic $T(x) = T(y)$, then they will lead to the same inference regarding θ.

Proposition 2.2. *(Conditionality Principle)* If two experiments on θ are available, and if exactly one of them is carried out with some probability p which is independent of θ, then the resulting inference on θ should only depend on the selected experiment and the resulting observation.

Informally, the conditionality principle can be taken as the claim that experiments which were not actually performed are statistically irrelevant.

Proposition 2.3. *(Likelihood Principle)* The information brought about by an observation x about θ is entirely contained in the likelihood function $L(\theta|x)$. Therefore, *if two observations x_1 and x_2 lead to proportional likelihood functions, then they shall lead to the same inference regarding θ.*

Why is likelihood so important? It comes up in two important applications:

1. Maximum Likelihood Estimation (MLE) – We covered this in the first quarter.
2. Bayesian Analysis: You start with a prior $\pi(\theta)$ which you update to get a posterior $\pi(\theta|x)$ by computing:

$$\pi\left(\theta|x\right) = \left(\frac{f(x|\theta)\pi(\theta)}{P(x)}\right) = \frac{L\left(\theta|x\right)\pi(\theta)}{\int_\Theta L\left(\theta|x\right)\pi(\theta)\mu(d\theta)}$$

which, alternatively put,

$$\pi(\theta|x) \propto L\left(\theta|x\right)\pi(\theta)$$

Note that we use the second expression way more often, usually in the log form:

$$\log \pi\left(\theta|x\right) = \log L\left(\theta|x\right) + \log \pi(\theta) - \log m\left(x\right)$$

The Bayesian and Frequentist approaches differ. Critically, the Bayesian treats the parameter θ_0 as random and x as fixed; the Frequentist treats the parameter as fixed and the observations as random.

An important consequence of the likelihood principle is the *Stopping Rule Principle:*

Proposition 2.4. *(Stopping Rule Principle)* If a sequence of experiments is directed by a stopping rule τ, which indicates when the experiments stop, then inference about θ shall depend on τ only through the resulting sample.

A stopping rule thus specifies when the researcher will stop gathering new data and commence analyzing what has been collected so far. Bayesians say that stopping rules are irrelevant to statistical inference; Frequentists take the contrary position. The Bayesian theory is inherently tied to the Likelihood Principle which entails that the stopping rule is irrelevant to the evidential force of the experimental outcome.

Example 2.1. *(The Conundrum of the Experimenter; Berger-Wolpert Ex. 19-1)* Suppose an experimenter has 100 observations $x_i \sim N(\theta, 1)$ with $\bar{x}_{100} = 0.2$ and wishes to test $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$ at the 5% level. Now consider the following two stopping rules:

1. Stopping Rule 1: Stop always. Then $\sqrt{100}(0.2) > 1.96$: reject
2. Stopping Rule 2: If $\sqrt{100} \cdot \bar{x}_{100} \geq c$, stop and reject; if not, take another 100 draws, reject if $\sqrt{200} \cdot \bar{x}_{200} \geq c$. For now, let $c = 2.18$ to take another 100 draws:
   - ▷ Suppose $1.96 < \sqrt{200} \cdot \bar{x}_{200} < 2.18$ so you don't reject. But the experimenter would have rejected if the experimenter had not "paused" half-way through.
   - ▷ Suppose $\sqrt{200} \cdot \bar{x}_{200} > 2.18$. But would the experimenter have kept going if not?

Such conundrum described above is avoided by the stopping rule principle.

Example 2.2. *(Significance Testing)* Consider the following data:

$x =$	0	1	2	3	4
$P(x\|\theta_0)$	.75	.14	.04	.037	.033
$P(x\|\theta_1)$	.70	.25	.04	.005	.005

Now you observe $x = 2$. If you do a one-sided hypothesis testing, you will find significant evidence against θ_1 at the 5% level, but not against θ_0. If you adhere to the likelihood principle, however, the evidence pro or against θ_0 is the same as pro or against θ_1. The bottom line is that many of the statistical tests we have been conducting violate the likelihood principle.

Example 2.3. *(Naive Stopping Rule)* Suppose you have observations $x_t \sim N(\theta, 1)$ with a stopping rule:

$$|\bar{x}_T| = \left| \frac{1}{T} \sum_{i=1}^{T} x_t \right| > \frac{1.96}{\sqrt{T}}$$

which means you collect data until you reject the null hypothesis. The frequentist shouldn't always reject H_0 because you need to adjust the critical value; you can't simply use the same critical value.

2.2 Admissibility and Bayes Estimators

The goal of this section is to connect the Frequentist and the Bayesian approach. Specifically, (1) Admissibility $\Leftrightarrow$ Bayesian and (2) , much to our dismay, are not always admissible.

2.2.1 Measures of Risk

We now augment the previous framework with components of decision. We now introduce a prior π with respect to $d\theta$ and a decision $\delta(x) \in \mathcal{D}$ accompanied by a loss function $\mathcal{L}(\theta, \delta(x))$. Once again, Frequentists and Bayesians diverge in their approaches.

1. *Bayesian Approach: Posterior Expected Loss:*

$$\rho(\pi, \delta(x)) = E_\pi \left[\mathcal{L}(\theta, \delta(x)) | x\right] = \int_\Theta \mathcal{L}(\theta, \delta(x)) \pi(\theta|x) d\theta$$

   where integration takes over the probability distribution of θ that is treated as random. (parameter space)

   ▷ This averages the loss according to the posterior distribution of the parameter θ conditional on the observed value x .

2. *Frequentist Approach:* Start by defining *Average Loss:*

$$\mathcal{R}(\theta, \delta) = E_\theta \left[\mathcal{L}(\theta, \delta(x))\right] = \int_X \mathcal{L}(\theta, \delta(x)) f(x|\theta) dx$$

   where integration takes place over the probability distribution of x that is treated as random. (sample space). Using this we further define *integrated risk:*

$$r(\delta) = E_\pi \left[\mathcal{R}(\theta, \delta)\right] = \int_\Theta \left(\int_X \mathcal{L}(\theta, \delta(x)) f(x|\theta) dx \right) \pi(\theta) d\theta = \int_X \rho(\pi, \delta(x)) m(x) dx$$

   where the last equality follows from Fubini's theorem.

   ▷ This is the frequentist risk averaged over the values of θ according to the prior distribution.

   ▷ This associates a real number with every esetimator, thereby incuding a total ordering on the set of estimators.

2.2.2 Frequentist ⇔ Bayesian

Definition 2.2. An estimator δ_0 is *admissible* if there exists no other estimator δ_1 which satisfies

$$\mathcal{R}(\theta, \delta_0) \geq \mathcal{R}(\theta, \delta_1)$$

and " $>$ " for at least one value of θ_0 where $\mathcal{R}$ is the risk function of the decision rule δ and the parameter θ. In other words, it is admissible (with respect to the loss function) if and only if no other rule dominates it.

Definition 2.3. A Bayes estimator associated with a prior π and a loss function $\mathcal{L}$ is any estimator δ^π which minimizes $r(\pi, \delta)$, the Bayesian posterior risk. The resulting value $r(\pi) = r(\pi, \delta^\pi)$ is called the *Bayes risk*.

Given these definitions, we argue (generally) that admissibility ⇔ Bayes estimators.

Proposition 2.5. *(Bayes Estimators are admissible)* If π is strictly positive on Θ with finite Bayes risk and the risk function $\mathcal{R}$ is a continuous function of θ for every δ, then the Bayes estimator δ^π is admissible. Furthermore, if the Bayes estimator associated with a prior π is unique, then it is admissible.

Proposition 2.6. *(Admissible estimators are Bayesian)* Suppose Θ is compact and $\mathcal{R}$ is convex. If all estimators have a continuous risk function, then, for every non-Bayes estimator δ', there is a Bayes estimator δ^π for some π, which dominates δ' i.e. the Bayes estimators constitute a complete class. Furthermore, under some mild conditions, all admissible estimators are limits of sequences of Bayes estimators.

The above proposition implies that you can produce all admissible estimators through the Bayesian approach. Frequentists should be very happy... or should they?

Proposition 2.7. *(MLEs are inadmissible)* Consider a problem where the vector θ is the unknown mean of a m-variate normally distributed random variable Y:

$$Y \sim N\left(\theta, \sigma^2 I\right)$$

with known covariance matrix. We are interested in obtaining an estimate $\hat{\theta}$ of θ based on a single observation y of Y. Furthermore, assume that measurements are corrupted by independent Gaussian noise. Since the noise has zero mean, it is very reasonable to use the measurements themselves as an estimate of the parameters. Stein (1960) showed that in terms of mean squared error, this approach is suboptimal and the James-Stein estimator always achieves lower MSE than the maximum likelihood estimator.

2.3 Exponential Families, Conjugacy, and Priors

We introduce the notion of *exponential families*. They unify many of the most important, widely-used statistical models such as the Normal, Binomial, Poisson, and Gamma into one framework. Furthermore, conjugate distributions are easy to write down.

2.3.1 Exponential Families

If we can write the log of a density in a "semi-linear" form, we call this density to be an element of the exponential family:

Definition 2.4. If there exist real-valued functions $c_1, ..., c_k, d$ of θ and real-valued functions $T_1, ..., T_k, S$ on $\mathbb{R}^n$ and a set $A \subset \mathbb{R}^n$ such that

$$f\left(x|\theta\right) = \exp\left(\sum_{i=1}^{k} c_i\left(\theta\right) T_i\left(x\right) + d\left(\theta\right) + S\left(x\right)\right) 1_A\left(x\right)$$

for all $\theta \in \Theta$, then $\{f\left(\cdot|\theta\right)|\theta \in \Theta\}$ is called a *k-parameter exponential family.* Note that the vector $T(x) = \{T_1(x), ..., T_k(x)\}$ is sufficient and is called the *natural sufficient statistic of the family.* Recall that if T is sufficient, then given the value of T, we can gain no more knowledge about θ from knowing more about the probability distribution of the xs. We can keep only T and throw away all the x_is without losing any information.

Many common probability distributions are exponential. A good example is the Normal distribution.

Example 2.4. (Normal Distribution) If we have $x \sim N\left(\mu, \sigma^2\right)$, then

$$f\left(x|\theta\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right)\right)$$

where $\theta = \left(\mu, \sigma^2\right)^T$ and

$$\begin{aligned} c_1(\theta) &= \frac{\mu}{\sigma^2}, \qquad T_1(x) = x \\ c_2(\theta) &= -\frac{1}{2\sigma^2}, \qquad T_2(x) = x^2 \\ d(\theta) &= -\frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right) \\ S(x) &= 0, \qquad A = \mathbb{R} \end{aligned}$$

Remark 2.1. Exponential families come with very nice properties:

- ▷ They unify many of the most important, widely-used statistical models such as the Normal, Binomial, Poisson, and Gamma into one framework.
- ▷ No matter how massive the data set is, there is a sufficient statistic of a fixed dimensionality. Under some regularity conditions (such as that the support does not depend on the parameter), this is only true for exponential families.
- ▷ You can easily see what the minimal sufficient statistic for the model is, and better yet it will be a complete sufficient statistic (under some regularity conditions). Usually completeness of a statistic is hard to prove, but in an exponential family you get it almost for free. This paves the way to be able to apply Basu's theorem, for example. Moreover, the complete sufficient statistic itself comes from an exponential family.
- ▷ Exponential families maximize entropy, among distributions satisfying certain natural constraints.
- ▷ Conjugate distributions are easy to write down, and the conjugate distributions come from an exponential family.
- ▷ Maximum likelihood estimation (MLE) behaves nicely in this setting, and has a very simple intuitive interpretation: set the observed value of the natural sufficient statistic equal to its expected value.

2.3.2 Conjugacy

Definition 2.5. If the prior π is a member of a parametric family of distributions so that the posterior $\pi\left(\theta|x\right)$ also belongs to that family, then this family is called *conjugate* to $\{f\left(\cdot|\theta\right)|\theta \in \Theta\}$.

Proposition 2.8. *(Conjugacy for Exponential Families)* Consider the following density

$$\pi\left(\theta; (t_1, ..., t_{k+1})\right) = \exp\left(\sum_{j=1}^{k} c_j\left(\theta\right) t_j + t_{k+1} d(\theta) - \log\omega\left(t_1, ..., t_{k+1}\right)\right)$$

which is a member of the $k+1$st parameter exponential family. This is conjugate to the exponential family

$$f\left(x|\theta\right) = \exp\left(\sum_{i=1}^{k} c_i\left(\theta\right) T_i\left(x\right) + d\left(\theta\right) + S\left(x\right)\right) 1_A\left(x\right)$$

Furthermore, the posterior is given by

$$\pi\left(\theta|x\right) = \pi\left(\theta; (t_1 + T_1(x), ..., t_k + T_k(x), t_{k+1} + 1)\right)$$

The usefulness of the proposition is best illustrated with an example.

