29 Linear Regression

The mean-square-error (MSE) problem of estimating a random variable $\boldsymbol{x}$ from observations of another random variable $\boldsymbol{y}$ seeks a mapping $c(\boldsymbol{y})$ that solves

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}=c(\boldsymbol{y})}{\operatorname{argmin}} \, \mathbb{E}\,(\boldsymbol{x} - c(\boldsymbol{y}))^2 \tag{29.1}$$

We showed in (27.18) that the optimal estimate is given by the conditional mean $\widehat{x} = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y)$. For example, for continuous random variables, the MSE estimate involves an integral computation of the form:

$$\widehat{x} \;=\; \int_{x\in\mathcal{X}} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \tag{29.2}$$

over the domain of realizations $x \in \mathcal{X}$. Evaluation of this solution requires knowledge of the conditional distribution, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. Even if $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ were available, computation of the integral expression is generally not possible in closed-form. In this chapter, we address this challenge by limiting $c(\boldsymbol{y})$ to the class of *affine* functions of $\boldsymbol{y}$. Assuming $\boldsymbol{y}$ is $M-$dimensional, affine functions take the following form:

$$c(\boldsymbol{y}) = \boldsymbol{y}^{\mathsf{T}} w - \theta \tag{29.3}$$

for some parameters $w \in \mathbb{R}^M$ and $\theta \in \mathbb{R}$; the latter is called the *offset* parameter. The problem of determining the MSE estimator $\widehat{\boldsymbol{x}}$ is then reduced to the problem of selecting optimal parameters (w, θ) to minimize the same mean-square-error in (29.1). Despite its apparent narrowness, this class of estimators leads to solutions that are tractable mathematically and deliver laudable performance in a wide range of applications.

29.1 REGRESSION MODEL

Although we can treat the inference problem in greater generality than below, by considering directly the problem of estimating a random *vector* $\boldsymbol{x}$ from another random *vector* $\boldsymbol{y}$, we will consider first the case of estimating a *scalar* $\boldsymbol{x}$ from a *vector* $\boldsymbol{y} \in \mathbb{R}^M$.

Let $\{\bar{x}, \bar{y}\}$ denote the first-order moments of the random variables $\boldsymbol{x} \in \mathbb{R}$ and

$\boldsymbol{y} \in \mathbb{R}^M$, i.e., their means:

$$\bar{x} = \mathbb{E}\,\boldsymbol{x}, \qquad \bar{y} = \mathbb{E}\,\boldsymbol{y} \tag{29.4a}$$

and let $\{\sigma_x^2, R_y, r_{xy}\}$ denote their second-order moments, i.e., their (co)-variances and cross-covariance vector:

$$\sigma_x^2 = \mathbb{E}\,(\boldsymbol{x} - \bar{x})^2, \qquad \text{(scalar)} \tag{29.4b}$$

$$R_y = \mathbb{E}\,(\boldsymbol{y} - \bar{y})(\boldsymbol{y} - \bar{y})^{\mathsf{T}}, \qquad (M \times M) \tag{29.4c}$$

$$r_{xy} = \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} = r_{yx}^{\mathsf{T}}, \qquad (1 \times M) \tag{29.4d}$$

The cross-covariance vector, r_{xy}, between $\boldsymbol{x}$ and $\boldsymbol{y}$ is a useful measure of the amount of information that one variable conveys about the other. We then pose the problem of determining the *linear* least-mean-square-error *estimator* (l.l.m.s.e., for short) of $\boldsymbol{x}$ given $\boldsymbol{y}$, namely, an estimator for the form

$$\widehat{\boldsymbol{x}}_{\text{LMSE}} = \boldsymbol{y}^{\mathsf{T}} w - \theta \;=\; w^{\mathsf{T}} \boldsymbol{y} - \theta, \qquad (\textbf{estimator for } \boldsymbol{x}) \tag{29.5}$$

where (w, θ) are determined by solving

$$\boxed{(w^o, \theta^o) = \underset{w,\theta}{\text{argmin}}\; \mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}}_{\text{LMSE}})^2} \tag{29.6}$$

The minus sign in front of the offset parameter θ in (29.5) is chosen for convenience, and the subscript LMSE refers to the "linear mean-square error estimator." Since in this chapter we will be dealing almost exclusively with linear estimators under the mean-square-error criterion, we will refrain from including the subscript and will simply write $\widehat{\boldsymbol{x}}$. It is customary to refer to model (29.5) as a *linear regression model* in the sense that the individual entries of $\boldsymbol{y}$ are being combined linearly, or a linear model is being fitted to the entries of $\boldsymbol{y}$, in order to estimate $\boldsymbol{x}$.

Theorem 29.1. (Linear estimators) *The solution (w^o, θ^o) to (29.6) satisfies the relations*

$$R_y w^o = r_{yx}, \qquad \theta^o = \bar{y}^{\mathsf{T}} w^o - \bar{x} \tag{29.7}$$

so that, when R_y is invertible, the estimator and the resulting minimum mean-square error (m.m.s.e.) are given by

$$\widehat{\boldsymbol{x}} - \bar{x} = r_{xy} R_y^{-1} (\boldsymbol{y} - \bar{y}) \tag{29.8a}$$

$$\text{m.m.s.e.} = \sigma_x^2 - r_{xy} R_y^{-1} r_{yx} \tag{29.8b}$$

Proof: We provide an algebraic proof. Let $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$. We expand the mean-square error (m.s.e.) to get

$$\begin{aligned} \text{m.s.e.} \;\triangleq\; \mathbb{E}\,\widetilde{\boldsymbol{x}}^2 &= \mathbb{E}\,(\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}} w + \theta)^2 \\ &= \mathbb{E}\,\boldsymbol{x}^2 - 2(\mathbb{E}\,\boldsymbol{x}\boldsymbol{y}^{\mathsf{T}}) w + 2\theta\bar{x} - 2\theta\bar{y}^{\mathsf{T}} w + w^{\mathsf{T}} (\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}) w + \theta^2 \end{aligned} \tag{29.9}$$

This m.s.e. is a quadratic function of w and θ. Differentiating with respect to w and θ and setting the derivatives to zero at the optimal solution gives:

$$\partial \text{ m.s.e.}/\partial\theta\Big|_{\theta=\theta^o, w=w^o} = 2\bar{x} - 2\bar{y}^{\mathsf{T}} w^o + 2\theta^o = 0 \tag{29.10}$$

$$\nabla_w \text{ m.s.e.}\Big|_{\theta=\theta^o, w=w^o} = -2\mathbb{E}\,\boldsymbol{x}\boldsymbol{y}^{\mathsf{T}} - 2\theta^o\bar{y}^{\mathsf{T}} + 2(w^o)^{\mathsf{T}}(\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}) = 0 \tag{29.11}$$

Solving for θ^o and w^o we find

$$\theta^o = \bar{y}^{\mathsf{T}} w^o - \bar{x} \quad \text{ and } \quad (\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}})w^o = \mathbb{E}\,\boldsymbol{y}\boldsymbol{x} + \theta^o\bar{y} \tag{29.12}$$

Replacing θ^o in the second expression for w^o and grouping terms gives

$$\left(\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} - \bar{y}\bar{y}^{\mathsf{T}}\right) w^o \;=\; \mathbb{E}\,\boldsymbol{y}\boldsymbol{x} - \bar{y}\bar{x} \iff R_y w^o = r_{yx} \tag{29.13}$$

which leads to (29.7). Substituting the expressions for (w^o, θ^o) into (29.9) leads to (29.8b). Finally, the Hessian matrix of the m.s.e. relative to w and θ is given by

$$H \;\triangleq\; \begin{bmatrix} \frac{\partial^2 \text{m.s.e.}}{\partial\theta^2} & \frac{\partial}{\partial\theta}\left(\nabla_w \text{ m.s.e.}\right) \\ \nabla_{w^{\mathsf{T}}}\left(\frac{\partial \text{m.s.e.}}{\partial\theta}\right) & \nabla_w^2 \text{ m.s.e.} \end{bmatrix} = \begin{bmatrix} 2 & -2\bar{y}^{\mathsf{T}} \\ -2\bar{y} & 2\,\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} \end{bmatrix} \tag{29.14}$$

This Hessian matrix is nonnegative-definite since its $(1,1)$ entry is positive and the Schur complement relative to this entry is nonnegative-definite:

$$\text{Schur complement} = 2\,\mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} - 2\bar{y}\bar{y}^{\mathsf{T}} \;=\; 2\,R_y \geq 0 \tag{29.15}$$

Since the m.s.e. cost is quadratic in the parameters (w, θ), we conclude that the solution (w^o, θ^o) corresponds to a global minimizer.

