EECS 225A Statistical Signal Processing	Lecture 2 - 1/23/2020

Lecture 2: Linear Estimation

Lecturer: Jiantao Jiao *Scribe: Dominic Carrano*

The problem of optimal linear estimation, or more precisely, optimal linear estimation under squared error loss, is one of the foundational pillars of statistical signal processing and Bayesian statistics with great theoretical depth and practical applications.

Abstractly, the objective is to find the best estimate of signal X using a *linear* function of the observation Y. Here X and Y do not need to have the same dimension, and could take values in general sets. We do assume that we know the *auto-correlation* and *cross-correlation* functions

$$R_X = \mathbb{E}[XX^*] \tag{1}$$
$$R_Y = \mathbb{E}[YY^*] \tag{2}$$
$$R_{XY} = \mathbb{E}[XY^*], \tag{3}$$

where we view X and Y as column vectors and $*$ denotes the conjugate transpose operation. Note that we *do not* need assumptions on the expectations of X and Y, and we *do not* need the knowledge of the joint distribution of X, Y.

The loss function we use would be the *squared error* loss function. This particular loss has deep connections with Hilbert space, which enables a general theory with beautiful formulas. Our treatment would be a bit unconventional, since we would use the notation $\langle \cdot, \cdot \rangle$ to denote a function whose output is a rectangular matrix, but the underlying theory is still classical Hilbert space theory. The benefits of this notation would appear apparent in the later parts of this course.

Mathematically, define the (possibly matrix valued) *inner product* between two random vectors Z_1, Z_2 that may not have the same dimension:

$$\langle Z_1, Z_2 \rangle \triangleq \mathbb{E}[Z_1 Z_2^*], \tag{4}$$

and denote

$$\|Z\|^2 \triangleq \langle Z, Z \rangle. \tag{5}$$

Note that it is consistent with our intuition of inner product for Euclidean spaces. We map a vector in the Euclidean space to a random variable, and a collection of vectors in the Euclidean space to a random vector.

It satisfies the following properties:

1. Linearity:

$$\langle \alpha_1 v_1 + \alpha_2 v_2, u \rangle = \alpha_1 \langle v_1, u \rangle + \alpha_2 \langle v_2, u \rangle \tag{6}$$

2. Reflexivity:

$$\langle u, v \rangle = \langle v, u \rangle^* \tag{7}$$

3. Nondegeneracy:

$$\|v\|^2 = 0 \Longleftrightarrow v = 0. \tag{8}$$

We would like to solve for the linear map $\hat{X}(Y)$ of Y that achieves

$$\min_{\hat{X}} \|\hat{X} - X\|^2. \tag{9}$$

Here we constrain that $\hat{X}(Y)$ to be a linear function of Y: in other words, it satisfies

$$\hat{X}(Y_1) + \hat{X}(Y_2) = \hat{X}(Y_1 + Y_2) \tag{10}$$

$$\hat{X}(\alpha Y) = \alpha \hat{X}(Y), \text{ for all } \alpha \in \mathbb{C} \tag{11}$$

Sometimes there exists an explicit matrix W such that $\hat{X}(Y) = WY$. Indeed, if Y is a finite dimensional random vector it is the case. But it is important to keep in mind that it is not always possible. One notable example is the optimal linear predictor of X_t given its causal history $\{X_j\}_{j \le t-1}$ when $X_t = V_t - V_{t-1}$, where $\{V_t\}_{t=-\infty}^{\infty}$ are mutually independent $\mathcal{N}(0,1)$ random variables [1, Remark 1].

On the face of it, it was not even clear (9) has a solution. Indeed, we are evaluating the error of estimating X with $\hat{X}$ using a matrix, which only has a partial order. The surprising/incredible property of Hilbert spaces guarantee that it is in fact solvable and has an unique solution.

Theorem 1. *(Orthogonality principle) The optimization problem (9) has a unique solution $\hat{X}(Y)$ such that for any linear function $\hat{X}'(Y)$ of Y,*

$$\|X - \hat{X}\|^2 \le \|X - \hat{X}'\|^2. \tag{12}$$

Further, $\hat{X}$ is the unique linear function of Y satisfying the orthogonality principle:

$$\langle X - \hat{X}, Y \rangle = 0, \tag{13}$$

Proof Following similar steps in [2, Theorem 3.2] one can show that there exists some $\hat{X}$ satisfying (13). Now we need to show that it is unique, and it solves the minimization problem. Note that (13) is equivalent to $\langle X - \hat{X}, \hat{X}'(Y) \rangle = 0$ for any linear function $\hat{X}'(Y)$ of Y.