Example 2.5. (Signal extraction in Normal Distribution) This is a setup where the variance σ^2 is known.

1. $f(x|\theta)$ is thus given by $N\left(\theta, \sigma^2\right)$, which is a member of a 1-parameter family. The conjugate family should thus have two parameters.
2. After some derivations, it can be shown that $\pi(\theta)$ is given by $N\left(\mu, \tau^2\right)$.
3. The posterior $\pi(\theta|x)$, obtained from the updating rules, is given by $N\left(\tilde{\mu}, \tilde{\tau}^2\right)$ where

$$\frac{1}{\tilde{\tau}^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$$
$$\tilde{\mu} = \frac{\sigma^{-2}}{\sigma^{-2}+\tau^{-2}}x + \frac{\tau^{-2}}{\sigma^{-2}+\tau^{-2}}\mu$$

The posterior mean is the precision-weighted average of the prior mean (μ) and the observations (x).

2.3.3 Some Distributions

1. Poisson $\mathcal{P}(\theta), \theta > 0 : E[x] = \theta$ with
$$f(x|\theta) = e^{-\theta}\frac{\theta^x}{x!}1_{\mathbb{N}}(x)$$
2. Gamma $\mathcal{G}(\alpha, \beta) : E[x] = \alpha/\beta$ with
$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)\,1_{[0,\infty)}(x)$$
Note that χ^2 is a member of this family.
3. Beta $Beta(\alpha, \beta), \alpha > 0, \beta > 0$: $E[x] = \alpha/(\alpha+\beta)$ with
$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}1_{[0,1]}(x)$$

These distributions are useful because it makes Bayesian updating very convenient:

$f(x\|\theta)$	π	$\pi(\theta\|x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $Be(\alpha, \beta)$	Beta $Be(\alpha + x, \beta + n - x)$
Normal $N(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}\left(\alpha + 0.5, \beta + (\mu - x)^2/2\right)$
Gamma $\mathcal{G}(\nu/2, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu/2, \beta + x)$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$

Definition 2.6. Jeffreys Prior is a prior proportional to square root of determinant of information matrix:

$$\pi^*(\theta) \propto \det(\mathcal{I}(\theta))^{1/2}$$

Note that Jeffreys prior is flat if $f(x|\theta)$ is $N\left(\theta, \sigma^2\right)$. It's very useful because it is invariant to reparametrizations.

Example 2.6. (Exponential as Gamma) Consider the exponential density $f(x|\theta) = \theta\exp(-\theta x) = \mathcal{G}(1, \theta)$ and suppose are given the conjugate prior $\pi(\theta) = \exp(-\theta) = \mathcal{G}(1, 1)$. Then the posterior is

$$f(x|\theta)\,\pi(\theta) = \mathcal{G}(1+1, 1+x) = \mathcal{G}(2+x)$$

2.4 TA Session (2019.01.11)

2.4.1 Sufficient Statistics

Example 2.7. (Sufficient statistics for normal distribution) We have a density $f_X(x)$ such that we can factorize it into

$$f_X(x) = \exp\left(g(\theta)K(x) + s(\theta) + h(x)\right)$$

then $K(x)$ is your sufficient statistic. Once you pin down the value of $K(x)$ then you don't have to care about the part where θ and x interact. For example, the sufficient statistics for

$$x_1, ..., x_N \sim N(\mu, \sigma^2)$$
$$f_X\left(x_1, ..., x_N\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{\left(x_i - \mu\right)^2}{\sigma^2}\right) = \left(2\pi\sigma^2\right) \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{\sigma^2} + 2\sum_{i=1}^{n} \frac{\mu x_i}{\sigma^2} - \frac{n\mu^2}{\sigma^2}\right)$$

is $T(X) = \left(\sum X_i, \sum X_i^2\right)$.

To see this, note that

$$f_{X|T}\left(x|t\right) = \frac{f_{X,T}\left(x,t\right)}{f_T(t)} = \frac{f_X(x) \cdot 1\left\{t, x \text{ are compatible}\right\}}{f_T(t)}$$

Example 2.8. (Sufficient statistics for Poisson) Given $x_1, ..., x_N \sim Poisson(\lambda)$, we have

$$f_X(x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum x_i} \prod_{i=1}^{n} \frac{1}{x_i!}$$

In this case, $\sum x_i$ is the sufficient statistic. To see this more explicitly,

$$f_{X,T}(x,t) = f_X(x) \cdot 1\left\{t, x \text{ are compatible}\right\}$$

Furthermore, $T = \sum x_i \sim Poisson(n\lambda)$ so the marginal density should be

$$f_T(t) = e^{-n\lambda}(n\lambda)^t \frac{1}{t!}$$

and the conditional density is then obtained as

$$f_{X|T}(x|t) = \frac{f_{X,T}(x,t)}{f_T(t)} = \frac{1}{n^t} t! \prod_{i=1}^{n} \frac{1}{x_i!}$$

which is $Multinomial\left(\frac{1}{n}, ..., \frac{1}{n}, N\right)$

2.4.2 Conjugate Prior

Example 2.9. Prior $\theta \sim Beta\left(\alpha, \beta\right)$ and conditional $x|\theta \sim Binomial\left(p, \theta\right)$ yields posterior

$$\theta|x \sim Beta\left(\alpha + x, \beta + n - x\right)$$

where the density of a Beta distribution is given as

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} 1\left\{0 \le \theta \le 1\right\}$$

and the density of a Binomial distribution is given as

$$\frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}1\{x=0,1,...,n\}$$

Then

$$f_{\Theta|X}(\theta|x)=\frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{f_X(x)}$$

Note that $P(x=1)=P(x=0)=1/2$, so we have

$$f(\theta|x)=2\theta^x(1-\theta)^{1-x}1\{0\leq\theta\leq 1\}\sim Beta(1+x,2-x)$$

3 Numerical Methods for Bayesian Inference

Conjugate distributions are nice, but how should we deal with non-conjugate distributions? This section provides numerical methods as an answer to this question. Specifically, they are used to produce samples from a given dsitribution to (1) get an idea about the distribution, or (2) solve an integration /optimization problem related with f.

Usually, the object of interest is

$$E\left[g(\theta)\right] = \int_{\Theta} g(\theta)\pi\left(\theta\right)\lambda\left(d\theta\right) \qquad \text{[A]}$$

where π denotes the posterior probability.

3.1 Preliminaries

We must restrict our choice of T, the transition operator, to have the following properties:

1. (Ergodicity) For $m \to \infty$, the distribution $p\left(x^{(m)}|x^{(0)}\right)$ converges to the distribution $p^*(x)$ regardless of choice of $p\left(x^{(0)}\right)$. This is a way to ensure that T avoids traps and can visit everywhere in our state space.
2. (Detailed Balance) A sufficient condition to ensure that T has the equilibrium distribution π^* we want.
3. (Strong LLN) Averaging over simulations of samples from a Markov Chain with T and π^* average nicely.

3.1.1 Markov Chains

A Markov chain can be specified by the initial distribution and a stochastic transition function.

Ergodicity Ergodic implies states "mix over time." The following are examples of transition matrices that are *not ergodic*:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

whereas the following matrix is ergodic:

$$\begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

Formally, ergodic states are those that are *aperiodic* and *positive recurrent.*

- ▷ *Aperiodicity*: A state has period k if any return to k must occur in multiples of k steps.
- ▷ *Positive recurrence*: A state has finite expected return time.

If we define $T : S \to S$ to be ergodic as $T^{-1}(A) = A \Leftrightarrow A = \emptyset, A = S$ then $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is ergodic but periodic. In this case T^2 is not ergodic.

Detailed Balance We want to discipline our choice of the kernel $k\left(\theta'|\theta\right) \equiv P\left(\theta \to \theta'\right)$. Specifically, we want the transition kernel density to satisfy

$$k\left(\theta'|\theta\right)\pi\left(\theta\right) = k\left(\theta|\theta'\right)\pi\left(\theta'\right)$$

- ▷ Intuition: there is no net flux of probability between edge θ and θ' during one time step, provided that the chain is in the stationary distribution.

- ▷ These kernels in general do not satisfy the balance condition – you have to work hard to achieve this!
- ▷ The balance condition is *stronger* than that π is a stationary distribution.

Markov chains that satisfy this balance condition are called *reversible.* This is because the chain is statistically indistinguishable whether it is run forward or backward in time. The whole point of checking this detailed balance condition is that it is a sufficient condition for T having an equilibrium distribution.

Example 3.1. (2010 Final Exam) Suppose a posterior π for $\theta \in \mathbb{R}$ happens to be the normal distribution $N(0,1)$. A researcher seeks to calculate $E[\theta]$ for this posterior using the MCMC algorithm per

$$\bar{\theta}_n = \frac{1}{n}\sum_{m=1}^{n} \theta^{(m)}$$

where $\theta^{(m)}$ is a sequence of draws. She uses the following algorithm to generate these draws. She draws $\theta^{(1)}$ from π and proceeds recursively. Given the draws $\theta^{(m)}$, she draws $\xi \sim N(0,1)$ independently across m. She then sets:

$$\theta^{(m+1)} = \begin{cases} \xi & \text{with probability 0.2} \\ \theta^{(m)} & \text{with probability 0.8} \end{cases}$$

We are interested in the asymptotic distribution of $\bar{\theta}_n$. To derive this, the Markov Chain CLT states that

$$\sqrt{n}\left(\bar{\theta}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

where

$$\sigma^2 = Var\left(\theta^{(m)}\right) + 2\sum_{k=1}^{\infty} Cov\left(\theta^{(1)}, \theta^{(1+k)}\right)$$

In this example, we immediately see that $\mu = 0$. The tricky part is the terms to compute σ:

$$Var\left(\theta^{(m)}\right) = E\left[\left(\theta^{(m)}\right)^2\right] = 0.2E\left[\xi^2\right] + 0.8E\left[\left(\theta^{(m)}\right)^2\right]$$

$$\Rightarrow 0.2E\left[\left(\theta^{(m)}\right)^2\right] = 0.2$$

$$\Rightarrow E\left[\left(\theta^{(m)}\right)^2\right] = 1$$

Furthermore, the general covariance term is

$$Cov\left(\theta^{(1)}, \theta^{(1+1)}\right) = E\left[\theta^{(1)}\theta^{(1+1)}\right] = 0.8$$
$$Cov\left(\theta^{(1)}, \theta^{(1+2)}\right) = E\left[\theta^{(1)}\theta^{(1+2)}\right] = 0.8E\left[\theta^{(1)}\theta^{(2)}\right] = 0.8^2$$

Extending this process further, we see that

$$2\sum_{k=1}^{\infty} Cov\left(\theta^{(1)}, \theta^{(1+k)}\right) = 2\left(0.8 + 0.8^2 + 0.8^3 + \cdots\right) = 2\frac{1}{1-0.8} = 10$$

Thus, $\sigma^2 = 1 + 10 = 11$ and the asymptotic distribution is then

$$\sqrt{n}\left(\bar{\theta}_n\right) \xrightarrow{d} N(0, 11)$$

3.1.2 Stylized Facts

Here are some stylized facts on Bayesian inference that may seem counter-intuitive at first:

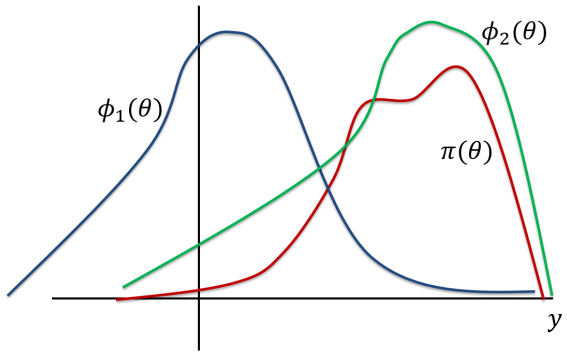
1. It turns out that we (= Matlab) only know how to directly sample from very few distributions such as uniform, Gaussian etc.
2. It's difficult to know when chains have "burned-in" so in reality, you don't care too much.

3.2 Plain Monte-Carlo (MC)

Monte Carlo is a stochastic way to approximate integrals. While algorithms usually evaluate the integrand at a regular grid, Monte Carlo randomly chooses points at which the integrand is evaluated.

3.2.1 MC #1: Importance Sampling

Importance sampling is *not* a method fo sample directly from π. Specifically, this refers to sampling from a distribution that overweights the *important* region, and having oversampled this region we have to adjust our estimate somehow to account for having sampled from this distribution. Graphically:



1. Choose a convenient approximating density $\phi(\theta)\lambda(d\theta)$. In the graph, ϕ_2 is a better choice than ϕ_1.
2. Take i.i.d. samples $\theta^{(j)}, j = 1, ..., n$ from $\phi(\theta)$.
3. Compute the weights $\omega_j = \pi\left(\theta^{(j)}\right)/\phi\left(\theta^{(j)}\right)$
4. Evaluate [A] empirically using

$$\bar{g}_n = \frac{\sum_{j=1}^{n} \omega_j g\left(\theta^{(j)}\right)}{\sum_{j=1}^{n} \omega_j}$$

Unfortunately, importance sampling does not work well in high dimensions.

