1 The basic setting

Throughout, $p$ and $k$ are arbitrary positive integers. Let the random parameter $\vartheta$ be modelled as an $\mathbb{R}^p$-valued rv, while the observation rv $Y$ 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

$$\mathbb{P}[Y \leq y | \vartheta = \theta] = F_\theta(y), \quad y \in \mathbb{R}^k, \theta \in \Theta.$$

We assume that the rv $\vartheta$ is a second-order rv, namely,

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

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

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

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

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

for some $p \times k$ matrix $A$ and a vector $b$ 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 $a = 0$. 
2 Minimum Mean Square Error (MMSE) Estimation

The MMSE problem can be formulated as follows: Find \( g^* \) in \( G_2(k; Y) \) such that

\[
\mathbb{E} \left[ \| \vartheta - g^*(Y) \|^2 \right] \leq \mathbb{E} \left[ \| \vartheta - g(Y) \|^2 \right], \quad g \in G_2(k; Y).
\]

Any estimator \( g^* \) in \( G_2(k; Y) \) which satisfies (1) is known as a MMSE estimator of \( \vartheta \) on the basis of \( Y \).

**Theorem 2.1** The estimator \( g^* \) in \( G_2(p; Y) \) satisfies

\[
\mathbb{E} \left[ \| \vartheta - g^*(Y) \|^2 \right] \leq \mathbb{E} \left[ \| \vartheta - g(Y) \|^2 \right], \quad g \in G_2(p; Y)
\]

if and only if the Orthogonality Principle

\[
\mathbb{E} \left[ (\vartheta - g^*(Y))^\prime h(Y) \right] = 0, \quad h \in G_2(p; Y)
\]

holds.

This characterization is geometric in nature, and points to \( g^* \) as the projection of \( \vartheta \) on the subspace of second-order rv's

\[
\{ g(Y) : g \in G_2(p; Y) \}.
\]

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 rv's. Orthogonality in \( L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p) \) is defined by

\[
\mathbb{E} [\xi^\prime \eta] = 0, \quad \xi, \eta \in L_2((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^p).
\]
The Orthogonality Principle (2) can be restated as saying that the error \( \vartheta - g^*(Y) \) is orthogonal to the subspace \( \{ g(Y) : g \in G_2(p; Y) \} \).

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

**Corollary 2.1** If \( g^*_1 \) and \( g^*_2 \) are estimators in \( G_2(p; Y) \) such that

\[
E \left[ \| \vartheta - g^*_i(Y) \|^2 \right] \leq E \left[ \| \vartheta - g(Y) \|^2 \right], \quad g \in G_2(p; Y)
\]

for each \( i = 1, 2 \), then we have

\[
P \left[ g^*_1(Y) = g^*_2(Y) \right] = 1.
\]

**Proof.** Use the Orthogonality Principle (2) with \( h = g^*_1 - g^*_2 \) for both \( g^*_1 \) and \( g^*_2 \).

\[ \blacksquare \]

**Basic ideas behind the proof of Theorem 2.1**

With estimator \( g \) in \( G_2(p; Y) \), note that

\[
E \left[ \| \vartheta - g(Y) \|^2 \right] = E \left[ (\vartheta - g(Y))' (\vartheta - g(Y)) \right] = E \left[ ((\vartheta - g^*(Y)) + (g^*(Y) - g(Y)))' ((\vartheta - g^*(Y)) + (g^*(Y) - g(Y))) \right] = E \left[ \| \vartheta - g^*(Y) \|^2 \right] + 2E \left[ (\vartheta - g^*(Y))' (g^*(Y) - g(Y)) \right] + E \left[ \| g^*(Y) - g(Y) \|^2 \right]
\]

so that

\[
E \left[ \| \vartheta - g(Y) \|^2 \right] - E \left[ \| \vartheta - g^*(Y) \|^2 \right] = E \left[ \| g^*(Y) - g(Y) \|^2 \right]
\]

(3)

If \( g^* \) in \( G_2(p; Y) \) satisfies the Optimality Principle (2), then the equality (3) implies

\[
E \left[ \| \vartheta - g(Y) \|^2 \right] - E \left[ \| \vartheta - g^*(Y) \|^2 \right] = E \left[ \| g^*(Y) - g(Y) \|^2 \right]
\]
since $g - g^*$ is an element of $G_2(p; Y)$ as both $g^*$ and $g$ are. It follows that
\[
\mathbb{E} \left[ \|\vartheta - g(Y)\|^2 \right] - \mathbb{E} \left[ \|\vartheta - g^*(Y)\|^2 \right] \geq 0, \quad g \in G_2(p; Y)
\]
and $g^*$ is an MMSE estimator.