We first show that it solves the minimization problem. We can write

$$\|X - \hat{X}'\|^2 = \|X - \hat{X} + \hat{X} - \hat{X}'\|^2 \tag{14}$$

$$= \|X - \hat{X}\|^2 + \|\hat{X} - \hat{X}'\|^2 + \langle X - \hat{X}, \hat{X} - \hat{X}' \rangle + \langle \hat{X} - \hat{X}', X - \hat{X} \rangle. \tag{15}$$

Since $\hat{X} - \hat{X}'$ is a linear function of Y, we know that the last two terms are zero (matrices). Hence, we have

$$\|X - \hat{X}\|^2 = \|X - \hat{X}'\|^2 - \|\hat{X} - \hat{X}'\|^2 \tag{16}$$

$$\preceq \|X - \hat{X}'\|^2. \tag{17}$$

Now we show the uniqueness. Suppose both $\hat{X}_1$ and $\hat{X}_2$ satisfy the orthogonality condition. Then,

$$\|\hat{X}_1 - \hat{X}_2\|^2 = \langle \hat{X}_1 - \hat{X}_2, \hat{X}_1 - \hat{X}_2 \rangle \tag{18}$$

$$= \langle \hat{X}_1 - \hat{X}_2, \hat{X}_1 - X + X - \hat{X}_2 \rangle \tag{19}$$

$$= \langle \hat{X}_1 - \hat{X}_2, X - \hat{X}_2 \rangle - \langle \hat{X}_1 - \hat{X}_2, X - \hat{X}_1 \rangle \tag{20}$$

$$= 0, \tag{21}$$

where in the last step we used that $\hat{X}_1 - \hat{X}_2$ is a linear function of Y. □

Now we apply Theorem 1 to solve (9) assuming that we can write $\hat{X}(Y)$ as WY. The orthogonality principle says that

$$\langle X - WY, Y \rangle = 0, \tag{22}$$

which is equivalent to

$$\langle X, Y\rangle = W\langle Y, Y\rangle. \tag{23}$$

Since we have defined (2) and (3), it can also be written as

$$WR_Y = R_{XY}, \tag{24}$$

known as the *normal equation*.

A first question is, does (24) always have a solution? The abstract Theorem 1 guarantees this even when R_Y is singular, and clearly if R_Y is invertible we have

$$W = R_{XY}R_Y^{-1}. \tag{25}$$

We now show that no matter what solution W to the normal equation we pick, the corresponding estimator WY is always the same. Indeed, suppose $W_1R_Y = R_{XY}, W_2R_Y = R_{XY}$. Then

$$\begin{aligned}
\|W_1Y - W_2Y\|^2 &= \langle W_1Y - W_2Y, W_1Y - W_2Y\rangle && (26)\\
&= (W_1 - W_2)\langle Y, Y\rangle(W_1 - W_2)^* && (27)\\
&= (W_1 - W_2)R_Y(W_1 - W_2)^* && (28)\\
&= W_1R_Y(W_1 - W_2)^* - W_2R_Y(W_1 - W_2)^* && (29)\\
&= R_{XY}(W_1 - W_2)^* - R_{XY}(W_1 - W_2)^* && (30)\\
&= 0. && (31)
\end{aligned}$$

If R_Y is singular, then one solution of W (not unique) could be $W = R_{XY}R_Y^\dagger$, where $\dagger$ denotes the pseudoinverse. If we do eigenvalue decomposition for $R_Y = U_r\Lambda_rU_r^*$, where R_Y has rank r, and Λ_r is a invertible $r \times r$ matrix. The columns of U_r are orthonormal. Then the pseudoinverse of R_Y can be defined by

$$R_Y^\dagger = U_r\Lambda_r^{-1}U_r^*. \tag{32}$$

Also we have the representation

$$Y = \sum_{i=1}^{r} u_i(u_i^*Y), \tag{33}$$

since Y has zero second moment on directions orthogonal to those in U_r. Here $\{u_i\}_{1\le i\le r}$ are the columns of U_r.