3.2.2 MC #2: Rejection Sampling

Now we actually generate samples from π instead of some approximating density ϕ. Again, assume we know ϕ only and a constant c such that

$$c\phi(\theta) \geq p(\theta), \forall x$$

Then rejection sampling generates a sample in two steps:

1. Take a sample $\theta^{(j)}$ from $\phi(\theta)$.
2. Sample a from Uniform$\left[0, c\phi\left(\theta^{(j)}\right)\right]$. If $a \leq \phi\left(\theta^{(j)}\right)$, then $\theta^{(j)}$ is accepted and becomes a sample. Otherwise, θ is rejected and no sample is generated.

Note that there is high efficiency gain if algorithm accepts with high probability.

3.3 Markov-Chain Monte-Carlo (MCMC)

MCMC works by constructing and simulating a Markov chain whose equilibrium distribution is the distribution of interest. It create samples from a possibly multi-dimensional continuous random variable, with probability density proportional to a known function. These samples can be used to evaluate an integral over that variable, as its expected value or variance.

3.3.1 Algorithm

The key insight of MCMC is that these samples need not be i.i.d.! Instead, we want to find a Markov sequence $\theta^{(j)}$ with ergodic distribution $\pi(\theta)$:

1. Find a Markov sequence $\theta^{(j)}$ with ergodic distribution $\pi(\theta)$.
2. Evaluate [A] empirically using

$$\bar{g}_n = \frac{1}{n}\sum_{j=1}^{n} g\left(\theta^{(j)}\right)$$

3.4 MCMC #1: Metropolis-Hastings Algorithm

In M-H algorithm, we require a proposal distribution $q(y|x)$. Unlike importance or rejection sampling, however, q can be quite different from p. Like rejection sampling, M-H is a two-step procedure; unlike rejection sampling, the algorithm *always* generates a sample and needs the previously generated sample in the process.

3.4.1 Algorithm

Suppose that a Markov chain is in position x. Denote $\pi(\theta)$ as the target distribution and choose a proposal distribution $q(\theta'|\theta)$. The M-H algorithm is as follows:

1. Start from any θ_0.
2. Given $\theta^{(m)}$, generate $\xi \sim q\left(\xi|\theta^{(m)}\right)$
3. Compute the ratio

$$\rho\left(\xi|\theta^{(m)}\right) = \frac{\pi(\xi)q\left(\theta^{(m)}|\xi\right)}{\pi(\theta^{(m)})q\left(\xi|\theta^{(m)}\right)}$$

4. Accept the proposed move with probability $\alpha = \min\{1, r\}$; otherwise, remain at $\theta^{(m)}$ i.e. $\theta^{(m+1)} = \theta^{(m)}$.

3.4.2 Implementation Details

1. The proposal distribution q can be absolutely anything.

   ▷ This gives a lot of freedom to choose proposal dynamics based on what is convenient for the problem at hand. Choosing a good q, however, is an art.

2. You don't actually need to know π but rather only a function $g \propto \pi$ – you don't need to know the normalization constant.

3. The chain may take some time from its starting point before it actually reaches the stationary distribution. The length of the initial transient data is sometimes referred to as the *burn-in* time.

4. A popular proposal distribution is random walk: $\xi = \theta^{(m)} + \epsilon$ where ϵ has a symmetric distribution around zero. Since it's symmetric, then the ratio simplifies to

$$\rho\left(\xi|\theta^{(m)}\right) = \min\left\{1, \frac{\pi\left(\xi\right)}{\pi\left(\theta^{(m)}\right)}\right\}$$

3.5 MCMC #2: Gibbs Sampling

The basic idea is to split the multidimensional θ into blocks and sample each block separately, conditional on the most recent values of other blocks. The beauty of Gibbs sampling is that it simplifies a complex high-dimensional problem by breaking it down into simple, low-dimensional problems. Note that Gibbs sampling is a special case of the M-H algorithm where every candidate is accepted with probability 1.

3.5.1 Density Splitting

Gibss sampling is useful for the case when $\theta = (\theta_1, ..., \theta_r)$ such that the conditionals

$$\pi_j\left(\theta_j|\theta_i, i \neq j\right), j = 1, ..., r$$

are easy to draw from. If the conditional density for θ_j is not easy to draw from, one my instead draw by taking a single Metropolis-Hastings step with that conditional density as target distribution.

3.5.2 Algorithm

Let $\theta^{(t)}$ consist of $\theta_1^{(t)}, ..., \theta_k^{(t)}$ at iteration (t).

1. Draw $\theta_1^{(t+1)}$ from $p\left(\theta_1|\theta_2^{(t)}, ..., \theta_k^{(t)}\right)$. Then draw $\theta_2^{(t+1)}$ from $p\left(\theta_2|\theta_1^{(t+1)}, ..., \theta_k^{(t)}\right)$. And so forth.

   ▷ This completes one iteration of the Gibbs sampler, thereby producing one draw $\theta^{(t+1)}$. Above process is then repeated many times.

3.6 Dynare (not covered in lecture)

Skipped.

Dynare is a set of Matlab codes that are used to solve and simulate DSGE models. It can also be used to estimate parameters of DSGE models via Bayesian Maximum Likelihood.

4 Time-Series: A Bayesian Approach

4.1 Kalman Filter

Kalman Filtering process seeks to discover an underlying set of state variables $\{\xi_t\}$ given a set of measurements $\{Y_t\}$ where the process and measurement equations are both *linear*:

$$Y_t = H_t\xi_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_t)$$
$$\xi_{t+1} = F_{t+1}\xi_t + \eta_{t+1}, \quad \eta_{t+1} \sim N(0, \phi_{t+1})$$

4.1.1 Preliminaries

We introduce two useful lemmas that are used in the algorithm.

Lemma 4.1. *(Conditional Normal Densities #1)* Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(0, \begin{bmatrix} S_{XX'} & S_{XY'} \\ S_{YX'} & S_{YY'} \end{bmatrix}\right)$$

Then the conditional distribution is given as

$$X|Y \sim N\left(AY, S_{XX'|Y}\right)$$

where $A = S_{XY'}S_{YY'}^{-1}$ which is simply the β from a linear regression, and $S_{XX'|Y} = S_{XX'} - S_{XY'}S_{YY'}^{-1}S_{YX'} = S_{XX'} - AS_{YY'}A'$.

Example 4.1. (2009 Final Exam) Suppose that $y, \xi_1, \xi_2 \in \mathbb{R}$. Given ξ_1, ξ_2, suppose that $y = \xi_1 + \xi_2 + \epsilon, \epsilon \sim N(0, 2)$ and that the prior for $[\xi_1, \xi_2]'$ is given by

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

1. Compute the posterior distribution upon observing y: Note that we can rewrite the distribution as

$$\begin{bmatrix} \xi_1 \\ \xi_2 \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & E[\xi_1 y] \\ 0 & 1 & E[\xi_2 y] \\ E[\xi_1 y] & E[\xi_2 y] & 4 \end{bmatrix}\right) = \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{bmatrix}\right)$$

   And thus the A matrix is simply $\begin{bmatrix} 1/4 & 1/4 \end{bmatrix}'$,which gives us the posterior

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} | y \sim N\left(\begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix} y, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 4\begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix}\begin{bmatrix} 1/4 & 1/4 \end{bmatrix}\right) = N\left(\begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix} y, \begin{bmatrix} 3/4 & -1/4 \\ -1/4 & 3/4 \end{bmatrix}\right)$$

2. Define $d = \xi_1 - \xi_2$ and calculate its prior and posterior distribution: The posterior mean is simply 0 from the distribution above. The posterior variance of d is

$$E\left[\xi_1^2 - 2\xi_1\xi_2 + \xi_2^2 | y\right] = 3/4 - 2 \cdot (-1/4) + 3/4 = 2$$

   The prior mean and the variance are also identical, so both are distributed $N(0, 2)$. This is not surprising, since y is informative only about the orthogonal linear combination $[1, 1]\,\xi$.

Lemma 4.2. *(Conditional Normal Densities #2)* Let

$$Y|H,\xi \sim N\left(H\xi,\Sigma\right), \qquad \xi|H \sim N\left(\hat{\xi},\Omega\right)$$

Then the conditional distribution is given as

$$\left[\begin{array}{c} \xi \\ Y \end{array}\right]|H \sim N\left(\left[\begin{array}{c} \hat{\xi} \\ H\hat{\xi} \end{array}\right],\left[\begin{array}{cc} S_{\xi\xi'} & S_{\xi Y'} \\ S_{Y\xi'} & S_{YY'} \end{array}\right]\right) = N\left(\left[\begin{array}{c} \hat{\xi} \\ H\hat{\xi} \end{array}\right],\left[\begin{array}{cc} \Omega & \Omega H' \\ \Omega H' & H\Omega H'+\Sigma \end{array}\right]\right)$$

Furthermore,

$$\xi|Y,H \sim N\left(\hat{\xi}+G\hat{\epsilon},\Omega - GS_{YY'}G'\right)$$

where $G = S_{\xi Y'}S_{YY'}^{-1}$ and $\hat{\epsilon} = Y - H\hat{\xi}$.

4.1.2 Intuition

Given the horrendous math underlying Kalman Filtering, we first introduce intuition for how it works. Here are the key points:

- ▷ Three different problems should be distinguished:
  - * Prediction: Estimate ξ_t in terms of $y_1, ..., y_{t-1}$
  - * Filtering: Estimate ξ_t in terms of $y_1, ..., y_t$. This is real-time, given data so far.
  - * Smoothing: Estimate ξ_t in terms of all observations $y_1, ..., y_T$. This is post-processing, given all data.
- ▷ To determine the best estimate, Kalman filtering uses the MSE criterion, under which the best estimate is the conditional expectation. Since the expectation is generally non-linear and difficult to find, we confine ourselves to linear filters.
- ▷ Kalman gain tells you how much I want to change my estimate given a measurement.
- ▷ Kalman filter's useful for learning in macroeconomics.
- ▷ The Kalman filter is a uni-modal, recursive estimator. Only the (1) estimated state from the previous time step and (2) current measurement is required to make a prediction for the current state.

4.1.3 Algorithm

To recap, the relevant equations are:

$$\begin{aligned} \text{[Observation]}: & Y_t = H_t\xi_t + \epsilon_t, \quad \epsilon_t \sim N\left(0,\Sigma_t\right) \\ \text{[State]}: & \xi_{t+1} = F_{t+1}\xi_t + \eta_{t+1}, \quad \eta_{t+1} \sim N\left(0,\phi_{t+1}\right) \end{aligned}$$

- ▷ Notation:
  - * $\{\xi_t\}$ is the state process (unobservable) and $\{Y_t\}$ is the measurement process (observable).
  - * $\{\phi_t\}$ is the covariance matrix for the states
  - * $\{\Omega_t\}$ represents the a posteriori error covariance matrix, a measure of the estimated accuracy of the state estimate.

▷ F_{t+1} is the prediction matrix, which gives us the next state and the updated covariance matrix.

The algorithm is summarized as the following. Given a fuzzy idea about the current location, $\xi_t \sim N\left(\hat{\xi}_{t|t-1}, \Omega_{t|t-1}\right)$:

1. Forecast Y_t:
$$Y_t \sim N\left(\hat{Y}_t, S_{YY'|t}\right) = N\left(H_t\hat{\xi}_{t|t-1}, \Sigma_t + H_t\Omega_{t|t-1}H_t'\right)$$
which is also just a straightforward result from the observation state equation.

2. Update inference for ξ_t after observing Y_t: (= *a posteriori* state estimate):
$$\xi_t \sim N\left(\hat{\xi}_{t|t}, \Omega_{t|t}\right) = N\left(\hat{\xi}_{t|t-1} + G_t\hat{\epsilon}_t, \Omega_{t|t-1} - S_{\xi Y'|t}S_{YY'|t}^{-1}S_{Y\xi'|t}\right)$$
so compute
$$\hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} + \left(H_t\Omega_{t|t-1}\left(\Sigma_t + H_t\Omega_{t|t-1}H_t'\right)^{-1}\right)\hat{\epsilon}_t$$
$$\Omega_{t|t} = \Omega_{t|t-1} - H_t\Omega_{t|t-1}\left(\Sigma_t + H_t\Omega_{t|t-1}H_t'\right)^{-1}\Omega_{t|t-1}H_t'$$
Notice that
$$S_{Y\xi'|t} = \Omega_{t|t-1}H_t' = S_{\xi Y'|t}'$$
$$S_{YY'|t} = \Sigma_t + H_t\Omega_{t|t-1}H_t'$$
$$\hat{\epsilon}_t = Y_t - \hat{Y}_t$$

3. Forecast ξ_{t+1} at date t (= *a priori* state estimate):
$$\xi_t \sim N\left(\hat{\xi}_{t+1|t}, \Omega_{t+1|t}\right) = N\left(F_{t+1}\hat{\xi}_{t|t}, F_{t+1}\Omega_{t|t}F_{t+1}' + \phi_{t+1}\right)$$
which is also just a straightforward result from our state equation.