Conversely, if $g^*$ is an MMSE estimator, then (3) implies
\[
\mathbb{E} \left[ \|g^*(Y) - g(Y)\|^2 \right] + 2\mathbb{E} \left[ (\vartheta - g^*(Y))' (g^*(Y) - g(Y)) \right] \geq 0
\]
for every $g$ element of $G_2(p; Y)$. Thus, with $h$ an arbitrary element of $G_2(p; Y)$ and $t$ in $\mathbb{R}$, consider the estimator $g_t : \mathbb{R}^k \rightarrow \mathbb{R}^p$ given by
\[
g_t(y) = g^*(t) + th(y), \quad y \in \mathbb{R}^k.
\]
If is plain that $g_t$ is also an element of $G_2(p; Y)$. Applying the last inequality with $g = g_t$ we conclude that
\[
t^2 \mathbb{E} \left[ \|g(Y)\|^2 \right] + 2t \mathbb{E} \left[ (\vartheta - g^*(Y))' h(Y) \right] \geq 0, \quad t \in \mathbb{R}
\]
and Fact 1.1 immediately leads to the Optimality Principle (2).

To identify the MMSE estimator we focus on the following problem: With $\xi$ be a second-order $\mathbb{R}^p$-valued rv, we seek $a^*$ in $\mathbb{R}^p$ such that
\[
\mathbb{E} \left[ \|\xi - a^*\|^2 \right] \leq \mathbb{E} \left[ \|\xi - a\|^2 \right], \quad a \in \mathbb{R}^p.
\]
The solution to this problem is well known to be unique, and is given by
\[
a^* = \mathbb{E} [\xi].
\]

Returning to the MMSE problem, we recall that
\[
\mathbb{E} \left[ \|\vartheta - g(Y)\|^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ \|\vartheta - a\|^2 | Y = y \right]_{Y = Y, a = g(Y)} \right]
\]
for every estimator $g$ in $G_2(p; Y)$. This fact readily leads to concluding
\[
g^*(y) = \mathbb{E} [\vartheta | Y = y], \quad y \in \mathbb{R}^k.
\]
It is customary to write
\[
g^*(Y) = \mathbb{E} [\vartheta | Y]
\]
where the right handside is understood as the conditional expectation of the rv $\vartheta$ given the $\sigma$-field generated by the rv $Y$. 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 $Y$ 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 $\ell^*$ in $L(k; p)$ such that

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

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

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

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

if and only if the Orthogonality Principle

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

holds.

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

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

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 $\vartheta - g^*(Y)$ is orthogonal to the subspace $\{ \ell(Y) : \ell \in L(k; p) \}$.

**Basic ideas behind the proof of Theorem 3.1**

With affine estimator $\ell$ in $L(k; p)$, this time we note that

$$\mathbb{E} \left[ \| \vartheta - \ell(Y) \|^2 \right] = \mathbb{E} \left[ (\vartheta - \ell(Y))' (\vartheta - \ell(Y)) \right] = \mathbb{E} \left[ \| \vartheta - \ell^*(Y) \|^2 \right] + 2 \mathbb{E} \left[ (\vartheta - \ell^*(Y))' (\ell^*(Y) - \ell(Y)) \right] + \mathbb{E} \left[ \| \ell^*(Y) - \ell(Y) \|^2 \right]$$
so that
\[
\mathbb{E} \left[\| \vartheta - \ell(Y) \|^2 \right] - \mathbb{E} \left[\| \vartheta - \ell^*(Y) \|^2 \right] = \mathbb{E} \left[\| \ell^*(Y) - \ell(Y) \|^2 \right] + 2 \mathbb{E} \left[ (\vartheta - \ell^*(Y))' (\ell^*(Y) - \ell(Y)) \right].
\]
(5)

If \( \ell^* \) in \( L_2(k; p) \) satisfies the Optimality Principle (4), then the equality (5) implies
\[
\mathbb{E} \left[\| \vartheta - \ell(Y) \|^2 \right] - \mathbb{E} \left[\| \vartheta - \ell^*(Y) \|^2 \right] = \mathbb{E} \left[\| \ell^*(Y) - \ell(Y) \|^2 \right] \geq 0, \quad \ell \in L(k; p)
\]
and \( \ell^* \) is an LMSE estimator.

