The projection theorem

Hans Karlsen

UiB: 2021-02-10 22:09:12

Motivation for the projection theorem

1. Introduction

The projection theorem is a central theoretical result which you have in high school, collage and at the university. It turns out that the proof is not difficult, but the context is somewhat abstract. However we will start by some examples before we present the restult. The projection theorem is important for defining the one-step predictor and the corresponding one step prediction error. These concepts help us to find the optimal [correct] linear models and they are also involved in estimation algorthms giving both estimators and their basic properties.

Table 1: Discrete - continuous

time \state space	discrete statespace	continuous statespace
discrete time	Markov chain	time series
continuous time	jump processes	Brownian motion

Table 2: Data and models

type	case	# points	sample	models	variables
univariate	x	n	$x_1, \ldots, x_n$	F_X	X
bivariate	(x, y)	n	$(x_1, y_1), \ldots, (x_n, y_n)$	$F_{X,Y}$	(X, Y)
multivariate	$\mathbf{x}$	n	$\mathbf{x}_1, \ldots, \mathbf{x}_n$	$F_{\mathbf{X}}$	$(X_1, \ldots, X_p)$
time series	$\{x_t,\, t \in \mathbb{Z}\}$	1	$x_1, \ldots, x_n$	?	$\{X_t,\, t \in \mathbb{Z}\}$

It is possible to observe a panel of with relatively many short pieces of many time series or a sample of few long stretches of time series (longitudinell study). We also have spatial series. In recent (last 20 years times there has been an increasing interest in discrete valued time series. An innovative aspect here compared to more classical Markov chain models is definition of analogues to the ARMA models which central for continuous case.

A finite dimensional vector space of stochastic variables

Let stochastic vector $\boldsymbol{W} = (W_1, \ldots, W_{n+1})' = (\mathbf{X}', Y)$ be defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the covariance matrix is finite and with full rank and $\mathbb{E}\,\boldsymbol{W} = \mathbf{0}$,

$$\mathbb{\Sigma}_W = \begin{bmatrix} \mathbb{\Sigma}_X & \mathbb{\Sigma}_{XY} \\ \mathbb{\Sigma}_{YX} & \mathbb{\Sigma}_Y \end{bmatrix}$$

Let

$$\mathcal{V} = \text{span}\{W_1, \ldots, W_{n+1}\}$$

$$= \{V: \ V = \sum_{j=1}^{n+1} a_j W_j \ , \quad \mathbf{a} \in \mathbb{R}^{n+1}\}$$

It is a vector space

Then

1. $\mathbf{0} \in \mathcal{V}$.
2. If $V \in V$ and $c \in \mathbb{R}$ then $cV \in \mathcal{V}$.
3. For U and V are in $\mathcal{V}$ then $U + V = V$.
4. For U and V are in $\mathcal{V}$ then $U + V = V + U$.
5. The distributional law holds ;$c(U + V) = cU + cV$.
6. For $U, V \in \mathcal{V}$, $U - V = U + (-1)V$.
7. Any vector U has an inverse $-U$ with respect to the vector addition.

Hence $\mathcal{V}$ is a vector space. The dimension is equal to the rank of the covariance matrix for $\mathbf{X}$. This implies that this vector space is finite dimensional with dimension $n + 1$.

Inner product , length and distance

We want to have a dot product [inner product] on this space. Define

$$\langle U, V \rangle = \mathbb{E}\, UV = \mathrm{Cov}(U, V) \qquad \text{when means are zero}$$

The inner product also give a distance;

$$\|U - V\| = \langle U - V, U - V \rangle^{1/2}$$

and length

$$\|U\| = \langle U, U \rangle^{1/2}$$

REMARK 1. We see that the distance here is the mean square distance (square root of) and the length equals the standard deviation.

A finite set of stochastic variables

EXAMPLE 1. Let

$$\mathcal{V}_0 = \text{span}\{X_1, \ldots, X_n\}$$

$$= \{U : \ U = \sum_{j=1}^{n} b_j X_j \ , \quad \mathbf{b} \in \mathbb{R}^n\}$$

Then $\mathcal{V}_0 \ \subset \mathcal{V}$ and we also see that $\mathcal{V}_0$ is a subspace of $\mathcal{V}$.

MSE

Let $V \in \mathcal{V}$ that is not contained in $\mathcal{V}_0$. Can we find a vector $\widehat{V} \in \mathcal{V}_0$ that is closest to V, i.e.

$$\widehat{V} = \underset{U \in \mathcal{V}_0}{\operatorname{argmin}} \|V - U\|^2 \tag{1}$$

The problem is to find a (the) point in $\mathcal{V}_0$ that is closest to a given point in $\mathcal{V}$. Since $\widehat{V} \in \mathcal{V}_0$ we have that

$$\widehat{V} = \sum_{j=1}^{n} \beta_j U_j$$

Can we find $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_n\big)$ such that (1) holds?

