ENEE 621
SPRING 2016
DETECTION AND ESTIMATION THEORY

MEAN-SQUARE ESTIMATION

1 The basic setting

Throughout, p and k are arbitrary positive integers. Let the random parameter ϑ be modelled as an $\mathbb{R}^p$-valued rv, while the observation rv $\boldsymbol{Y}$ is an $\mathbb{R}^k$-valued rv. Here the family of distributions $\{F_\theta,\ \theta \in \Theta\}$ on $\mathbb{R}^k$ is interpreted as conditional distributions in the sense that

$$\mathbb{P}\left[\boldsymbol{Y} \leq \boldsymbol{y} | \vartheta = \theta\right] = F_\theta(\boldsymbol{y}), \qquad \begin{array}{c} \boldsymbol{y} \in \mathbb{R}^k \\ \theta \in \Theta. \end{array}$$

We assume that the rv ϑ is a *second-order* rv, namely,

$$\mathbb{E}\left[|\vartheta_i|^2\right] < \infty, \quad i = 1, \ldots, p.$$

We shall use $\mathcal{B}(k;p)$ to denote the collection of all Borel mappings $\mathbb{R}^k \to \mathbb{R}^p$. With $r \geq 1$, let $\mathcal{G}_r(p;\boldsymbol{Y})$ denote the collection of all estimators for ϑ on the basis of $\boldsymbol{Y}$ with finite r^{th} moment. Formally,

$$\mathcal{G}_r(p;\boldsymbol{Y}) = \left\{g \in \mathcal{B}(k;p) :\ \mathbb{E}\left[|g_i(\boldsymbol{Y})|^r\right] < \infty, \quad i = 1, \ldots, p\right\}.$$

We shall also introduce $\mathcal{L}(k;p)$ as the collection of *affine* estimators for ϑ on the basis of $\boldsymbol{Y}$. Thus, the estimator g in $\mathcal{B}(k;p)$ is an affine estimator in $\mathcal{L}(k;p)$ if it takes the form

$$g(\boldsymbol{y}) = A\boldsymbol{y} + \boldsymbol{b}, \quad \boldsymbol{y} \in \mathbb{R}^k$$

for some $p \times k$ matrix A and a vector $\boldsymbol{b}$ in $\mathbb{R}^p$.

The following easy fact will be found handy in a number of places.

Fact 1.1 *With scalars a and b, if*

$$at + bt^2 \geq 0, \quad t \in \mathbb{R},$$

then necessarily $a = 0$.

Proof. It is plain that

$$a + bt \geq 0, \quad t > 0$$

and

$$-a + b|t| \geq 0, \quad t < 0.$$

Letting t go to zero in both sets of inequalities we find that $a \geq 0$ and $a \leq 0$ both hold, hence $a = 0$. ■

2 Minimum Mean Square Error (MMSE) Estimation

The MMSE problem can be formulated as follows: Find $g^\star$ in $\mathcal{G}_2(k; \boldsymbol{Y})$ such that

(1) $$\mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] \leq \mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right], \quad g \in \mathcal{G}_2(k; \boldsymbol{Y}).$$

Any estimator $g^\star$ in $\mathcal{G}_2(k; \boldsymbol{Y})$ which satisfies (1) is known as a MMSE estimator of ϑ on the basis of $\boldsymbol{Y}$.

Theorem 2.1 *The estimator $g^\star$ in $\mathcal{G}_2(p; \boldsymbol{Y})$ satisfies*

$$\mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] \leq \mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right], \quad g \in \mathcal{G}_2(p; \boldsymbol{Y})$$

if and only if the Orthogonality Principle

(2) $$\mathbb{E}\left[(\vartheta - g^\star(\boldsymbol{Y}))' \, h(\boldsymbol{Y})\right] = 0, \quad h \in \mathcal{G}_2(p; \boldsymbol{Y})$$

holds.

This characterization is geometric in nature, and points to $g^\star$ as the projection of ϑ on the subspace of second-order rvs

$$\{g(\boldsymbol{Y}) : \; g \in \mathcal{G}_2(p; \boldsymbol{Y})\}.$$

This is a subspace of $L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p)$, the space of all second-order $\mathbb{R}^p$-valued rvs. Orthogonality in $L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p)$ is defined by

$$\mathbb{E}\left[\boldsymbol{\xi}'\boldsymbol{\eta}\right] = 0, \quad \boldsymbol{\xi}, \boldsymbol{\eta} \in L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p).$$

The Orthogonality Principle (2) can be restated as saying that the error $\vartheta - g^\star(\boldsymbol{Y})$ is orthogonal to the subspace $\{g(\boldsymbol{Y}) : \ g \in \mathcal{G}_2(p;\boldsymbol{Y})\}$.

As an immediate consequence of Theorem 2.1 we have the following uniqueness result.