■

It is sufficient for our purposes to assume that $R_y > 0$. Observe from the statement of the theorem that the solution to the *linear* regression problem only requires knowledge of the first and second-order moments $\{\bar{x}, \bar{y}, \sigma_x^2, R_y, r_{xy}\}$; there is no need to know the full conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ as was the case with the optimal conditional mean estimator (27.18). Moreover, the observation, $\boldsymbol{y}$, does not appear in the m.m.s.e. expression. This means that we can assess beforehand, even before receiving the observation, the performance level that will be expected from the solution.

Remark 29.1. (Linear model) Consider two random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ and assume they are related by a linear model of the form:

$$\boldsymbol{x} = \boldsymbol{y}^{\mathsf{T}} z^o + \boldsymbol{v} \tag{29.16}$$

for some unknown parameter $z^o \in \mathbb{R}^M$, and where $\boldsymbol{v}$ has zero mean and is orthogonal to $\boldsymbol{y}$, i.e., $\mathbb{E}\,\boldsymbol{v}\boldsymbol{y}^{\mathsf{T}} = 0$. Taking expectation of both sides gives $\bar{x} = \bar{y}^{\mathsf{T}} z^o$ so that

$$\boldsymbol{x} - \bar{x} = (\boldsymbol{y} - \bar{y})^{\mathsf{T}} z^o + \boldsymbol{v} \tag{29.17}$$

Multiplying both sides by $(\boldsymbol{y} - \bar{y})$ from the left and taking expectations again gives

$$r_{yx} = R_y z^o \tag{29.18}$$

This is the same equation satisfied by the solution w^o in (29.7). We therefore conclude that when the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ happen to be related by a linear model as in (29.16),

then the estimator, $\widehat{\boldsymbol{x}}_{\rm LMSE}$, is able to recover the *exact* model z^o. Put in another way, when we solve a linear mean-square error problem, we are implicitly assuming that the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ satisfy a linear model of the above form.

■

Two properties

The linear least-mean-squares estimator satisfies two useful properties. First, the estimator is unbiased since by taking expectations of both sides of (29.8a) we get

$$\mathbb{E}\,(\widehat{\boldsymbol{x}} - \bar{x}) \;=\; r_{xy}R_y^{-1}\underbrace{\mathbb{E}\,(\boldsymbol{y} - \bar{y})}_{=0} \;=\; 0 \tag{29.19}$$

so that,

$$\boxed{\mathbb{E}\,\widehat{\boldsymbol{x}} = \bar{x}} \qquad (\textbf{unbiased estimator}) \tag{29.20}$$

Moreover, using (29.8a) again, we find that the variance of the estimator $\widehat{\boldsymbol{x}}$ is given by

$$\begin{aligned}
\sigma_{\widehat{x}}^2 &= \mathbb{E}\,(\widehat{\boldsymbol{x}} - \bar{x})^2 \\
&\overset{(29.8a)}{=} r_{xy}R_y^{-1}\underbrace{\big(\mathbb{E}\,(\boldsymbol{y} - \bar{y})(\boldsymbol{y} - \bar{y})^{\mathsf{T}}\big)}_{=R_y}R_y^{-1}r_{yx} \\
&= r_{xy}R_y^{-1}r_{yx}
\end{aligned} \tag{29.21}$$

so that expression (29.8b) for the m.m.s.e. can be equivalently rewritten as

$$\text{m.m.s.e.} \;=\; \mathbb{E}\,\widetilde{\boldsymbol{x}}^2 \;=\; \sigma_x^2 - \sigma_{\widehat{x}}^2 \tag{29.22}$$

Second, the linear least-mean-square error estimator satisfies an important orthogonality condition, namely, it is is uncorrelated with the observation:

$$\boxed{\mathbb{E}\,\widetilde{\boldsymbol{x}}\boldsymbol{y}^{\mathsf{T}} \;=\; 0} \qquad (\textbf{orthogonality principle}) \tag{29.23}$$

Proof of (29.23): From expression (29.8a) we note that

$$\begin{aligned}
\mathbb{E}\,\widetilde{\boldsymbol{x}}(\boldsymbol{y} - \bar{y})^{\mathsf{T}} &= \mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} \\
&\overset{(29.8a)}{=} \mathbb{E}\,\big(\boldsymbol{x} - \bar{x} - r_{xy}R_y^{-1}(\boldsymbol{y} - \bar{y})\big)(\boldsymbol{y} - \bar{y})^{\mathsf{T}} \\
&= \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} \;-\; r_{xy}R_y^{-1}\mathbb{E}\,(\boldsymbol{y} - \bar{y})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} \\
&= r_{xy} \;-\; r_{xy}R_y^{-1}R_y \\
&= 0
\end{aligned} \tag{29.24}$$

But since $\mathbb{E}\,\widetilde{\boldsymbol{x}} = 0$, we conclude that (29.23) holds.

■

It is because of the orthogonality property (29.23) between the estimation error

and the observation that equations (29.7) for w^o are referred to as the *normal equations*:

$$\boxed{R_y w^o = r_{yx}} \qquad \textbf{(normal equations)} \tag{29.25}$$

It can be verified that these equations are always consistent, i.e., a solution w^o always exists independent of whether R_y is invertible or not — see Appendix 29.A. Moreover, the orthogonality condition (29.23) plays a critical role in characterizing linear estimators.

Theorem 29.2. (Orthogonality principle) *An unbiased linear estimator is optimal in the least-mean-square-error sense if, and only if, its estimation error satisfies the orthogonality condition (29.23).*

Proof: One direction of the argument has already been proven prior to the statement, namely, the linear least-mean-square-error estimator, $\widehat{\boldsymbol{x}}$, given by (29.8a), satisfies (29.23). With regards to the converse statement, assume now that we are given an unbiased linear estimator for $\boldsymbol{x}$ of the form

$$\widehat{\boldsymbol{x}}_u = \boldsymbol{y}^{\mathsf{T}} w_u - \theta_u \tag{29.26}$$

and that the corresponding estimation error satisfies the orthogonality condition (29.23), i.e., $\widetilde{\boldsymbol{x}}_u \perp \boldsymbol{y}$. We verify that the parameters $\{w_u, \theta_u\}$ must coincide with the optimal parameters $\{w^o, \theta^o\}$ given by (29.7).

Indeed, the fact that $\widehat{\boldsymbol{x}}_u$ is unbiased means that $\mathbb{E}\,\widehat{\boldsymbol{x}}_u = \bar{x}$ and, hence, θ_u satisfies

$$\theta_u = -(\bar{x} - \bar{y}^{\mathsf{T}} w_u) \tag{29.27}$$

so that, by substituting into (29.26), we get that $\widehat{\boldsymbol{x}}_u$ satisfies

$$\widehat{\boldsymbol{x}}_u - \bar{x} = (\boldsymbol{y} - \bar{y})^{\mathsf{T}} w_u \tag{29.28}$$

Now using the assumed orthogonality condition $\widetilde{\boldsymbol{x}}_u \perp \boldsymbol{y}$ we must have

$$\mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}}_u)\boldsymbol{y}^{\mathsf{T}} = 0 \tag{29.29}$$

which is equivalent to

$$\mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}}_u)(\boldsymbol{y} - \bar{y})^{\mathsf{T}} = 0 \tag{29.30}$$

since, by assumption, $\mathbb{E}\,\widehat{\boldsymbol{x}}_u = \mathbb{E}\,\boldsymbol{x}$. Substituting expression (29.28) for $\widehat{\boldsymbol{x}}_u$ into (29.30), we find that w_u must satisfy

$$\mathbb{E}\left(\boldsymbol{x} - \bar{x} - (\boldsymbol{y} - \bar{y})^{\mathsf{T}} w_u\right)(\boldsymbol{y} - \bar{y})^{\mathsf{T}} = 0 \tag{29.31}$$

which leads to

$$r_{yx} = R_y w_u \tag{29.32}$$

so that w^o and w_u satisfy the same normal equations. Substituting w_u by w^o into expression (29.27) we get that $\theta_u = \theta^o$.

