LINEAR ESTIMATION

Linear estimators are an important class of estimators

- simplicity
- dependence on first and second order moments only
- ease of implementation

This last point is especially important for filtering problems where we must process data real-time.

Our discussion of filtering will rely on Bayesian linear estimators, but for completeness we begin with classical linear estimators.

We make the following distinctions:

<u>constant</u>: $\underline{\hat{\theta}}(\underline{x}) = \underline{b}$

<u>linear</u>: $\underline{\hat{\theta}}(\underline{x}) = H\underline{x}$

<u>affine</u>: $\underline{\hat{\theta}}(\underline{x}) = H\underline{x} + \underline{b}$

where $H \in \mathbb{R}^{p \times N}$, $\underline{b} \in \mathbb{R}^{p}$. We'll study all three cases.

Best Linear Unbiased Estimation

The MVUE is often not computable.

- CRLB or Rao-Blackwell not appliable
- intractable mathematical model
- randomness is only known up to first and second order moments.

In such cases, we must be content with a suboptimal estimator. One approach is to compute

Definition | The best linear unbiased estimator (BLUE) is the linear estimator

$$\underline{\hat{\theta}}(\underline{x}) = H\underline{x} \quad , \quad H \in \mathbb{R}^{p \times N}$$

with smallest variance among all linear, unbiased estimators.

Note that for $\underline{\hat{\theta}}$ to be unbiased we must have

$$\underline{\theta} = E\{\underline{\hat{\theta}}\} = E\{H\underline{X}\} = HE\{\underline{X}\}$$

Therefore the mean of the data must obey a linear relationship with the true parameter.

This relationship will not always be true, so even the suboptimal BLUE isn't always feasible.

However, there is an important class of problems where it does hold.

Linear Model

Suppose $\underline{X}$ and $\underline{\theta}$ are related through

$$\underline{X} = A\underline{\theta} + \underline{W}$$

$\underline{\theta}$ fixed but unknown

where

A is $N \times p$ and known, full rank

$\underline{W}$ is random

$E[\underline{W}] = \underline{0}$

$E[\underline{W}\,\underline{W}^T] =: R$ is known, positive definite

Note: $\underline{W}$ is not necessarily Gaussian

Question: Can you give an H such that

$H\,E[\underline{X}] = \underline{\theta}$?

Taking $H = (A^T A)^{-1} A^T$ we find

$$H\, E[\underline{X}] = \underbrace{(A^T A)^{-1} A^T A}_{I}\, \underline{\theta} + (A^T A)^{-1} A^T \underbrace{E[\underline{W}]}_{\underline{0}}$$

$$= \underline{\theta}.$$

So the pseudo inverse is <u>an</u> unbiased estimator. But is its variance minimal?

<u>Theorem</u> (Gauss-Markov Theorem)

In the linear model described above, the BLUE is

$$\hat{\underline{\theta}}(\underline{x}) = (A^T R^{-1} A)^{-1} A^T R^{-1} \underline{x}$$

and its covariance matrix is

$$R_{\hat{\theta}} = E[(\hat{\underline{\theta}} - \underline{\theta})(\hat{\underline{\theta}} - \underline{\theta})^T] = (A^T R^{-1} A)^{-1}$$

Proof 1 | $\hat{\underline{\theta}}$ is unbiased $\Longleftrightarrow$ $E\{H\underline{x}\} = \underline{\theta}$ $\forall \underline{\theta}$.

Now $E\{H\underline{x}\} = HE\{x\} = HE\{A\theta + \underline{w}\} = HA\underline{\theta}$.

Thus $\hat{\underline{\theta}}$ is unbiased $\Longleftrightarrow$ $HA\underline{\theta} = \underline{\theta}$ $\forall \underline{\theta}$ $\Longleftrightarrow$ $HA = I_{p\times p}$.