Corollary 2.1 *If $g_1^\star$ and $g_2^\star$ are estimators in $\mathcal{G}_2(p;\boldsymbol{Y})$ such that*

$$\mathbb{E}\left[\|\vartheta - g_i^\star(\boldsymbol{Y})\|^2\right] \leq \mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right], \quad g \in \mathcal{G}_2(p;\boldsymbol{Y})$$

for each $i = 1, 2$, then we have

$$\mathbb{P}\left[g_1^\star(\boldsymbol{Y}) = g_2^\star(\boldsymbol{Y})\right] = 1.$$

Proof. Use the Orthogonality Principle (2) with $h = g_1^\star - g_2^\star$ for both $g_1^\star$ and $g_2^\star$. ■

Basic ideas behind the proof of Theorem 2.1

With estimator g in $\mathcal{G}_2(p;\boldsymbol{Y})$, note that

$$\begin{aligned}
&\mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right] \\
&\quad = \mathbb{E}\left[(\vartheta - g(\boldsymbol{Y}))'(\vartheta - g(\boldsymbol{Y}))\right] \\
&\quad = \mathbb{E}\left[((\vartheta - g^\star(\boldsymbol{Y})) + (g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})))'((\vartheta - g^\star(\boldsymbol{Y})) + (g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})))\right] \\
&\quad = \mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - g^\star(\boldsymbol{Y}))'(g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y}))\right] \\
&\qquad\quad + \mathbb{E}\left[\|g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})\|^2\right]
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] \\
&= \mathbb{E}\left[\|g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - g^\star(\boldsymbol{Y}))'(g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y}))\right].
\end{aligned} \tag{3}$$

If $g^\star$ in $\mathcal{G}_2(p;\boldsymbol{Y})$ satisfies the Optimality Principle (2), then the equality (3) implies

$$\mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] = \mathbb{E}\left[\|g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})\|^2\right]$$

since $g - g^\star$ is an element of $\mathcal{G}_2(p; \boldsymbol{Y})$ as both $g^\star$ and g are. It follows that

$$\mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - g^\star(\boldsymbol{Y})\|^2\right] \geq 0, \quad g \in \mathcal{G}_2(p; \boldsymbol{Y})$$

and $g^\star$ is an MMSE estimator.

Conversely, if $g^\star$ is an MMSE estimator, then (3) implies

$$\mathbb{E}\left[\|g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - g^\star(\boldsymbol{Y}))'\left(g^\star(\boldsymbol{Y}) - g(\boldsymbol{Y})\right)\right] \geq 0$$

for every g element of $\mathcal{G}_2(p; \boldsymbol{Y})$. Thus, with h an *arbitrary* element of $\mathcal{G}_2(p; \boldsymbol{Y})$ and t in $\mathbb{R}$, consider the estimator $g_t : \mathbb{R}^k \to \mathbb{R}^p$ given by

$$g_t(\boldsymbol{y}) = g^\star(\boldsymbol{t}) + th(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

If is plain that g_t is also an element of $\mathcal{G}_2(p; \boldsymbol{Y})$. Applying the last inequality with $g = g_t$ we conclude that

$$t^2\mathbb{E}\left[\|g(\boldsymbol{Y})\|^2\right] + 2t\mathbb{E}\left[(\vartheta - g^\star(\boldsymbol{Y}))' h(\boldsymbol{Y})\right] \geq 0, \quad t \in \mathbb{R}$$

and Fact 1.1 immediately leads to the Optimality Principle (2).

To identify the MMSE estimator we focus on the following problem: With $\boldsymbol{\xi}$ be a second-order $\mathbb{R}^p$-valued rv, we seek $\boldsymbol{a}^\star$ in $\mathbb{R}^p$ such that

$$\mathbb{E}\left[\|\boldsymbol{\xi} - \boldsymbol{a}^\star\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{\xi} - \boldsymbol{a}\|^2\right], \quad \boldsymbol{a} \in \mathbb{R}^p.$$

The solution to this problem is well known to be unique, and is given by

$$\boldsymbol{a}^\star = \mathbb{E}\left[\boldsymbol{\xi}\right].$$

Returning to the MMSE problem, we recall that

$$\mathbb{E}\left[\|\vartheta - g(\boldsymbol{Y})\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\vartheta - \boldsymbol{a}\|^2 | \boldsymbol{Y} = \boldsymbol{y}\right]_{\boldsymbol{y}=\boldsymbol{Y}, \boldsymbol{a}=g(\boldsymbol{Y})}\right]$$

for every estimator g in $\mathcal{G}_2(p; \boldsymbol{Y})$. This fact readily leads to concluding

$$g^\star(\boldsymbol{y}) = \mathbb{E}\left[\vartheta | \boldsymbol{Y} = \boldsymbol{y}\right], \quad \boldsymbol{y} \in \mathbb{R}^k.$$

It is customary to write

$$g^\star(\boldsymbol{Y}) = \mathbb{E}\left[\vartheta | \boldsymbol{Y}\right]$$

where the right handside is understood as the conditional expectation of the rv ϑ given the σ-field generated by the rv $\boldsymbol{Y}$. The reason to proceed via the Orthogonality Principle is to show the parallel with the next problem where only affine estimators are considered.

3 Linear Mean Square Error (LMSE) Estimation

Assume that the observation rv $\boldsymbol{Y}$ is also a second-order rv, i.e.,

$$\mathbb{E}\left[|Y_j|^2\right] < \infty, \quad j = 1, \ldots, k.$$

The LMSE problem can be formulated as follows: Find $\ell^\star$ in $\mathcal{L}(k;p)$ such that

$$\mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] \le \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right], \quad \ell \in \mathcal{L}(k;p).$$

We refer to this affine estimator $\ell^\star$ in $\mathcal{L}(k;p)$ as the *Linear Mean Square Error* (LMSE) estimator of ϑ on the basis of $\boldsymbol{Y}$. It is characterized by the following version of the *Orthogonality Principle*.

Theorem 3.1 *The estimator $\ell^\star$ in $\mathcal{L}(k;p)$ satisfies*

$$\mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] \le \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right], \quad \ell \in \mathcal{L}(k;p)$$

if and only if the Orthogonality Principle

$$\mathbb{E}\left[(\vartheta - \ell^\star(\boldsymbol{Y}))' h(\boldsymbol{Y})\right] = 0, \quad h \in \mathcal{L}(k;p) \tag{4}$$

holds.

