Definition

We want to estimate a random variable $x : \Omega \to \mathbf{R}^n$ from a random variable $y : \Omega \to \mathbf{R}^n$ using an estimator $\phi : \mathbf{R}^m \to \mathbf{R}^n$ which is affine.[1] In other words, $\phi(\xi) = A\xi + b$ for some $A \in \mathbf{R}^{n \times m}$ and $b \in \mathbf{R}^n$. We will use the mean squared error cost.

We want to find A and b to minimize

$$\mathbf{E}\|Ax + b - y\|^2.$$

Proof. Express $\mathbf{E}(\|Ax + b - y\|^2)$ as $\mathbf{E}((Ax + b - y)^\top (Ax + b - y))$

$$
\begin{array}{rrrr}
+ & \operatorname{tr}(A\mathbf{E}(xx^\top)A^\top) & + \quad \mathbf{E}(x)^\top A^\top b & - \quad \operatorname{tr}(A^\top \mathbf{E}(yx^\top)) \\
+ & b^\top A\mathbf{E}(x) & + \quad b^\top b & - \quad b^\top \mathbf{E}(y) \\
- & \operatorname{tr}(A\mathbf{E}(xy^\top)) & - \quad \mathbf{E}(y)^\top b & + \quad \mathbf{E}(yy^\top)
\end{array}
$$

The gradients with respect to b are

$$
\begin{array}{rrrr}
+ & 0 & + \quad A\mathbf{E}(x) & - \quad 0 \\
+ & A\mathbf{E}(x) & + \quad 2b & - \quad \mathbf{E}(y) \\
- & 0 & - \quad \mathbf{E}(y) & + \quad 0
\end{array}
$$

so $2A\mathbf{E}(x) + 2b - 2\mathbf{E}(y)$. The gradients with respect to A are

$$
\begin{array}{rrrr}
+ & \mathbf{E}(xx^\top)A^\top + \mathbf{E}(xx^\top)^\top A^\top & + \quad \mathbf{E}(x)b^\top & - \quad \mathbf{E}(yx^\top)^\top \\
+ \quad \mathbf{E}(x)b^\top & & + \quad 0 & - \quad 0 \\
- \quad \mathbf{E}(xy^\top) & & - \quad 0 & + \quad 0
\end{array}
$$

so $2\mathbf{E}(xx^\top)A^\top + 2\mathbf{E}(x)b^\top - 2\mathbf{E}(xy^\top)$. We want A and b solutions to

$$A\mathbf{E}(x) + b - \mathbf{E}(y) = 0$$
$$\mathbf{E}(xx^\top)A^\top + \mathbf{E}(x)b^\top - \mathbf{E}(xy^\top) = 0$$

so first get $b = \mathbf{E}(y) - A\mathbf{E}(x)$. Then express

$$\mathbf{E}(xx^\top)A^\top + \mathbf{E}(x)(\mathbf{E}(y) - A\mathbf{E}(x))^\top - \mathbf{E}(xy^\top) = 0.$$
$$\mathbf{E}(xx^\top)A^\top + \mathbf{E}(x)\mathbf{E}(y)^\top - \mathbf{E}(x)\mathbf{E}(x)^\top A^\top - \mathbf{E}(xy^\top) = 0.$$
$$(\mathbf{E}(xx^\top) - \mathbf{E}(x)\mathbf{E}(x)^\top)A^\top = \mathbf{E}(xy^\top) - \mathbf{E}(x)\mathbf{E}(y)^\top.$$
$$\operatorname{cov}(x,x)A^\top = \operatorname{cov}(x,y).$$

---

[1] Actually, the development flips this. Future editions will correct.

So $A^\top = (\text{cov}(x, x))^{-1} \text{cov}(x, y)$ means $A = \text{cov}(y, x)(\text{cov}(x, x))^{-1}$ is a solution. Then $b = \mathbf{E}(y) - \text{cov}(y, x) \text{cov}(x, x)^{-1}\mathbf{E}(x)$. So to summarize, the estimator $\phi(x) = Ax + b$ is

$$\text{cov}(y, x)(\text{cov } x, x)^{-1}x + \mathbf{E}(y) - \text{cov}(y, x)(\text{cov}(x, x))^{-1}\mathbf{E}(x)$$

or

$$\mathbf{E}(y) + \text{cov}(y, x)(\text{cov } x, x)^{-1}(x - \mathbf{E}(x))$$

$\square$