Orthogonal complement

Suppose that $\widehat{V}$ satisfies (1),

$$V = \widehat{V} + (V - \widehat{V}) = \widehat{V} + e, \text{ say}$$

where $e = (V - \widehat{V})$ is the residual, the part of V that cannot be approximated by any vector in $\mathcal{V}_0$. It is easy show that e is uncorrelated with all $U \in \mathcal{V}_0$. This motivates

$$\begin{aligned} \mathcal{V}_0^{\perp} &\stackrel{\text{def}}{=} \{V \in \mathcal{V}\colon\ \langle V, U\rangle \equiv 0, \quad U \in \mathcal{V}_0\} \\ &= \{\text{The set of vectors that are orthogonal to all vectors } \mathcal{V}_0\} \end{aligned}$$

$\mathcal{V}_0^{\perp}$ is called the orthognal complement of $\mathcal{V}_0$ and it is subspace also when $\mathcal{V}_0$ is just a set of fixed vectors.

Predictions equations

If we can choose $\widehat{V}$ so that $e \in \mathcal{V}_0^{\perp}$ then (1) holds. In order to do that we must solve the prediction equations;

$$V - \widehat{V} \in \mathcal{V}_0^{\perp}$$

$$\Downarrow$$

$$\mathrm{Cov}\big(V - \sum_{j=1}^{n} \beta_j X_j,\, X_k\big) = 0, \quad k = 1, \ldots, n$$

$$\Downarrow$$

$$\mathrm{Cov}(V, X_k) = \sum_{j=1}^{n} \beta_j \mathrm{Cov}(X_j, X_k), \qquad k = 1, \ldots, n$$

Solution

The solution is

$$\boldsymbol{\beta} = \mathbb{\Sigma}_X^{-1}\mathbb{\Sigma}_{X,V}$$

Suppose that $V = Y$,

$$\boldsymbol{\beta} = \mathbb{\Sigma}_X^{-1}\mathbb{\Sigma}_{X,Y}$$

REMARK 2. If $\boldsymbol{W}$ is multinormal distributed then

$$Y = \sum_{j=1}^{n} \beta_j X_j + e$$

where with $\mathbf{X}$ and e indepedent. This is model version of the multiple regression formula.

The predictor variance

$$\begin{aligned}\mathrm{Var}(e) &= \mathbb{E}\, e^2 = \mathrm{Cov}(e,e) = \|e\|^2 = \|V - \widehat{V}\|^2 \\ &= \|V\|^2 - \|\widehat{V}\|^2\end{aligned}$$

and here we have that

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^n}{\operatorname{argmin}} \|e\|^2 \tag{2}$$

Hilbert space

DEFINITION 1. A Hilbert space, $\mathcal{H}$, is a vector space over $\mathbb{R}$ (or eventually over $\mathbb{C}$) with an inner product, $\langle \cdot, \cdot \rangle$ such Cauchy sequences are convergent.

DEFINITION 2. The inner product satisfies

$$\begin{aligned} \langle v, v \rangle &\geq 0 && \text{it is zero iff } v = 0 \\ \langle u + cw, v \rangle &= \langle u, v \rangle + c\langle u, w \rangle && \text{linarity} \\ \langle u, w \rangle &= \langle w, u \rangle && \text{symmetry} \end{aligned}$$

REMARK 3. In the complex case the symmetry is replaced by $\langle u, w \rangle = \overline{\langle w, u \rangle}$.

The inner product is bilinear

REMARK 4. The inner product is bilinear

$$\left\langle \sum_{i=1}^{m} u_i, \sum_{j=1}^{n} v_j \right\rangle = \sum_{i=1}^{m} \left\langle u_i, \sum_{i=1}^{n} v_j \right\rangle = \sum_{i=1}^{m} \overline{\left\langle \sum_{i=1}^{n} v_j, u_i \right\rangle}$$
$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \overline{\langle v_j, u_i \rangle} = \sum_{i=1}^{m} \sum_{j=1}^{n} \langle u_i, v_j \rangle$$

and

$$\langle au, bv \rangle = a\,\overline{b}\,\langle u, v \rangle$$

This is for the complex case and therefore also true in real case.

The squared distance of a difference

PROBLEM 1.1. Prove that

$$\|a-b\|^2 = \|a\|^2 - \langle a,b\rangle - \langle b,a\rangle + \|b\|^2$$

PROBLEM 1.2. Verify the parallelogram law

$$\|a-b\|^2 + \|a+b\|^2 = 2\|a\|^2 + 2\|b\|^2.$$

Hint: The two cross terms in the first term on left hand side cancel with the two cross terms in the second term on the same side.

Cauchy sequences

EXAMPLE 2. A Cauchy sequence of real numbers $\{a_n,\, n \geq 1\}$ satisfies

$$\lim_{n,m\to\infty} |a_n - a_m| = 0$$

By definition of the real numbers a Cauchy sequence is convergent, i.e, there exist an $a \in \mathbb{R}$ such that

$$\lim_{n\to\infty} a_n = a$$

REMARK 5. A convergent sequence of real numbers is a Cauchy sequence.