■

Example 29.1 **(Multiple noisy measurements of a binary signal)** Consider a signal $\boldsymbol{x}$ that assumes the values ± 1 with probability $1/2$ each. We collect N noisy measurements

$$\boldsymbol{y}_\ell = \boldsymbol{x} + \boldsymbol{v}_\ell, \qquad \ell = 1, \ldots, N \tag{29.33}$$

where $\boldsymbol{v}_\ell$ is zero-mean noise of unit-variance and independent of $\boldsymbol{x}$. We introduce the observation vector $\boldsymbol{y} = \mathrm{col}\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$. Then, say, for $N = 5$, it is straightforward to find that

$$r_{xy} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \qquad R_y = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix} \tag{29.34}$$

so that

$$\widehat{\boldsymbol{x}} \;=\; r_{xy} R_y^{-1} \boldsymbol{y} \;=\; \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}^{-1} \right) \boldsymbol{y} \tag{29.35}$$

We need to evaluate R_y^{-1}. Due to the special structure of R_y, its inverse can be evaluated in closed form for any N. Observe that, for any N, the matrix R_y can be expressed as $R_y = I_N + \mathbb{1}\mathbb{1}^{\mathsf{T}}$, where I_N is the $N \times N$ identity matrix and $\mathbb{1}$ is the $N \times 1$ column vector with unit entries, $\mathbb{1} = \mathrm{col}\{1, 1, 1, \ldots, 1\}$. In other words, R_y is a rank-one modification of the identity matrix. This is a useful observation since the inverse of every such matrix has a similar form (see Prob. 1.10). Specifically, it can be verified that, for any column vector $a \in \mathbb{R}^N$,

$$\left(I_N + aa^{\mathsf{T}}\right)^{-1} \;=\; I_N - \frac{aa^{\mathsf{T}}}{1 + \|a\|^2} \tag{29.36}$$

where $\|a\|^2 = a^{\mathsf{T}}a$ denotes the squared Euclidean norm of a. Using this result with $a = \mathbb{1}$, we find that

$$r_{xy} R_y^{-1} = \mathbb{1}^{\mathsf{T}} \left(I_N - \frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N+1} \right) \;=\; \mathbb{1}^{\mathsf{T}} - \frac{N}{N+1}\mathbb{1}^{\mathsf{T}} \;=\; \frac{\mathbb{1}^{\mathsf{T}}}{N+1} \tag{29.37}$$

so that

$$\widehat{\boldsymbol{x}} = \frac{1}{N+1}\,\mathbb{1}^{\mathsf{T}}\boldsymbol{y} \;=\; \frac{1}{N+1}\sum_{\ell=1}^{N} \boldsymbol{y}_\ell \tag{29.38}$$

Recall that in this problem the variable $\boldsymbol{x}$ is discrete and assumes the values ± 1. The estimator $\widehat{\boldsymbol{x}}$ in (29.38) will generally assume real-values. If one wishes to use $\widehat{\boldsymbol{x}}$ to decide whether $\boldsymbol{x} = +1$ or $\boldsymbol{x} = -1$, then one may consider examining the sign of $\widehat{\boldsymbol{x}}$ and use the sub-optimal estimator:

$$\widehat{\boldsymbol{x}}_{\mathrm{sub}} \;=\; \mathrm{sign}\left(\frac{1}{N+1}\sum_{n=1}^{N} \boldsymbol{y}_n \right) \tag{29.39}$$

where the sign function was defined earlier in (27.36).

Example 29.2 (**Learning a regression model from data**) Assume we can estimate the price of a house in some neighborhood $\mathbb{A}$, measured in units of $\times 1000$ USD, in some affine manner from the surface area s (measured in m^2) and the unit's age a (measured in years), say, as:

$$\widehat{P} = \alpha s + \beta a - \theta \tag{29.40}$$

for some unknown scalar parameters (α, β, θ). Here, the scalar θ denotes an *offset* parameter and the above relation represents the equation of a plane mapping values

(s, a) into an estimate for the house price, denoted by $\widehat{P}$. We collect the attributes $\{s, a\}$ into an observation vector:

$$y = \begin{bmatrix} s \\ a \end{bmatrix} \tag{29.41}$$

and the unknown parameters $\{\alpha, \beta\}$ into a column vector:

$$w = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{29.42}$$

Then, the mapping from y to $\widehat{P}$ can be written more compactly as:

$$\widehat{P} = y^{\mathsf{T}} w - \theta \tag{29.43}$$

The price P plays the role of the variable $\boldsymbol{x}$ that we wish to estimate from observations of $\boldsymbol{y}$. If we happen to know the first and second-order moments of the price and observation variables, then we could estimate (w, θ) by using

$$w^o = R_y^{-1} r_{yP}, \quad \theta^o = \bar{y}^{\mathsf{T}} w^o - \bar{P} \tag{29.44}$$

where $\bar{P}$ is the average price of houses in neighborhood $\mathbb{A}$. Often, in practice, these statistical moments are not known beforehand. They can, however, be estimated from measurements. Assume we have available a list of N houses from neighborhood $\mathbb{A}$ with their prices, size, and age; obviously, the house whose price we are interested in estimating should not be part of this list. We denote the available information by $\{P_n, y_n\}$ where $n = 1, 2, \ldots, N$. Then, we can estimate the first and second-order moments from this data by using the sample averages:

$$\widehat{\bar{P}} = \frac{1}{N} \sum_{n=1}^{N} P_n, \qquad \widehat{\bar{y}} = \frac{1}{N} \sum_{n=1}^{N} y_n \tag{29.45a}$$

and

$$\widehat{r}_{yP} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{\bar{y}})(P_n - \widehat{\bar{P}}) \tag{29.46a}$$

$$\widehat{R}_y = \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{\bar{y}})(y_n - \widehat{\bar{y}})^{\mathsf{T}} \tag{29.46b}$$

The parameters (w^o, θ^o) would be approximated by

$$w^\star = \widehat{R}_y^{-1} \widehat{r}_{yP} \tag{29.47a}$$

$$\theta^\star = \widehat{\bar{y}}^{\mathsf{T}} w^\star - \widehat{\bar{P}} \tag{29.47b}$$

where we are using the star notation to refer to parameters estimated directly from data measurements; this will be a standard convention in our treatment.

Figure 29.1 illustrates these results by means of a simulation. The figure shows the scatter diagram for $N = 500$ points (P_n, y_n) representing the triplet (price, area, age) for a collection of 500 houses. The price is measured in units of $\times 1000$ USD, the area in units of m^2, and the age in units of years. The spheres represent the measured data. The flat plane represents the regression plane that results from the above calculations, namely,

$$\widehat{P} = y^{\mathsf{T}} w^\star - \theta^\star \tag{29.48}$$

with

$$w^\star = \begin{bmatrix} 3.02 \\ -2.04 \end{bmatrix}, \quad \theta^\star = 1.26 \tag{29.49}$$

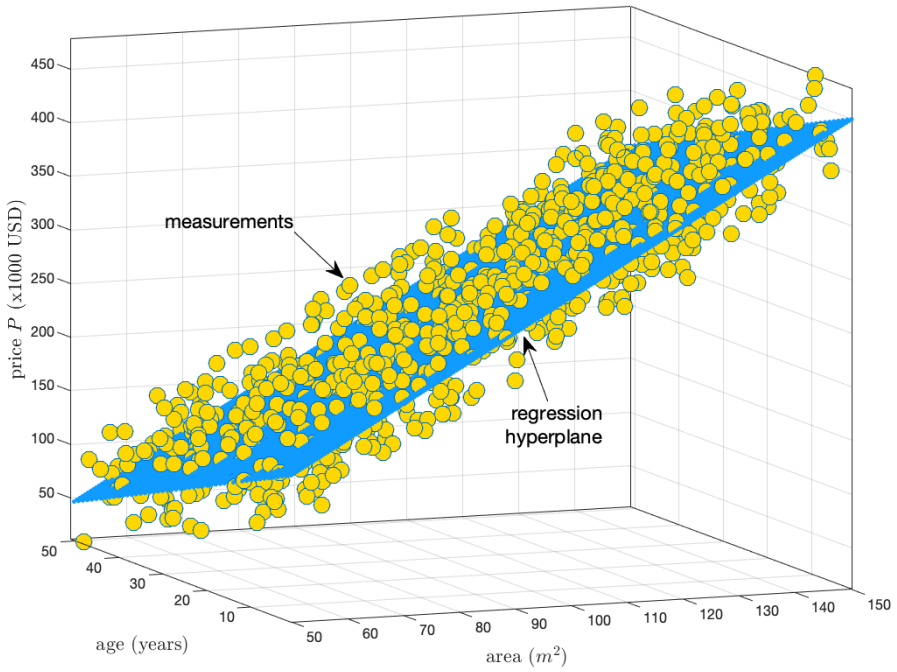

Figure 29.1 Scatter diagram of $N = 500$ points (P_n, y_n) representing the triplet (price, area, age) for a collection of 500 houses. The price is measured in units of $\times 1000$ USD, the area in units of m^2, and the age in units of years. The spheres represent the measured data. The flat plane represents the fitted regression plane (29.48).