Hence,

$$\begin{aligned}
R_{XY}R_Y^\dagger R_Y &= R_{XY}U_r\Lambda_r^{-1}U_r^*U_r\Lambda_rU_r^* && (34)\\
&= R_{XY}U_rU_r^*. && (35)
\end{aligned}$$

It suffices to show that

$$R_{XY}U_rU_r^* = R_{XY}. \tag{36}$$

It is true since

$$\begin{aligned}
R_{XY} &= \langle X, \sum_{i=1}^{r} u_i(u_i^*Y)\rangle && (37)\\
&= \sum_{i=1}^{r}\langle X, Y\rangle u_iu_i^* && (38)\\
&= \langle X, Y\rangle U_rU_r^*. && (39)
\end{aligned}$$

1 The error matrix

If $R_Y > 0$,

$$\begin{aligned}
\langle \hat{X}, \hat{X} \rangle &= \langle WY, WY \rangle && (40)\\
&= W\langle Y, Y \rangle W^* && (41)\\
&= W R_Y W^* && (42)\\
&= R_{XY} R_Y^{-1} R_Y R_Y^{-1} R_{YX} && (43)\\
&= R_{XY} R_Y^{-1} R_{YX}. && (44)
\end{aligned}$$

Meanwhile, since

$$X = \hat{X} + (X - \hat{X}), \tag{45}$$

and $\langle \hat{X}, X - \hat{X} \rangle = 0$, we have

$$\langle X, X \rangle = \langle \hat{X}, \hat{X} \rangle + \langle X - \hat{X}, X - \hat{X} \rangle, \tag{46}$$

which is equivalent to

$$R_X = R_{\hat{X}} + R_e, \tag{47}$$

where $e \triangleq X - \hat{X}$ is the *error*.

2 When do we need assumptions on the mean?

Note that we do not impose assumptions on the mean (expectation) of X, Y in (24). That is because we are *only allowed* to use a linear function of Y, but not a linear function of $[Y; 1]$ to estimate X. If we are allowed to do the latter, then the optimal estimator of X should be given by

$$\hat{X} = W(Y - \mu_Y) + \mu_X, \tag{48}$$

where $W\Sigma_Y = \Sigma_{XY}$. Here

$$\begin{aligned}
\mu_X &\triangleq \mathbb{E}[X] && (49)\\
\mu_Y &\triangleq \mathbb{E}[Y] && (50)\\
\Sigma_Y &\triangleq \mathbb{E}[(Y - \mu_Y)(Y - \mu_Y)^*] && (51)\\
\Sigma_X &\triangleq \mathbb{E}[(X - \mu_X)(X - \mu_X)^*] && (52)\\
\Sigma_{XY} &\triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^*]. && (53)
\end{aligned}$$

Now consider the case that $\Sigma_Y > 0$. We show that (48) can be derived from (24). To do this, we define a new vector $Z = [Y; 1]$, and compute R_{XZ} and R_Z.

$$\begin{aligned}
R_{XZ} &= \mathbb{E}[XZ^*]\\
&= \mathbb{E}[X(Y^*, 1)]\\
&= [\mathbb{E}[XY^*], \mu_X]\\
&= [\Sigma_{XY} + \mu_X \mu_Y^*, \mu_X].
\end{aligned}$$

$$
\begin{aligned}
R_Z &= \mathbb{E}[ZZ^*] \\
&= \mathbb{E}\left[\begin{bmatrix} YY^* & Y \\ Y^* & 1 \end{bmatrix}\right] \\
&= \begin{bmatrix} \Sigma_Y + \mu_Y \mu_Y^* & \mu_Y \\ \mu_Y^* & 1 \end{bmatrix}
\end{aligned}
$$

Recall the block matrix inversion formula

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C) & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}.
$$

and apply it to R_Z, we obtain that

$$
R_Z^{-1} = \begin{bmatrix} \Sigma_Y^{-1} & -\Sigma_Y^{-1}\mu_Y \\ -\mu_Y^*\Sigma_Y^{-1} & 1 + \mu_Y^*\Sigma_Y^{-1}\mu_Y \end{bmatrix}.
$$