Conversely, if \( \ell^* \) is an LMSE estimator, then (5) implies
\[
\mathbb{E} \left[\| \ell^*(Y) - \ell(Y) \|^2 \right] + 2 \mathbb{E} \left[ (\vartheta - \ell^*(Y))' (\ell^*(Y) - \ell(Y)) \right] \geq 0
\]
for every \( \ell \) element of \( L(k; p) \). Thus, with \( h \) an arbitrary element of \( L(k; p) \) and \( t \) in \( \mathbb{R} \), consider the estimator \( \ell_t : \mathbb{R}^k \rightarrow \mathbb{R}^p \) given by
\[
\ell_t(y) = \ell^*(t) + th(y), \quad y \in \mathbb{R}^k.
\]
If is plain that \( g_t \) is also an element of \( L(k; p) \). Applying the last inequality with \( \ell = \ell_t \) we conclude that
\[
t^2 \mathbb{E} \left[\| h(Y) \|^2 \right] + 2t \mathbb{E} \left[ (\vartheta - \ell^*(Y))' h(Y) \right] \geq 0, \quad t \in \mathbb{R}
\]
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

\[
\mu_Y = \mathbb{E} [Y] \quad \text{and} \quad \mu_\vartheta = \mathbb{E} [\vartheta].
\]
Next, the appropriate covariance matrices are given by

\[ \Sigma_{\vartheta Y} = \text{Cov} \left[ \vartheta, Y \right] = \mathbb{E} \left[ (\vartheta - \mu_\vartheta) (Y - \mu_Y)^T \right] \]

and

\[ \Sigma_Y = \text{Cov} \left[ Y \right] = \mathbb{E} \left[ (Y - \mu_Y) (Y - \mu_Y)^T \right]. \]

The matrices \( \Sigma_{\vartheta Y} \) and \( \Sigma_Y \) are \( p \times k \) and \( k \times k \) matrices, respectively.

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

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

With such an estimator \( \ell^* : \mathbb{R}^k \to \mathbb{R}^p \) being given by

\[ \ell^*(y) = A^* y + b^*, \quad y \in \mathbb{R}^k \]

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

(6) \[ A^* \Sigma_Y = \Sigma_{\vartheta Y} \]

and

(7) \[ b^* = \mu_\vartheta - A^* \mu_Y. \]

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

**Corollary 4.1** If \( \Sigma_Y \) is invertible, then the LMSE estimator \( \ell^* \) is uniquely determined by

\[ \ell^*(y) = \mu_\vartheta + \Sigma_{\vartheta Y} \Sigma_Y^{-1} (y - \mu_Y), \quad y \in \mathbb{R}^k. \]

If \( \Sigma_Y \) is not invertible, there is is still uniqueness in the following sense; see analogy with Corollary 2.1.

**Corollary 4.2** Let \( \ell^*_1 \) and \( \ell^*_2 \) be affine estimators in \( \mathcal{L}(k; p) \) such that

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

for each \( i = 1, 2 \), then we have

\[ \mathbb{P} \left[ \ell^*_1(Y) = \ell^*_2(Y) \right] = 1. \]
In analogy with the notation used for MMSE estimators, we shall write
\[ \ell^*(y) = \hat{E}[\nu | Y = y], \quad y \in \mathbb{R}^k \]
and
\[ \ell^*(Y) = \hat{E}[\nu | Y]. \]
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 \( b = b(A) \) in \( \mathbb{R}^p \). For any \( p \times k \) matrix \( A \) and vector \( \nu \) in \( \mathbb{R}^p \), note that