This characterization is also geometric in nature, pointing to the LMSE estimator to $\ell^\star$ as the projection of ϑ on the subspace of second-order rvs

$$\{\ell(\boldsymbol{Y}) : \ \ell \in \mathcal{L}(k;p))\}.$$

This is a also subspace of $L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p)$, with the Orthogonality Principle (4) stating that the error $\vartheta - g^\star(\boldsymbol{Y})$ is orthogonal to the subspace $\{\ell(\boldsymbol{Y}) : \ \ell \in \mathcal{L}(k;p))\}$.

Basic ideas behind the proof of Theorem 3.1

With affine estimator ℓ in $\mathcal{L}(k;p)$, this time we note that

$$\begin{aligned}
&\mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] \\
&= \mathbb{E}\left[(\vartheta - \ell(\boldsymbol{Y}))' (\vartheta - \ell(\boldsymbol{Y}))\right] \\
&= \mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - \ell^\star(\boldsymbol{Y}))' (\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y}))\right] \\
&\qquad + \mathbb{E}\left[\|\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\|^2\right]
\end{aligned}$$

so that

$$
\begin{aligned}
& \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] \\
(5) \qquad = \; & \mathbb{E}\left[\|\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - \ell^\star(\boldsymbol{Y}))'\left(\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\right)\right].
\end{aligned}
$$

If $\ell^\star$ in $\mathcal{L}_2(k;p)$ satisfies the Optimality Principle (4), then the equality (5) implies

$$
\mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] = \mathbb{E}\left[\|\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\|^2\right]
$$

since $\ell - \ell^\star$ is an element of $\mathcal{L}(k;p)$ as both $\ell^\star$ and ℓ are. It follows that

$$
\mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] - \mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] \geq 0, \quad \ell \in \mathcal{L}(k;p)
$$

and $\ell^\star$ is an LMSE estimator.

Conversely, if $\ell^\star$ is an LMSE estimator, then (5) implies

$$
\mathbb{E}\left[\|\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\|^2\right] + 2\mathbb{E}\left[(\vartheta - \ell^\star(\boldsymbol{Y}))'\left(\ell^\star(\boldsymbol{Y}) - \ell(\boldsymbol{Y})\right)\right] \geq 0
$$

for every ℓ element of $\mathcal{L}(k;p)$. Thus, with h an *arbitrary* element of $\mathcal{L}(k;p)$ and t in $\mathbb{R}$, consider the estimator $\ell_t : \mathbb{R}^k \to \mathbb{R}^p$ given by

$$
\ell_t(\boldsymbol{y}) = \ell^\star(\boldsymbol{t}) + th(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.
$$

If is plain that g_t is also an element of $\mathcal{L}(k;p)$. Applying the last inequality with $\ell = \ell_t$ we conclude that

$$
t^2\mathbb{E}\left[\|h(\boldsymbol{Y})\|^2\right] + 2t\mathbb{E}\left[(\vartheta - \ell^\star(\boldsymbol{Y}))' h(\boldsymbol{Y})\right] \geq 0, \quad t \in \mathbb{R}
$$

and Fact 1.1 immediately leads to the Optimality Principle (4).

4 Algebraic characterization of the LMSE estimators

The main results concerning the *existence* and *algebraic characterization* of the LMSE estimators are given next. First some notation:We shall write

$$
\boldsymbol{\mu}_Y = \mathbb{E}\left[\boldsymbol{Y}\right] \quad \text{and} \quad \boldsymbol{\mu}_\vartheta = \mathbb{E}\left[\vartheta\right].
$$

Next, the appropriate covariance matrices are given by

$$\Sigma_{\vartheta Y} = \mathrm{Cov}\left[\vartheta, \boldsymbol{Y}\right] = \mathbb{E}\left[\left(\vartheta - \boldsymbol{\mu}_{\vartheta}\right)\left(\boldsymbol{Y} - \boldsymbol{\mu}_{Y}\right)'\right]$$

and

$$\Sigma_{Y} = \mathrm{Cov}\left[\boldsymbol{Y}\right] = \mathbb{E}\left[\left(\boldsymbol{Y} - \boldsymbol{\mu}_{Y}\right)\left(\boldsymbol{Y} - \boldsymbol{\mu}_{Y}\right)'\right].$$

The matrices $\Sigma_{\vartheta Y}$ and Σ_Y are $p \times k$ and $k \times k$ matrices, respectively.

Theorem 4.1 *There always exists an affine estimator $\ell^\star$ in $\mathcal{L}(k;p)$ which satisfies*

$$\mathbb{E}\left[\|\vartheta - \ell^\star(\boldsymbol{Y})\|^2\right] \leq \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right], \quad \ell \in \mathcal{L}(k;p).$$

With such an estimator $\ell^\star : \mathbb{R}^k \to \mathbb{R}^p$ being given by

$$\ell^\star(\boldsymbol{y}) = A^\star \boldsymbol{y} + \boldsymbol{b}^\star, \quad \boldsymbol{y} \in \mathbb{R}^k$$

for some $p \times k$ matrix $A^\star$ and a vector $\boldsymbol{b}^\star$ in $\mathbb{R}^p$, then $A^\star$ and $\boldsymbol{b}^\star$ satisfy the normal equations

$$A^\star \Sigma_Y = \Sigma_{\vartheta Y} \tag{6}$$

and

$$\boldsymbol{b}^\star = \boldsymbol{\mu}_{\vartheta} - A^\star \boldsymbol{\mu}_Y. \tag{7}$$

The normal equations (6)-(7) have a unique solution when Σ_Y is invertible.