The values of the parameters (α, β, θ) are measured in units of $\times 1000$ USD. Now given a 17−year old house with area $102\, m^2$, we can use the above parameter values to estimate/predict its price as follows:

$$\widehat{P} = (3.02 \times 102) - (2.04 \times 17) - 1.26 = 272.1\text{K USD} \tag{29.50}$$

29.2 CENTERING AND AUGMENTATION

We describe in this section two useful pre-processing steps that are common in inference and learning implementations in order to remove the need for the offset parameter, θ.

29.2.1 Centering

We start with centering. One useful fact to note is that the mean values $\{\bar{x}, \bar{y}\}$ appear subtracted from both sides of (29.8a). We say that the random variables

$\{\boldsymbol{x},\boldsymbol{y}\}$ are being centered, by subtracting their means, and that the estimation problem actually amounts to estimating one centered variable from another centered variable, i.e., to estimating $\boldsymbol{x}_c = \boldsymbol{x} - \bar{x}$ from $\boldsymbol{y}_c = \boldsymbol{y} - \bar{y}$ in a linear manner. Indeed, note that the variables $\{\boldsymbol{x},\boldsymbol{y}\}$ and $\{\boldsymbol{x}_c,\boldsymbol{y}_c\}$ have the same second-order moments since

$$R_{y_c} \triangleq \mathbb{E}\,\boldsymbol{y}_c\boldsymbol{y}_c^{\mathsf{T}} = \mathbb{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} \triangleq R_y \tag{29.51a}$$

and

$$r_{x_cy_c} \triangleq \mathbb{E}\,\boldsymbol{x}_c(\boldsymbol{y}_c)^{\mathsf{T}} = \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} \triangleq r_{xy} \tag{29.51b}$$

so that estimating $\boldsymbol{x}_c$ from $\boldsymbol{y}_c$ leads to the same relation

$$\widehat{\boldsymbol{x}}_c = r_{xy}R_y^{-1}\boldsymbol{y}_c \tag{29.52}$$

For this reason, and without loss in generality, it is customary in linear estimation problems to assume that the variables $\{\boldsymbol{x},\boldsymbol{y}\}$ have zero means (or have already been centered) and to solve the estimation task under this condition.

The fact that centering arises in the context of linear least-mean-square-error estimation can also be seen directly from (29.5) if we were to impose the requirement that the estimator should be unbiased. In that case, we would conclude from (29.5) that θ must satisfy

$$\bar{x} = \bar{y}^{\mathsf{T}}w - \theta \implies \theta = \bar{y}^{\mathsf{T}}w - \bar{x} \tag{29.53}$$

Substituting this condition for θ into (29.5) we find that, in effect, the estimator we are seeking should be of the form

$$\widehat{\boldsymbol{x}} - \bar{x} = (\boldsymbol{y}-\bar{y})^{\mathsf{T}}w \tag{29.54}$$

with centered variables appearing on both sides of the expression.

29.2.2 Augmentation

There is a second equivalent construction to centering, and which will be used extensively in later chapters. We refer to the affine model (29.5) and introduce the extended vectors:

$$\boldsymbol{y}' \triangleq \begin{bmatrix} 1 \\ \boldsymbol{y} \end{bmatrix}, \quad w' \triangleq \begin{bmatrix} -\theta \\ w \end{bmatrix}, \quad (M+1)\times 1 \tag{29.55}$$

where we have added the scalars 1 and $-\theta$ as leading entries on top of $\boldsymbol{y}$ and w, respectively. Then, the affine estimator (29.5) can be rewritten in the *linear* (rather than affine) form:

$$\widehat{\boldsymbol{x}} = (\boldsymbol{y}')^{\mathsf{T}}w' \tag{29.56}$$

and the design problem becomes one of determining a parameter vector $(w')^o$ that minimizes $\mathbb{E}\,\widetilde{\boldsymbol{x}}^2$. We already know from the statement of Theorem 29.1 applied to this extended problem that $(w')^o$ should satisfy the linear equations:

$$R_{y'}(w')^o = r_{y'x} \tag{29.57}$$

where the first and second-order moments are computed as follows:

$$\bar{y}' \triangleq \mathbb{E}\,\boldsymbol{y}' = \begin{bmatrix} 1 \\ \bar{y} \end{bmatrix} \tag{29.58a}$$

while

$$\begin{aligned} r_{y'x} &\triangleq \mathbb{E}\,(\boldsymbol{y}' - \bar{y}')(\boldsymbol{x} - \bar{x}) \\ &= \mathbb{E}\begin{bmatrix} 0 \\ (\boldsymbol{y} - \bar{y}) \end{bmatrix}(\boldsymbol{x} - \bar{x}) \\ &= \begin{bmatrix} 0 \\ r_{yx} \end{bmatrix} \end{aligned} \tag{29.58b}$$

and

$$\begin{aligned} R_{y'} &\triangleq \mathbb{E}\,\boldsymbol{y}'(\boldsymbol{y}')^{\mathsf{T}} - \bar{y}'(\bar{y}')^{\mathsf{T}} \\ &= \mathbb{E}\begin{bmatrix} 1 & \boldsymbol{y}^{\mathsf{T}} \\ \boldsymbol{y} & \boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} \end{bmatrix} - \begin{bmatrix} 1 & \bar{y}^{\mathsf{T}} \\ \bar{y} & \bar{y}\bar{y}^{\mathsf{T}} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & R_y \end{bmatrix} \end{aligned} \tag{29.58c}$$

If we denote the individual entries of $(w')^o$ by

$$(w')^o \triangleq \begin{bmatrix} -\theta^o \\ w^o \end{bmatrix} \tag{29.59}$$

then it follows from the expressions for $R_{y'}$ and $r_{y'x}$ and from (29.57) that the w^o component satisfies $R_y w^o = r_{yx}$. This result agrees with (29.7). Again, if we impose the unbiasedness condition that $\mathbb{E}\,\widehat{\boldsymbol{x}} = \bar{x}$ at the optimal solution $(w')^o$, then we conclude from (29.56) that θ^o must satisfy

$$\begin{aligned} \bar{x} &= (\bar{y}')^{\mathsf{T}}(w')^o \\ &= \begin{bmatrix} 1 & \bar{y}^{\mathsf{T}} \end{bmatrix}\begin{bmatrix} -\theta^o \\ w^o \end{bmatrix} \\ &= -\theta^o + \bar{y}^{\mathsf{T}} w^o \end{aligned} \tag{29.60}$$

from which we conclude that $\theta^o = \bar{y}^{\mathsf{T}} w^o - \bar{x}$, which again agrees with (29.7). For this reason, it is customary to assume that the variables have been extended to $(\boldsymbol{y}', w')$ as in (29.55), and to seek a linear (as opposed to affine) model as in (29.56). It is a matter of convenience whether we assume that the variables $(\boldsymbol{x}, \boldsymbol{y})$ are centered and replaced by $(\boldsymbol{x}_c, \boldsymbol{y}_c)$ or that $(\boldsymbol{y}, w)$ are extended and replaced by $(\boldsymbol{y}', w')$. The net effect in both cases is that we can assume that we are dealing with an *offset-free* problem that estimates $\boldsymbol{x}_c$ from $\boldsymbol{y}_c$ or $\boldsymbol{x}$ from $\boldsymbol{y}'$ in a linear (rather than affine) manner.