\[
\begin{align*}
\mathbb{E} \left[ \| \nu - (AY + b) \|^2 \right] &= \mathbb{E} \left[ \| (\nu - \mu_{\nu}) - A(Y - \mu_Y) + (\mu_{\nu} - (A\mu_Y + b)) \|^2 \right] \\
&= \mathbb{E} \left[ \| (\nu - \mu_{\nu}) - A(Y - \mu_Y) \|^2 \right] + \| \nu - (A\mu_Y + b) \|^2 \\
&\geq \mathbb{E} \left[ \| (\nu - \mu_{\nu}) - A(Y - \mu_Y) \|^2 \right]
\end{align*}
\]
(8)

if we select \( b = b(A) \) with
\[ b(A) = \mu_{\nu} - A\mu_Y, \]
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 \( \ell^* : \mathbb{R}^k \to \mathbb{R}^p \) is of the form
\[ \ell^*(y) = A^*y + b^*, \quad y \in \mathbb{R}^k \]
with \( b^* \) necessarily given by
\[ b^* = \mu_{\nu} - A^*\mu_Y. \]

The Orthogonality Principle states that \( A^* \) and \( b^* \) are completely characterized by
\[ \mathbb{E} \left[ (\nu - (A^*Y + b^*))' (CY + c) \right] = 0 \]
for every $p \times k$ matrix $C$ and and every $c$ in $\mathbb{R}^p$. This last relation is equivalent to

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

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

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

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

$$
\mathbb{E} \left[ (\vartheta - (A^* Y + b^*))' C Y \right] = \mathbb{E} \left[ \text{Tr} \left((\vartheta - (A^* Y + b^*))' C Y \right) \right]
$$

$$
= \mathbb{E} \left[ \text{Tr} \left( C Y (\vartheta - (A^* Y + b^*))' \right) \right]
$$

$$
= \mathbb{E} \left[ \text{Tr} \left( ((\vartheta - (A^* Y + b^*)) (C Y)') \right) \right]
$$

$$
= \text{Tr} \left( \mathbb{E} \left[ ((\vartheta - (A^* Y + b^*)) Y' C' \right] \right)
$$

$$
= \text{Tr} \left( \mathbb{E} \left[ (\vartheta - A^* (Y - \mu_y)) Y' \right] C' \right)
$$

$$
= \text{Tr} \left( (\Sigma_{\vartheta Y} - A^* \Sigma Y) C' \right)
$$

(9)

whence

$$
\text{Tr} \left( (\Sigma_{\vartheta Y} - A^* \Sigma Y) C' \right) = 0.
$$

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

$$
\Sigma_{\vartheta Y} - A^* \Sigma Y = O_{p \times k}
$$

and the normal equations (6) are now established.

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

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

If $\Sigma_{Y}$ is not invertible, then proceed as follows: Since $\Sigma_{Y}$ 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
\[ H' = I_k \quad (\text{hence } H' = H^{-1}) \]
and \( D \) is a \( k \times k \) diagonal matrix. Therefore, \( \Sigma_Y = HDH^{-1} \), and the normal equations can now be rewritten as
\[ A^* \Sigma_Y = A^* (HDH^{-1}) = \Sigma_{\theta Y}, \]
or equivalently,
\[ A^* HD = \Sigma_{\theta Y} H. \]
Thus, with arbitrary \( i = 1, \ldots, p \) and \( j = 1, \ldots, k \), entrywise we have
\[ ((A^*H)D)_{ij} = (\Sigma_{\theta Y} H)_{ij}, \]
whence
\[ \sum_{\ell=1}^{k} (A^*H)_{i \ell} D_{\ell j} \delta_{ij} = (\Sigma_{\theta Y} H)_{ij}. \]
Thus,
\[ (A^*H)_{ij} D_{jj} = (\Sigma_{\theta Y} H)_{ij}. \]
(10)

If \( D_{jj} \neq 0 \), it follows that
\[ (a_i^* h_j) = \frac{1}{D_{jj}} (\Sigma_{\theta Y} H)_{ij}. \]

6 Properties of LMSE estimators

These properties are easy consequences of the Orthogonality Principle.

