ECON 7335 INFORMATION, LEARNING AND EXPECTATIONS IN MACRO

LECTURE 1: BASICS

KRISTOFFER P. NIMARK

1. Bayes' Rule

Definition 1. *Bayes' Rule. The probability of event A occurring conditional on the event B having occurred is given by*

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \tag{1.1}$$

as long as $p(B) \neq 0$.

It can be derived from the definition of conditional probability

$$p(A \mid B) \equiv \frac{p\left(A \cap B\right)}{p(B)} \tag{1.2}$$

and that the probability of A and B occurring is the same as the probability of B and A occurring implying

$$p(A \mid B)p(B) = p(B \mid A)p(A) \tag{1.3}$$

which can be rearranged to yield Bayes' Rule.

Bayes Rule is general and applies whether A and B are discrete or continuous random variables. It can be used to update a prior belief $p(x)$ about the latent variable x conditional on the signal y. Bayes Rule then gives the posterior belief

$$p\left(x \mid y\right) = \frac{p(y \mid x)p(x)}{p(y)} \tag{1.4}$$

Date: February 11, 2021.

1.1. **Example with binary variables and signals (adapted from O'Hara 1995).** Consider a market maker that holds inventory of an asset that can have either a High or Low value. The market maker's prior belief that the value of the asset is high is denoted $p(H) = \frac{1}{2}$. There are two types of traders, informed and uninformed and the market maker is equally likely to meet either type. An informed trader always buys when the value is high and always sells when the value is low. An uninformed trader is equally likely to buy or sell. What is the posterior probability that the value is high if the first trade is a sale (S)? Bayes' Rule states that

$$p(H \mid S) = \frac{p(S \mid H)p(H)}{p(S)}.$$

Start by finding the probability of a sale conditional on the value being high

$$p(S \mid H) = \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \tag{1.5}$$

and low

$$p(S \mid L) = \frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4} \tag{1.6}$$

We also need to find the (unconditional) probability of sale

$$p(S) = p(S \mid H)p(H) + p(S \mid L)p(L) \tag{1.7}$$

Plug into Bayes' Rule to get the posterior conditional on a sale

$$p\left(H \mid S\right) = \frac{p(S \mid H)p(H)}{p(S \mid H)p(H) + p(S \mid L)p(L)} \tag{1.8}$$

$$= \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{1}{4} \times \frac{1}{2} + \frac{3}{4} \times \frac{1}{2}} = \frac{1}{4} \tag{1.9}$$

The same steps can be used to find the posterior probability that the value is high if the first transaction is a buy (B)

$$p\,(H \mid B) = \frac{p(B \mid H)p(H)}{p(B \mid H)p(H) + p(B \mid L)p(L)} \tag{1.10}$$

$$= \frac{\frac{3}{4} \times \frac{1}{2}}{\frac{1}{4} \times \frac{1}{2} + \frac{3}{4} \times \frac{1}{2}} = \frac{3}{4} \tag{1.11}$$

1.1.1. *Updating beliefs when new information arrives.* If the arrival of traders is independent over time, it is straightforward to update the posterior after a second transaction takes place by simply using the posterior after the first observation as the prior in the update. If both the first and second transaction are sales, the posterior probability is then given by

$$p\,(H \mid S, S) = \frac{\frac{1}{4} \times \frac{1}{4}}{\frac{1}{4} \times \frac{1}{4} + \frac{3}{4} \times \frac{3}{4}} = \frac{1}{10} \tag{1.12}$$

However, if the first transaction is a sale and the second a buy, we get

$$p\,(H \mid S, B) = \frac{\frac{3}{4} \times \frac{1}{4}}{\frac{3}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{3}{4}} = \frac{1}{2} \tag{1.13}$$

i.e. the second signal cancels the first and posterior equals the original prior belief.

Useful fact: Under quite general conditions, beliefs that are updated using Bayes' Rule converge almost surely to the truth.

1.2. **Bayes' Rule and jointly normally distributed variables.** To illustrate the usefulness of Bayes' Rule, we can apply it to a simple bivariate setting. Let the prior about the latent variable x be normally distributed

$$x \sim N(0, \sigma_x^2) \tag{1.14}$$

and the signal y be the sum of the true x plus a normally distributed noise term so that

$$y = x + \eta : \eta \sim N(0, \sigma_\eta^2) \tag{1.15}$$

We then have all the ingredients need to apply Bayes Rule to find the posterior $p(x \mid y)$:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \tag{1.16}$$

$$p(y) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_\eta^2}} e^{-\frac{1}{2}\frac{(x+\eta)^2}{(\sigma_x^2+\sigma_\eta^2)}} \tag{1.17}$$

$$p(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma_\eta} e^{-\frac{1}{2}\frac{(y-x)^2}{\sigma_\eta^2}} \tag{1.18}$$