Corollary 4.1 *If Σ_Y is invertible, then the LMSE estimator $\ell^\star$ is uniquely determined by*

$$\ell^\star(\boldsymbol{y}) = \boldsymbol{\mu}_{\vartheta} + \Sigma_{\vartheta Y}\Sigma_Y^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_Y\right), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

If Σ_Y is not invertible, there is is still uniqueness in the following sense; see analogy with Corollary 2.1.

Corollary 4.2 *Let $\ell_1^\star$ and $\ell_2^\star$ be affine estimators in $\mathcal{L}(k;p)$ such that*

$$\mathbb{E}\left[\|\vartheta - \ell_i^\star(\boldsymbol{Y})\|^2\right] \leq \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right], \quad \ell \in \mathcal{L}(k;p)$$

for each $i = 1, 2$, then we have

$$\mathbb{P}\left[\ell_1^\star(\boldsymbol{Y}) = \ell_2^\star(\boldsymbol{Y})\right] = 1.$$

In analogy with the notation used for MMSE estimators, we shall write

$$\ell^\star(\boldsymbol{y}) = \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}=\boldsymbol{y}\right], \quad \boldsymbol{y} \in \mathbb{R}^k$$

and

$$\ell^\star(\boldsymbol{Y}) = \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right].$$

This last rv is unambiguously defined (in the a.s. sense) in view of Corollary 4.2.

5 A proof of Theorem 4.1

The proof has three parts:

Part 1: Given the $p \times k$ matrix A, there is always a best vector $\boldsymbol{b} = \boldsymbol{b}(A)$ in $\mathbb{R}^p$ For any $p \times k$ matrix A and vector $\boldsymbol{b}$ in $\mathbb{R}^p$, note that

$$\begin{aligned} & \mathbb{E}\left[\|\vartheta - (A\boldsymbol{Y} + \boldsymbol{b})\|^2\right] \\ &= \mathbb{E}\left[\|(\vartheta - \boldsymbol{\mu}_\vartheta) - A(\boldsymbol{Y} - \boldsymbol{\mu}_Y) + (\boldsymbol{\mu}_\vartheta - (A\boldsymbol{\mu}_Y + \boldsymbol{b}))\|^2\right] \\ &= \mathbb{E}\left[\|(\vartheta - \boldsymbol{\mu}_\vartheta) - A(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\|^2\right] + \|\boldsymbol{\mu}_\vartheta - (A\boldsymbol{\mu}_Y + \boldsymbol{b})\|^2 \\ &\geq \mathbb{E}\left[\|(\vartheta - \boldsymbol{\mu}_\vartheta) - A(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\|^2\right] \end{aligned} \tag{8}$$

if we select $\boldsymbol{b} = \boldsymbol{b}(A)$ with

$$\boldsymbol{b}(A) = \boldsymbol{\mu}_\vartheta - A\boldsymbol{\mu}_Y,$$

so that (7) holds.

Part 2: The optimal $p \times k$ matrix A is characterized by the normal equations (6)-(7) Part 1 shows that any LMSE estimator $\ell^\star : \mathbb{R}^k \to \mathbb{R}^p$ is of the form

$$\ell^\star(\boldsymbol{y}) = A^\star\boldsymbol{y} + \boldsymbol{b}^\star, \quad \boldsymbol{y} \in \mathbb{R}^k$$

with $\boldsymbol{b}^\star$ necessarily given by

$$\boldsymbol{b}^\star = \boldsymbol{\mu}_\vartheta - A^\star\boldsymbol{\mu}_Y.$$

The Orthogonality Principle states that $A^\star$ and $\boldsymbol{b}^\star$ are completely characterized by

$$\mathbb{E}\left[(\vartheta - (A^\star\boldsymbol{Y} + \boldsymbol{b}^\star))'(C\boldsymbol{Y} + \boldsymbol{c})\right] = 0$$

for every $p \times k$ matrix C and and every $\boldsymbol{c}$ in $\mathbb{R}^p$. This last relation is equivalent to

$$\mathbb{E}\left[(\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))' C\boldsymbol{Y}\right] = 0$$

since $\mathbb{E}[\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star)] = \boldsymbol{0}_p$ – This fact is just the fact, established later (in Section 6) that the LMSE estimator is unbiased in the sense that

$$\mathbb{E}\left[\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}]\right] = \mathbb{E}[\vartheta].$$

But, by elementary properties of the trace operator, we get

$$\begin{aligned}
\mathbb{E}\left[(\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))' C\boldsymbol{Y}\right] &= \mathbb{E}\left[\mathrm{Tr}\left((\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))' C\boldsymbol{Y}\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(C\boldsymbol{Y}(\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))'\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left((\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))(C\boldsymbol{Y})'\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left((\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))\boldsymbol{Y}'C'\right)\right] \\
&= \mathrm{Tr}\left(\mathbb{E}\left[(\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))\boldsymbol{Y}'C'\right]\right) \\
&= \mathrm{Tr}\left(\mathbb{E}\left[(\vartheta - (A^\star \boldsymbol{Y} + \boldsymbol{b}^\star))\boldsymbol{Y}'\right]C'\right) \\
&= \mathrm{Tr}\left(\mathbb{E}\left[(\vartheta - \boldsymbol{\mu}_\vartheta - A^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y))\boldsymbol{Y}'\right]C'\right) \\
&= \mathrm{Tr}\left((\Sigma_{\vartheta Y} - A^\star \Sigma_Y)C'\right) \qquad (9)
\end{aligned}$$

whence

$$\mathrm{Tr}\left((\Sigma_{\vartheta Y} - A^\star \Sigma_Y)C'\right) = 0.$$

Since the $p \times k$ matrix C is *arbitrary*, we conclude that

$$\Sigma_{\vartheta Y} - A^\star \Sigma_Y = \boldsymbol{O}_{p\times k}$$

and the normal equations (6) are now established.