Hence, the optimal linear estimator of X given Z is

$$
\begin{aligned}
R_{XZ}R_Z^{-1}Z &= [\Sigma_{XY} + \mu_X\mu_Y^*, \mu_X] \begin{bmatrix} \Sigma_Y^{-1} & -\Sigma_Y^{-1}\mu_Y \\ -\mu_Y^*\Sigma_Y^{-1} & 1 + \mu_Y^*\Sigma_Y^{-1}\mu_Y \end{bmatrix} \begin{bmatrix} Y \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma_{XY}\Sigma_Y^{-1} & -\Sigma_{XY}\Sigma_Y^{-1}\mu_Y + \mu_X \end{bmatrix} \begin{bmatrix} Y \\ 1 \end{bmatrix} \\
&= \Sigma_{XY}\Sigma_Y^{-1}(Y - \mu_Y) + \mu_X.
\end{aligned}
$$

In this case, we have the error matrices:

$$
\Sigma_{\hat{X}} = \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} \tag{54}
$$

$$
\Sigma_X = \Sigma_{\hat{X}} + \Sigma_e, \tag{55}
$$

where $e = X - \hat{X}$.

3 The scalar case

To get a more intuitive understanding of this result, let's consider the case that both X, Y are scalar random variables. Let ρ be the Pearson correlation coefficient, defined as

$$
\rho(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^*]}{\sqrt{\mathbb{E}[\|X - \mathbb{E}[X]\|^2]\mathbb{E}[\|Y - \mathbb{E}[Y]\|^2]}}, \tag{56}
$$

we know from Cauchy–Schwarz that $0 \leq |\rho| \leq 1$. Then (48) reduces to

$$
\hat{X} = \frac{\rho\sigma_X}{\sigma_Y}(Y - \mu_Y) + \mu_X, \tag{57}
$$

where σ_X, σ_Y are standard deviations of X, Y, respectively. This formula can be intuitively explained as skewing/standardizing $Y - \mu_Y$ by σ_Y, then reskewing by σ_X and scaling by correlation coefficient ρ to obtain the estimate $\hat{X} - \mu_X$.

4 The scalar additive noise case

Consider the noise model $Y = X + Z$, where $X, Y, Z \in \mathbb{C}$ are scalar-valued random variables. We'll assume that:

- The signal power is P: $\mathbb{E}[|X|^2] = P$.

- The noise power is N: $\mathbb{E}[|Z|^2] = N$.

- The signal and noise are uncorrelated: $\mathbb{E}[XZ^*] = 0$.

Note that we *do not* assume the means of X, Z. Now we compute

$$R_{XY} = \mathbb{E}[XY^*] = \mathbb{E}[X(X+Z)^*] = \mathbb{E}[|X|^2 + XZ^*] = \mathbb{E}[|X|^2] + \mathbb{E}[XZ^*] = P \tag{58}$$

and

$$R_Y = \mathbb{E}[|Y|^2] = \mathbb{E}[(X+Z)(X+Z)^*] = \mathbb{E}[|X|^2 + |Z|^2 + 2XZ^*] = P + N \tag{59}$$

Then the optimal linear estimator of X given Y is

$$\hat{X}(Y) = \frac{P}{P+N} Y. \tag{60}$$

Intuitively, we shrink the observation Y down based on how strong the signal is compared to the noise. In the extreme case of $P = 0$, there is no signal, and $\hat{X} = 0$. At the other extreme of $P/N \to \infty$, the signal-to-noise-ratio is as strong as it can be, and we obtain $\hat{X} = Y$.

In the special case of $X \sim \mathcal{N}(0, P), Z \sim \mathcal{N}(0, N)$, and X is independent of Z, the formula (60) gives the *optimal Bayes estimator* of X given Y (not constraining ourselves to linear estimators) under squared error loss. In that case, $P = \infty$ corresponds to the maximum likelihood estimator, which corresponds to the case of no prior information on X. As a final note for this example, we remark that the MSE is

$$\mathbb{E}[|\hat{X} - X|^2] = \frac{PN}{P+N} \tag{61}$$

and as $P \to \infty$, the MSE approaches N.

We observe that

$$\frac{PN}{P+N} \leq \min\{P, N\}, \tag{62}$$

which admits an interesting interpretation: P is the MSE achieved by using constant zero as an estimator, and N is the MSE achieved by using Y as the estimator. The optimal error, clearly should not be bigger than any of these two numbers.