EXAMPLE 3. A Cauchy sequence of rational numbers, $\mathbb{Q}$, is not necessarily convergent in $\mathbb{Q}$.

$\mathbb{R}^n$

EXAMPLE 4. The set of real n-tuples $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbb{R}^n$, is a finite dimensional Hilbert space with,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{j=1}^{n} x_j y_j$$

EXAMPLE 5. The set of complex n-tuples $\mathbf{z} = (z_1, \ldots, z_n)'$, $\mathbb{C}^n$, is a finite dimensional Hilbert space with $\langle \mathbf{z}, \mathbf{w} \rangle = \mathbf{z}^T \overline{\mathbf{w}}$. This means that

$$\langle \mathbf{z}, \mathbf{w} \rangle = \sum_{j=1}^{n} z_j \overline{w} j$$

Orthogonal projection on a vector

PROPOSITION 1. Let $a, b \in \mathcal{H}$. Then the (orthogonal) projection of a onto b is given by

$$a_b \stackrel{\text{def}}{=} \frac{\langle a, b \rangle}{\|b\|^2} b$$

and pythagoras holds,

$$\|a\|^2 = \|a_b\|^2 + \|a - a_b\|^2$$

Proof of Pythagoras

PROOF.
We can write

$$a = a_b + (a - a_b)$$

and we must prove that the cross term vanish;

$$\langle a - a_b, b\rangle = \langle a, b\rangle - \langle a_b, b\rangle = \langle a, b\rangle - \frac{\langle a, b\rangle}{\|b\|^2}\langle b, b\rangle = 0$$

□

REMARK 6. Note that $\|a\|^2 > \|a_b\|^2$ unless a is parallel with b, i.e. $a_b = 0$.

CS

COROLLARY. [Cauchy-Schwartz]

$$\langle a, b\rangle^2 \leq \|a\|^2 \|b\|^2$$

The inequality is strict unless b is parall with a.

PROOF.
By Proposition

$$\|a_b\|^2 \leq \|a\|^2$$
$$\Downarrow$$
$$\frac{\langle a, b\rangle^2}{\|b\|^4}\|b\|^2 \leq \|a\|^2$$
$$\Downarrow$$
$$\langle a, b\rangle^2 \leq \|a\|^2 \|b\|^2$$

□

triangle inequality

PROBLEM 1.3. Show that the triangle inequality holds;

$$\|a+b\|^2 \leq \|a\|^2 + \|b\|^2$$

Motivation

i) The Durbin Levins [DL] algorithm which gives an operational formula for the partial autocovariance function and a recursive efficient method to calculate PACF.

ii) DL also opens up for study the asymptotic behaviour of PACF.

iii) The innovation algorithm [IA] is related to moving average in an analogous way as DL is related to a autoregressive process.

iv) Any weakly stationary process which has an ACF that is nonzero for only a finite number of lags has a moving average representation.

v) The autoregressive model that comes from DL is causal.

vi) The IA algorithm for a moving average needs only a fixed number of lags and the use of memory is bounded.

vii) A causal autoregressive model has residual process with minimum variance.

Motivation II

viii) An invertible moving average model has a residual variance of maximum size among all possible representations.

ix) Wold- decomposition of a time series.

x) The causal representation of an autoregressive model can be found by application of the innovation algorithm.

xi) Convergence of IA.

xii) Convergence of DL.

Other questions

2. Other problems

xiii) The Yule Walker equation for an causal autoregressive model gives the regression parameter of order p as a function of the first p lags of the ACF. However, these equations can be turned around and that is a programming challenge.

xiv) The covariance structure of a causal and invertible ARMA model follows quite easily in terms af an efficient recursive system.

xv) An irreducible arma model with a unit root in the autoregressive part does not have any stationary solution. A proof for $p = 1$ is close to be trivial, but the proof for p strictly greater than p is more demanding.

xvi) Unit roots in the moving average part of an ARMA model do not affect the existence of a unique solution causal and extended invertible solution, but they cannot be eleminated.

xvii) How does unit roots in a moving avrerage model influence the convergence rate of an application of IA?

Frequency perspective

3. The Spectral distribution

xviii) The covariance structure can be described by the ACVF or by the spectral distribution.

xix) The spectral distribution is a distribution when it is properly scaled.

xx) ARMA models have a rational spectral density.

xxi) Spectral methods are needed for establishing the uniqueness of a causal an invertible ARMA model.

xxii) The empirical version of the spectral distribution is the periodogram.

xxiii) Nonparametric estimation of a spectral density is closely related to kernal estimation of an ordinary density.

xxiv) Peridogram analysis is related for discrete Fourier analyse. In particular it is related to Fast Fourier Transform.