4.1.4 Initialization

To initialize the Kalman Filter, we have the following options:

1. Known starting point.
2. Flat prior (nearly flat: $\hat{\xi}_{1|0} = 0, \Omega_{1|0} = \omega I_r$ with ω very large)
3. Stationary distribution

We skip detailed expositions for now.

Example 4.2. (2010 Final Exam) Consider the $MA(1)$
$$y_t = \epsilon_t + 3\epsilon_{t-1}, \epsilon_{t-1} \sim N(0,1)$$
and denote u_t as the one-step ahead prediction error.

1. Wold decomposition for y_t: writing the provided equation as $y_t = (1+3L)\,\epsilon_t$ and flipping the roots, we have
$$y_t = (1+3L)\,\epsilon_t = \left(1 + \frac{1}{3}L\right)u_t, \qquad Var[u_t] = 9$$

2. Representation as Kalman Filtering: Suppose data $y_t, t = 1, ..., T$ is given. Treat $\xi_t = [\epsilon_t, \epsilon_{t-1}]'$ as an unobserved state with the initial prior $\xi_1 \sim N\left(\hat{\xi}_{1|0}, \Omega_{1|0}\right)$ where

$$\hat{\xi}_{1|0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Omega_{1|0} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}, \sigma_1^2 > 0$$

Then the state equation and the observation equations are:

$$y_t = \underbrace{\begin{bmatrix} 1 & 3 \end{bmatrix}}_{=H} \begin{bmatrix} \epsilon_t \\ \epsilon_{t-1} \end{bmatrix}$$

$$\begin{bmatrix} \epsilon_{t+1} \\ \epsilon_t \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}}_{=F} \begin{bmatrix} \epsilon_t \\ \epsilon_{t-1} \end{bmatrix} + \underbrace{\begin{bmatrix} \epsilon_{t+1} \\ 0 \end{bmatrix}}_{=\eta_{t+1}}, \eta_{t+1} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}}_{=\Phi} \right)$$

3. Computing $\Omega_{t+1|t}$ for all t. We will assume that $\Omega_{t|t-1}$ has the same form as $\Omega_{1|0}$ and calculate the recursion for $\sigma_t^2 = \left[\Omega_{t|t-1}\right]_{22}$.

   (a) Step 1: Forecast y_1

$$y_1 \sim N\left(\begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma + \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right) = N\left(0, 1 + 9\sigma_1^2\right)$$

   (b) Step 2: Update inference for $\xi_{1|1}$ given y_1, H: Since $\hat{\xi}_{1|1} = \hat{\xi}_{1|0} + G\epsilon_1$, we have:

$$\hat{\xi}_{1|1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \left(\begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right)^{-1} (y_1 - 0)$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{y_1}{1 + 9\sigma_1^2} \begin{bmatrix} 1 \\ 3\sigma_1^2 \end{bmatrix} = \begin{bmatrix} 1/(1 + 9\sigma_1^2) \\ 3\sigma_1^2/\left(1 + 9\sigma_1^2\right) \end{bmatrix} y_1$$

   and

$$\hat{\Omega}_{1|1} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} - \begin{bmatrix} 1 \\ 3\sigma_1^2 \end{bmatrix} \frac{1}{1 + 9\sigma_1^2} \begin{bmatrix} 1 & 3\sigma_1^2 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} - \frac{1}{1 + 9\sigma_1^2} \begin{bmatrix} 1 & 3\sigma_1^2 \\ 3\sigma_1^2 & 9\sigma_1^4 \end{bmatrix}$$
$$= \frac{1}{1 + 9\sigma_1^2} \begin{bmatrix} 9\sigma_1^2 & -3\sigma_1^2 \\ -3\sigma_1^2 & 1 \end{bmatrix}$$

   and thus

$$\xi_{1|1}|y_1 \sim N\left(\begin{bmatrix} 1/(1 + 9\sigma_1^2) \\ 3\sigma_1^2/\left(1 + 9\sigma_1^2\right) \end{bmatrix} y_1, \frac{1}{1 + 9\sigma_1^2} \begin{bmatrix} 9\sigma_1^2 & -3\sigma_1^2 \\ -3\sigma_1^2 & 1 \end{bmatrix} \right)$$

   (c) Step 3: Forecast $\xi_{2|1}$: This is straightforward

$$\hat{\xi}_{2|1} = F\hat{\xi}_{1|1} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1/(1 + 9\sigma_1^2) \\ 3\sigma_1^2/\left(1 + 9\sigma_1^2\right) \end{bmatrix} y_1 = \begin{bmatrix} 0 \\ 1/\left(1 + 9\sigma_1^2\right) \end{bmatrix} y_1$$

and

$$\begin{aligned}\hat{\Omega}_{2|1} &= \Phi + F\hat{\Omega}_{1|1}F' \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{1+9\sigma_1^2}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 9\sigma_1^2 & -3\sigma_1^2 \\ -3\sigma_1^2 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 9\sigma_1^2/(1+9\sigma_1^2) \end{bmatrix}\end{aligned}$$

(d) Step 4: Deduce the form for $\Omega_{t+1|t}$: Thus, we guess that

$$\Omega_{t+1|t} = \begin{bmatrix} 1 & 0 \\ 0 & 9\sigma_t^2/(1+9\sigma_t^2) \end{bmatrix}$$

4. Focusing on σ_t^2, note that

$$\sigma_t^2 = \frac{9\sigma_{t-1}^2}{1+9\sigma_{t-1}^2} = 1 - \frac{1}{1+9\sigma_{t-1}^2}$$

As $T \to \infty$, we then have

$$\sigma^2\left(1+9\sigma^2\right) = 9\sigma^2 \Rightarrow \sigma^2 = \frac{8}{9}$$

assuming things don't blow up.

5. Now instead, consider the Kalman Smoother. Let $\sigma_{1|T}^2 = \left(\Omega_{1|T}\right)_{22}$ be the entry of the diagonal of $\Omega_{1|T}$ for ϵ_0. Since we are using all information, we will be able to obtain the true variance of ϵ_0 which is equal to 1.

4.2 Bayesian Vector Autoregressions (BVAR) (not covered in lecture)

BVAR uses Bayesian methods to estimate a vector autoregression (VAR). The difference with standard VAR models is the fact that the model parameters are treated as random variables, and prior probabilities are assigned to them.

4.2.1 Issues with Plain Vector Autoregressions (VAR)

Skipped.

4.2.2 BVAR per Kalman Filtering

Skipped.

4.2.3 BVAR per Normal-Wishart Distributions

Skipped.

5 Univariate Time-Series: Non-Bayesian Approach

5.1 Lag Operator Calculus

This section covers the most basic tools in univariate time-series. Everyone should know this.

5.1.1 AR, MA, and ARMA

First, $AR(1)$ is defined as $y_t = \rho y_{t-1} + \epsilon_t$ or $(1 - \rho(L))\, y_t = \epsilon_t$ using the Lag operator. Furthermore, $AR(m)$ is defined as

$$y_t = \sum_{j=1}^{m} \rho_j y_{t-j} + \epsilon_t$$

$AR(m)$ can be re-written as $VAR(1)$ with $x_t = Bx_{t-1} + A\epsilon_t$ where

$$x_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-m+1} \end{bmatrix} = \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_{m-1} & \rho_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} x_{t-1} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \epsilon_t$$

Second, $MA(1)$ is defined as $y_t = \theta_0 \epsilon_t + \theta_1 \epsilon_{t-1}$. Furthermore, $MA(n)$ is defined as

$$y_t = \sum_{j=0}^{n} \theta_j \epsilon_{t-j}$$

Finally, $ARMA$ combines AR and MA to yield:

$$y_t - \sum_{j=1}^{m} \rho_j y_{t-j} = \sum_{j=0}^{n} \theta_j \epsilon_{t-j}$$

Uhlig's Comment: ARMAs are pretty messy and we won't be covering it too much in this class.

5.1.2 Autocovariances

The kth autocovariance $\Gamma_k = E\left[y_t y_{t-k}'\right]$ is defined as the covariance between y_t and y_{t-k}. If the series is univariate, we use γ_k to denote the scalar-valued quantity. We also assume that these variables are demeaned. A stochastic sequence is called *covariance stationary* if the mean and autocovariances of y_t are finite and do not depend on t.

- ▷ For $AR(1)$, we use this to derive, for $|\rho| < 1$: $\gamma_0 = \sigma^2/(1-\rho^2)$ which is just the variance and $\gamma_k = \rho^k \sigma^2 / \left(1 - \rho^2\right)$ for k time units apart.
- ▷ For $AR(m)$, use stacked $VAR(1)$ instead: $x_t = Bx_{t-1} + A\epsilon_t$, $E\left[\epsilon_t \epsilon_t'\right] = \Omega$. Note that x_t's are column vectors.
  - * First, observe the Yule-Walker Equation: $\Gamma_k = B\Gamma_{k-1}$.
  - * Using this equation and the fact that x_t and ϵ_t are uncorrealted, we can compute:

$$\begin{aligned} \Gamma_0 = E\left[x_t x_t'\right] &= BE\left[x_{t-1} x_{t-1}'\right] B' + AE\left[\epsilon_t \epsilon_t'\right] A' \\ \Rightarrow \Gamma_0 &= B\Gamma_0 B + A\Omega A' \end{aligned}$$

* To facilitate computation, we employ the vec notation:

$$vec\left(\Gamma_0\right) = \left(B \otimes B\right) vec\left(\Gamma_0\right) + vec\left(A\Omega A'\right)$$

* Note the following useful results:
  - $vec\left(C_{p\times q}\right)$ is a stacked *vector* obtained by stacking the q columns of the C matrix. $\otimes$ is the Kronecker Product, which is the outer product of matrices.
  - The result we use here is the following: $vec\left(DEF\right) = \left(F' \otimes D\right) vec\left(E\right)$ where D, E, F are any matrices that can be multiplied in the following form.
* If B has only eigenvalues smaller than unity in absolute value, then we have a further decomposition that

$$vec\left(\Gamma_0\right) = \left(I_{m^2} - B \otimes B\right)^{-1} vec\left(A\Omega A'\right)$$

▷ For $MA(n)$ with $E\left[\epsilon_t, \epsilon_t'\right] = \Omega$: Note that y_t and y_{t-k} can be re-written as

$$y_t = \sum_{j=0}^{n} \theta_j \epsilon_{t-j}, \qquad y_{t-k} = \sum_{j=k}^{n+k} \theta_{j-k} \epsilon_{t-j}$$

and thus the covariance would be computed as

$$\Gamma_k = \sum_{j=\max\{0,k\}}^{\min\{n,n+k\}} \theta_j \Omega \theta_{j-k}'$$

This is intuitive since non-overlapping shocks are uncorrelated with each other.

* For example for MA(2), we saw that $\Gamma_0 = \theta_0^2 + \theta_1^2$ and $\Gamma_1 = \theta_1\theta_0$.

Note that $AR(1)$ is covariance stationary if and only if $|\rho| < 1$. The maintained assumption is that $\{\epsilon_t\}$ is a martingale difference sequence with constant variance, i.e. its expectation with respect to its past is zero.

Remark 5.1. *(BDS Test for Non-linear Time Series) The BDS test is a test developed by Brock, Dechert, and Scheinkman (1987) originally designed to test for the null hypothesis of independent and identical distribution (IID) for the purpose of detecting non-random chaotic dynamics. In particular, when applied to residuals from a fitted linear time series model, the BDS test can be used to detect remaining dependence and the presence of omitted non-linear structure.*

5.1.3 Characteristic Polynomial

Recall that for $AR(m)$, we had

$$y_t = \sum_{j=1}^{m} \rho_j y_{t-j} + \epsilon_t$$

or alternately, the following equation: $\left(1 - \rho\left(L\right)\right) y_t = \epsilon_t$. Denote $\tilde{p}(L) = 1 - \rho(L)$. Then the characteristic polynomial is defined as

$$p\left(\lambda\right) = \lambda^m \tilde{p}\left(\lambda^{-1}\right) = \lambda^m - \rho_1\lambda^{m-1} - \rho_2\lambda^{m-2} - \cdots - \rho_m$$

where $\lambda \in \mathbb{C}$. The solution to $p\left(\lambda\right) = 0$ are called the *roots* of the characteristic polynomial.