Now assume $\hat{\underline{\theta}}$ is unbiased and let's compute

$$\begin{aligned} Var(\hat{\underline{\theta}}) &= E\{(\hat{\underline{\theta}} - \underline{\theta})^T(\hat{\underline{\theta}} - \underline{\theta})\} \\ &= E\{(H\underline{x} - \underline{\theta})^T(H\underline{x} - \underline{\theta})\} \\ &= E\{(HA\underline{\theta} + H\underline{w} - \underline{\theta})^T(HA\underline{\theta} + H\underline{w} - \underline{\theta})\} \\ &= E\{(H\underline{w})^T(H\underline{w})\}. \end{aligned}$$

Denote

$$H = [\underline{h}_1 \cdots \underline{h}_p]^T = \begin{bmatrix} \underline{h}_1^T \\ \vdots \\ \underline{h}_p^T \end{bmatrix} \qquad (p \times N)$$

Then

$$\begin{aligned} Var(\hat{\underline{\theta}}) &= E\left\{\sum_{i=1}^{p} (\underline{h}_i^T \underline{w})^2\right\} \\ &= \sum_{i=1}^{p} E\{(\underline{h}_i^T \underline{w})^2\} \\ &= \sum_{i=1}^{p} E\{(\underline{h}_i^T \underline{w})(\underline{w}^T \underline{h}_i)\} \\ &= \sum_{i=1}^{p} \underline{h}_i^T R\, \underline{h}_i \end{aligned}$$

Thus, we need to solve the constrained optimization problem

$$\min_{H} \sum_{i=1}^{p} \underline{h}_i^T R \, \underline{h}_i$$

$$\text{s.t} \quad H \cdot A = I$$

By the theory of Lagrange multipliers, it suffices to solve the unconstrained problem

$$\min_{H, \underline{\lambda}} \quad L(H, \underline{\lambda})$$

where L is the Lagrangian and $\underline{\lambda}$ is a vector of real numbers called Lagrange multipliers, one for each equality constraint.

The Lagrangian is

$$L = \sum_{i=1}^{p} \underline{h}_i^T R \underline{h}_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_j^{(i)} \left(\underline{h}_i^T \underline{a}_j - \delta_{ij} \right)$$

Taking derivatives

$$\frac{\partial L}{\partial \underline{h}_i} = 2 R \underline{h}_i + \sum_{j=1}^{p} \lambda_j^{(i)} \underline{a}_j$$

$$= 2 R \underline{h}_i + A \underline{\lambda}^{(i)}$$

where $\underline{\lambda}^{(i)} = [\lambda_1^{(i)} \; \cdots \; \lambda_p^{(i)}]^T.$

$$\Longrightarrow \quad \hat{\underline{h}}_i = -\frac{1}{2} R^{-1} A \underline{\lambda}^{(i)}$$

From the constraint we know

$$A^T \hat{\underline{h}}_i = \underline{e}_i = [0 \; \cdots \; \overset{\text{ith position}}{1} \; \cdots \; 0]$$

$$\Longrightarrow \quad \underline{e}_i = -\frac{1}{2} A^T R^{-1} A \underline{\lambda}^{(i)}$$

$$\Longrightarrow \quad \underline{\lambda}^{(i)} = -2 \left(A^T R^{-1} A \right)^{-1} \underline{e}_i$$

$$\Longrightarrow \quad \hat{\underline{h}}_i = R^{-1} A \left(A^T R^{-1} A \right) \underline{e}_i$$

Assembling these scalar results, we have

$$\hat{H} = [\hat{h}_1 \; \cdots \; \hat{\underline{h}}_p]^T$$

$$= \left(R^{-1}A\,(A^TR^{-1}A)^{-1} \cdot I_{p\times p} \right)^T$$

$$= (A^TR^{-1}A)^{-1}A^TR$$

The covariance matrix of $\hat{\underline{\theta}}$ is

$$Cov(\hat{\underline{\theta}}) =$$

Proof 2| From proof 1, we know that the BLUE depends only on the first and second order moments of $\underline{X}$ (or equivalently, of $\underline{W}$).