The projection theorem

Let $\mathcal{H}$ be a Hilbert space and let $\mathcal{H}_0$ be a closed subspace of $\mathcal{H}$. This means that $\mathcal{H}_0$ itself is a Hilbert space. We use this notation.

THEOREM. [Projection theorem]

There exists a unique linear projection $\mathcal{P}$ from $\mathcal{H}$ onto $\mathcal{H}_0$. It satisfies.

$\mathcal{P}\colon \mathcal{H} \rightsquigarrow \mathcal{H}_0$	the range is contained in $\mathcal{H}_0$
$\mathcal{P}(u+av) = \mathcal{P}(u) + a\,\mathcal{P}(v)$	it is a linear operator
$\mathcal{P}(u) = u$ on $\mathcal{H}_0$	it is dempotent

What does theorem tell us?

i) Given u and a subspace $\mathcal{H}_0$. Then there exists a unique $\widehat{u}$ in the subspace $\mathcal{H}_0$ that is closed to u among all vectors in the subspace. We could say that the distance between $\widehat{u}$ and u is the distance between u an the subspace.

ii) The map $u \rightsquigarrow \widehat{u}$ is linear.

EXAMPLE 6. Let $\{X_t,\, t \in \mathbb{Z}\}$ be a stationary time series with zero mean.

$$\mathcal{P}_n \stackrel{\text{def}}{=} \mathcal{P}_{\mathcal{H}_n}, \qquad \mathcal{H}_n \stackrel{\text{def}}{=} \text{span}\big\{X_1, \ldots, X_n\big\},$$
$$\widehat{X}_{n+1} \stackrel{\text{def}}{=} \mathcal{P}_n(X_{n+1})$$

is the one-step predictor based on $\{X_1, \ldots, X_n\}$. It should satisfy

i) $\widehat{X}_{n+1} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \cdots + \phi_{nn}X_1$.

ii) $\langle X_{n+1} - \widehat{X}_{n+1},\, X_{n+1-j}\rangle \equiv 0$.

The equations for ϕ_n

We insert $\widehat{X}_{n+1} = \sum_{k=1}^{n} \phi_{nk} X_{n+1-k}$ in the prediction equations.

$$
\begin{aligned}
0 &= \langle X_{n+1} - \widehat{X}_{n+1},\, X_{n+1-j} \rangle \\
&= \langle X_{n+1},\, X_{n+1-j} \rangle - \langle \widehat{X}_{n+1},\, X_{n+1-j} \rangle \\
&= \langle X_{n+1},\, X_{n+1-j} \rangle - \langle \sum_{k=1}^{n} \phi_{nk} X_{n+1-k},\, X_{n+1-j} \rangle \\
&= \langle X_{n+1},\, X_{n+1-j} \rangle - \sum_{k=1}^{n} \phi_{nk} \langle X_{n+1-k},\, X_{n+1-j} \rangle \\
&= \gamma\big(n+1-(n+1-j)\big) - \sum_{k=1}^{n} \phi_{nk} \gamma\big(n+1-k-(n+1-j)\big) \\
&= \gamma(j) - \sum_{k=1}^{n} \phi_{nk} \gamma(j-k), \quad j = 1, \ldots, n
\end{aligned}
$$

Matrix form

Let

$$\boldsymbol{\gamma}_n \stackrel{\text{def}}{=} \begin{bmatrix} \gamma(1) \\ \vdots \\ \gamma(n) \end{bmatrix}, \qquad \boldsymbol{\phi}_n = \begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{bmatrix}$$

$$\mathbb{\Gamma}_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(n-1) & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(n-3) & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(n-2) & \gamma(n-3) & \dots & \gamma(0) & \gamma(1) \\ \gamma(n-1) & \gamma(n-2) & \dots & \gamma(1) & \gamma(0) \end{bmatrix}$$

The matrix form for the one-step predictor is

$$\boldsymbol{\gamma}_n = \mathbb{\Gamma}_n \boldsymbol{\phi}_n$$

The predictor

The predictor equations solves to

$$\begin{aligned}\boldsymbol{\phi}_n &= \mathbb{\Gamma}_n^{-1}\boldsymbol{\gamma}_n \\ &= \mathbb{R}_n^{-1}\boldsymbol{\rho}_n\end{aligned}$$

The predictor can be written in compact form as

$$\widehat{X}_{n+1} = \boldsymbol{\phi}_n^T \mathbf{X}_n^{\text{rev}}$$

R notation

REMARK 7. In R notation

```
    # We need \gamma(0) \ldots  \gamma(n)

    mPhi<-matrix(0,n,n)
    mGa <-matrix(0,n+1,n+1)
    bga<- mGa[2:(n+1),1]

#__________Assume that we have values for mGa________

#_____________loop for phi_______
for(k in 1:n) mPhi[k,1:k] <- solves(mGa[1:k,1:k] ,\bga[1:k] )
```

Proof of Theorem .