▷ The roots of $p\left(\lambda\right)$ are the eigenvalues of the stacked matrix B and vice versa.

* This is obvious since $\det\left(\lambda I - B\right) = p\left(\lambda\right)$.

▷ Fundamental Theorem of Algebra: $p(\lambda)$ of mth degree has always exactly m roots, provided one permits multiplicity of roots of the same value. This implies that given solutions to the CP $\lambda_1, ..., \lambda_m$, we can write;

$$(1 - \rho(L)) = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_m L)$$

* To see this, note that

$$\begin{aligned} & p(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_m) \\ \Leftrightarrow & \lambda^m \left(1 - \rho\left(\lambda^{-1}\right)\right) = \lambda^m \left(1 - \lambda_1 \lambda^{-1}\right) \cdots \left(1 - \lambda_m \lambda^{-1}\right) \\ \Leftrightarrow & 1 - \rho\left(\lambda^{-1}\right) = \left(1 - \lambda_1 \lambda^{-1}\right) \cdots \left(1 - \lambda_m \lambda^{-1}\right) \end{aligned}$$

and replacing $L = \lambda^{-1}$ yields out desired equation.

* Is it okay to replace λ^{-1}, which is a complex number, with L, which is an operator? Yes (apparently).

Using this machinery, we can show that an $AR(m)$ process is covariance stationary if all roots of the CP are smaller than 1 in absolute value.

Example 5.1. Take $AR(2)$ for example. We can transform it into an $AR(1)$ process:

$$(1 - \lambda_1 L)(1 - \lambda_2 L) y_t = \epsilon_t \Leftrightarrow (1 - \lambda_1 L) x_t = \epsilon_t$$

which is covariance stationary if $|\lambda_1| < 1$. Furthermore, we can write

$$(1 - \lambda_2 L) y_t = x_t$$

which is also covariance stationary if $|\lambda_2| < 1$. The logic applies more generally to $m > 2$ as well.

5.1.4 Wold Decomposition

The Wold representation is the unique linear representation where the innovations are linear forecast errors. Its significance is that the dynamic of any non-deterministic covariance-stationary process can be arbitrarily well approximated by an ARMA process. First, the theorem:

Theorem 5.1. *(Wold Decomposition)* Any covariance stationary time series can be represented as

$$y_t = \mu_t + \sum_{j=0}^{\infty} c_j u_{t-j}$$

with $c_0 = 1$ and u_t are the one-step ahead forecast errors for y_t, given information on lagged values of y_{t-j}. Furthermore, two processes with the same autocovariances, they will have the same coefficients c_j in their Wold decomposition and vice versa. Note that u_t is $y_t - P(y_t | y_{t-1}, y_{t-2}, ...)$ so not exactly $u_t = \epsilon_t$.

The above result implies that the non-deterministic process can be written as a linear combination of lagged values of a white noise process $(MA(\infty))$ representation, and this can be approximated by a ratio of two finite-lag polynomials:

$$c(L) = \frac{\theta(L)}{\phi(L)}$$

which implies that y_t can be accurately approximated by a ARMA process

$$y_t^* = \frac{\theta(L)}{\phi(L)} u_t$$

Example 5.2. (General Approach for Decomposing $AR(m)$) Take an $AR(m)$ process. Then using the result from above, we can write:

$$\begin{aligned}
(1-\rho(L))\, y_t &= \epsilon_t \\
\Rightarrow y_t &= \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)\cdots(1-\lambda_m L)}\epsilon_t \\
&= \frac{1}{1-\lambda_1 L}\cdot\frac{1}{1-\lambda_2 L}\cdots\frac{1}{1-\lambda_m L}\epsilon_t \\
&= \left(\sum_{j=0}^{\infty}(\lambda_1 L)^j\right)\cdot\left(\sum_{j=0}^{\infty}(\lambda_2 L)^j\right)\cdots\left(\sum_{j=0}^{\infty}(\lambda_m L)^j\right)\epsilon_t \\
&= \left(1+\lambda_1 L+\lambda_1^2 L^2+\cdots\right)\left(1+\lambda_2 L+\lambda_2^2 L^2+\cdots\right)\cdots\left(1+\lambda_m L+\lambda_m^2 L^2+\cdots\right)\epsilon_t \\
\Rightarrow y_t &= \sum_{j=0}^{\infty}\theta_j \epsilon_{t-j}
\end{aligned}$$

where θ_js are derived from the convolution above. Of course, an alternate way to do this is to assume $VAR(1)$ with stable eigenvalues:

$$(1-BL)\, x_t = A\epsilon_t \qquad E\left[\epsilon_t \epsilon_t'\right] = \Omega$$

Thus,

$$x_t = B^k x_{t-k} + \sum_{j=0}^{k-1} B^j A \epsilon_{t-j}$$

and therefore:

$$x_t = \sum_{j=0}^{\infty} B^j A \epsilon_{t-j}$$

and the coefficients of the Wold decomposition for y_t are given by $\left(B^j A\right)_{11}$. If B is diagonalizable, $B = VDV^{-1} \Rightarrow B^j = VD^jV^{-1}$ so computing is much easier in this case. Same result applies for Jordan-Form decomposition.

Example 5.3. (Decomposing $AR(2)$) Consider $y_t = \epsilon_t - 1.5\epsilon_{t-1} - \epsilon_{t-2}$ with $\epsilon_t \sim N(0,1)$ iid. Since

$$y_t = \left(1-1.5L-L^2\right)\epsilon_t = (1-2L)(1+0.5L)\,\epsilon_t$$

Flip the roots to obtain:

$$y_t = \left(1-\frac{1}{2}L\right)\left(1+\frac{1}{2}L\right)u_t, \qquad Var[u_t] = 4Var[\epsilon_t] = 4$$

Example 5.4. (Decomposing $MA(2)$ #1) Consider $y_t = \epsilon_t - 1.5\epsilon_{t-1} - \epsilon_{t-2}$ with $\epsilon_t \sim N(0,1)$ iid. Since

$$y_t = \left(1-1.5L-L^2\right)\epsilon_t = (1-2L)(1+0.5L)\,\epsilon_t$$

Flip the roots to obtain:

$$y_t = \left(1-\frac{1}{2}L\right)\left(1+\frac{1}{2}L\right)u_t, \qquad Var[u_t] = 4Var[\epsilon_t] = 4$$

Example 5.5. (Decomposing $MA(2)$ #2) Consider

$$y_t = \epsilon_t + \begin{bmatrix} 0.5 & 1.5 \\ 0 & 2 \end{bmatrix} \epsilon_{t-1}, \qquad \epsilon_t = \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \sim N(0, \Sigma)\, iid, \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

First, find the eigenvalues of the transition matrix:

$$\det\left(\begin{bmatrix} \lambda - 0.5 & 1.5 \\ 0 & \lambda - 2 \end{bmatrix}\right) = (\lambda - 0.5)(\lambda - 2) = 0 \Rightarrow \lambda_1 = \frac{1}{2}, v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \lambda_2 = 2, v_2 = v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

So we have

$$y_t = \epsilon_t + VDV^{-1}\epsilon_{t-1} = \epsilon_t + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \epsilon_{t-1}$$

Multiplying each side by V^{-1}:

$$V^{-1}y_t = V^{-1}\epsilon_t + DV^{-1}\epsilon_{t-1}$$
$$\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} y_t = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \epsilon_t + \begin{bmatrix} 0 & 2 \\ 0.5 & -0.5 \end{bmatrix} \epsilon_{t-1}$$

Note that $V^{-1}y_t$ has a diagonal covariance matrix:

$$\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \Sigma \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}' + \begin{bmatrix} 0 & 2 \\ 0.5 & -0.5 \end{bmatrix} \Sigma \begin{bmatrix} 0 & 2 \\ 0.5 & -0.5 \end{bmatrix}' = \begin{bmatrix} 5 & 0 \\ 0 & 1.25 \end{bmatrix}$$

which means we can separate the two shocks! Writing out the the two linear equations:

$$y_{t,2} = \epsilon_{t,2} + 2\epsilon_{t-1,2}$$
$$y_{t,1} - y_{t,2} = \epsilon_{t,1} - \epsilon_{t,2} + 0.5(\epsilon_{t-1,1} - \epsilon_{t-1,2})$$

For the first one, we have

$$y_{t,2} = (1 + 2L)\epsilon_{t,2} = (1 + 0.5L)u_{t,2}$$

and for the second we have

$$y_{t,1} - y_{t,2} = (1 + 0.5L)(\epsilon_{t,1} - \epsilon_{t,2})$$

and thus

$$y_{t,1} = y_{t,2} + (1 + 0.5L)(\epsilon_{t,1} - \epsilon_{t,2}) = (1 + 0.5L)u_{t,1}$$

.

Forecasting with a Wold Decomposition Suppose the coefficients μ_t, c_j in the Wold decomposition are known:

$$y_{t+k} = \mu_{t+k} + \sum_{j=0}^{\infty} c_j u_{t+k-j}$$

Assuming a finite-order AR, finite-order MA, the one-step ahead prediction errors u_t can be calculated from available data, and the best linear forecast is given as

$$P(y_{t+k}|y_t, y_{t-1}, ...) = \mu_{t+k} + \sum_{j=k}^{\infty} c_j u_{t+k-j}$$

If the decomposition is unknown, then the c_js have to be estimated. We defer this topic for now.

5.1.5 Impulse Responses

In signal processing, the impulse response, or impulse response function (IRF), of a dynamic system is its output when presented with a brief input signal, called an impulse. More generally, an impulse response is the reaction of any dynamic system in response to some external change.

- ▷ The impulse *response* simply refers to y_t. The Wold Decomposition tells us that it can be represented as a sum of the deterministic part and the stochastic part. But the stochastic part has to be a linear combination of prediction errors, not just any errors.
- ▷ For $AR(m)$, the impulse response can be calculated efficiently by transforming it into a $VAR(1)$.
- ▷ Note that when you're plotting the standard errors of your impulse response, be careful when scaling. This is because usually the initial impulse is scaled to be one standard-deviation, and Professor Uhlig commented that depending on the time point at which you scale, the graph could look really different.

How do we compute an impulse-response function? There are broadly two ways: (1) using the Wold decomposition and (2) using the recursive simulation. To illustrate each method, consider the case of a univariate $AR(1)$ process:

$$x_t = \phi x_{t-1} + \epsilon_t$$

where $\phi < 1$ and u_t is a scalar random disturbance with mean 0.

Computation using the Wold Decomposition Since the above process is stationary, we can find the infinite moving average representation:

$$(1 - \phi L)\, x_t = \epsilon_t \Rightarrow x_t = \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots$$

so now we have a $VAR(1)$ instead of an $AR(1)$. We are, however, interested in the evolution of x_t after a structural shock rather than an innovation in u_t. If we think of ϵ_t as reduced-form innovations that are mixed combination of some structural shocks u_t, we can assume the following relationship:

$$\epsilon_t = Bu_t$$

where B is a $n \times n$ matrix and u is a column vector containing the structural shocks. Then we can write the above MA representation as

$$x_t = Bu_t + \phi Bu_{t-1} + \phi^2 Bu_{t-2} + \cdots = \sum_{j=0}^{\infty} \phi^j Bu_{t-j} = \sum_{j=0}^{\infty} C_j u_{t-j}$$

The above representation is very important since the coefficients of the moving average representation (Defined as $C_j = B\phi^j$) are the responses of variables contained in x to impulses in these structural shocks. An impulse response function is simply a plot of $\frac{\partial x_{t+j}}{\partial \epsilon_t} = \phi^j B$ so all we need to plot this graph is an estimate of ϕ (obtained from the reduced-form regression) and B (obtained after imposing some identifying restrictions)

Computation using the Recursive Simulation Using the previous results, write:

$$x_t = \phi x_{t-1} + Bu_t$$

and we can equivalent compute the impulse response function through a recursive simulation of the system. Note that the recursion yields the same impulse response:

$$x_1 = \phi x_0 + B(1) = B$$
$$x_2 = \phi x_1 + B(0) = \phi B$$

and so forth.

Interpretation Suppose you took an original time series specification $y_t = \phi y_{t-1} + \epsilon_t$ and transformed it into $y_t = \mu_t + \sum_{j=0}^{\infty} C_j u_{t-j}$.