29.3 VECTOR ESTIMATION

The results from Sec. 29.1 can be easily extended to the case of estimating a *vector* (as opposed to a scalar) variable $\boldsymbol{x}$ from multiple measurements. Thus, consider a column vector $\boldsymbol{x}$ with entries

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_M \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_M \end{bmatrix}, \quad (M \times 1) \tag{29.61}$$

We wish to estimate $\boldsymbol{x}$ from multiple observations, say,

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}, \quad \bar{y} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_N \end{bmatrix}, \quad (N \times 1) \tag{29.62}$$

The solution to this problem can be deduced from the results of the previous sections. First, according to (29.8a), we express the estimate for an arbitrary $m-$th entry of $\boldsymbol{x}$ from all observations as follows:

$$(\widehat{\boldsymbol{x}}_m \; - \; \bar{x}_m) = w_m^{\mathsf{T}}(\boldsymbol{y} - \bar{y}) \tag{29.63}$$

for some column vector $w_m \in \mathbb{R}^N$, and where the variables have been centered around their respective means. The vectors $\{w_m\}$ are then determined by minimizing the aggregate mean-square-error:

$$\{w_m^o\} \triangleq \underset{\{w_m \in \mathbb{R}^N\}_{m=1}^M}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} \mathbb{E}\,(\boldsymbol{x}_m - \widehat{\boldsymbol{x}}_m)^2 \right\} \tag{29.64}$$

29.3.1 Error Covariance Matrix

We can rewrite the cost in an equivalent form. Assume we collect the estimators into a column vector,

$$\widehat{\boldsymbol{x}} = \begin{bmatrix} \widehat{\boldsymbol{x}}_1 \\ \widehat{\boldsymbol{x}}_2 \\ \vdots \\ \widehat{\boldsymbol{x}}_M \end{bmatrix} \tag{29.65}$$

and introduce the error vector:

$$\widetilde{\boldsymbol{x}} \triangleq \boldsymbol{x} - \widehat{\boldsymbol{x}} \tag{29.66}$$

Then, the mean-squared Euclidean norm of $\widetilde{\boldsymbol{x}}$ agrees with the cost appearing in (29.64), i.e.,

$$\mathbb{E}\,\|\widetilde{\boldsymbol{x}}\|^2 \;=\; \sum_{m=1}^{M} \mathbb{E}\,(\boldsymbol{x}_m - \widehat{\boldsymbol{x}}_m)^2 \tag{29.67}$$

If we further let $R_{\tilde{x}}$ denote the covariance matrix of the error vector, namely,

$$R_{\tilde{x}} \;\triangleq\; \mathbb{E}\,\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\mathsf{T}} \tag{29.68}$$

then we also have that

$$\mathbb{E}\,\|\widetilde{\boldsymbol{x}}\|^2 \;=\; \mathrm{Tr}\,(R_{\tilde{x}}) \tag{29.69}$$

so that, in effect, problem (29.64) is seeking the weight vectors $\{w_m^o\}$ that minimize the trace of the error covariance matrix:

$$\{w_m^o\} \;\triangleq\; \underset{\{w_m \in \mathbb{R}^N\}_{m=1}^M}{\mathrm{argmin}} \Big\{\mathrm{Tr}\,(R_{\tilde{x}})\Big\} \tag{29.70}$$

We refer to the trace of $R_{\tilde{x}}$, which appears in the above expression, as the mean-square-error in the vector case. We also refer to the matrix $R_{\tilde{x}}$ as the mean-square-error matrix. In this way, for vector estimation problems, when we write m.s.e. we may be referring either to the scalar quantity $\mathrm{Tr}(R_{\tilde{x}})$ or to the matrix quantity $R_{\tilde{x}}$ depending on the context. It is common to use the matrix representation for the m.s.e. in the vector case.

29.3.2 Normal Equations

Continuing with (29.64), we observe that the cost consists of the sum of M non-negative separable terms, with each term depending on the respective w_m. Therefore, we can determine the optimal coefficients $\{w_m,\ m = 1, \ldots, M\}$ by minimizing each term separately:

$$w_m^o \;=\; \underset{w_m \in \mathbb{R}^N}{\mathrm{argmin}}\ \mathbb{E}\,(\boldsymbol{x}_m - \widehat{\boldsymbol{x}}_m)^2 \tag{29.71}$$

This is the same problem we studied before: estimating a scalar variable $\boldsymbol{x}_m$ from multiple observations $\{\boldsymbol{y}_\ell,\ \ell = 1, \ldots, N\}$. We already know that the solution is given by the normal equations:

$$R_y w_m^o = r_{yx_m}, \quad m = 1, 2, \ldots, M \tag{29.72}$$

where $r_{x_m y}$ is the cross-covariance vector of $\boldsymbol{x}_m$ with $\boldsymbol{y}$:

$$r_{x_m y} \;\triangleq\; \mathbb{E}\,(\boldsymbol{x}_m - \bar{x}_m)(\boldsymbol{y} - \bar{y})^{\mathsf{T}} \tag{29.73}$$

Moreover, the resulting estimation error satisfies the orthogonality condition:

$$\mathbb{E}\,\widetilde{\boldsymbol{x}}_m \boldsymbol{y}^{\mathsf{T}} \;=\; 0, \quad m = 1, \ldots, M \tag{29.74}$$

where $\widetilde{\boldsymbol{x}}_m = \boldsymbol{x}_m - \widehat{\boldsymbol{x}}_m$. Note that the $\{r_{x_m y}\}$ are the rows of the $M \times N$ cross-covariance matrix R_{xy} between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$R_{xy} = \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})^{\mathsf{T}} = \begin{bmatrix} r_{x_1 y} \\ r_{x_2 y} \\ \vdots \\ r_{x_M y} \end{bmatrix}, \quad (M \times N) \tag{29.75}$$

Therefore, by collecting the w_m^o from (29.72) as columns into a matrix W^o, for all $k = 1, 2, \ldots, M$:

$$W^o = \begin{bmatrix} w_1^o & w_2^o & \ldots & w_M^o \end{bmatrix}, \quad (N \times M) \tag{29.76}$$

and by noting that $R_{yx} = R_{xy}^{\mathsf{T}}$, we find that W^o satisfies the normal equations

$$\boxed{R_y W^o = R_{yx}} \qquad (\textbf{normal equations}) \tag{29.77}$$

It follows that the optimal estimator is given by

$$\widehat{\boldsymbol{x}} - \bar{x} = (W^o)^{\mathsf{T}}(\boldsymbol{y} - \bar{y}) \tag{29.78}$$

or, equivalently,

$$\boxed{\widehat{\boldsymbol{x}} - \bar{x} = R_{xy} R_y^{-1}(\boldsymbol{y} - \bar{y})} \tag{29.79}$$

In view of (29.74), the resulting estimation error vector satisfies the orthogonality condition:

$$\boxed{\mathbb{E}\,\widetilde{\boldsymbol{x}}\boldsymbol{y}^{\mathsf{T}} = 0} \qquad (\textbf{orthogonality principle}) \tag{29.80}$$

where $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$. The corresponding minimum mean-square error matrix is given by

$$\begin{aligned} \text{m.m.s.e.} &= \mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}})^{\mathsf{T}} \\ &= \mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}})(\boldsymbol{x} - \bar{x} + \bar{x} - \widehat{\boldsymbol{x}})^{\mathsf{T}} \\ &= \mathbb{E}\,(\boldsymbol{x} - \widehat{\boldsymbol{x}})(\boldsymbol{x} - \bar{x})^{\mathsf{T}} \\ &= \mathbb{E}\,((\boldsymbol{x} - \bar{x}) - (\widehat{\boldsymbol{x}} - \bar{x}))(\boldsymbol{x} - \bar{x})^{\mathsf{T}} \\ &= \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}} - \mathbb{E}\,(\widehat{\boldsymbol{x}} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}} \end{aligned} \tag{29.81}$$

and we conclude from (29.79) that

$$\boxed{\text{m.m.s.e.} = R_x - R_{xy} R_y^{-1} R_{yx}} \tag{29.82}$$

in terms of the covariance and cross-covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$.