Step 1 Existence

We first prove that for any $v \in \mathcal{H}$ there exists a unique $\widehat{v}$ such that

$$\widehat{v} = \operatorname{argmin} \left\{ \|v - u\|^2, \quad u \in \mathcal{H}_0 \right\}.$$

Let

$$d^2 = \inf_{u \in \mathcal{H}_0} \|v - u\|^2.$$

Greatest lower bound

REMARK 8. By definition of the real numbers, each subset of A of real numbers that is bounded from below has unique greatest lower bound. You get that bound by writing $x = \inf A$. If $x \in A$, then $x = \min(A)$. In any case we have by definition of x that for any $\epsilon > 0$ there exist an $a \in A$ such that $0 \geq a - x \geq \epsilon$. Therefore we can choose a sequence $\{a_n\} \subseteq A$ such that $a_n \downarrow x$. Here

$$A = \left\{ \|v - u\|^2 \colon u \in \mathcal{H}_0 \right\}.$$

The lower bound for A is zero.

Choose a sequence $\{u_n\} \in \mathcal{H}_0$ such that $\lim_{n\to\infty} \|v - u_n\|^2 = d^2$. This is possible by definition of d^2.

Use of the parallogram law

We use the parallelogram law

$$\begin{aligned} &\|a-b\|^2+\|a+b\|^2=2\|a\|^2+2\|b\|^2 \\ &\qquad\qquad\qquad\Downarrow \\ &\|a-b\|^2=2\|a\|^2+2\|b\|^2-\|a+b\|^2 \end{aligned} \tag{3}$$

Let $a=u_n-v$ and $b=u_m-v$ for any $n,m\geq 1$. Then

$$\begin{aligned} a-b&=(u_n-v)-(u_m-v)=u_n-u_m \\ a+b&=(u_n-v)+(u_m-v)=u_n+u_m-2v=2(2^{-1}(u_n+v_m)-v) \end{aligned}$$

Since $x\stackrel{\text{def}}{=}2^{-1}(u_n+v_m)\in\mathcal{H}_0$ we have that the squared distance $\|x-v\|^2\geq d^2$,

$$\begin{aligned} \|a-b\|^2&=\|u_n-u_m\|^2 \\ \|a+b\|^2&=\|2(2^{-1}(u_n+u_m-v))\|^2=4\|2^{-1}(u_n+u_m)-v\|^2\geq 4d^2 \end{aligned} \tag{4}$$

Use of the parallelogram law II

By (3) and (4),

$$\begin{aligned}\|u_n - u_m\|^2 &= 2\|u_n - v\|^2 + 2\|u_m - v\|^2 - 4\|2^{-1}(u_n + u_m) - v\|^2 \\ &\le 2\|u_n - v\|^2 + 2\|u_m - v\|^2 - 4d^2 \qquad \text{by (4)} \\ &= 2d^2 + o(1) + 2d^2 + o(1) - 4d^2 \\ &= o(1)\end{aligned}$$

when n, m goes to infinity. Hence $\{u_n\}$ is a Cauchy sequence in $\mathcal{H}_0$ and therefore there exits a unique $\widehat{v} \in \mathcal{H}_0$,

$$\lim_{n\to\infty} u_n = \widehat{v}$$

The residual is residual

STEP 2 THE RESIDUAL IS ORTHOGONAL TO $\mathcal{H}_0$

We know that $\widehat{u}$ gives a minimum value. Let $e \stackrel{\text{def}}{=} v - \widehat{v}$ be the residual. If e is correlated with any vector in $\mathcal{H}_0$ then we can improve this minimum by adjusting for such a correlation. However, we are not allowed to improve the minimum obtained and therefore the residual has to be orthogonal to $\mathcal{H}_0$. We give the details below.

Suppose that $\langle e, u\rangle \neq 0$ for some $u \in \mathcal{H}_0$. Let $\widetilde{v} = \widehat{u} + e_u$.Then

$$\|v - \widetilde{v}\|^2 = \|v - \widehat{u} - e_u\|^2 = \|e - e_u\|^2 = \|e\|^2 - \|e_u\|^2 = d^2 - \|e_u\|^2$$

This is a contradiction and we can conclude that $\langle e, u\rangle = 0$. Hence $e \in \mathcal{H}_0^{\perp}$.

Uniqueness

STEP 3 UNIQUENESS

We now change the meaning of $\widetilde{v}$. Suppose that $\widetilde{v}$ is an alternative projection for v, i.e. it also satisfies $\|v - \widetilde{v}\|^2 = d^2$. Let $\widetilde{e}$ be its residual. Then

$$\widetilde{e} - e = (v - \widetilde{v}) - (v - \widehat{v}) = \widehat{v} - \widetilde{v}$$

The left hand side is in $\mathcal{H}_0^\perp$ and the right hand side is in $\mathcal{H}_0$. This is only possible if both vectors are zero. Hence $\widehat{v} = \widetilde{v}$ and we have uniqueness.