Therefore, we may assume

$$\underline{W} \sim \mathcal{N}(\underline{0}, R).$$

We have previously seen that

$$\hat{\theta}(\underline{x}) = (A^T R^{-1} A)^{-1} A^T R^{-1} \underline{x}$$

is MVUE. Since this estimator is already linear, it is also the BLUE ■

Remark| As the above discussion notes, when $\underline{W}$ is Gaussian, the BLUE is the MVUE for the linear model, i.e. the BLUE is optimal.

Linear Minimum Mean Squared Error Estimation

Let us now turn to a Bayesian setting. There are some important similarities/differences w.r.t. classical linear estimation:

- the optimal linear estimator depends only on first and second order moments, but now for $\underline{X}$ and $\underline{\theta}$
- the optimal linear estimator always exists in the Bayesian setting
- Bayesian linear estimation has important geometric interpretations in terms of orthogonal projections in Hilbert space

Definition $\hat{\underline{\theta}}(\underline{x}) = \hat{H}\underline{x}$ is the LMMSE estimator if $\hat{H}$ minimizes

$$\text{BMSE}(H) := E_{X,\theta}\left[(\underline{\theta} - H\underline{X})^T(\underline{\theta} - H\underline{X})\right]$$

Introduce the notation

$$R_{\theta\theta} = E\left\{ (\underline{\theta} - E\underline{\theta})(\underline{\theta} - E\underline{\theta})^T \right\}$$

$$(p \times p)$$

auto-covariance

$$R_{\theta X} = E\left\{ (\underline{\theta} - E\underline{\theta})(\underline{X} - E\underline{X})^T \right\}$$

$$(p \times N)$$

cross-covariance

$$R_{XX} = E\left\{ (\underline{X} - E\underline{X})(\underline{X} - E\underline{X})^T \right\}$$

$$(N \times N)$$

<u>Theorem</u> If $E\underline{\theta} = \underline{0}$ and $E\underline{X} = \underline{0}$ then the LMMSE is

$$\hat{\underline{\theta}}(\underline{x}) = R_{\theta X} R_{XX}^{-1} \underline{x}$$

provided R_{XX} is positive definite.

Proof :

$$\begin{aligned} BMSE(H) &= E\left[(H\underline{x} - \underline{\theta})^T (H\underline{x} - \underline{\theta}) \right] \\ &= E\left[\sum_{i=1}^{p} \left(\underline{h}_i^T \underline{x} - \theta_i \right)^2 \right] \\ &= \sum_{i=1}^{p} E\left[\left(\underline{h}_i^T \underline{x} - \theta_i \right)^2 \right] \end{aligned}$$

To minimize this term w.r.t $H = [\underline{h}_1 \cdots \underline{h}_p]^T$, it suffices to optimize each $\underline{h}_i$ independently. That is, the vector linear estimation problem reduces to the scalar linear estimation problem.

Thus, let's drop the subscript i and focus on minimizing $E\left[(\underline{h}^T \underline{x} - \theta) \right]$ w.r.t. $\underline{h}$.

Now

$R_{\theta\theta} = Var(\theta)$ is 1×1

$R_{\theta x}$ is $1 \times N$

Now

$$\text{BMSE}(\underline{h}) = E_{\underline{x},\theta}\left[(\theta - \underline{h}^T \underline{x})^2 \right]$$

$$= E[\theta^2] - 2E[\underline{h}^T \underline{X} \cdot \theta] + E[(\underline{h}^T \underline{X})^2]$$

$$= E[\theta^2] - 2\underline{h}^T E[\underline{X} \cdot \theta] + E[\underline{h}^T \underline{X} \cdot \underline{X}^T \cdot \underline{h}]$$

$$= R_{\theta\theta} - 2\underline{h}^T R_{X\theta} + \underline{h}^T R_{XX} \underline{h} \qquad \circledast$$

By completing the square, we have

$$\text{BMSE}(\underline{h}) = (\underline{h} - R_{XX}^{-1} R_{X\theta})^T R_{XX} (\underline{h} - R_{XX}^{-1} R_{X\theta})$$

$$- \underbrace{R_{X\theta}^T R_{XX}^{-1} R_{X\theta} + R_{\theta\theta}}_{\text{independent of } \underline{h}}$$

Since R_{XX} is positive definite, the unique minimizer is

$$\hat{\underline{h}} = R_{XX}^{-1} R_{X\theta}$$

$$\Updownarrow$$

$$\hat{\underline{h}}^T = R_{\theta X} R_{XX}^{-1} .$$

Alternatively, you could apply vector calculus to minimize $\circledast$.