Property A (LMSE estimators are unbiased)

We have
\[ \mathbb{E} \left[ \mathbb{E} [\theta | Y] \right] = \mathbb{E} [\theta]. \]

With \( v \) arbitrary in \( \mathbb{R}^p \), apply the Orthogonality Principle with the (degenerate) affine estimator \( h_v \) in \( \mathcal{L}(k; p) \) given by
\[ h_v(y) = v, \quad y \in \mathbb{R}^k. \]
This yields

\begin{equation}
0 = E \left[ \left( \vartheta - \widehat{E}[\vartheta|Y] \right)^\prime v \right] = \left( E[\vartheta] - E \left[ \widehat{E}[\vartheta|Y] \right] \right)^\prime v.
\end{equation}

The result follows since \( v \) is arbitrary.

**Property B (Marginalization)**

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

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

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

We have

\[ \ell(y) = Ay + b, \quad y \in \mathbb{R}^k \]

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

\[ A = \begin{bmatrix} a_1' \\ \vdots \\ a_p' \end{bmatrix} \]

with \( a_1, \ldots, a_p \) elements of \( \mathbb{R}^k \), we note that

\[ \ell_i(y) = (Ay)_i + b_i = a_i'y + b_i, \quad y \in \mathbb{R}^k. \]

It follows

\begin{equation}
E \left[ ||\vartheta - \ell(Y)||^2 \right] = \sum_{i=1}^{p} E\left[ ||\vartheta_i - a_i'Y - b_i||^2 \right] \geq \sum_{i=1}^{p} E\left[ ||\vartheta_i - \widehat{E}[\vartheta_i|Y]||^2 \right]
\end{equation}

and the desired result is straightforward by uniqueness.
Property C (Matrix version of the Orthogonality Principle)

With \( q \) a positive integer, it holds that

\[
E \left[ \left( \vartheta - \hat{E}[\vartheta|Y] \right) \ell(Y)' \right] = O_{p \times q}, \quad \ell \in \mathcal{L}(k; q)
\]

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

\[
\ell(y) = By + c, \quad y \in \mathbb{R}^k
\]

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

Property D

If \( \vartheta \) is a.s. constant, then

\[
\hat{E}[\vartheta|Y] = \vartheta \quad \mathbb{P} \text{ - a.s.}
\]

An easy consequence of the Orthogonality Principle as we note that

\[
\vartheta - \hat{E}[\vartheta|Y] = \ell(Y) \quad \mathbb{P} \text{ - a.s.}
\]

for some \( \ell \) in \( \mathcal{L}(k; p) \).

Property E (Linearity)

With positive integer \( q \), we have

\[
\hat{E}[M\vartheta + m|Y] = M\hat{E}[\vartheta|Y] + m \quad \mathbb{P} \text{ - a.s.}
\]

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

Property F

If the rvs \( \vartheta \) and \( Y \) are related through

\[
\vartheta = CY + c \quad \mathbb{P} \text{ - a.s.}
\]

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

\[
\hat{E}[\vartheta|Y] = \vartheta \quad \mathbb{P} \text{ - a.s.}
\]
Property G
For every $d$ in $\mathbb{R}^k$, we have
\[
\hat{E}[\vartheta|Y + d] = \hat{E}[\vartheta|Y] \quad P - a.s.
\]

This property is an immediate consequence of the Orthogonality Principle upon noting the following equivalence: For every $\ell$ in $L(k;p)$, there exists a unique $\tilde{\ell}$ in $L(k;p)$ such that
\[
\ell(y + d) = \tilde{\ell}(y), \quad y \in \mathbb{R}^k.
\]
Conversely, for every $\tilde{\ell}$ in $L(k;p)$, there exists a unique $\ell$ in $L(k;p)$ such that
\[
\tilde{\ell}(y) = \ell(y + d), \quad y \in \mathbb{R}^k.
\]
Just take
\[
\ell(y) = \tilde{\ell}(y - d), \quad y \in \mathbb{R}^k.
\]

Property H
With $D$ an invertible $k \times k$ matrix, we have
\[
\hat{E}[\vartheta|DY] = \hat{E}[\vartheta|Y] \quad P - a.s.
\]

It follows from Properties G and H that
\[
\hat{E}[\vartheta|DY + d] = \hat{E}[\vartheta|Y] \quad P - a.s.
\]
for any invertible $k \times k$ matrix $D$ and every $d$ in $\mathbb{R}^k$.