If $v \in \mathcal{H}_0$ then it clear that $\widehat{v} = v$. We also se that if $v = u + e$ with $u \in \mathcal{H}_0$ then $\widehat{v} = u$.

Linearity

STEP 3 LINEARITY

It is enough to show that $\widehat{v_1 + v_2} = \widehat{v}_1 + \widehat{v}_2$. This is true if $a \stackrel{\text{def}}{=} (v_1 + v_2) - (\widehat{v}_1 + \widehat{v}_2)$ is in $\mathcal{H}_0^\perp$. But $a = (v_1 - \widehat{v}_1) + (v_2 - \widehat{v}_2) = e_1 + e_2 \in \mathcal{H}_0^\perp$ since $\mathcal{H}_0^\perp$ is a subspace.

Definition of $\mathcal{P}$

STEP 4 DEFINITION OF $\mathcal{P}$

The projection operator is defined by: $\mathcal{P}(v) = \widehat{v}$. □

REMARK 9. Note that $\mathcal{P}(v) \equiv 0$ on $\mathcal{H}_0^{\perp}$.

REMARK 10. The proof does not use any basis of the vector space.

Prediction equations

The definition of the projection is essential an existence theorem and as such it is not computable. However we have the following more operational criterium.

COROLLARY. [prediction-equations]
Let $v \in \mathcal{H}$ and $\mathcal{H}_0$ be given. Then $\widehat{v}$ is the projection of v onto this subspace iff

i) $\widehat{v} \in \mathcal{H}_0$.

ii) The predictions equations hold: $e = v - \widehat{v} \in \mathcal{H}_0^{\perp}$.

PROOF.
We can write $v = \widehat{v} + e$.

□

Empirical multiple regression

Suppose that we have data $\{(\mathbf{x}_j, y_j),\, j = 1, \ldots, n\}$ with the $\mathbf{x}_j$s as a p dimensional vectors. Let $\mathbb{x}$ be the $n \times p$ matrix with $\mathbf{x}_j$ as row vector number j. The empirical regression model,

$$y_j = \sum_{i=1}^{p} b_i x_{ji} + e_i, \quad j = 1, \ldots, n$$

$$e_j \stackrel{\text{def}}{=} y_j - \sum_{i=1}^{p} b_i x_{ji}$$

In vector form, this is

$$\mathbf{y} = \mathbb{x}\boldsymbol{\beta} + \mathbf{e}$$

The vector space is $\mathbb{R}^n$ with the standard inner product, $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$. Let $\mathcal{V} = \mathbb{R}^n$, $\mathcal{V}_0$ be the column space of $\mathbb{X}$ and $\mathcal{P}$ the projection onto $\mathcal{V}_0$. Let $\widehat{\mathbf{y}} = \mathcal{P}(\mathbf{y})$, then

$$\langle \mathbf{y} - \widehat{\mathbf{y}}, \mathbf{x}^i \rangle = 0, \quad i = 1, \ldots, p$$

since

$$\mathcal{V}_0 = \text{span}\{\mathbf{x}^i,\, i = 1, \ldots, n\}$$

The set predictions equations are equal to

$$\mathbb{X}^T(\mathbf{y} - \widehat{\mathbf{y}}) = 0$$

so that

$$\mathbb{X}^T \widehat{\mathbf{y}} = \mathbb{X}^T \mathbf{y}$$

since $\widehat{\mathbf{y}} \in \mathcal{V}_0$ there exists a $\boldsymbol{\beta}$ so that

$$\widehat{\mathbf{y}} = \sum_{i=1}^{p} \beta_i \mathbf{x}^i = \mathbb{X}\boldsymbol{\beta}$$

Hence

$$\mathbb{x}^T \mathbb{x} \boldsymbol{\beta} = \mathbb{x}^T \mathbf{y}$$

If $\mathbb{x}$ has full rank,

$$\boldsymbol{\beta} = \left(\mathbb{x}^T \mathbb{x}\right)^{-1} \mathbb{x}^T \mathbf{y}$$

and

$$\mathcal{P}(\mathbf{y}) = \widehat{\mathbf{y}} = \mathbb{x} \boldsymbol{\beta} = \mathbb{x} \left(\mathbb{x}^T \mathbb{x}\right)^{-1} \mathbb{x}^T \mathbf{y}$$

This shows that

$$\mathcal{P} = \mathbb{x} \left(\mathbb{x}^T \mathbb{x}\right)^{-1} \mathbb{x}^T$$

and the left hand side is a projection matrix,

i) $\mathcal{P} = \mathcal{P}^T$; symmetric.

ii) $\mathcal{P} = \mathcal{P}^2$; idempotent.