Now the vector LMMSE estimator is obtained by stacking these scalar estimators. ▨

Nonzero Means and Affine Estimators

When $E\underline{\theta} \neq \underline{0}$ or $E\underline{X} \neq \underline{0}$, the LMMSE is generalized by the affine MMSE estimator (AMMSE), which is defined to be the estimator $\hat{\underline{\theta}}(\underline{x}) = \hat{H}\underline{x} + \hat{\underline{b}}$

minimizing

$$BMSE(H, \underline{b}) = E\left[(\underline{\theta} - H\underline{X} - \underline{b})^T(\underline{\theta} - H\underline{X} - \underline{b})\right].$$

Theorem| The AMMSE is

$$\hat{\underline{\theta}}(\underline{x}) = E\underline{\theta} + R_{\theta x} \cdot R_{xx}^{-1}(\underline{x} - E\underline{X})$$

i.e.

$$\hat{H} = R_{\theta x} R_{xx}^{-1}$$

$$\hat{\underline{b}} = E\underline{\theta} - \hat{H} E\underline{X}$$

Proof| Introduce the variables

Skip in lecture

$$\underline{\theta}' = \underline{\theta} - E\underline{\theta}$$

$$\underline{X}' = \underline{X} - E\underline{X}$$

Then $BMSE(H, \underline{b})$

$$= E\Big[(\underline{\theta}' - H\underline{X}' - (\underline{b} - E\underline{\theta} + HE\underline{X}))^T \cdot (\underline{\theta}' - H\underline{X}' - (\underline{b} - E\underline{\theta} + HE\underline{X}))\Big]$$

$$= E\left[(\underline{\theta}' - H\underline{X}')^T(\underline{\theta}' - H\underline{X}')\right]$$

$$-2E\left[(\underline{\theta}' - H\underline{X}')^T(\underline{b} - E\underline{\theta} + HE\underline{X})\right]$$

$$+ E\left[(\underline{b} - E\underline{\theta} + HE\underline{X})^T(\underline{b} - E\underline{\theta} + HE\underline{X})\right]$$

The second term is zero since $\underline{b} - E\underline{\theta} + HE\underline{X}$ is constant and $\underline{\theta}' - H\underline{X}'$ is zero mean.

For fixed H, the optimal $\underline{b}$ must minimize the third term, which can be made zero by taking

$$\underline{b} = E\underline{\theta} - HE\underline{X}. \qquad \circledast$$

Therefore, H is obtained by minimizing the first term. Since $\underline{\theta}'$, $\underline{X}'$ are zero mean, $\hat{H}$ is the LMMSE estimator for $\underline{\theta}'$, $\underline{X}'$, i.e.

$$\hat{H} = R_{\theta' X'} R_{X'X'}^{-1} = R_{\theta X} R_{XX}^{-1} \qquad \square$$

<u>Remark 1</u> From the above argument and $\circledast$, we see that the optimal constant estimator $(H = 0)$ is

$$\hat{\underline{b}} = E\underline{\theta},$$

the prior mean.

Connection to the Jointly Gaussian Case

The LMMSE and AMMSE estimators look exactly like the MMSE estimators for the case where $\underline{\theta}, \underline{X}$ are jointly Gaussian. In fact, we can use the Gaussian case to give an alternate derivation of the LMMSE/AMMSE.