Part 3: The existence of the optimal $p \times k$ matrix $A^\star$ If Σ_Y is invertible, then the normal equations can be solved. Just take

$$A^\star = \Sigma_{\vartheta Y}(\Sigma_Y)^{-1}.$$

If Σ_Y is not invertible, then proceed as follows: Since Σ_Y is a covariance matrix, it is symmetric and positive semi-definite, hence it can always be diagonalized: There exists a $k \times k$ matrix H such that

$$H'\Sigma_Y H = D$$

where

$$H'H = \boldsymbol{I}_k \quad \text{(hence } H' = H^{-1}\text{)}$$

and D is a $k \times k$ diagonal matrix. Therefore, $\Sigma_Y = HDH^{-1}$, and the normal equations can now be rewritten as

$$A^\star \Sigma_Y = A^\star \left(HDH^{-1}\right) = \Sigma_{\vartheta Y},$$

or equivalently,

$$A^\star HD = \Sigma_{\vartheta Y} H.$$

Thus, with arbitrary $i = 1, \ldots, p$ and $j = 1, \ldots, k$, entrywise we have

$$\left((A^\star H)D\right)_{ij} = \left(\Sigma_{\vartheta Y} H\right)_{ij},$$

whence

$$\sum_{\ell=1}^{k} (A^\star H)_{i\ell} D_{\ell\ell} \delta_{\ell j} = (\Sigma_{\vartheta Y} H)_{ij}.$$

Thus,

$$(A^\star H)_{ij} D_{jj} = (\Sigma_{\vartheta Y} H)_{ij}. \tag{10}$$

If $D_{jj} \neq 0$, it follows that

$$(\boldsymbol{a}_i^\star)' \boldsymbol{h}_j = \frac{1}{D_{jj}} (\Sigma_{\vartheta Y} H)_{ij}.$$

6 Properties of LMSE estimators

These properties are easy consequences of the Orthogonality Principle.

Property A (LMSE estimators are unbiased)

We have

$$\mathbb{E}\left[\widehat{\mathbb{E}}\left[\vartheta | \boldsymbol{Y}\right]\right] = \mathbb{E}\left[\vartheta\right].$$

With $\boldsymbol{v}$ arbitrary in $\mathbb{R}^p$, apply the Orthogonality Principle with the (degenerate) affine estimator $h_{\boldsymbol{v}}$ in $\mathcal{L}(k;p)$ given by

$$h_{\boldsymbol{v}}(\boldsymbol{y}) = \boldsymbol{v}, \quad \boldsymbol{y} \in \mathbb{R}^k.$$

This yields

$$0 = \mathbb{E}\left[\left(\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\right)'\boldsymbol{v}\right] = \left(\mathbb{E}\left[\vartheta\right] - \mathbb{E}\left[\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\right]\right)'\boldsymbol{v}. \tag{11}$$

The result follows since $\boldsymbol{v}$ is arbitrary.

Property B (Marginalization)

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]_i = \widehat{\mathbb{E}}\left[\vartheta_i|\boldsymbol{Y}\right], \quad i = 1, \ldots, p.$$

For each affine estimator ℓ in $\mathcal{L}(k;p)$, we have the decomposition

$$\mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] = \sum_{i=1}^{p} \mathbb{E}\left[|\vartheta_i - \ell_i(\boldsymbol{Y})|^2\right].$$

We have

$$\ell(\boldsymbol{y}) = A\boldsymbol{y} + \boldsymbol{b}, \quad \boldsymbol{y} \in \mathbb{R}^k$$

where A is a $p \times k$ matrix and $\boldsymbol{b}$ an element of $\mathbb{R}^p$. Therefore, writing

$$A = \begin{bmatrix} \boldsymbol{a}_1' \\ \vdots \\ \boldsymbol{a}_p' \end{bmatrix}$$

with $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ elements of $\mathbb{R}^k$, we note that

$$\ell_i(\boldsymbol{y}) = (A\boldsymbol{y})_i + b_i = \boldsymbol{a}_i'\boldsymbol{y} + b_i, \quad \boldsymbol{y} \in \mathbb{R}^k.$$

It follows

$$\begin{aligned} \mathbb{E}\left[\|\vartheta - \ell(\boldsymbol{Y})\|^2\right] &= \sum_{i=1}^{p} \mathbb{E}\left[|\vartheta_i - \boldsymbol{a}_i'\boldsymbol{Y} - b_i|^2\right] \\ &\geq \sum_{i=1}^{p} \mathbb{E}\left[\left|\vartheta_i - \widehat{\mathbb{E}}\left[\vartheta_i|\boldsymbol{Y}\right]\right|^2\right] \end{aligned} \tag{12}$$

and the desired result is straightforward by uniqueness.

Property C (Matrix version of the Orthogonality Principle)
With q a positive integer, it holds that

$$\mathbb{E}\left[\left(\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\right)\ell(\boldsymbol{Y})'\right] = \boldsymbol{O}_{p\times q}, \quad \ell \in \mathcal{L}(k;q)$$

Any ℓ in $\mathcal{L}(k;q)$ can be written as

$$\ell(\boldsymbol{y}) = B\boldsymbol{y} + \boldsymbol{c}, \quad \boldsymbol{y} \in \mathbb{R}^k$$

with $q \times k$ matrix B and vector $\boldsymbol{c}$ in $\mathbb{R}^q$

Property D
If ϑ is a.s. constant, then

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] = \vartheta \quad \mathbb{P}-a.s.$$

An easy consequence of the Orthogonality Principle as we note that

$$\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] = \ell(\boldsymbol{Y}) \quad \mathbb{P}-a.s.$$

for some ℓ in $\mathcal{L}(k;p)$.