Note that $\mathcal{P}$ is uniquely defined also when $\mathbb{X}$ does not have full rank. In that case any $\boldsymbol{\beta}$ in an affine space of dimension $p - \text{rank}(\mathbb{x})$ works,

$$\beta = \beta' + v, \quad \widehat{\mathbf{y}} = \mathbb{x}\,\beta' \text{ and } v \text{ in the null space of } \mathbb{x}$$

The projection $\mathcal{P}$ is a $n \times n$ matrix and $\text{rank}(\mathcal{P}) = \text{rank}(\mathbb{X})$. We also have that

$$\boldsymbol{\beta} = \underset{\mathbf{b}}{\text{argmin}} \, \|\mathbf{y} - \mathbb{x}\,\mathbf{b}\|^2$$

REMARK 11. No distributional properties in this derivation of $\boldsymbol{\beta}$ are used. We can write this empirical model in terms of stochastic variables but still keeping the geomtric vectors. The most common notation is then

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\boldsymbol{e}$ is stochastic and $\mathbb{X}$ is fixed explanatory variables or stochastic regressors. A third variant is the mutliple regression model without data. In

that case the projection corresponds to in the empirical setting to prediction of a new Y given a new vector of regressors.

Preparing for LS estimering in an AR(p) model

Consider the following stochastic empirical mutiple regression model;

$$\mathbf{Y} = \mathbb{X}_0 \boldsymbol{\xi}_0 + \mathbf{e}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$ and $\mathbb{X}_0$ is $m \times (p+1)$ matrix and

$$\mathbb{X}_0 = \begin{bmatrix} \mathbf{1}_m & \mathbb{X} \end{bmatrix}$$

allowing for a constant mean value for the dependent variable. Hence

$$\mathbf{Y} = \mathbf{1}_m \xi_0 + \mathbb{X}\boldsymbol{\xi} + \mathbf{e}$$

We want to substract the column means from the $\mathbb{X}$,

$$\mathbb{W} \stackrel{\text{def}}{=} \mathbb{X} - m^{-1}\mathbf{1}_m \otimes \mathbf{1}_m^T \mathbb{X}$$
$$W_{ji} = X_{ji} - \overline{X}_{\cdot i}$$

$$\begin{aligned}\mathbf{Y} &= \big(\mathbf{1}_m\xi_0 + \mathbf{1}_m \otimes n^{-1}\mathbf{1}_m^T\mathbb{X}\boldsymbol{\xi}\big) + \big(\mathbb{X} - m^{-1}\mathbf{1}_m \otimes \mathbf{1}_m^T\mathbb{X}\big)\boldsymbol{\xi} + \mathbf{e} \\ &= \mathbf{1}_m\beta_0 + \mathbb{W}\boldsymbol{\beta}\end{aligned}$$

REMARK 12. This means that

$$\mathbb{W}^T\mathbb{W}[r,s] = \sum_{j=1}^{m}\big(X_{j,r} - \overline{X}_{\cdot r}\big)\big(X_{js} - \overline{X}_{\cdot s}\big)$$

PROBLEM 3.1. Let $\mathcal{P} = \mathbb{I} - m^{-1}\mathbf{1}_m \otimes \mathbf{1}_m^T$

i) Show that $\mathcal{P}$ is projection matrix.

ii) Show that $\mathbb{W} = \mathcal{P}\,\mathbb{X}$ and $\mathbb{W}^T\mathbb{W} = \mathbb{X}^{\mathrm{T}}\,\mathcal{P}\,\mathbb{X}$.

iii) Suppose that $\mathbb{X}$ has full rank. What is the rank of $\mathcal{P}$?

Time series and Hilbert spaces

4. Weak stationarity

The vector space approach to time series is important for linear models.

Definition 3. A time series $\{X_t,\, t \in \mathbb{Z}\}$ is weakly stationary if

i) For all t, $\mathbb{E}\, X_t^2 < \infty$.

ii) $\mathbb{E}\, X_t \equiv \mu$.

iii) For all h, $\mathrm{Cov}(X_{t+h}, X_t)$ does not depend on t.

Remark 13. The expection plays minor role.

Definition 4. [white noise]
The time series $\{Z_t,\, t \in \mathbb{Z}\}$ is white noise if

i) For all t, $\mathbb{E}\, Z_t^2 < \infty$.

ii) $\mathbb{E}\, Z_t \equiv 0$.

iii) For all h, $\mathrm{Cov}(Z_{t+h}, Z_t) = \delta_{0,h}\sigma_Z^2$ does not depend on t.

EXAMPLE 7. Let $\{Z_t,\, t \in \mathbb{Z}\}$ be independent standard normal distributed variables. The $\{Z_t,\, t \in \mathbb{Z}\}$ is a white noise with unit variance.

4.1. ACVF AND ACVF

DEFINITION 5. [autocovariance and autocorrelation function] Let $\{X_t,\, t \in \mathbb{Z}\}$ be a weakly stationary time series. Then

$$\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t)$$
$$\rho(h) = \mathrm{Corr}(X_{t+h}, X_t)$$

Acronyms for γ and ρ are ACVF and ACF, respectively.