Let's consider the AMMSE. From the definition of the Bayesian MSE

$$BMSE(H, \underline{b}) = E\left[(\underline{\theta} - H\underline{X} - \underline{b})^T (\underline{\theta} - H\underline{X} - \underline{b}) \right]$$

is can be seen that this criterion depends only on the first and second order moments of $\underline{\theta}$ and $\underline{X}$. Therefore, we can assume the higher order moments are whatever we want. So let's assume $\underline{\theta}$ & $\underline{X}$ are jointly Gaussian with the given means and covariances:

$$\begin{bmatrix} \underline{X} \\ \underline{\theta} \end{bmatrix} \sim N\left(\begin{bmatrix} E\underline{X} \\ E\underline{\theta} \end{bmatrix}, \begin{bmatrix} R_{XX} & R_{X\theta} \\ R_{\theta X} & R_{\theta\theta} \end{bmatrix} \right)$$

We know the MMSE estimator is the posterior mean,
which is

$$E\underline{\theta} + R_{\theta x} R_{xx}^{-1} (\underline{x} - E\underline{x})$$

by the Gaussian conditioning principle.

Since this estimator is affine, it is the AMMSE ■

Theorem | Bayesian Gauss-Markov Theorem

Assume

$$\underline{X} = A\underline{\theta} + \underline{W}$$

where

A is full rank, $N \times p$, known

$E\underline{\theta}$, $R_{\theta\theta}$ known

$E\underline{W} = \underline{0}$, R_{ww} known

$R_{\theta W} = 0_{p\times N}$ ($\underline{\theta}$ & $\underline{W}$ uncorrelated)

Then the affine MMSE estimator is

$$\underline{\hat{\theta}}(\underline{x}) = E\underline{\theta} + R_{\theta\theta}A^T(AR_{\theta\theta}A^T + R_{ww})^{-1}(\underline{x} - AE\underline{\theta})$$

$$= E\underline{\theta} + (R_{\theta\theta}^{-1} + A^T R_{ww}^{-1} A)^{-1} A^T R_{ww}^{-1}(\underline{x} - AE\underline{\theta})$$

How would you prove this?

Proofs

1. Show $R_{\theta X} =$

ⓐ

$R_{XX} =$

and apply the AMMSE result.

2. Argue that the BMSE depends only on the first and second order moments of $\underline{\theta}$ and $\underline{W}$. Then assume $\underline{\theta}, \underline{W}$ are jointly Gaussian. The MMSE estimator has the stated form, and since it is affine it is the AMMSE.

Exercise Suppose

$$X_i = A + W_i \;, \qquad i = 1, \ldots, N$$

where

$$\left.\begin{array}{l} A \sim \text{unif}(-A_0, A_0) \\ W_i \overset{iid}{\sim} N(0, \sigma_w^2) \end{array}\right\} \text{independent}$$

1. Find the posterior $f(A/\underline{x})$
2. Can you compute $E[A/\underline{x}]$?
3. Determine the LMMSE estimator of A.

Solution

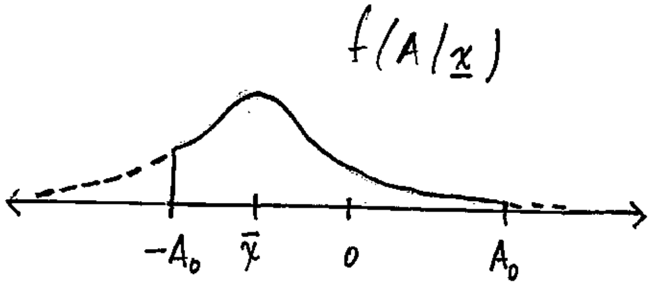
1. $$f(A|\underline{x}) \propto f(\underline{x}|A)\cdot f(A)$$

$$= \underbrace{\prod_{i=1}^{N} \phi(x_i \mid A, \sigma^2)} \cdot \frac{1}{2A_0} I_{[-A_0, A_0]}(A)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - A)^2\right\}$$

$$= \exp\left\{\frac{-1}{2\sigma^2}\left[\Sigma(x_i - \bar{x})^2 + N(\bar{x} - A)^2\right]\right\}$$

$$\Longrightarrow f(A|\underline{x}) \propto \exp\left\{\frac{(A-\bar{x})^2}{2(\sigma^2/N)}\right\}\cdot I_{[-A_0, A_0]}(A)$$



Truncated normal.