Plugging these expressions into Bayes' Rule gives

$$p(x \mid y) = \frac{\frac{1}{\sqrt{2\pi}\sigma_\eta} e^{-\frac{1}{2}\frac{(y-x)^2}{\sigma_\eta^2}} \times \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}}}{\frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_\eta^2}} e^{-\frac{1}{2}\frac{(x+\eta)^2}{(\sigma_x^2+\sigma_\eta^2)}}} \tag{1.19}$$

which can be simplified to

$$p(x \mid y) = \frac{1}{\sqrt{2\pi}\left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_\eta^2}\right)^{-\frac{1}{2}}} e^{-\frac{1}{2}\frac{\left(x-\frac{\sigma_x^2}{(\sigma_x^2+\sigma_\eta^2)}y\right)^2}{\left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_\eta^2}\right)^{-1}}} \tag{1.20}$$

The posterior distribution $p(x \mid y)$ is thus normally distributed with mean $\frac{\sigma_x^2}{(\sigma_x^2+\sigma_\eta^2)}y$ and variance $\left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_\eta^2}\right)^{-1}$. This illustrates two points: First, normally distributed priors combined with normally distributed noise in the signal result in normally distributed posteriors. This is an extremely useful property of normal distributions. Second, the posterior variance of x is lower than the prior variance, i.e.

$$\left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_\eta^2}\right)^{-1} < \sigma_x^2 \tag{1.21}$$

More information thus reduces uncertainty about x.

2. The Projection Theorem

This section explains how orthogonal projections can be used to find least squares predictions of random variables. We start by defining some concepts needed for stating the projection theorem. For more details about the projection theorem, see for instance Chapter 2 of Brockwell and Davis (2006) or Chapter 3 in Luenberger (1969).

Definition 2. *(Inner Product Space) A real vector space $\mathcal{H}$ is said to be an inner product space if for each pair of elements x and y in $\mathcal{H}$ there is a number $\langle x, y\rangle$ called the inner product of x and y such that*

$$\langle x, y\rangle = \langle y, x\rangle \tag{2.1}$$

$$\langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle \text{ for all } x, y, z \in \mathcal{H} \tag{2.2}$$

$$\langle \alpha x, y\rangle = \alpha \langle x, y\rangle \text{ for all } x, y \in \mathcal{H} \text{ and } \alpha \in \mathbb{R} \tag{2.3}$$

$$\langle x, x\rangle \geq 0 \text{ for all } x \in \mathcal{H} \tag{2.4}$$

$$\langle x, x\rangle = 0 \text{ if and only if } x = \mathbf{0} \tag{2.5}$$

Definition 3. *(Norm) The norm of an element x of an inner product space is defined to be*

$$\|x\| = \sqrt{\langle x, x\rangle} \tag{2.6}$$

Definition 4. *(Cauchy Sequence) A sequence $\{x_n, n = 1, 2, ...\}$ of elements of an inner product space is said to be Cauchy sequence if*

$$\|x_n - x_m\| \to 0 \text{ as } m, n \to \infty$$

i.e. for every $\varepsilon > 0$ there exists a positive integer $N(\varepsilon)$ such that

$$\|x_n - x_m\| < \varepsilon \text{ as } m, n > N(\varepsilon)$$

Definition 5. *(Hilbert Space) A Hilbert space $\mathcal{H}$ is an inner product space which is complete, i.e. every Cauchy sequence $\{x_n\}$ converges in norm to some element $x \in \mathcal{H}$.*

Theorem 1. *(The Projection Theorem) If $\mathcal{M}$ is a closed subspace of the Hilbert Space $\mathcal{H}$ and $x \in \mathcal{H}$, then*
(i) there is a unique element $\widehat{x} \in \mathcal{M}$ such that

$$\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$$

and
(ii) $\widehat{x} \in \mathcal{M}$ and $\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\widehat{x} \in \mathcal{M}$ and $(x - \widehat{x}) \in \mathcal{M}^{\perp}$ where $\mathcal{M}^{\perp}$ is the orthogonal complement to $\mathcal{M}$ in $\mathcal{H}$.

The element $\widehat{x}$ is called the orthogonal projection of x onto $\mathcal{M}$.

Proof. We first show that if $\widehat{x}$ is a minimizing vector then $x - \widehat{x}$ must be orthogonal to $\mathcal{M}$. Suppose to the contrary that there is an element $m \in \mathcal{M}$ which is not orthogonal to the error $x - \widehat{x}$. Without loss of generality we may assume that $\|m\| = 1$ and that $\langle x - \widehat{x}, m \rangle = \delta \neq 0$. Define the vector $m_1 \in \mathcal{M}$

$$m_1 \equiv \widehat{x} + \delta m \tag{2.7}$$

We then have that

$$\begin{aligned}
\|x - m_1\|^2 &= \|x - \widehat{x} - \delta m\|^2 &(2.8)\\
&= \|x - \widehat{x}\|^2 - \langle x - \widehat{x}, \delta m \rangle - \langle \delta m, x - \widehat{x} \rangle + |\delta|^2 &(2.9)\\
&= \|x - \widehat{x}\|^2 - |\delta|^2 &(2.10)\\
&< \|x - \widehat{x}\|^2 &(2.11)
\end{aligned}$$