Property E (Linearity)
With positive integer q, we have

$$\widehat{\mathbb{E}}\left[M\vartheta + \boldsymbol{m}|\boldsymbol{Y}\right] = M\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] + \boldsymbol{m} \quad \mathbb{P}-a.s.$$

where M is a $q \times p$ matrix and $\boldsymbol{m}$ is an element of $\mathbb{R}^q$.

Property F
If the rvs ϑ and $\boldsymbol{Y}$ are related through

$$\vartheta = C\boldsymbol{Y} + \boldsymbol{c} \quad \mathbb{P}-a.s.$$

where C is a $p \times k$ matrix and $\boldsymbol{C}$ is an element of $\mathbb{R}^p$, then

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] = \vartheta \quad \mathbb{P}-a.s.$$

Property G

For every $\boldsymbol{d}$ in $\mathbb{R}^k$, we have

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}+\boldsymbol{d}\right]=\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\quad \mathbb{P}-a.s.$$

This property is an immediate consequence of the Orthogonality Principle upon noting the following equivalence: For every ℓ in $\mathcal{L}(k;p)$, there exists a unique $\tilde{\ell}$ in $\mathcal{L}(k;p)$ such that

$$\ell(\boldsymbol{y}+\boldsymbol{d})=\tilde{\ell}(\boldsymbol{y}),\quad \boldsymbol{y}\in\mathbb{R}^k.$$

Conversely, for every $\tilde{\ell}$ in $\mathcal{L}(k;p)$, there exists a unique ℓ in $\mathcal{L}(k;p)$ such that

$$\tilde{\ell}(\boldsymbol{y})=\ell(\boldsymbol{y}+\boldsymbol{d}),\quad \boldsymbol{y}\in\mathbb{R}^k.$$

Just take

$$\ell(\boldsymbol{y})=\tilde{\ell}(\boldsymbol{y}-\boldsymbol{d}),\quad \boldsymbol{y}\in\mathbb{R}^k.$$

Property H

With D an invertible $k\times k$ matrix, we have

$$\widehat{\mathbb{E}}\left[\vartheta|D\boldsymbol{Y}\right]=\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\quad \mathbb{P}-a.s.$$

It follows from Properties G and H that

$$\widehat{\mathbb{E}}\left[\vartheta|D\boldsymbol{Y}+\boldsymbol{d}\right]=\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\quad \mathbb{P}-a.s.$$

for any invertible $k\times k$ matrix D and every $\boldsymbol{d}$ in $\mathbb{R}^k$.

Property H is an immediate consequence of the following equivalence: For every ℓ in $\mathcal{L}(k;p)$, the estimator $\tilde{\ell}:\mathbb{R}^k\to\mathbb{R}^p$ given by

$$\tilde{\ell}(\boldsymbol{y})=\ell(D\boldsymbol{y}),\quad \boldsymbol{y}\in\mathbb{R}^k$$

is an affine estimator in $\mathcal{L}(k;p)$. Conversely, for every $\tilde{\ell}$ in $\mathcal{L}(k;p)$, there exists a unique affine estimator ℓ in $\mathcal{L}(k;p)$ such that

$$\tilde{\ell}(\boldsymbol{y})=\ell(D\boldsymbol{y}),\quad \boldsymbol{y}\in\mathbb{R}^k.$$

Just take

$$\ell(\boldsymbol{y}) = \tilde{\ell}(D^{-1}\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^k.$$

Property I

If the rvs ϑ and $\boldsymbol{Y}$ are uncorrelated, i.e.,

$$\Sigma_{\vartheta Y} = \boldsymbol{O}_{p\times k},$$

then

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] = \mathbb{E}\left[\vartheta\right] \quad \mathbb{P}-a.s.$$

The Orthogonality Principle states that

$$\mathbb{E}\left[\left(\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\right)'\ell(\boldsymbol{Y})\right] = 0, \quad \ell \in \mathcal{L}(k;p).$$

We note that

$$\begin{aligned} \mathbb{E}\left[\vartheta'\ell(\boldsymbol{Y})\right] &= \mathbb{E}\left[\left(\vartheta - \mathbb{E}\left[\vartheta\right]\right)'\ell(\boldsymbol{Y})\right] + \mathbb{E}\left[\mathbb{E}\left[\vartheta\right]'\ell(\boldsymbol{Y})\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\vartheta\right]'\ell(\boldsymbol{Y})\right] \end{aligned} \tag{13}$$

because the the rvs ϑ and $\boldsymbol{Y}$ are uncorrelated. The Orthogonality Principle now takes the form

$$\mathbb{E}\left[\left(\mathbb{E}\left[\vartheta\right] - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right]\right)'\ell(\boldsymbol{Y})\right] = 0, \quad \ell \in \mathcal{L}(k;p)$$

and the conclusion follows.

The next three properties involve the $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$ and the $\mathbb{R}^m$-valued rv $\boldsymbol{Z}$ with k and m arbitrary positive integers. Both rvs are *second-order* rvs.