29.4 LINEAR MODELS

We apply the mean-square-error estimation theory of the previous sections to the important case of linear models, which arises often in applications. Thus, assume that zero-mean random vectors $\{\boldsymbol{x}, \boldsymbol{y}\}$ are related via a linear model of the form:

$$\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{v} \tag{29.83}$$

for some $N \times M$ known matrix H. We continue to assume that $\boldsymbol{y}$ is $N \times 1$ and $\boldsymbol{x}$ is $M \times 1$ so that we are estimating a vector variable from another vector variable. We explained earlier that the zero-mean assumption is not restrictive since the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ can be assumed to have been centered. In the above model, the variable $\boldsymbol{v}$ denotes a zero-mean random additive noise vector with known covariance matrix, $R_v = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}$. The covariance matrix of $\boldsymbol{x}$ is also assumed to be known, say, $R_x = \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}$. Both $\{\boldsymbol{x}, \boldsymbol{v}\}$ are uncorrelated with each other, i.e., $\mathbb{E}\,\boldsymbol{x}\boldsymbol{v}^{\mathsf{T}} = 0$, and we further assume that

$$R_x > 0, \quad R_v > 0 \tag{29.84}$$

Two equivalent representations

According to (29.79), when $R_y > 0$, the linear least-mean-square-error estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\widehat{\boldsymbol{x}} = R_{xy}R_y^{-1}\boldsymbol{y} \tag{29.85}$$

Because of (29.83), the covariances $\{R_{xy}, R_y\}$ can be determined in terms of the linear model parameters $\{H, R_x, R_v\}$. Indeed, the uncorrelatedness of $\{\boldsymbol{x}, \boldsymbol{v}\}$ gives

$$R_y = \mathbb{E}\,\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}} = \mathbb{E}\,(H\boldsymbol{x} + \boldsymbol{v})(H\boldsymbol{x} + \boldsymbol{v})^{\mathsf{T}} \;=\; HR_xH^{\mathsf{T}} + R_v \tag{29.86}$$

$$R_{xy} = \mathbb{E}\,\boldsymbol{x}\boldsymbol{y}^{\mathsf{T}} = \mathbb{E}\,\boldsymbol{x}(H\boldsymbol{x} + \boldsymbol{v})^{\mathsf{T}} \;=\; R_xH^{\mathsf{T}} \tag{29.87}$$

Moreover, since $R_v > 0$ we immediately get $R_y > 0$. Expression (29.85) for $\widehat{\boldsymbol{x}}$ then becomes

$$\widehat{\boldsymbol{x}} = R_xH^{\mathsf{T}}\left(R_v + HR_xH^{\mathsf{T}}\right)^{-1}\boldsymbol{y} \tag{29.88}$$

This expression can be rewritten in an equivalent form by using the *matrix inversion formula* (1.81). The result states that for arbitrary matrices $\{A, B, C, D\}$ of compatible dimensions, if A and C are invertible, then

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{29.89}$$

Applying this identity to the matrix $\left(R_v + HR_xH^{\mathsf{T}}\right)^{-1}$ in (29.88), with the identifications

$$A \leftarrow R_v, \quad B \leftarrow H, \quad C \leftarrow R_x, \quad D \leftarrow H^{\mathsf{T}} \tag{29.90}$$

we obtain

$$\begin{aligned}\widehat{\boldsymbol{x}} &= R_x H^{\mathsf{T}}\Big\{R_v^{-1} - R_v^{-1}H(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\Big\}\boldsymbol{y}\\ &= \Big\{R_x(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H) - R_xH^{\mathsf{T}}R_v^{-1}H\Big\}(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y}\\ &= \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \qquad (29.91)\end{aligned}$$

where in the second equality we factored out $(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y}$ from the right. Hence,

$$\widehat{\boldsymbol{x}} = \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \qquad (29.92)$$

This alternative form is useful in several contexts. Observe, for example, that when H happens to be a column vector (i.e., when $\boldsymbol{x}$ is a scalar), the quantity $(R_v + HR_xH^{\mathsf{T}})$ that appears in (29.88) is a matrix, while the quantity $(R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H)$ that appears in (29.92) is a scalar. In this case, the representation (29.92) leads to a simpler expression for $\widehat{\boldsymbol{x}}$. In general, the convenience of using (29.88) or (29.92) depends on the situation at hand.

It further follows that the m.m.s.e. matrix is given by

$$\begin{aligned}\text{m.m.s.e.} = \mathbb{E}\,\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\mathsf{T}} &= \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} - \mathbb{E}\,\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^{\mathsf{T}}\\ &= \mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})\boldsymbol{x}^{\mathsf{T}}, \quad \text{since } \widetilde{\boldsymbol{x}}\perp\widehat{\boldsymbol{x}}\\ &= R_x - \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1}H^{\mathsf{T}}R_v^{-1}HR_x\\ &= \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1} \qquad (29.93)\end{aligned}$$

where in the last equality we used the matrix inversion lemma again. That is,

$$\text{m.m.s.e.} = \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1} \qquad (29.94)$$

Lemma 29.1. (Equivalent linear estimators) *Consider two zero-mean random vectors $\{\boldsymbol{x},\boldsymbol{y}\}$ that are related via a linear model of the form $\boldsymbol{y} = H\boldsymbol{x}+\boldsymbol{v}$. The variables $\{\boldsymbol{x},\boldsymbol{v}\}$ have zero mean, positive-definite covariance matrices R_x and R_v, respectively, and are uncorrelated with each other. The linear least-mean-square-error estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ can be computed by either of the following equivalent expressions:*

$$\widehat{\boldsymbol{x}} = R_xH^{\mathsf{T}}\big(R_v+HR_xH^{\mathsf{T}}\big)^{-1}\boldsymbol{y} \qquad (29.95a)$$

$$= \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \qquad (29.95b)$$

with the resulting m.m.s.e. value given by

$$\text{m.m.s.e.} = \big(R_x^{-1}+H^{\mathsf{T}}R_v^{-1}H\big)^{-1} \qquad (29.96)$$

29.5 DATA FUSION

We illustrate one application of the theory for linear models to the important problem of fusing information from several sources in order to enhance the accuracy of the estimation process. Thus, assume that we have a collection of N sensors that are distributed over some region in space. All sensors are interested in estimating the same zero-mean vector $\boldsymbol{x}$ with covariance matrix $R_x = \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}$. For example, the sensors could be tracking a moving object and their objective is to estimate the speed and direction of motion of the target.

Assume each sensor k collects a measurement vector $\boldsymbol{y}_k$ that is related to $\boldsymbol{x}$ via a linear model, say,

$$\boldsymbol{y}_k = H_k\boldsymbol{x} + \boldsymbol{v}_k \tag{29.97}$$

where H_k is the model matrix that maps $\boldsymbol{x}$ to $\boldsymbol{y}_k$ at sensor k and $\boldsymbol{v}_k$ is zero-mean measurement noise with covariance matrix $R_k = \mathbb{E}\,\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{T}}$. In general, the quantity $\boldsymbol{y}_k$ is a vector, say, of size $L_k \times 1$. If the vector $\boldsymbol{x}$ has size $M \times 1$, then H_k is $L_k \times M$. We assume $L_k \geq M$ so that each sensor k has at least as many measurements as the size of the unknown vector $\boldsymbol{x}$.

29.5.1 Fusing Raw Data

In the first fusion method, each sensor k transmits its measurement vector $\boldsymbol{y}_k$ and its model parameters $\{H_k, R_k, R_x\}$ to a remote fusion center. The latter collects all measurements from across the nodes, $\{\boldsymbol{y}_k, k = 1, 2, \ldots, N\}$, and all model parameters $\{(H_k, R_k), k = 1, 2, \ldots, N\}$. The collected data satisfies the model:

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_N \end{bmatrix}}_{H} \boldsymbol{x} + \underbrace{\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}}_{\boldsymbol{v}} \tag{29.98}$$

which is a linear model of the same form as (29.83). We assume that the noises $\{\boldsymbol{v}_k\}$ across all nodes are uncorrelated with each other so that the covariance matrix of the aggregate noise vector, $\boldsymbol{v}$, is block diagonal:

$$R_v = \text{blkdiag}\Big\{R_1, R_2, \ldots, R_N\Big\} \tag{29.99}$$

Now, using (29.92), the fusion center can determine the estimator of $\boldsymbol{x}$ that is based on all measurements $\{\boldsymbol{y}_k\}$ as follows:

$$\widehat{\boldsymbol{x}} = \left(R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H\right)^{-1} H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \tag{29.100}$$

The resulting m.m.s.e. would be

$$P = \left(R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H\right)^{-1} \tag{29.101}$$

where we are denoting the m.m.s.e. by the letter P; it is a matrix of size $M \times M$. We note that we are deliberately using form (29.92) for the estimator of $\boldsymbol{x}$ because, as we are going to see shortly, this form will allow us to derive a second more efficient fusion method.