2. $E[A \mid \underline{x}] = \int_{-\infty}^{\infty} A f(A/\underline{x})\, dA$

$$= \frac{\int_{-A_0}^{A_0} A\, \phi(A \mid \bar{x}, \frac{\sigma^2}{N})\, dA}{\int_{-A_0}^{A_0} \phi(A \mid \bar{x}, \frac{\sigma^2}{N})\, dA} \leftarrow \boxed{\text{Normalization Constant}}$$

Numerator : closed form solution

Denominator : no closed form solution

Note: $E[A \mid \underline{x}]$ is non linear.

3. $\underline{X} = \underline{1} \cdot A + \underline{W}, \quad \underline{W} \sim N(\underline{0}, \sigma_w^2 I)$

$$\sigma_A^2 = \int_{-A_0}^{A_0} \frac{x^2}{2A_0}\, dx = \frac{1}{6A_0} x^3 \Big|_{-A_0}^{A_0} = \frac{A_0^2}{3}$$

$$\Longrightarrow \hat{A} = \left((\sigma_A^2)^{-1} + (\underline{1}^T \sigma_w^2 I\, \underline{1})^{-1}\right)^{-1} (\sigma_w^2 I)^{-1} \underline{1}^T \underline{x}$$

$$= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}} \bar{x}$$

$$= \frac{A_0^2/3}{A_0^2/3 + \sigma^2/N} \bar{x}$$

Remarks

a) To compute the LMMSE estimator of A, we didn't need to know it was uniform, just its mean and variance.

b) Estimator is the same as the MMSE estimator for the prior

$$A \sim N\left(0, \frac{A_0^2}{3}\right)$$

The Orthogonality Principle

The LMMSE estimator satisfies

$$R_{\theta X} = \underline{\hat{h}}^T R_{XX}$$

or equivalently

$$\begin{aligned}\underline{0} &= R_{\theta X} - \underline{\hat{h}}^T R_{XX} \\ &= E\left[\theta \underline{X}^T - \underline{\hat{h}}^T \underline{X}\,\underline{X}^T\right] \\ &= E\left[(\theta - \underline{\hat{h}}^T \underline{X})\underline{X}^T\right] \\ &= E\left[(\theta - \hat{\theta})\underline{X}^T\right]\end{aligned}$$

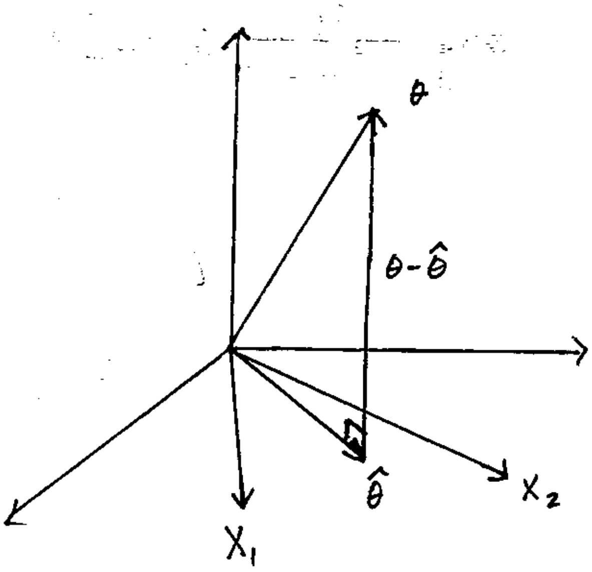
In words, this says that the prediction error is orthogonal to every measurement variable,

$$\theta - \hat{\theta} \perp X_i \qquad \forall i$$

where orthogonality is with respect to the inner product given by expectation.