- ▷ ϵ_t is the "reduced-form innovations" that are mixed combination of some structural shocks u_t.
  - * This is why in our impulse functions, we apply shock $u_0 = 1$ instead of $\epsilon_0 = 1$. We want to apply a fundamental shock to the economy, we want an innovation in u, not ϵ.
- ▷ To apply this shock then, we need a representation of the economy (y_t) as a function of structural shocks (u_t).
  - * For $AR(m)$, it's already in a form such that ϵ_t is the only stochastic element. Therefore, $AR(m)$ is its own Wold Decomposition and we can obtain the impulse response by letting $\epsilon_0 = 1$.
  - * For $MA(m)$, however, it is expressed as a linear combination of ϵ's so we need to transform it into a linear combination of the u's.

5.2 Spectral Theory

This section doesn't necessarily provide the econometric tools, but it does provide a useful set of terminologies that are resued in the literature.

5.2.1 Fourier Transforms (FT)

In general, we talk about three kinds of FTs:

1. (Continuous) Fourier Transforms (FT): Continous input ⇒ Continuous output
2. Discrete-time Fourier Transforms (DTFT): Discrete input ⇒ Continuous & Periodic output
3. Discrete Fourier Transforms (DFT): Discrete & Periodic input ⇒ Discrete & Periodic output

FT and DTFT are analysis tools, whereas DFT is a computational tool. For this class, we are interested in DTFT since we want to be able to apply this technique to discrete (and stationary) time series. In traditional FT, we decompose a deterministic function into combinations of sinusoids; here, we decompose a stationary time series $\{X_t\}$ into a combination of time series with random (and uncorrelated) coefficients.

Discrete-time Fourier Transform (for covariance-stationary time series) Given a sequence $\{x_t\}$ with $\sum_{j=-\infty}^{\infty} |x_t| < \infty$, define the Fourier Transform

$$\tilde{x}(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} x_t e^{-it\omega}$$

and the corresponding Inverse Fourier Trasform

$$x_t = \int_{-\pi}^{\pi} \tilde{x}(\omega) e^{-it\omega} d\omega$$

A few important remarks regarding the expression:

- ▷ The functions $\left\{e^{-it\omega}\right\}$ are called harmonics and constitute an orthonomrla base.
- ▷ The Riesz-Fisher Theorem and its converse assures that the Fourier Trassform is a bijection.

▷ The mapping is also an isometric isomorphism since it preserves linearity and distance: for any two series $\{x_t\}$ and $\{y_t\}$ with corresponding Fourier transforms, we have

$$x(\omega) + y(\omega) = \sum_{t=-\infty}^{\infty} (x_n + y_n) e^{-it\omega}, \qquad \alpha x(\omega) = \sum_{t=-\infty}^{\infty} \alpha x_t e^{-it\omega}$$

▷ There is a limit in the amount of information in a given time series, and we either have concentration in the original series or concentration in the Fourier transform.

Usefulness The biggest advantage of using Fourier Transforms is that it simplies computation since convolutions are now multplications. To see this, consider

$$y_t = h(L)x_t = \sum_{j=-\infty}^{\infty} h_j x_{t-j}$$

Then taking the Fourier Transform on the LHS

$$\begin{aligned}
\tilde{y_t} &= \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} y_t e^{-i\omega t} \\
&= \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} h_j x_{t-j} \right) e^{-i\omega t} \\
&= \left(\frac{1}{2\pi} \sum_{j=-\infty}^{\infty} h_j e^{-i\omega j} \right) \left[\sum_{t=-\infty}^{\infty} x_{t-j} e^{-i\omega(t-j)} \right] \\
&= \tilde{h}(\omega) \left[2\pi \tilde{x}(\omega) \right]
\end{aligned}$$

and noting the fact that

$$\tilde{h}(\omega) = \frac{1}{2\pi} \underbrace{\sum_{j=-\infty}^{\infty} h_j e^{-i\omega j}}_{} = \frac{1}{2\pi} \underbrace{h\left(e^{-i\omega}\right)}_{}$$

$$\tilde{y}_t = h\left(e^{-i\omega}\right) \tilde{x}(\omega) = 2\pi \tilde{h}(\omega) \tilde{x}(\omega)$$

which stems from our lag operator notation:

$$(1 - \rho(L)) y_t = \epsilon_t, \qquad \rho(L) = \sum_{j=1}^{m} \rho_j L^j$$

Similarly, we have

$$\begin{aligned}
&AR(m) : (1 - \rho(L)) y_t = \epsilon_t \Leftrightarrow \left(1 - \rho\left(e^{-i\omega}\right)\right) \tilde{y}(\omega) = \tilde{\epsilon}(\omega) \\
&MA(n) : y_t = \theta(L) \epsilon_t \Leftrightarrow \tilde{y}(\omega) = \theta\left(e^{-i\omega}\right) \tilde{\epsilon}(\omega)
\end{aligned}$$

5.2.2 Spectral Analysis

Let x_t be covariance stationary with mean zero. Denote $\gamma_j = E[x_t x_{t-j}]$. The *population spectrum* is defined as the Fourier Transformation of γ:

$$s_x(\omega) = \tilde{\gamma}(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ij\omega} \approx E\left[\tilde{x}(\omega) \overline{\tilde{x}(\omega)} \right]$$

i.e. the Fourier Transform of the covariance of the processes, divided by 2π to normalize the integral of $s_x(\cdot)$ to 1. Note that $s_x(\omega) = s_x(-\omega)'$.

- ▷ The spectrum and the autocovariances are equivalent; there is no information in one that is not presented in other.
- ▷ If the information is equivalent, what is useful about this alternate representation?
  - * Some characteristics of the series, such as the serial correlation, are easier to grasp with the autocovariances, while others such as its unobserved components (as the different fluctuations that compose the series) are much easier to understand in the spectrum.
- ▷ Why is $s_x(\omega) \approx E\left[\tilde{x}(\omega)\overline{\tilde{x}(\omega)}\right]$?

Let's use this result for a more complicated process,

$$y_t = \sum_{-\infty}^{\infty} h_j \epsilon_{t-j} = h(L)\epsilon_t$$

where θ is absolutely summable, then we have the following relationship:

$$s_y(\omega) = \left|h\left(e^{-i\omega}\right)\right|^2 s_x(\omega)$$

To see this heuristically:

$$\begin{aligned} s_y(\omega) &\approx E\left[\tilde{y}(\omega)\overline{\tilde{y}(\omega)}\right] \\ &\approx h\left(e^{-i\omega}\right) E\left[\tilde{x}(\omega)\overline{\tilde{x}(\omega)}\right] h\left(e^{i\omega}\right) \\ &= h\left(e^{-i\omega}\right) h\left(e^{i\omega}\right) s_x(\omega) \\ &= \left|h\left(e^{-i\omega}\right)\right|^2 s_x(\omega) \end{aligned}$$

and for multivariate, we have

$$s_y(\omega) = h\left(e^{-i\omega}\right) s_x(\omega) \overline{h\left(e^{-i\omega}\right)}$$

For a more rigorous proof, start by writing out Γ_j, the covariance function for y_t:

$$\begin{aligned} \Gamma_j &= E[y_t y_{t-j}] \\ &= E\left[\left(\sum_{\ell=-\infty}^{\infty} h_\ell \epsilon_{t-\ell}\right)\left(\sum_{k=-\infty}^{\infty} h_k \epsilon_{t-k-j}\right)\right] \\ &= \sum_{\ell,k=-\infty}^{\infty} h_\ell h_k E[\epsilon_{t-\ell}\epsilon_{t-k-j}] \\ &= \sum_{\ell,k=-\infty}^{\infty} h_\ell h_k \gamma_{k+j-\ell} \end{aligned}$$

Taking the Fourier Trasform of Γ_j:

$$\begin{aligned} s_y(\omega) &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} e^{-ij\omega}\Gamma_j \\ &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} e^{-ij\omega} \sum_{\ell,k=-\infty}^{\infty} h_\ell h_k \gamma_{k+j-\ell} \end{aligned}$$

Denoting $m = k + j - \ell$, note that

$$s_x(\omega) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} e^{-im\omega} \gamma_m$$

so rewriting $s_y(\omega)$:

$$\begin{aligned} s_y(\omega) &= \sum_{\ell=-\infty}^{\infty} h_\ell e^{-ij\omega} \sum_{j=-\infty}^{\infty} h_j e^{ij\omega} \left(\frac{1}{2\pi} \sum_{m=-\infty}^{\infty} e^{-im\omega} \gamma_m \right) \\ &= h\left(e^{-i\omega}\right) h\left(e^{i\omega}\right) s_x(\omega) \end{aligned}$$

and we have the desired result.

Below, we explore some examples. The key result that gets re-used often is the fact that $1-\rho L$ becomes $\left(1-\rho e^{i\omega}\right)\left(1-\rho e^{-i\omega}\right)$ upon the transform.

Example 5.6. (White Noise) Suppose $x_t = \epsilon_t$. We know that $\gamma_0 = \sigma^2$ and $\gamma_t = 0, \forall t > 0$. Thus, $s_x(\omega) = \sigma^2/2\pi$ which is a flat line.

Example 5.7. Spectrum of $AR(1)$: Suppose $\epsilon_t = (1-\rho L) y_t$ so

$$\begin{aligned} \frac{\sigma^2}{2\pi} &= \left(1-\rho e^{i\omega}\right)\left(1-\rho e^{-i\omega}\right) s_y(\omega) \\ &= \left(1 - 2\rho\cos\omega + \rho^2\right) s_y(\omega) \\ \Rightarrow s_y(\omega) &= \frac{1}{1-2\rho\cos\omega+\rho^2} \frac{\rho^2}{2\pi} \end{aligned}$$

Example 5.8. Spectrum of $AR(m)$: Suppose $\epsilon_t = (1-\rho(L)) y_t$ with all stable roots: so

$$\begin{aligned} \epsilon_t &= (1-\lambda_1 L)\cdot(1-\lambda_2 L)\cdots(1-\lambda_m L) y_t \\ \Leftrightarrow \frac{\sigma^2}{2\pi} &= s_y(\omega) \prod_{j=1}^{m} \left(1 - 2\lambda_j \cos\omega + \lambda_j^2\right) \\ \Rightarrow s_y(\omega) &= \frac{\sigma^2}{2\pi} \prod_{j=1}^{m} \frac{1}{\left(1 - 2\lambda_j \cos\omega + \lambda_j^2\right)} \end{aligned}$$

Note that for complex roots, we have

$$s_y(\omega) = \frac{\sigma^2}{2\pi} \prod_{j=1}^{m} \left| \frac{1}{\left(1-\lambda_j e^{i\omega}\right)\left(1-\lambda_j e^{-i\omega}\right)} \right|$$

and since they always come in complex-conjugate pairs, we can simplify above to:

$$s_y(\omega) = \frac{\sigma^2}{2\pi} \prod_{j=1}^{m} \frac{1}{\left(1-\lambda_j e^{-i\omega}\right)\left(1-\bar{\lambda}_j e^{i\omega}\right)}$$

Example 5.9. Spectrum of $MA(1)$: Suppose $y_t = (\theta_0 + \theta_1 L)\,\epsilon_t$ so

$$\begin{aligned} s_y(\omega) &= \left(\theta_0 + \theta_1 e^{i\omega}\right)\left(\theta_0 + \theta_1 e^{-i\omega}\right)\frac{\sigma^2}{2\pi} \\ \Rightarrow s_y(\omega) &= \left(\theta_0^2 + 2\theta_0\theta_1\cos\omega + \theta_1^2\right)\frac{\sigma^2}{2\pi} \end{aligned}$$

Example 5.10. Spectrum of $MA(n)$: Suppose $y_t = \theta(L)\,\epsilon_t$ so

$$s_y(\omega) = \theta\left(e^{i\omega}\right)\theta\left(e^{-i\omega}\right)\frac{\sigma^2}{2\pi}$$

Blaschke Factor and Fundamental Representation The Blaschke factor for $0 \neq \lambda \in \mathbb{C}$ is defined as

$$B_\lambda(z) = \frac{z-\lambda}{1-\lambda z} = -\lambda\frac{1-\lambda^{-1}z}{1-\lambda z}$$

Note that

$$B_\lambda\left(e^{-i\omega}\right) = \left(B_\lambda\left(e^{i\omega}\right)\right)^{-1}$$

Why is this factor useful? It allows us to construct a new polynomial and take its spectrum, which will be equivalent to the spectrum of the original representation. Specifically, let $y_t = \theta(L)\,\epsilon_t$ and $p(\lambda) = \lambda^n\theta\left(\lambda^{-1}\right)$. Define $\hat{y}_t = \hat{\theta}(L)\,\epsilon_t = B_\lambda(L)\,\theta(L)\,\epsilon_t$. Then we have that

$$s_{\hat{y}}(\omega) = s_y(\omega)$$

i.e. the spectrums of $\hat{y}$ and y are identical. Identical spectrum implies identical autocorrelations, and identical autocorrelations imply identical Wold decomposition as the original representation. This allows us to multiply y_t by the Blaschke fator for λ and preserve the Wold decomposition.