The solution method (29.100) requires all sensors to transmit $\{\boldsymbol{y}_k, H_k, R_k\}$ to the fusion center; this amounts to a total of

$$L_k + L_k M + \frac{1}{2}L_k^2 \quad \text{entries to be transmitted per sensor} \tag{29.102}$$

where the last term $(L_k^2/2)$ arises from the transmission of (half of) the entries of the $L_k \times L_k$ symmetric matrix R_k.

29.5.2 Fusing Processed Data

In the fusion method (29.100), the fusion center fuses the *raw* data $\{\boldsymbol{y}_k, H_k, R_k\}$ that are collected at the sensors. A more efficient data fusion method is possible and leads to a reduction in the amount of communication resources that are necessary between the sensors and the fusion center. This alternative method is based on the sensors performing some local processing first and then sharing the results of the processing step with the fusion center. The two fusion modes are illustrated in Fig. 29.2.

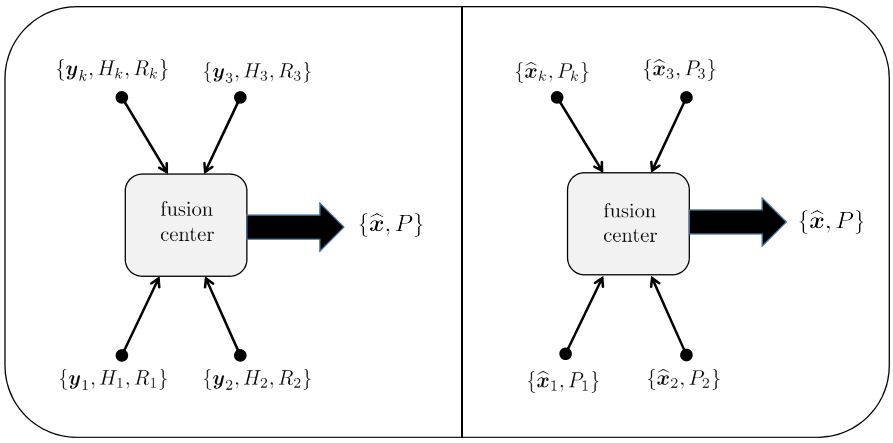

Figure 29.2 Two modes for data fusion: (a) on the left, the sensors share their raw data $\{\boldsymbol{y}_k, H_k, R_k\}$ with the fusion center, (b) while on the right, the sensors share their local estimation results $\{\widehat{\boldsymbol{x}}_k, P_k\}$ with the fusion center.

Specifically, assume that each node estimates $\boldsymbol{x}$ using its own data $\boldsymbol{y}_k$. We denote the resulting estimator by $\widehat{\boldsymbol{x}}_k$ and it is given by:

$$\widehat{\boldsymbol{x}}_k = \left(R_x^{-1} + H_k^{\mathsf{T}} R_k^{-1} H_k\right)^{-1} H_k^{\mathsf{T}} R_k^{-1} \boldsymbol{y}_k \tag{29.103}$$

The corresponding m.m.s.e. is

$$P_k = \left(R_x^{-1} + H_k^{\mathsf{T}} R_k^{-1} H_k\right)^{-1} \tag{29.104}$$

We assume now that the nodes share the processed data $\{\widehat{\boldsymbol{x}}_k, P_k\}$ with the fusion center rather than the raw data $\{\boldsymbol{y}_k, H_k, R_k\}$. It turns out that the desired global quantities $\{\widehat{\boldsymbol{x}}, P\}$ can be recovered from these processed data.

To begin with, observe that we can rework expression (29.101) for the global m.m.s.e. as follows:

$$\begin{aligned} P^{-1} &= R_x^{-1} + H^{\mathsf{T}} R_v^{-1} H \\ &= R_x^{-1} + \sum_{k=1}^{N} H_k^{\mathsf{T}} R_k^{-1} H_k \\ &= \sum_{k=1}^{N} \left(R_x^{-1} + H_k^{\mathsf{T}} R_k^{-1} H_k\right) \;-\; (N-1) R_x^{-1} \\ &= \sum_{k=1}^{N} P_k^{-1} \;-\; (N-1) R_x^{-1} \end{aligned} \tag{29.105}$$

This expression allows the fusion center to determine P^{-1} directly from knowledge of the quantities $\{P_k^{-1}, R_x^{-1}\}$. Note further that the global expression (29.100) can be rewritten as

$$P^{-1}\widehat{\boldsymbol{x}} = H^{\mathsf{T}} R_v^{-1} \boldsymbol{y} \tag{29.106}$$

which, using H and R_v from (29.98)–(29.99), leads to

$$P^{-1}\widehat{\boldsymbol{x}} = \sum_{k=1}^{N} H_k^{\mathsf{T}} R_k^{-1} \boldsymbol{y}_k \;=\; \sum_{k=1}^{N} P_k^{-1} \widehat{\boldsymbol{x}}_k \tag{29.107}$$

Therefore, we arrive at the following conclusion to fuse the data from multiple sensors.

Lemma 29.2. (Data fusion) *Consider a collection of N linear measurements of the form $\boldsymbol{y}_k = H_k \boldsymbol{x} + \boldsymbol{v}_k$, where $\boldsymbol{x}$ and $\boldsymbol{v}_k$ have zero mean, positive-definite covariance matrices R_x and R_k, respectively, and are uncorrelated with each other. Let $\widehat{\boldsymbol{x}}$ denote the linear least-mean-square-error estimator of $\boldsymbol{x}$ given the N observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ with error covariance matrix $P = \mathbb{E}\,\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\mathsf{T}}$. Let $\widehat{\boldsymbol{x}}_k$ denote the linear least-mean-square-error estimator of the same $\boldsymbol{x}$ given only $\boldsymbol{y}_k$ with error covariance matrix $P_k = \mathbb{E}\,\widetilde{\boldsymbol{x}}_k\widetilde{\boldsymbol{x}}_k^{\mathsf{T}}$. It holds that*

$$P^{-1} = \left(P_1^{-1} + P_2^{-1} + \ldots + P_N^{-1}\right) \;-\; (N-1) R_x^{-1} \tag{29.108a}$$

$$P^{-1}\widehat{\boldsymbol{x}} = P_1^{-1}\widehat{\boldsymbol{x}}_1 + P_2^{-1}\widehat{\boldsymbol{x}}_2 + \ldots + P_N^{-1}\widehat{\boldsymbol{x}}_N \tag{29.108b}$$

Observe from (29.108b) that the individual estimators are scaled by the inverses of their m.m.s.e matrices so that more accurate estimators are given more weight.

In this method, the sensors need to send to the fusion center the processed information $\{\widehat{\boldsymbol{x}}_k, P_k\}$. This amounts to a total of

$$M + M^2/2 \text{ entries to be transmitted per node} \tag{29.109}$$

which is smaller than (29.102) given that $L_k \geq M$.

29.6 MINIMUM-VARIANCE UNBIASED ESTIMATION

In the previous section we examined the linear model (29.83) where the unknown, $\boldsymbol{x}$, is modeled as a *random* variable with covariance matrix, R_x. We encountered one instance of this model earlier in Example 29.1, which dealt with the problem of estimating a zero-mean scalar random variable, $\boldsymbol{x}$, from a collection of noisy measurements, $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$. The model relating the variables in that example is a special case of (29.83) since it amounts to

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\triangleq \boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\triangleq H} \boldsymbol{x} + \underbrace{\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}}_{\triangleq \boldsymbol{v}} \tag{29.110}$$

where

$$H = \text{col}\{1, 1, \ldots, 1\} \tag{29.111a}$$

$$R_x = \sigma_x^2 \tag{29.111b}$$

$$R_v = \sigma_v^2 I_N \tag{29.111c}$$

Note that, for generality, we are using generic values for σ_x^2 and σ_v^2 rather than set them equal to one, as was the case in Example 29.1. We can recover the solution (29.38) by appealing to expression (29.92), which gives

$$\widehat{\boldsymbol{x}} = \left(\frac{1}{\sigma_x^2} + \frac{N}{\sigma_v^2}\right)^{-1} \frac{1}{\sigma_v^2} \sum_{\ell=1}^{N} \boldsymbol{y}_\ell \;=\; \frac{1}{N + \frac{1}{\text{SNR}}} \sum_{\ell=1}^{N} \boldsymbol{y}_\ell \tag{29.112}$$

in terms of the signal-to-noise ratio defined as

$$\text{SNR} \;\triangleq\; \sigma_x^2/\sigma_v^2 \tag{29.113}$$

In (29.110), the variable $\boldsymbol{x}$ is assumed to have been selected at random and, subsequently, N noisy measurements of this same value are collected. The observations are used to estimate $\boldsymbol{x}$ according to (29.112). Observe that the solution does *not* correspond to computing the sample mean of the observations. Expression (29.112) would reduce to the sample mean only when $\text{SNR} \to \infty$.