Property J

If the $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$ and the $\mathbb{R}^m$-valued rv $\boldsymbol{Z}$ are uncorrelated , then

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y},\boldsymbol{Z}\right] = \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}\right] + \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Z}\right] \quad \mathbb{P}-a.s.$$

whenever

$$\mathbb{E}\left[\vartheta\right] = \mathbf{0}_p.$$

Any affine estimator ℓ in $\mathcal{L}(k+m;p)$ is of the form

$$\ell(\boldsymbol{y},\boldsymbol{z}) = A_y\boldsymbol{y} + A_z\boldsymbol{z} + \boldsymbol{b}, \qquad \begin{array}{c}\boldsymbol{y}\in\mathbb{R}^k\\ \boldsymbol{z}\in\mathbb{R}^m\end{array}$$

where A_y and A_z are $p\times k$ and $p\times m$ matrices, and $\boldsymbol{b}$ an element in $\mathbb{R}^p$. In particular,

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y},\boldsymbol{Z}\right] = \ell^\star(\boldsymbol{Y},\boldsymbol{Z}) \quad \mathbb{P}-a.s.$$

with affine estimator $\ell^\star$ in $\mathcal{L}(k+m;p)$ of the form

$$\ell^\star(\boldsymbol{y},\boldsymbol{z}) = A_y^\star\boldsymbol{y} + A_z^\star\boldsymbol{z} + \boldsymbol{b}^\star, \qquad \begin{array}{c}\boldsymbol{y}\in\mathbb{R}^k\\ \boldsymbol{z}\in\mathbb{R}^m\end{array}$$

where $A_y^\star$ and $A_z^\star$ are $p\times k$ and $p\times m$ matrices, and $\boldsymbol{b}^\star$ an element in $\mathbb{R}^p$. Since $\boldsymbol{\mu}_\vartheta$ we recall that $\boldsymbol{b}^\star$ is given by

$$\boldsymbol{b}^\star = -A_y^\star\boldsymbol{\mu}_Y - A_z^\star\boldsymbol{\mu}_Z.$$

so that

$$\ell^\star(\boldsymbol{y},\boldsymbol{z}) = A_y^\star\left(\boldsymbol{y}-\boldsymbol{\mu}_Y\right) + A_z^\star\left(\boldsymbol{z}-\boldsymbol{\mu}_Z\right), \qquad \begin{array}{c}\boldsymbol{y}\in\mathbb{R}^k\\ \boldsymbol{z}\in\mathbb{R}^m.\end{array}$$

The Orthogonality Principle will read

$$\mathbb{E}\left[\left(\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y},\boldsymbol{Z}\right]\right)'\ell(\boldsymbol{Y},\boldsymbol{Z})\right] = 0, \quad \ell\in\mathcal{L}(k+m;p). \tag{14}$$

With the notation introduced earlier we see that

$$\begin{aligned}
& \left(\vartheta - \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y},\boldsymbol{Z}\right]\right)'\ell(\boldsymbol{Y},\boldsymbol{Z}) \\
= & \left(\vartheta - A_y^\star\boldsymbol{Y} - A_z^\star\boldsymbol{Z} - \boldsymbol{b}^\star\right)'\left(A_y\boldsymbol{Y} + A_z\boldsymbol{Z} + \boldsymbol{b}\right) \\
= & \left(\vartheta - A_y^\star\left(\boldsymbol{Y}-\boldsymbol{\mu}_Y\right) - A_z^\star\left(\boldsymbol{Z}-\boldsymbol{\mu}_Z\right)\right)'\left(A_y\boldsymbol{Y} + A_z\boldsymbol{Z} + \boldsymbol{b}\right) \\
= & \left(\vartheta - A_y^\star\left(\boldsymbol{Y}-\boldsymbol{\mu}_Y\right) - A_z^\star\left(\boldsymbol{Z}-\boldsymbol{\mu}_Z\right)\right)'\left(A_y\boldsymbol{Y} + \boldsymbol{b}\right) \\
& \quad + \left(\vartheta - A_y^\star\left(\boldsymbol{Y}-\boldsymbol{\mu}_Y\right) - A_z^\star\left(\boldsymbol{Z}-\boldsymbol{\mu}_Z\right)\right)' A_z\boldsymbol{Z} \qquad (15)
\end{aligned}$$

Next, upon taking $A_z = \boldsymbol{O}_{p\times m}$ in (15) and using the resulting (14), we conclude that

$$0 \;=\; \mathbb{E}\left[\left(\vartheta - A_y^\star\left(\boldsymbol{Y}-\boldsymbol{\mu}_Y\right) - A_z^\star\left(\boldsymbol{Z}-\boldsymbol{\mu}_Z\right)\right)'\left(A_y\boldsymbol{Y} + \boldsymbol{b}\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\vartheta - A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\right)'(A_y\boldsymbol{Y} + \boldsymbol{b})\right] \\
&\quad - \mathbb{E}\left[(A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z))'(A_y\boldsymbol{Y} + \boldsymbol{b})\right] \\
&= \mathbb{E}\left[\left(\vartheta - A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\right)'(A_y\boldsymbol{Y} + \boldsymbol{b})\right] \qquad (16)
\end{aligned}
$$

as we make use of the fact that the rvs $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are uncorrelated. It follows that

$$\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] = A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y) \quad \mathbb{P}-a.s.$$

by the Orthogonality Principle characterizing the LMMSE estimator of ϑ on the basis of $\boldsymbol{Y}$.

To proceed, take $A_y = \boldsymbol{O}_{p\times k}$ and $\boldsymbol{b} = \boldsymbol{0}_p$ in (15) and use the resulting (14). This gives

$$
\begin{aligned}
0 &= \mathbb{E}\left[\left(\vartheta - A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y) - A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z)\right)' A_z\boldsymbol{Z}\right] \\
&= \mathbb{E}\left[(\vartheta - A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z))' A_z\boldsymbol{Z}\right] - \mathbb{E}\left[\left(A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y)\right)' A_z\boldsymbol{Z}\right] \\
&= \mathbb{E}\left[(\vartheta - A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z))' A_z\boldsymbol{Z}\right] \qquad (17)
\end{aligned}
$$

as we make use of the fact that the rvs $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are uncorrelated. It follows that

$$\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Z}] = A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z) \quad \mathbb{P}-a.s.$$

by the Orthogonality Principle characterizing the LMMSE estimator of ϑ on the basis of $\boldsymbol{Z}$.