As a consequence,

$$\theta - \hat{\theta} \perp \underline{h}^T \underline{X} \qquad \forall \underline{h}.$$



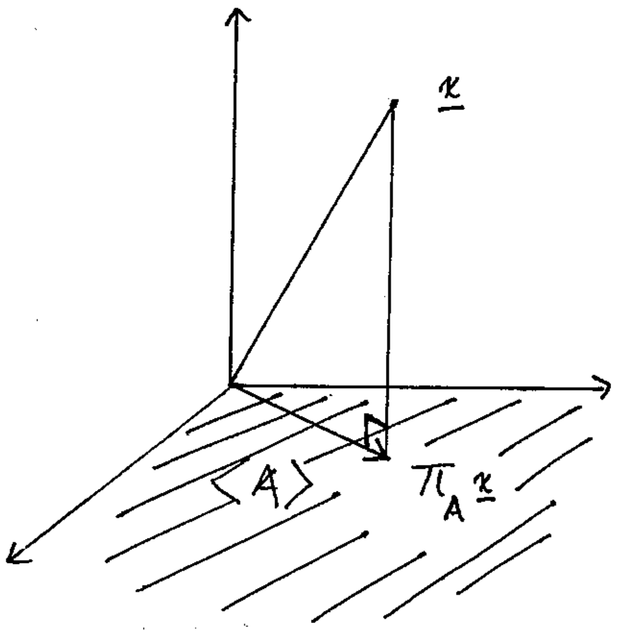
Therefore, the prediction error is orthogonal to the entire subspace of linear estimators.

This is one manifestation of the orthogamility principle.

The orthogonality principle is also manifested in nonstatistical projections

In particular, if $\underline{x} \in \mathbb{R}^N$ and $\langle A \rangle \subseteq \mathbb{R}^N$, then

$$\underline{x} - \Pi_A \underline{x} \quad \perp \quad \underline{u} \qquad \forall \underline{u} \in \langle A \rangle$$



Here the inner product is the standard dot product.

To prove this fact, suppose $\underline{u} \in \langle A \rangle$. Then $\underline{u} = A\underline{\phi}$ for some $\underline{\phi}$,

$$\begin{aligned} \langle \underline{x} - \Pi_A \underline{x}, \underline{u} \rangle &= \underline{u}^T (\underline{x} - \Pi_A \underline{x}) \\ &= \underline{\phi}^T \underline{A}^T \left(\underline{x} - A(A^TA)^{-1}A^T\underline{x} \right) \\ &= \underline{\phi}^T A^T \underline{x} - \underline{\phi}^T A^T \underline{x} = 0 \end{aligned}$$

In full generality...

Definition | Let $\mathcal{H}$ be an inner product space, and let $S \subseteq \mathcal{H}$ be a linear subspace. The orthogonal complement of S is

$$S^{\perp} = \{ \underline{v} \in \mathcal{H} \mid \underline{v} \perp \underline{u} \quad \forall \underline{u} \in S \}$$

Theorem | Let $\mathcal{H}$ be a Hilbert space and $S \subseteq \mathcal{H}$ a closed linear subspace.

1. Projection Theorem : $\mathcal{H} = S \oplus S^{\perp}$, i.e. $\forall \underline{w} \in \mathcal{H}$, $\exists$ unique $\underline{u} \in S$, $\underline{v} \in S^{\perp}$ s.t. $\underline{w} = \underline{u} + \underline{v}$. Thus we can define the (orthogonal) projection onto S, $\Pi_S(\underline{w}) = \underline{u}$.

2. Closest point property : $\Pi_S \underline{w}$ is the unique solution of $\min\limits_{\underline{u} \in S} \| \underline{w} - \underline{u} \|$ where $\| \cdot \|$ is the norm induced by $\langle \cdot , \cdot \rangle$.

3. Orthogonality principle : For all $\underline{w} \in \mathcal{H}$,

$$\underline{w} - \Pi_S \underline{w} \perp \underline{u} \qquad \forall \underline{u} \in S$$

i.e., $\underline{w} - \Pi_S \underline{w} \in S^{\perp}$.

Note | 2 + 3 follow easily from 1. See Moon and Stirling for details.