This is useful for deriving the *fundamental representation* defined as the following. Consider $y_t = \theta(L)\,\epsilon_t = \theta_0(1-\lambda_1 L)\cdots(1-$ and suppose that

$$|\lambda_1| > ... > |\lambda_r| > 1 > |\lambda_{r+1}| \cdots > |\lambda_n|$$

We will flip the explosive roots to do:

$$y_t = \theta(L)\,\epsilon_t \Rightarrow y_t = C(L)u_t$$

where

$$\begin{aligned} C(L) &= B_{\lambda_1}(L)\cdots B_{\lambda_r}(L)\,\theta(L)\,\frac{1}{(-\lambda_1)...(-\lambda_r)\,\theta_0} \\ &= \left(1-\lambda_1^{-1}L\right)\cdots\left(1-\lambda_r^{-1}L\right)(1-\lambda_{r+1}L)\cdots(1-\lambda_n L) \\ Var(u_t) &= (\lambda_1\cdots\lambda_r\theta_0)^2\,Var(\epsilon_t) \end{aligned}$$

Note that the fundamental representation is invertible, so we can back out u_t from y_t: $u_t = C(L)^{-1}\,y_t$. Furthermore, we can back out ϵ_t by treating it as a hidden state and applying the Kalman Filter using the observations u_t.

Example 5.11. In one of the problem sets, we started with

$$y_t = \nu_t + 7\nu_{t-1} + 10.75\nu_{t-2} + 3.75\nu_{t-3}$$

and arrived at the Wold decomposition of

$$y_t = \left(1 + \frac{1}{2}\lambda\right)\left(1 + \frac{2}{3}\lambda\right)\left(1 + \frac{1}{5}\lambda\right)\epsilon_t$$

where we flipped the root $\lambda = -5$ and $\lambda = -3/2$. According to the result above, it must be that

$$Var\left(\epsilon_t\right) = \left((-5)(-3/2)\right)^2 = \frac{225}{4}$$

Example 5.12. Recall the MA(1) example:

$$\begin{aligned} y_t &= \epsilon_t + 2\epsilon_{t-1}, E\left[\epsilon_t^2\right] = 1 \\ &= \left(1 - \lambda L\right)\epsilon_t, \lambda = -2 \end{aligned}$$

Since this root is explosive, flip it to obtain the Wold decomposition:

$$y_t = \left(1 - \lambda^{-1}L\right)u_t = u_t + 0.5u_{t-1}, Var\left(u_t\right) = \lambda^2 Var\left(\epsilon_t\right) = 4$$

5.3 Unit Roots

5.3.1 Terminology

We introduce some basic definitions:

- ▷ *Integrated*; An $AR(m)$ process with roots smaller than 1 in absolute value or exactly equal to 1, with at least one root exactly equal to 1, is called *integrated* or is said to *have a unit root*. Furthermore, let r be the number of roots exactly equal to one. The process is then said to be *integrated of order r* or denoted as $I\left(r\right)$.
  - * Thus, "the process is covariance stationary" is equivalent to saying that "the process is $I\left(0\right)$."
- ▷ *Seasonally integrated*; An extension of this definition is the following: an $AR\left(m\right)$ process with roots smaller than or equal to 1 in absolute value, with at least one root equal to 1 in absolute value, is called *seasonally integrated*.

Now define the difference operator Δ per

$$\Delta y_t = y_t - y_{t-1} = \left(1 - L\right)y_t$$

and obviously multiple differencing is indicated by powers, for example:

$$\Delta^2 y_t = \Delta\left(\Delta y_t\right) = \Delta\left(y_t - y_{t-1}\right) = y_t - 2y_{t-1} + y_{t-2}$$

Note that we have the following result:

Proposition 5.1. *Let y_t an $AR(m)$ process be $I\left(r\right)$. Then $\Delta^r y_t$ is a covariance stationary $AR\left(m - r\right)$ process.*

We can verify this proposition with the following $AR(4)$ process which is $I(2)$:

$$\begin{aligned} &\left(1 - 2L + .75L^2 + .5L^3 - .25L^4\right)y_t = \epsilon_t \\ \Leftrightarrow &\left(1 - .5L\right)\left(1 + .5L\right)\left(1 - L\right)\left(1 + L\right)y_t = \epsilon_t \\ \Leftrightarrow &\left(1 - .5L\right)\left(1 + .5L\right)\Delta^2 y_t = \epsilon_t \end{aligned}$$

where $x_t := \Delta^2 y_t = y_t - 2y_{t-1} + y_{t-2} = y_t\left(1 - 2L + L^2\right) = y_t\left(1 - L\right)^2$ is stationary.

5.3.2 Estimating an AR(1)

Intuitively, we can think of using OLS to estimate the following AR(1) process:

$$y_t = \rho y_{t-1} + \epsilon_t$$

For $|\rho| < 1$, there is an established asymptotic result:

$$\sqrt{T}\left(\hat{\rho} - \rho\right) \xrightarrow{d} N\left(0, 1 - \rho^2\right)$$

For $\rho = 1$, the result needs to be modified a little bit. In fact, we have that it converges at a faster rate – T instead of $\sqrt{T}$– to a non-normal distribution. In fact, we can show that under the null hypothesis of $\rho = 1$:

$$T\left(\hat{\rho}_T - 1\right) \xrightarrow{d} \frac{1}{2} \frac{W\left(1\right)^2 - 1}{\int W\left(s\right)^2 ds}$$

where the RHS is some random variable defined as a continuous functional of a Brownian motion. We call this "super-consistent" since the rate of convergence is higher than $\sqrt{T}$.

5.3.3 A Bayesian Perspective

We can apply a Bayesian approach to forecasting with $AR(1)$. Specifically, find a posterior distribution for y_{t+k} given $y_0, ..., y_t$:

$$y_{t+k}|y_0, ..., y_t \sim \rho^k y_t + \sum_{j=0}^{k-1} \rho^j \epsilon_{t+k-j}, \epsilon_s \sim N\left(0, \sigma^2\right)$$

Therefore we can (1) draw (ρ, σ) from the posterior, (2) draw $\epsilon_{t+1}, ..., \epsilon_{t+k}$ and (3) combine a generate a draw for y_{t+k}.

6 Multivariate Time-Series: Non-Bayesian Approach

6.1 Roots, Impulse Responses, and Cointegration

This section covers the application of multivariate time-series in a representative economic context.

6.1.1 From $VAR(n)$ to $VAR(1)$

Consider $VAR(n)$ in $y_t \in \mathbb{R}^m$ which are vectors:

$$y_t = \sum_{j=1}^{n} B_j y_{t-j} + A\epsilon_t, \epsilon_t \sim N(0, I_m) \Rightarrow (I_m - B(L)) y_t = A\epsilon_t$$

Then the one-step ahead prediction error u_t can be defined as $u_t = A\epsilon_t$ with the covariance matrix $\Sigma = AA'$.

- ▷ Under this representation, the ϵ_ts are now the structural shocks that are assumed to be independent of each other. It is the independence assumption that allows us to specify its variance as a diagonal covariance matrix and a normalized one.
- ▷ You cannot get A directly, so you need some identifying assumptions.

Since we can transform this into $VAR(1)$, we now exclusively focus on analyzing the $VAR(1)$ process where

$$x_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-n+1} \end{bmatrix} = \mathcal{B}x_{t-1} + \mathcal{A}\epsilon_t, \quad \mathcal{B} = \begin{bmatrix} B_1 & \cdots & B_{n-1} & B_n \\ I_m & \cdots & 0 & 0 \\ \vdots & \ddots & & \vdots \\ 0 & & I_m & 0 \end{bmatrix}, \mathcal{A} = \begin{bmatrix} A \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The characteristic polynomial is defined analogously as before:

$$p(\lambda) = \det(\lambda I_{mn} - \mathcal{B})$$

Here we assume that $\mathcal{B}$ is diagonalizable. If it is not, we can proceed similarly with the Jordan decomposition form. Prof. Uhlig mentioned in-class that we almost always deal with a diagonlizable $\mathcal{B}$. Proceeding with the diagonalization, we can write:

$$\mathcal{B} = VDV^{-1}, \qquad D = diag(\lambda_1, ..., \lambda_{nm})$$

Notice that this process simplifies the VAR process (assuming $|\lambda_j| < 1, \forall j$ since:

$$x_t = \sum_{j=0}^{\infty} \mathcal{B}^j \mathcal{A}\epsilon_{t-j} = V \sum_{j=0}^{\infty} diag\left(\lambda_1^j, ..., \lambda_{nm}^j\right) V^{-1} \mathcal{A}\epsilon_{t-j}$$

To see its usefulness, write from our original $VAR(n)$ formulation:

$$A\epsilon_t = u_t = y_t - \sum_{j=1}^{n} B_j y_{t-j} = y_t (I_m - B(L)) \Rightarrow y_t = (I_m - B(L))^{-1} u_t$$

Then first m rows of the computed x_t provides an explicit expression for the Wold decomposition.

6.1.2 Impulse Response

Once again, the impulse response $r_a(k)$ to a vector a is similarly defined as the forecast revision for y_k given a shock $u_0 = a$:

$$r_a(k) = E[y_k | u_0 = a, y_t, t < 0] - E[y_k | y_t, t < 0], k \geq 0$$

Usually, we set $y_t = 0$ for convenience. More generally, we set $y_t = E[y_t]$ to use $E[y_k | y_t, t < 0] = E[y_k]$.

Generally for $VAR(1)$ of the form $y_t = By_{t-1} + u_t$, the impulse response (giving a shock $u_0 = 1$) will simply be $r_a(k) = B^k a$ for obvious reasons (you can see this if you work the recursion out). The difficult part is actually computing B^k but we can use the diagonalization result earlier to see that

$$r_a(k) = B^k a = VD^k V^{-1} a$$

Note that a can be a vector! We discuss more details in the next section.

Introducing Cointegration We say that the vector time series y_t is co-integrated of rank r if each of the series taken individually is $I(1)$ i.e. has one unit root, while some linear combination of the series $\beta' y_t$ is stationary for some co-integrating matrix $\beta_{m \times r}$ of rank r. Generally, y_t can be non-stationary. Well-known examples of such linear combination include: the consumption-output ratio, Lettau-Ludgvigson CAY series, and the dividend-price ratio.

Decomposing the B matrix Why do we care about cointegration? The ultimate goal is to decompose the dynamics into a stationary and non-stationary component. To do this, take a $VAR(1)$

$$Y_t = BY_{t-1} + u_t$$

where B is diagonalizable $B = VDV^{-1}$ and define the following matrixes $\beta, \beta_*, \nu, \nu_*$:

$$V = \begin{bmatrix} | & | & \vdots & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & \vdots & | \end{bmatrix} = \begin{bmatrix} \nu_{m \times r} & \nu^*_{m \times (m-r)} \end{bmatrix}$$

$$V^{-1} = \begin{bmatrix} -- & w_1 & -- \\ -- & w_2 & -- \\ \vdots & \vdots & \cdots \\ -- & w_m & -- \end{bmatrix} \Rightarrow \left(V^{-1}\right)' = \begin{bmatrix} \beta_{m \times r} & \beta^*_{m \times (m-r)} \end{bmatrix} = \begin{bmatrix} | & | & \vdots & | & | & \vdots & | \\ w_1' & w_2' & & w_r' & w_{r+1}' & \cdots & w_m' \\ | & | & \vdots & | & | & \vdots & | \end{bmatrix}$$

with $z_1, ..., z_r$ being the stationary roots and $z_{r+1}, ..., z_m = 1$ are the unit roots.

- ▷ β is the cointegrating matrix. Essentially, with β you're taking the eigenvectors corresponding only to the stationary roots.
- ▷ Since β is the cointegrating matrix, by definition applying it to y_t should yield a stationary series, i.e. $\beta' y_t$ is stationary.

This allows us to re-express the dynamics as

$$Y_t = BY_{t-1} + u_t = VDV^{-1}Y_{t-1} + u_t = \begin{bmatrix} \nu & \nu^* \end{bmatrix} \begin{bmatrix} D_{r \times r} & 0_{r \times (m-r)} \\ 0_{(m-r) \times r} & I_{(m-r) \times (m-r)} \end{bmatrix} \begin{bmatrix} \beta & \beta^* \end{bmatrix}^T y_{t-1} + u_t$$

Computing the Impulse Response Now take an initial shock vector a and denote $a_r = \beta' a$ and $a^* = (\beta^*)' a$. Then the impulse response with initial shock a at the kth time step is:

$$r_a(k) = \nu D^k_{r \times r} a_r + \nu^* a^*$$

where the first part is transitory (they die out since the roots are stationary) and the second part is permanent. In the long-run, $r_a(\infty) = \nu^* a^*$.