To conclude the proof we note that

$$
\begin{aligned}
\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}, \boldsymbol{Z}] &= \ell^\star(\boldsymbol{Y}, \boldsymbol{Z}) \\
&= A_y^\star(\boldsymbol{Y} - \boldsymbol{\mu}_Y) + A_z^\star(\boldsymbol{Z} - \boldsymbol{\mu}_Z) \\
&= \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] + \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Z}] \quad \mathbb{P}-a.s. \qquad (18)
\end{aligned}
$$

as desired.

Property K

If the $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$ and the $\mathbb{R}^m$-valued rv $\boldsymbol{Z}$ are uncorrelated , then

$$\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}, \boldsymbol{Z}] = \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] + \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Z}] - \mathbb{E}[\vartheta] \quad \mathbb{P}-a.s.$$

Property F applied to the zero-mean rv $\vartheta - \mathbb{E}[\vartheta]$ gives

$$\widehat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta]\,|\boldsymbol{Y}, \boldsymbol{Z}] = \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}, \boldsymbol{Z}] - \mathbb{E}[\vartheta] \quad \mathbb{P}-a.s.$$

while Property J applied to the zero-mean rv $\vartheta - \mathbb{E}[\vartheta]$ yields

$$\begin{aligned}\widehat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta]|\boldsymbol{Y},\boldsymbol{Z}] &= \widehat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta]|\boldsymbol{Y}] + \widehat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta]|\boldsymbol{Z}] \\ &= \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] - \mathbb{E}[\vartheta] + \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Z}] - \mathbb{E}[\vartheta] \\ &= \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] + \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Z}] - 2\mathbb{E}[\vartheta] \quad \mathbb{P}-a.s. \end{aligned} \tag{19}$$

where the last step follows by Property F . Comparing we get the result.

Property L

More generally, with arbitrary $\mathbb{R}^k$-valued rv $\boldsymbol{Y}$ and $\mathbb{R}^m$-valued rv $\boldsymbol{Z}$, we have

$$\widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y},\boldsymbol{Z}] = \widehat{\mathbb{E}}[\vartheta|\boldsymbol{Y}] + \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Z} - \widehat{\mathbb{E}}[\boldsymbol{Z}|\boldsymbol{Y}]\right] - \mathbb{E}[\vartheta] \quad \mathbb{P}-a.s.$$

The rv $\boldsymbol{Z} - \widehat{\mathbb{E}}[\boldsymbol{Z}|\boldsymbol{Y}]$ is known as the *(linear) innovations* in $\boldsymbol{Z}$ with respect to $\boldsymbol{Y}$. The rvs $\boldsymbol{Y}$ and $\boldsymbol{Z} - \widehat{\mathbb{E}}[\boldsymbol{Z}|\boldsymbol{Y}]$ are always uncorrelated.

We start by noting that

$$\widehat{\mathbb{E}}[\boldsymbol{Z}|\boldsymbol{Y}] = A^\star \boldsymbol{Y} + \boldsymbol{b}^\star$$

for some $m \times k$ matrix $A^\star$ and an element $\boldsymbol{b}^\star$ of $\mathbb{R}^m$.

Thus, with

$$\boldsymbol{V} \equiv \boldsymbol{Z} - \widehat{\mathbb{E}}[\boldsymbol{Z}|\boldsymbol{Y}],$$

it holds that

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{V} \end{bmatrix} = D \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0}_k \\ -\boldsymbol{b}^\star \end{bmatrix}$$

with $(m+k) \times (m+k)$ matrix R given by

$$D = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{O}_{k\times m} \\ -A^\star & \boldsymbol{I}_m \end{bmatrix}.$$

Observe that the equation

$$\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{O}_{k\times m} \\ -A^\star & \boldsymbol{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0}_k \\ \boldsymbol{0}_m \end{bmatrix}$$

implies

$$\boldsymbol{I}_k \boldsymbol{y} + \boldsymbol{O}_{k\times m} \boldsymbol{z} = \boldsymbol{0}_k$$

and

$$-A^\star \boldsymbol{y} + \boldsymbol{I}_m \boldsymbol{z} = \boldsymbol{0}_m.$$

The first equation implies $\boldsymbol{y} = \mathbf{0}_k$; replacing this fact into the second equation we get $\boldsymbol{z} = \mathbf{0}_m$. In other words, $\mathrm{Ker}(D)$ is reduced to the zero vector in $\mathbb{R}^{k+m}$, and is therefore invertible.

As a result,

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} = D^{-1} \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{V} \end{bmatrix} + D^{-1} \begin{bmatrix} \mathbf{0}_k \\ -\boldsymbol{b}^\star \end{bmatrix}.$$

Invoking Property G and Property H we conclude that

$$\widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}, \boldsymbol{Z}\right] = \widehat{\mathbb{E}}\left[\vartheta|\boldsymbol{Y}, \boldsymbol{V}\right]$$

and the desired conclusion now follows by Property K since rvs $\boldsymbol{Y}$ and $\boldsymbol{Z} - \widehat{\mathbb{E}}\left[\boldsymbol{Z}|\boldsymbol{Y}\right]$ are always uncorrelated (as an immediate consequence of the Orthoganility Principle).

7 The Gaussian case