Property H is an immediate consequence of the following equivalence: For every $\ell$ in $L(k;p)$, the estimator $\tilde{\ell} : \mathbb{R}^k \to \mathbb{R}^p$ given by
\[
\tilde{\ell}(y) = \ell(Dy), \quad y \in \mathbb{R}^k
\]
is an affine estimator in $L(k;p)$. Conversely, for every $\tilde{\ell}$ in $L(k;p)$, there exists a unique affine estimator $\ell$ in $L(k;p)$ such that
\[
\tilde{\ell}(y) = \ell(Dy), \quad y \in \mathbb{R}^k.
\]
Just take
\[ \ell(y) = \tilde{\ell}(D^{-1}y), \quad y \in \mathbb{R}^k. \]

**Property I**

If the rvs \( \vartheta \) and \( Y \) are uncorrelated, i.e.,
\[ \Sigma_{\vartheta Y} = O_{p \times k}, \]
then
\[ \hat{E}[\vartheta|Y] = E[\vartheta] \quad \mathbb{P} - a.s. \]

The Orthogonality Principle states that
\[ E\left[ (\vartheta - \hat{E}[\vartheta|Y])' \ell(Y) \right] = 0, \quad \ell \in \mathcal{L}(k; p). \]

We note that
\[
E[\vartheta'\ell(Y)] = E\left[ ((\vartheta - E[\vartheta])' \ell(Y)) + E[\vartheta]' \ell(Y) \right] \\
= E\left[ E[\vartheta]' \ell(Y) \right]
\]
(13)
because the rvs \( \vartheta \) and \( Y \) are uncorrelated. The Orthogonality Principle now takes the form
\[ E\left[ \left( E[\vartheta] - \hat{E}[\vartheta|Y] \right)' \ell(Y) \right] = 0, \quad \ell \in \mathcal{L}(k; p) \]
and the conclusion follows.

---

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

**Property J**

If the \( \mathbb{R}^k \)-valued rv \( Y \) and the \( \mathbb{R}^m \)-valued rv \( Z \) are uncorrelated, then
\[ \hat{E}[\vartheta|Y, Z] = \hat{E}[\vartheta|Y] + \hat{E}[\vartheta|Z] \quad \mathbb{P} - a.s. \]
whenever
\[ E[\vartheta] = 0_p. \]
Any affine estimator $\ell$ in $L(k + m; p)$ is of the form

$$\ell(y, z) = A_y y + A_z z + b, \quad y \in \mathbb{R}^k, \quad z \in \mathbb{R}^m$$

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

$$\hat{E}[\vartheta|Y, Z] = \ell^*(Y, Z), \quad \mathbb{P} - a.s.$$ with affine estimator $\ell^*$ in $L(k + m; p)$ of the form

$$\ell^*(y, z) = A_y^* y + A_z^* z + b^*, \quad y \in \mathbb{R}^k, \quad z \in \mathbb{R}^m$$

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

$$b^* = -A_y^* \mu_Y - A_z^* \mu_Z.$$ so that

$$\ell^*(y, z) = A_y^* (y - \mu_Y) + A_z^* (z - \mu_Z), \quad y \in \mathbb{R}^k, \quad z \in \mathbb{R}^m.$$ The Orthogonality Principle will read

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

With the notation introduced earlier we see that

$$\left(\vartheta - \hat{E}[\vartheta|Y, Z]\right)'\ell(Y, Z)$$

$$= \left(\vartheta - A_y^* Y - A_z^* Z - b^*\right)'(A_y Y + A_z Z + b)$$

$$= \left(\vartheta - A_y^* (Y - \mu_Y) - A_z^* (Z - \mu_Z)\right)'(A_y Y + A_z Z + b)$$

$$= \left(\vartheta - A_y^* (Y - \mu_Y) - A_z^* (Z - \mu_Z)\right)'(A_y Y + b)$$

$$+ \left(\vartheta - A_y^* (Y - \mu_Y) - A_z^* (Z - \mu_Z)\right)'A_z Z$$

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

$$0 = \mathbb{E}\left[(\vartheta - A_y^* (Y - \mu_Y) - A_z^* (Z - \mu_Z))' (A_y Y + b)\right]$$
6 PROPERTIES OF LMSE ESTIMATORS