PROPOSITION 2. Basic properties for γ and ρ,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$
$$|\gamma(h)| \leq \gamma(0)$$
$$\gamma(h) = \gamma(-h)$$

PROOF OF PROPOSITION **??**.
The first one follows from the definitons. By Cauchy-Schwartz we get second one. The symmetry is a consequence of the weak stationarity and the symmetry of covariance, i.e. use $t = t - h$. □

EXAMPLE 8. Let $\{X_t,\, t \in \mathbb{Z}\}$ be a stationary time series and let $a_S = \{a_j,\, j \in S\}$ be a finite vector. Then

$$\begin{aligned} \mathrm{Var}(\sum_{j\in S} a_j X_j) &= \sum_{i\in S}\sum_{j\in S} a_i \gamma(i-j) a_j = a_S^T \Gamma_S a_S \geq 0 \\ \Gamma_S &= \{\gamma(i-j),\ \ (i,j) \in S \times S\} \end{aligned} \tag{5}$$

For $S = \{2, 4, 9\}$,

$$\Gamma_S = \begin{bmatrix} \gamma(0) & \gamma(2) & \gamma(7) \\ \gamma(2) & \gamma(0) & \gamma(5) \\ \gamma(7) & \gamma(5) & \gamma(0) \end{bmatrix}$$

REMARK 14. In multivariate notation $\mathbf{X}_S = \{X_t,\, t \in S\}$, $\mathbf{a}_S = \{a_j,\, t \in S\}$, and

$$\begin{aligned} \mathrm{Var}(\sum_{j \in S} a_j X_j) &= \mathrm{Var}(\mathbf{a}_S^T \mathbf{X}_S) = \mathrm{Cov}\big(\mathbf{a}_S^T \mathbf{X}_S,\, \mathbf{a}_S^T \mathbf{X}_S\big) = \mathbf{a}_S^T \mathrm{Cov}(\mathbf{X}_S, \mathbf{X}_S)\mathbf{a}_S \\ &= \mathbf{a}_S^T \Gamma_S \mathbf{a}_S \end{aligned}$$

4.2. NND AND PD

A sequence defined on $\mathbb{Z}$ can be nonnegative definite [nnd]. The concept is borrowed from a matrix being nonnegative definite.

DEFINITION 6. A real valued sequence $\{\lambda(h),\, h \in \mathbb{Z}\}$ is nonnegative definite if is symmetric and for all finite $S \subset \mathbb{Z}$ and all real vectors $\{a_j,\, j \in S\}$

$$\sum_{i \in S} \sum_{j \in S} a_i \lambda(i-j) a_j \geq 0 \tag{6}$$

If the inequality is strict for all nontrivial bano vectors then λ is positive definite.

REMARK 15. A real valued sequence $\{\lambda(h),\, h \in \mathbb{Z}\}$ is nonnegative definite iff for all finite $S \in \mathbb{Z}$, the matrix $\mathbb{\Lambda}_S = \{\lambda(i-j),\ (i,j) \in S \times S\}$ is nonnegative definite.

PROPOSITION 3. The autocovariance function is nonnegative definite.

PROOF OF PROPOSITION 3.
We have that γ is symmetric and by Example 8 we see that (6) holds. □

PROPOSITION 4. A sequence $\{\lambda(h),\, h \in \mathbb{Z}\}$ defined on $\mathbb{Z}$ is an ACVF iff it is nnd.

PROOF OF PROPOSITION 4.

Let λ be a nonnegative definite sequence defined on $\mathbb{Z}$. Define

$$\mathbb{\Lambda}_S = \{\lambda(i-j),\ (i,j) \in S \times S\}$$

for all finite subset of $\mathbb{Z}$. Then $\mathbb{\Lambda}_S$ is covariance matrix for all S. Let F_S be the multinormal distribution function of $\mathcal{N}(0, \mathbb{\Lambda}_S)$. Then the family $\{F_S \colon S \subset \mathbb{Z}$ and S is finite$\}$ is consistent. This means that for disjoint S and S' we have

$$F_S(\mathbf{x}_S) = \lim_{\mathbf{x}_{S'} \to \infty} F_{S \cup S'}(\mathbf{x}_S, \mathbf{x}_{S'})$$

since $Y_{S \cup S'} \sim \mathcal{N}(0, \mathbb{\Lambda}_{S \cup S'})$ implies that $Y_S \sim \mathcal{N}(0, \mathbb{\Lambda}_S)$. Confer Stat210 or Stat201 (or the old Stat310). By Kolmogorov's existence theorem (Stat321) there exists a Gaussian time series $\{X_t,\ t \in \mathbb{Z}\}$ such that $\gamma = \lambda$. The family of mulitnormal distributions have densities if also λ is postive definite [pd]. $\square$

REMARK 16. A stationary Gaussian time series is completely described by its ACVF and apart form a scaling factor, it is completely described by its ACF.