Application to LMMSE Estimation

Consider the space $\mathcal{H}$ of all scalar random variables with zero mean and finite variance. It can be shown that $\mathcal{H}$ is a Hilbert space with inner product

$$\langle V_1 , V_2 \rangle = E\{V_1 \cdot V_2\}$$

Let $X_1, \ldots, X_N$ be random measurements and define the subspace

$$S = \{ \underline{h}^T \underline{X} \mid \underline{h} \in \mathbb{R}^N \}$$

where $\underline{X} = [X_1 \cdots X_N]^T$.

If $\theta \in \mathcal{H}$ is a scalar parameter of interest, the LMMSE estimator is the <u>projection</u>

$$\begin{aligned} \hat{\theta} &= \Pi_S \, \theta \\ &= \operatorname*{argmin}_{\hat{\theta} \in S} \| \theta - \hat{\theta} \|^2 \\ &= \operatorname*{argmin}_{\underline{h} \in \mathbb{R}^N} E\left[(\theta - \underline{h}^T \underline{X})^2 \right] \end{aligned}$$

By the orthogonality principle,

$$\theta - \hat{\theta} \perp u \qquad \forall u \in S.$$

Writing $\hat{\theta} = \underline{\hat{h}}^T \underline{X}$ and taking $u = X_i$, $i = 1, \ldots, N$ we have

$$\theta - \hat{\theta} \perp X_i \iff \theta - \underline{\hat{h}}^T \underline{X} \perp X_i$$

$$\iff E\left[(\theta - \underline{\hat{h}}^T \underline{X}) X_i\right] = 0$$

Applying this for $i = 1, \ldots, N$ we have

$$E\left[(\theta - \underline{\hat{h}}^T \underline{X}) \underline{X}^T\right] = [0 \cdots 0]$$

$$\Updownarrow$$

$$R_{\theta X} = \underline{\hat{h}}^T R_{XX}$$

$$\Updownarrow$$

$$\underline{\hat{h}}^T = R_{\theta X} R_{XX}^{-1}$$

Conclusion: The orthogonality principle gives us another proof of the form of the LMMSE estimator.

We will apply the orthogonality principle often in our study of filtering. Furthermore, the orthogonality principle applies when there are an infinite number of equations.

Terminology

The equations

$$R_{xx} \underline{\hat{h}} = R_{x\theta}$$

are called the Wiener-Hopf or normal equations.

The optimal $\underline{\hat{h}}$ is also called a Wiener estimator, after Norbert Wiener, especially in the context of estimating a signal in additive noise.

Summary

- Classical and Bayesian linear estimation
  - only depends on first and second order moments
  - optimal in Gaussian case
  - proof: show that solution determined by 1st & 2nd order moments, assume Gaussianity, and invoke Gaussian results
- classical BLUE
  - solution of constrained quadratic minimization, doesn't always exist
- Bayesian LMMSE
  - solution of unconstrained quadratic minimization, always exists
  - projection in Hilbert space, obeys orthogonality principle

Key

a. $$R_{\theta x} = E\left[(\underline{\theta} - E\underline{\theta})(\underline{x} - E\underline{x})^T\right]$$

$$= E\left[(\underline{\theta} - E\underline{\theta})(A\underline{\theta} + \underline{w} - AE\underline{\theta})^T\right]$$

$$= E\left[(\underline{\theta} - E\underline{\theta})\underline{w}^T\right] + E\left[(\underline{\theta} - E\underline{\theta})(\underline{\theta} - E\underline{\theta})^T A^T\right]$$

$$= 0 + R_{\theta\theta}A^T = R_{\theta\theta}A^T$$

$$R_{xx} = E\left[(\underline{X} - E\underline{X})(\underline{X} - E\underline{X})^T\right]$$

$$= E\left[(A\underline{\theta} + \underline{w} + AE\underline{\theta})(A\underline{\theta} + \underline{w} + AE\underline{\theta})^T\right]$$

$$= A R_{\theta\theta} A^T + R_{ww}$$

where we use the assumption that $\underline{\theta}$, $\underline{w}$ are uncorrelated, i.e. $E[(\underline{\theta} - E\underline{\theta})\underline{w}^T] = 0$.