THE PROJECTION THEOREM

KRISTOFFER P. NIMARK

These notes 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 1. *(Inner Product Space) An real vector space H is said to be an inner product space if for each pair of elements x and y in H there is a number $\langle x, y\rangle$ called the inner product of x and y such that*

$$\langle x, y\rangle = \langle y, x\rangle \tag{0.1}$$

$$\langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle \text{ for all } x, y, z \in \mathcal{H} \tag{0.2}$$

$$\langle \alpha x, y\rangle = \alpha \langle x, y\rangle \text{ for all } x, y \in \mathcal{H} \text{ and } \alpha \in \mathbb{R} \tag{0.3}$$

$$\langle x, x\rangle \geq 0 \text{ for all } x \in \mathcal{H} \tag{0.4}$$

$$\langle x, x\rangle = 0 \text{ if and only if } x = \mathbf{0} \tag{0.5}$$

Definition 2. *(Norm) The norm of an element x of an inner product space is defined to be*

$$\|x\| = \sqrt{\langle x, x\rangle} \tag{0.6}$$

Definition 3. *(Cauchy Sequence) A sequence $\{x_n, n = 1, 2, ...\}$ of elements of an inner product space is said to be Cauchy sequence if*

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

Date: June 14, 2014.

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

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

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

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

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

Proof. We first show that if $\widehat{x}$ is a minimizing vector then $x - \widehat{x}$ must be orthogonal to M. Suppose to the contrary that there is an element $m \in M$ 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 $m_1 \in M$

$$m_1 \equiv \widehat{x} + \delta m \tag{0.7}$$

We then have that

$$\begin{aligned}
\|x - m_1\|^2 &= \|x - \widehat{x} - \delta m\|^2 && (0.8)\\
&= \|x - \widehat{x}\|^2 - \langle x - \widehat{x}, \delta m\rangle - \langle \delta m, x - \widehat{x}\rangle + |\delta|^2 && (0.9)\\
&= \|x - \widehat{x}\|^2 - |\delta|^2 && (0.10)\\
&< \|x - \widehat{x}\|^2 && (0.11)
\end{aligned}$$

where the third line comes from the fact that $\langle x-\widehat{x},\delta m\rangle=\langle \delta m,x-\widehat{x}\rangle=|\delta|^2$ when $\|m\|=1$. 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 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 M then it is the unique minimizing vector. For any $m\in M$ we have that

$$\begin{aligned}\|x-m\|^2 &= \|x-\widehat{x}+\widehat{x}-m\|^2 &(0.12)\\ &= \|x-\widehat{x}\|^2+\|\widehat{x}-m\|^2 &(0.13)\\ &> \|x-\widehat{x}\|^2 \text{ for } \widehat{x}\neq m &(0.14)\end{aligned}$$

□

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

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

$$x=P_{\mathcal{M}}x+(I-P_{\mathcal{M}})x \tag{0.15}$$

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

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

(iv) $M_1\subseteq M_2$ 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.

(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

$$EX^2 = \int_\Omega X^2(\omega) P\,(d\omega) < \infty \tag{0.16}$$

and define the inner product of this space as

$$\langle X, Y \rangle = E\,(XY) \text{ for any } X, Y \in C \tag{0.17}$$

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 (0.17) 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

$$\|x - \widehat{x}\| = \sqrt{E\,(x - \widehat{x})\,(x - \widehat{x})'} \tag{0.18}$$

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

$$\begin{aligned} \langle x - \beta y, y \rangle &= E\,[(x - \beta y)\,y'] \\ &= 0 \end{aligned} \tag{0.19}$$

and solve for β

$$\beta = E\,(xy')\,[E\,(yy')]^{-1} \tag{0.20}$$

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.