5 The vector additive noise case

Consider the setup $Y = HX + Z$, where $Y \in \mathbb{C}^n, X \in \mathbb{C}^d, H \in \mathbb{C}^{n \times d}$, and $Z \in \mathbb{C}^n$. We assume the matrix H has full column rank. This is a generalization of the previous example; now we consider the vector case and add in a matrix H that performs a linear transformation (linear regression). Again, our goal is to estimate X from Y, and similar to before, we assume the signal and noise are uncorrelated: $\mathbb{E}[XZ^*] = 0$. From these assumptions, we find that

$$\begin{aligned} R_Y &= \mathbb{E}[(HX+Z)(HX+Z)^*] \\ &= H\mathbb{E}[XX^*]H^* + \mathbb{E}[ZZ^*] \\ &= HR_XH^* + R_Z, \end{aligned}$$

and

$$R_{XY} = \mathbb{E}[X(HX+Z)^*] = \mathbb{E}[XX^*H^*] = R_X H^*.$$

Hence, we obtain

$$W = R_X H^*(HR_X H^* + R_Z)^{-1} \tag{63}$$

$$\hat{X} = WY. \tag{64}$$

and if we denote by P_X the error auto-correlation matrix,

$$P_X = R_X - R_{XY}R_Y^{-1}R_{YX} \tag{65}$$

$$= R_X - R_X H^*(HR_X H^* + R_Z)^{-1}HR_X. \tag{66}$$

If both R_X and R_Z are invertible, we do have another set of expressions that are usually called the *information form* since they involve the inverses of autocorrelation matrices [3, Chapter 3.4].

Indeed, one can easily verify the following expression for W when both R_X and R_Z are invertible: [1]

$$W = (H^*R_Z^{-1}H + R_X^{-1})^{-1}H^*R_Z^{-1}. \tag{67}$$

Recall the Sherman–Morrison–Woodbury formula:

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{68}$$

and apply it to $R_X^{-1} + H^*R_Z^{-1}H$, we have

$$(R_X^{-1} + H^*R_Z^{-1}H)^{-1} = R_X - R_X H^*(HR_X H^* + R_Z)^{-1}HR_X \tag{69}$$

$$= P_X, \tag{70}$$

which implies that $W = P_X H^* R_Z^{-1}$, or equivalently

$$P_X^{-1}\hat{X} = H^*R_Z^{-1}Y. \tag{71}$$

Note that the RHS of (71) does not depend on R_X.

The equation (70) reduces to

$$P_X = \frac{1}{\frac{1}{P} + \frac{1}{N}} \tag{72}$$

in the scalar additive noise case.

The information form has an interesting advantage in deriving formulas combining two estimators together as shown by the following lemma.

Lemma 2. *[3, Lemma 3.4.1] Let Y_a and Y_b satisfy*

$$Y_a = H_a X + Z_a \tag{73}$$

$$Y_b = H_b X + Z_b, \tag{74}$$

where $\langle X, Z_a\rangle = \langle X, Z_b\rangle = \langle Z_a, Z_b\rangle = 0$. Denote by $\hat{X}_a$ the optimal linear estimator of X given only Y_a, and $\hat{X}_b$ the optimal linear estimator of X given only Y_b. Denote the estimation error matrices by $P_a = \langle X - \hat{X}_a, X - \hat{X}_a\rangle, P_b = \langle X - \hat{X}_b, X - \hat{X}_b\rangle$. Then $\hat{X}$, the optimal linear estimator of X given both Y_a, Y_b can be found as

$$P^{-1}\hat{X} = P_a^{-1}\hat{X}_a + P_b^{-1}\hat{X}_b, \tag{75}$$

where $P = \langle X - \hat{X}, X - \hat{X}\rangle$ satisfies

$$P^{-1} = P_a^{-1} + P_b^{-1} - R_X^{-1}. \tag{76}$$

[1] Exercise: show it.

5.1 Connections to the scalar case

Assume

$$R_X = PI_{d\times d}, R_Z = NI_{n\times n}, \tag{77}$$

where $P, N \in \mathbb{R}$. Then,

$$W = PH^*(pHH^* + NI_{n\times n})^{-1} \tag{78}$$

$$HW = PHH^*(PHH^* + NI_{n\times n})^{-1}. \tag{79}$$

Now, using the eigendecomposition of HH^* as

$$HH^* = UDU^* = U\text{diag}(\lambda_1^2, ..., \lambda_d^2, 0, ..., 0)U^*, \tag{80}$$

we can simplify (79) to

$$\begin{aligned} HW &= PUDU^*(PUDU^* + NI)^{-1} \\ &= PUDU^*(U(PD + NI)U^*)^{-1} \\ &= U[PD(PD + NI)^{-1}]U^*. \end{aligned}$$