where the second line follows from (2.2) and the definition of the norm, the third line comes from the fact that $\langle x - \widehat{x}, \delta m \rangle = \langle \delta m, x - \widehat{x} \rangle = |\delta|^2$. The inequality on the last line follows from the fact that $|\delta|^2 > 0$. We then have a contradiction: $\widehat{x}$ cannot be the element in $\mathcal{M}$ that minimizes the norm of the error if $\delta \neq 0$ since $\|x - m_1\|^2$ then is smaller than $\|x - \widehat{x}\|^2$.

We now show that if $x - \widehat{x}$ is orthogonal to $\mathcal{M}$ then it is the unique minimizing vector. For any $m \in \mathcal{M}$ we have that

$$\|x - m\|^2 = \|x - \widehat{x} + \widehat{x} - m\|^2 \tag{2.12}$$

$$= \|x - \widehat{x}\|^2 + \|\widehat{x} - m\|^2 \tag{2.13}$$

$$> \|x - \widehat{x}\|^2 \text{ for } \widehat{x} \neq m \tag{2.14}$$

$\square$

Properties of Projection Mappings. Let $\mathcal{H}$ be a Hilbert space and let $P_{\mathcal{M}}$ be a projection mapping onto a closed subspace $\mathcal{M}$. Then

(i) each $x \in \mathcal{H}$ has a unique representation as a sum of an element in $\mathcal{M}$ and an element in $\mathcal{M}^{\perp}$, i.e.

$$x = P_{\mathcal{M}}x + (I - P_{\mathcal{M}})x \tag{2.15}$$

(ii) $x \in \mathcal{M}$ if and only if $P_{\mathcal{M}}x = x$

(iii) $x \in \mathcal{M}^{\perp}$ if and only if $P_{\mathcal{M}}x = 0$

(iv) $\mathcal{M}_1 \subseteq \mathcal{M}_2$ if and only if $P_{\mathcal{M}1}P_{\mathcal{M}2}x = P_{\mathcal{M}1}$

(v) $\|x\|^2 = \|P_{\mathcal{M}}x\|^2 + \|(I - P_{\mathcal{M}})\,x\|^2$

The definitions and the proofs above refer to Hilbert spaces in general. We now define the space relevant for most of time series analysis.

Definition 6. *(The space $L^2(\Omega, F, P)$) We can define the space $L^2(\Omega, F, P)$ as the space consisting of all collections C of random variables X defined on the probability space (Ω, F, P) satisfying the condition*

$$EX^2 = \int_{\Omega} X(\omega)P\,(d\omega) < \infty \tag{2.16}$$

and define the inner product of this space as

$$\langle X, Y \rangle = E(XY) \text{ for any } X, Y \in C \tag{2.17}$$

Least squares estimation via the projection theorem. The inner product space L^2 satisfies all of the axioms above. Noting that the inner product definition (2.17) corresponds to a covariance means that we can use the projection theorem to find the minimum variance estimate of a vector of random variables with finite variances as a function of some other random variables with finite variances. That is, both the information set and the variables we are trying to predict must be elements of the relevant space and since $\langle X, Y \rangle = E(XY)$ implies that an estimate $\widehat{x}$ that minimizes the norm of the estimation error $\|x - \widehat{x}\|$ also minimizes the variance since

$$\|x - \widehat{x}\| = \sqrt{E(x - \widehat{x})(x - \widehat{x})'} \tag{2.18}$$

To find the estimate $\widehat{x}$ as a linear function of y simply use that

$$\begin{aligned} \langle x - \beta y, y \rangle &= E\left[(x - \beta y) y'\right] \\ &= 0 \end{aligned} \tag{2.19}$$

and solve for β

$$\beta = E(xy') \left[E(yy')\right]^{-1} \tag{2.20}$$

The advantage of this approach is that once you have made sure that the variables y and x are in a well defined inner product space, there is no need to minimize the variance directly. The projection theorem ensures that an estimate with orthogonal errors is the (linear) minimum variance estimate.

Two useful properties of linear projections.

- If two random variables X and Y are Gaussian, then the projection of Y onto X coincides withe the conditional expectation $E(Y \mid X)$.
- If X and Y are not Gaussian, the linear projection of Y onto X is the minimum variance linear prediction of Y given X.

References

[1] Brockwell, P.J. and R.A. Davis, 2006, *Time Series: Theory and Methods*, Springer-Verlag.

[2] Luenberger, D., (1969), *Optimization by Vector Space Methods*, John Wiley and Sons, Inc., New York.

[3] O'Hara, M., 1995. *Market microstructure theory*, Blackwell, Cambridge, MA.