Error-Correction Representation We consider an alternate representation of the $VAR(1)$. Define $\alpha = \nu\left(I_{r\times r} - D_{r\times r}\right)$ and β same as before. Then

$$\begin{aligned} Y_t &= BY_{t-1} + u_t \\ \Leftrightarrow Y_t - Y_{t-1} &= \left(B - I\right) Y_{t-1} + u_t \end{aligned}$$

Notice that

$$B - I = \left[\begin{array}{cc} \nu & \nu^* \end{array}\right] \left[\begin{array}{cc} D_{r\times r} - I_{r\times r} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{array}\right] \left[\begin{array}{cc} \beta & \beta^* \end{array}\right]^T$$

which implies that $-\alpha\beta' = B - I$. (The dimension of $\alpha\beta'$ must be equal to the dimension of the original system) Replacing this into the above equation, we have

$$\Delta Y_t = -\alpha\beta' Y_{t-1} + u_t = \left(B - I\right) Y_{t-1} + u_t$$

The idea here is that ΔY_t , the change of Y_t from one period to the next, is driven by a (1) convergence of the stationary component $(\beta' Y_{t-1})$ bakc to zero, and (2) the news shocks. We know that $\beta' Y_{t-1}$ is stationary since β is the co-integrating matrix.

Notice that Engel and Granger (1987) have a result that a $VAR(k)$ cointegrated of rank r can be given an error correction representation

$$A^*\left(L\right)\Delta y_t = -\alpha\beta' y_{t-1} + u_t$$

where $A^*\left(L\right)$ has only stable roots.

Example 6.1. (Simple Numerical Example): Suppose

$$B = \left[\begin{array}{cc} 0.8 & -0.6 \\ 0.6 & 0.8 \end{array}\right] \left[\begin{array}{cc} 0.5 & 0 \\ 0 & 1 \end{array}\right] \left[\begin{array}{cc} 0.8 & 0.6 \\ -0.6 & 0.8 \end{array}\right]$$

Then it yields

$$\nu = \left[\begin{array}{c} 0.8 \\ 0.6 \end{array}\right], v^* = \left[\begin{array}{c} -0.6 \\ 0.8 \end{array}\right], \beta = \left[\begin{array}{c} 0.8 \\ 0.6 \end{array}\right], \beta^* = \left[\begin{array}{c} -0.6 \\ 0.8 \end{array}\right], \qquad \alpha = \left[\begin{array}{c} 0.4 \\ 0.3 \end{array}\right], \alpha\beta' = \left[\begin{array}{cc} 0.32 & 0.24 \\ 0.24 & 0.18 \end{array}\right] = I_2 - B$$

Example 6.2. (2008 Final): Consider the VAR in $X_t \in \mathbb{R}^2$:

$$X_t = BX_{t-1} + \epsilon_t, B = \left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \end{array}\right], \epsilon_t = \left[\begin{array}{c} \epsilon_{t,1} \\ \epsilon_{t,2} \end{array}\right] \sim N\left(0, I_2\right) iid$$

where B is a matrix with stable eigenvalues. Suppose the first entry of X_t is the first difference (in logarithms) of consumption net of a constant, and the second entry is the log of the ratio of consumption to GNP net of a constant:

$$X_{t,1} = c_t - c_{t-1} - \mu_c, \qquad X_{t,2} = c_t - g_t - \mu_r$$

where c_t is the log of consumption, g_t is the log of GNP and μ_c, μ_r are the corresponding constants. Assume the constants are zero, and let $Y_t = [c_t, g_t]'$.

1. Is Y_t cointegrated? What is the cointegrating rank?: We know that c_t and g_t are both $I(1)$ but $c_t - g_t$ is stationary. Therefore Y_t is cointegrated with rank 1 and the co-integrating vector is $\beta = [1, -1]$.

2. Find the error correction representation for Y_t of the form $\Delta Y_t = -\alpha\beta' Y_{t-1} + B^* \Delta Y_{t-1} + A\epsilon_t$. In this case, it's easier to directly find the dynamics for ΔY_t:

$$\begin{bmatrix} c_t - c_{t-1} \\ g_t - g_{t-1} \end{bmatrix} = -\alpha\beta' \begin{bmatrix} c_{t-1} \\ g_{t-1} \end{bmatrix} + B^* \begin{bmatrix} c_{t-1} - c_{t-2} \\ g_{t-1} - g_{t-2} \end{bmatrix} + A\epsilon_t$$

To do so, note that

$$\begin{aligned} g_t - g_{t-1} &= -(c_t - g_t) + (c_{t-1} - g_{t-1}) + (c_t - c_{t-1}) \\ &= -(B_{21}(c_{t-1} - c_{t-2}) - B_{22}(c_{t-1} - g_{t-1})) + (c_{t-1} - g_{t-1}) \\ &\quad + (B_{11}(c_{t-1} - c_{t-2}) + B_{12}(c_{t-1} - g_{t-1})) + \epsilon_{1,t} - \epsilon_{2,t} \\ &= (B_{11} - B_{21})(c_{t-1} - c_{t-2}) + (1 + B_{12} - B_{22})(c_{t-1} - g_{t-1}) + \epsilon_{1,t} - \epsilon_{2,t} \end{aligned}$$

and

$$c_t - c_{t-1} = B_{11}(c_{t-1} - c_{t-2}) + B_{12}(c_{t-1} - g_{t-1})$$

Therefore:

$$\begin{bmatrix} c_t - c_{t-1} \\ g_t - g_{t-1} \end{bmatrix} = \begin{bmatrix} B_{12} & -B_{12} \\ 1 + B_{12} - B_{22} & B_{22} - 1 - B_{12} \end{bmatrix} \begin{bmatrix} c_t \\ g_t \end{bmatrix} + \begin{bmatrix} B_{11} & 0 \\ B_{11} - B_{21} & 0 \end{bmatrix} \begin{bmatrix} c_{t-1} - c_{t-2} \\ g_{t-1} - g_{t-2} \end{bmatrix} + A\epsilon_t$$

where

$$-\alpha'\beta = \begin{bmatrix} B_{12} \\ 1 + B_{12} - B_{22} \end{bmatrix} [1, -1] = \begin{bmatrix} B_{12} & -B_{12} \\ 1 + B_{12} - B_{22} & B_{22} - 1 - B_{12} \end{bmatrix}, A = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$$

Example 6.3. (2016 Final): Let $y = By_{t-1} + u_t, y_t \in \mathbb{R}^2$ with Σ as the error covariance matrix and

$$B = \begin{bmatrix} 1 & 0 \\ -1 & 0.5 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

To find the error-correction representation of the process, diagonalize the B matrix:

$$B = VDV^{-1} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

and compute α, β:

$$\alpha = \begin{bmatrix} 0 \\ 1 \end{bmatrix} ([1] - [0.5]) = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}, \beta = \begin{bmatrix} 2 & 1 \end{bmatrix}$$

thus yielding

$$\Delta y_t = \begin{bmatrix} 0 & 0 \\ -1 & -0.5 \end{bmatrix} y_{t-1} + u_t$$

Note that the resulting matrix is equal to

$$B - I = \begin{bmatrix} 1 & 0 \\ -1 & 0.5 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1 & -0.5 \end{bmatrix}$$

which is good.

6.1.3 Addressing Unit Roots

Skipped.

6.2 Identification of the Shocks

"Finally, with 15 minutes left, we arrive at the actual time-series analysis." - Harald Uhlig

6.2.1 The Identification Problem

We don't really care about the responses to one-step ahead prediction errors. We care about the responses to a true structural shock. Thus we need a method to disentangle u_t into "structural shocks" such as monetary policy shocks or productivity shocks:

$$u_t = A\epsilon_t$$

where ϵ_t is the structural shock and u_t is the reduced form shock (a combination of the structural shocks).

Can we identify A nicely? Assuming $E\left[\epsilon_t \epsilon_t'\right] = I_m$, the only restriction on A at our disposal is $\Sigma = AA'$ so one needs at least $m(m-1)/2$ additional identifying assumptions to identify ϵ_t. Furthermore, one needs restrictions for the sign of the columns of A which doubles the required identifying assumptions. To circumvent this problem, we use economic theory in order to derive some restrictions on the effects of some shock on particular variables to fix the remaining $m\left(m-1\right)/2$.

6.2.2 Method #1: Zero Contemporaneous Restrictions via Cholesky Decomposition

- ▷ Additional restriction: Shock j has no contemporaneous impact on variables $i < j$.
- ▷ Technique: Cholesky Decomposition
  - * Given a matrix Σ, the Cholesky factor A is defined as the unique lower triangular matrix such that $AA' = \Omega$. For a two-by-two matrix, this is relatively easy to do.
- ▷ Remarks:
  - * This method implies ordering the variables into a contemporaneous causality, and by construction the result depends on the particular ordering of the variables.
  - * Whether this is a good idea or not cannot be judged by simply looking at the data, so identifying assumptions need to be motivated carefully and appear sensible on a priori basis.
  - * There are an infinite number of orthonormal matrices T with $TT' = I$ so consequently there are infinitely many ways to orthogonalize the reduced form errors ϵ_t that all fit the data equally well.

6.2.3 Method #2: Long-Run Identification

- ▷ Key Idea: There may be a subset of shocks that have permanent effects on some variables but not on others, and shocks that have no permanent effects on any variables. This splits reduced form shocks into permanent shocks and "everything else."
- ▷ Additional restriction: Shock j does not affect shock i in the long-run.
- ▷ Technique: Cholesky Decomposition
- ▷ Application: Most economists agree that demand shocks such as monetary policy shocks are neutral in the long-run, so this is applicable.john

Example 6.4. (Example Usage by Blanchard-Quah) In their model of unemployment and growth, they have essentially

$$\begin{bmatrix} a_{11}^0 & a_{12}^0 \\ a_{21}^0 & a_{22}^0 \end{bmatrix} \begin{bmatrix} U_t \\ \Delta Y_t \end{bmatrix} = \begin{bmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{bmatrix} \begin{bmatrix} U_{t-1} \\ \Delta Y_{t-1} \end{bmatrix} + \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix}$$

where U_t is unemployment and ΔY_t is the change in log GNP. The identifying assumption here is that u_t^d has no permanet effects on the level of GNP. To compute the long-run effects, write:

$$\begin{bmatrix} U_t \\ \Delta Y_t \end{bmatrix} = \Phi_1 \begin{bmatrix} U_{t-1} \\ \Delta Y_{t-1} \end{bmatrix} + A^{-1}\mathbf{u}_t$$

and summing all future changes in GNP, we have

$$E\left[\sum_{s=0}^{\infty} \begin{bmatrix} U_{t+s} \\ \Delta Y_{t+s} \end{bmatrix} | \mathbf{u}_t\right] = (I - \Phi_1)^{-1} A^{-1}\mathbf{u}_t$$

Imposing our assumption that $(I - \Phi_1)^{-1} A_0^{-1}$ is upper triangular, use the Choleski decomposition on the matrix

$$Q = (I - \Phi_1)^{-1} \Sigma \left((I - \Phi)^{-1}\right)'$$

Example 6.5. (Blanchard-Quah Decomposition of Σ) Let $y_t = By_{t-1} + u_t, \quad y_t \in \mathbb{R}^2$ with $E[u_t u_t'] = \Sigma$ and

$$B = \begin{bmatrix} 0.8 & 0 \\ -1 & 0.5 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

We want to find A such that $AA' = \Sigma$. In this case, we first find the Choleski decomposition of

$$QQ' = (I - B)^{-1} \Sigma \left((I - B)^{-1}\right)'$$

and then find A such that $A = (I - B)\, Q$. First, to compute the Choleski decomposition, note that

$$I - B = \begin{bmatrix} 0.2 & 0 \\ 1 & 0.5 \end{bmatrix} \Rightarrow (I - B)^{-1} = \begin{bmatrix} 5 & 0 \\ -10 & 2 \end{bmatrix}$$

and thus

$$QQ' = \begin{bmatrix} 5 & 0 \\ -10 & 2 \end{bmatrix} \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 5 & -10 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 100 & -180 \\ -180 & 328 \end{bmatrix} \Rightarrow Q = \begin{bmatrix} 10 & 0 \\ -18 & 2 \end{bmatrix}$$

Thus:

$$A = \begin{bmatrix} 0.2 & 0 \\ 1 & 0.5 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ -18 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$$

Indeed,

$$AA' = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

6.2.4 Method #3: Sign Restrictions

The previous two examples yield identification in the sense that the shocks are uniquely identified. Sign identification, on the other hand, is based on qualitative restriction involving the sign of some shocks on some variables.

▷ Remarks:

* In classical statistics approach, this does not deliver exact identification since there can be many A consistent with such a restriction. That is, for each parameter of the impulse response functions, we will have an dmissible set of values.
* Increasing the number of restrictions can be helpful in reducing the number of A consistent with such restrictions.