In the scalar case, we had

$$\hat{X} = \frac{P}{P+N}Y \tag{81}$$

and in the matrix case, we have

$$HW = U[PD(PD + NI)^{-1}]U^* \tag{82}$$

where

$$PD(PD + NI)^{-1} = \text{diag}\left(\frac{P\lambda_1^2}{P\lambda_1^2 + N}, ..., \frac{P\lambda_d^2}{P\lambda_d^2 + N}, 0, ..., 0\right). \tag{83}$$

This second formula is analogous to the first, where we

- Rotate the observation Y with U^* into the basis where the shrinkage operation can be performed element-wise.
- Shrink each dimension based on the signal power.
- Rotate back with U.

Again, as $P \to \infty$, we obtain

$$\lim_{P\to\infty} PD(PD + NI)^{-1} = \text{diag}\left(\underbrace{1, ..., 1}_{d}, \underbrace{0, ..., 0}_{n-d}\right) \tag{84}$$

and

$$HW = U_d U_d^* = \underbrace{H(H^*H)^{-1}H^*}_{\text{Projection into range}(H)} \tag{85}$$

where U_d is U with the last $n - d$ columns set to zero.

5.2 The Best Linear Unbiased Estimator

We can show that the estimator

$$\hat{X}_b = (H^*H)^{-1}H^*Y \tag{86}$$

is the *minimum variance unbiased* estimator for X if we assume X is deterministic. In other words, the previous example shows that assuming $R_X = PI$ with $P \to \infty$ is equivalent to assuming *no* knowledge about X, which is also equivalent to finding the best unbiased estimator.

Theorem 3. *(Gauss–Markov theorem)[3, Theorem 3.4.1] Consider model $Y = HX + Z$, where Z is a zero-mean random vector with covariance matrix identity $\langle Z, Z\rangle = I$, X is a deterministic vector, and H has full column rank. Then the estimator defined in (86) is the best linear unbiased estimator (BLUE) of X given Y in the sense that if $\hat{X}$ is another linear unbiased estimator of X given Y, then*

$$\langle \hat{X} - X, \hat{X} - X\rangle \geq \langle \hat{X}_b - X, \hat{X}_b - X\rangle. \tag{87}$$

By unbiasedness, we mean that the expectation of the estimator is always X no matter what the true X is.

Proof Assume $\hat{X} = KY$ is another linear estimator of X. In order for $\hat{X}$ to be unbiased we must require

$$\begin{aligned}
\mathbb{E}[\hat{X}] &= \mathbb{E}[K(HX + Z)] && (88)\\
&= KHX && (89)\\
&= X, && (90)
\end{aligned}$$

and since X could be arbitrary we know $KH = I$. We then write the covariance matrix of random vector $\hat{X}$ as

$$\begin{aligned}
\langle \hat{X} - X, \hat{X} - X\rangle &= \langle KY - X, KY - X\rangle && (91)\\
&= \langle K(HX + Z) - X, K(HX + Z) - X\rangle && (92)\\
&= \langle KZ, KZ\rangle && (93)\\
&= KK^*. && (94)
\end{aligned}$$

Hence, it suffices to show that for any K such that $KH = I$, we have

$$KK^* \geq (H^*H)^{-1}H^*H(H^*H)^{-1}. \tag{95}$$

It is indeed the case since

$$\begin{aligned}
KK^* - (H^*H)^{-1}H^*H(H^*H)^{-1} &= KK^* - (H^*H)^{-1} && (96)\\
&= KK^* - KH(H^*H)^{-1}H^*K^* && (97)\\
&= K(I - H(H^*H)^{-1}H^*)K^* && (98)\\
&\geq 0, && (99)
\end{aligned}$$

where in the last step we used the fact that $H(H^*H)^{-1}H^*$ is a projection matrix and we necessarily have $I - H(H^*H)^{-1}H^* \geq 0$. □

References

[1] H. Bierens, "The wold decomposition," 2012. [Online]. Available: \url{http://www.personal.psu.edu/hxb11/WOLD.PDF}

[2] B. Hajek, *Random processes for engineers.* Cambridge university press, 2015.

[3] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation.* Prentice Hall, 2000, no. BOOK.