\[
\begin{align*}
&= \mathbb{E} \left[ (\vartheta - A^*_y (Y - \mu_Y))' (A_y Y + b) \right] \\
&\quad - \mathbb{E} \left[ (A^*_z (Z - \mu_Z))' (A_y Y + b) \right] \\
(16) &\quad = \mathbb{E} \left[ (\vartheta - A^*_y (Y - \mu_Y))' (A_y Y + b) \right]
\end{align*}
\]

as we make use of the fact that the rvs \(Y\) and \(Z\) are uncorrelated. It follows that

\[
\hat{\mathbb{E}} [\vartheta | Y] = A^*_y (Y - \mu_Y) \quad P - a.s.
\]

by the Orthogonality Principle characterizing the LMMSE estimator of \(\vartheta\) on the basis of \(Y\).

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

\[
\begin{align*}
0 &= \mathbb{E} \left[ (\vartheta - A^*_y (Y - \mu_Y) - A^*_z (Z - \mu_Z))' A_z Z \right] \\
&= \mathbb{E} \left[ (\vartheta - A^*_z (Z - \mu_Z))' A_z Z \right] - \mathbb{E} \left[ (A^*_y (Y - \mu_Y))' A_z Z \right] \\
(17) &= \mathbb{E} \left[ (\vartheta - A^*_z (Z - \mu_Z))' A_z Z \right]
\end{align*}
\]

as we make use of the fact that the rvs \(Y\) and \(Z\) are uncorrelated. It follows that

\[
\hat{\mathbb{E}} [\vartheta | Z] = A^*_z (Z - \mu_Z) \quad P - a.s.
\]

by the Orthogonality Principle characterizing the LMMSE estimator of \(\vartheta\) on the basis of \(Z\).

To conclude the proof we note that

\[
\begin{align*}
\hat{\mathbb{E}} [\vartheta | Y, Z] &= \ell^*(Y, Z) \\
&= A^*_y (Y - \mu_Y) + A^*_z (Z - \mu_Z) \\
(18) &= \hat{\mathbb{E}} [\vartheta | Y] + \hat{\mathbb{E}} [\vartheta | Z] \quad P - a.s.
\end{align*}
\]

as desired.

**Property K**

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

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

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

\[
\hat{\mathbb{E}} [\vartheta - \mathbb{E} [\vartheta] | Y, Z] = \hat{\mathbb{E}} [\vartheta | Y, Z] - \mathbb{E} [\vartheta] \quad P - a.s.
\]
while Property J applied to the zero-mean rv $\vartheta - \mathbb{E}[\vartheta]$ yields

$$
\hat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta] \mid Y, Z] = \hat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta] \mid Y] + \hat{\mathbb{E}}[\vartheta - \mathbb{E}[\vartheta] \mid Z] = \hat{\mathbb{E}}[\vartheta \mid Y] - \mathbb{E}[\vartheta] + \hat{\mathbb{E}}[\vartheta \mid Z] - \mathbb{E}[\vartheta] \\
= \hat{\mathbb{E}}[\vartheta \mid Y] + \hat{\mathbb{E}}[\vartheta \mid Z] - 2\mathbb{E}[\vartheta] \quad \mathbb{P} - a.s. \quad (19)
$$

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

Property L

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

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

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

We start by noting that

$$
\hat{\mathbb{E}}[Z \mid Y] = A^* Y + b^*
$$

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

Thus, with

$$
V \equiv Z - \hat{\mathbb{E}}[Z \mid Y],
$$

it holds that

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

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

$$
D = \begin{bmatrix}
I_k & O_{k \times m} \\
-A^* & I_m
\end{bmatrix}.
$$

Observe that the equation

$$
\begin{bmatrix}
I_k & O_{k \times m} \\
-A^* & I_m
\end{bmatrix} \begin{bmatrix}
y \\
z
\end{bmatrix} = \begin{bmatrix}
0_k \\
0_m
\end{bmatrix}
$$

implies

$$
I_k y + O_{k \times m} z = 0_k
$$

and

$$
-A^* y + I_m z = 0_m.
$$
7 THE GAUSSIAN CASE

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

As a result,

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

Invoking Property G and Property H we conclude that

$$\hat{E}[^\vartheta|Y, Z] = \hat{E}[^\vartheta|Y, V]$$

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

7 The Gaussian case