We now consider an alternative formulation for estimating the unknown by modeling it as a *deterministic* unknown constant, say, x, rather than a *random*

quantity, $\boldsymbol{x}$. For this purpose, we replace the earlier linear model (29.83) by one of the form

$$\boldsymbol{y} = Hx + \boldsymbol{v} \tag{29.114}$$

where, compared with (29.83), we are replacing the boldface letter $\boldsymbol{x}$ by the normal letter x (remember that we reserve the boldface notation to random variables). The observation vector $\boldsymbol{y}$ in (29.114) continues to be random since the disturbance $\boldsymbol{v}$ is random. Any estimator for x that is based on $\boldsymbol{y}$ will also be a random variable itself. Given model (29.114), we now study the problem of designing an optimal linear estimator for x of the form

$$\widehat{\boldsymbol{x}} = W^{\mathsf{T}}\boldsymbol{y} \tag{29.115}$$

for some coefficient matrix $W^{\mathsf{T}} \in \mathbb{R}^{M\times N}$ to be determined. It will turn out that, for such problems, W is found by solving a *constrained* least-mean-square-error estimation problem, as opposed to the unconstrained estimation problem (29.70).

29.6.1 Problem Formulation

Thus, consider a zero-mean random noise variable $\boldsymbol{v}$ with a positive-definite covariance matrix $R_v = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} > 0$, and let $\boldsymbol{y}$ be a noisy measurement of Hx according to model (29.114) where x is the unknown constant vector that we wish to estimate. The dimensions of the data matrix H are denoted by $N \times M$ and it is assumed that $N \geq M$ and that H has full rank:

$$\text{rank}(H) = M, \quad N \geq M \tag{29.116}$$

That is, H is a tall matrix so that the number of available measurements is at least as many as the number of unknown entries in x. The full rank condition on H guarantees that the matrix product $H^{\mathsf{T}}R_v^{-1}H$ is positive-definite — recall result (1.59). The inverse of this matrix product will appear in the expression for the estimator.

We are interested in determining a linear estimator for x of the form $\widehat{\boldsymbol{x}} = W^{\mathsf{T}}\boldsymbol{y}$. The choice of W should satisfy two conditions:

(a) (**Unbiasedness**). That is, we must guarantee $\mathbb{E}\,\widehat{\boldsymbol{x}} = x$, which is the same as $W^{\mathsf{T}}\mathbb{E}\,\boldsymbol{y} = x$. But from (29.114) we have $\mathbb{E}\,\boldsymbol{y} = Hx$ so that W must satisfy $W^{\mathsf{T}}Hx = x$, no matter what the value of x is. This condition means that W should satisfy

$$W^{\mathsf{T}}H = I_M \tag{29.117}$$

(b) (**Optimality**). The choice of W should minimize the trace of the covariance matrix of the estimation error, $\widetilde{\boldsymbol{x}} = x - \widehat{\boldsymbol{x}}$. Using the condition $W^{\mathsf{T}}H = I_M$, we find that

$$\widehat{\boldsymbol{x}} = W^{\mathsf{T}}\boldsymbol{y} \;=\; W^{\mathsf{T}}(Hx+\boldsymbol{v}) \;=\; W^{\mathsf{T}}Hx + W^{\mathsf{T}}\boldsymbol{v} \;=\; x + W^{\mathsf{T}}\boldsymbol{v} \tag{29.118}$$

so that $\widetilde{\boldsymbol{x}} = -W\boldsymbol{v}$. This means that the error covariance matrix, as a function of W, is given by

$$\mathbb{E}\,\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\mathsf{T}} \;=\; W^{\mathsf{T}}R_vW \tag{29.119}$$

Combining (29.117) and (29.119), we conclude that the desired W can be found by solving the following constrained optimization problem:

$$W^o \;\triangleq\; \underset{W}{\operatorname{argmin}}\,\Big\{\operatorname{Tr}(W^{\mathsf{T}}R_vW)\Big\},\;\; \text{subject to}\;\; W^{\mathsf{T}}H = I_M,\;\; R_v > 0 \tag{29.120}$$

The estimator $\widehat{\boldsymbol{x}} = (W^o)^{\mathsf{T}}\boldsymbol{y}$ that results from the solution of (29.120) is known as the *minimum-variance-unbiased estimator*, or MVUE for short. It is also sometimes called the *best linear unbiased estimator* or BLUE.

Example 29.3 (**Guessing the solution**) Let us first try to guess the form of the solution to the constrained problem (29.120) by appealing to the solution of the linear least-mean-square-error estimation problem (29.83). In that formulation, the unknown, $\boldsymbol{x}$, is modeled as a random variable with covariance matrix R_x. From expression (29.92), we know that the linear estimator is given by

$$\widehat{\boldsymbol{x}} = (R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \tag{29.121}$$

Now assume that the covariance matrix of $\boldsymbol{x}$ has the particular form $R_x = \sigma_x^2 I$, with a sufficiently large positive scalar σ_x^2 (i.e., $\sigma_x^2 \to \infty$). That is, assume that the variance of each of the entries of $\boldsymbol{x}$ is "infinitely" large. In this way, the variable $\boldsymbol{x}$ can be "regarded" as playing the role of some unknown constant vector, x. Then, the above expression for $\widehat{\boldsymbol{x}}$ reduces to

$$\widehat{\boldsymbol{x}} = (H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \tag{29.122}$$

This conclusion suggests that the choice

$$(W^o)^{\mathsf{T}} = (H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1} \tag{29.123}$$

should solve the problem of estimating the unknown constant vector x for model (29.114). We establish this result more formally next.

29.6.2 Gauss-Markov Theorem

Result (29.123) is a manifestation of the Gauss-Markov Theorem.

Theorem 29.3. (Gauss-Markov theorem) *Consider a linear model of the form* $\boldsymbol{y} = Hx + \boldsymbol{v}$, *where* x *is an unknown* constant, $\boldsymbol{v}$ *has zero-mean and covariance matrix* $R_v > 0$, *and* H *is a tall full-rank matrix (with as least as many rows as columns). The minimum-variance unbiased estimator for* x, *the one that solves (29.120), is given by*

$$\widehat{\boldsymbol{x}}_{\text{MVUE}} = (H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}\boldsymbol{y} \tag{29.124}$$

Equivalently, the optimal W *in (29.120) is*

$$(W^o)^{\mathsf{T}} = (H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1} \tag{29.125}$$

Moreover, the resulting minimum mean-square error is

$$R_{\widetilde{x}} \triangleq \mathbb{E}\,\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\mathsf{T}} = (H^{\mathsf{T}}R_v^{-1}H)^{-1} \tag{29.126}$$

Proof: Let $\mathcal{J}(W) = W^{\mathsf{T}}R_vW$ denote the cost function that appears in (29.120). Some straightforward algebra shows that $\mathcal{J}(W)$ can be expressed as

$$\mathcal{J}(W) = (W - W^o)^{\mathsf{T}}R_v(W - W^o) + (W^o)^{\mathsf{T}}R_vW^o \tag{29.127}$$

This is because, using $W^{\mathsf{T}}H = I$,

$$\begin{aligned} W^{\mathsf{T}}R_vW^o &= W^{\mathsf{T}}R_v\left(R_v^{-1}H(H^{\mathsf{T}}R_v^{-1}H)^{-1}\right) \\ &= W^{\mathsf{T}}H(H^{\mathsf{T}}R_v^{-1}H)^{-1} = (H^{\mathsf{T}}R_v^{-1}H)^{-1} \end{aligned} \tag{29.128}$$

Likewise, $(W^o)^{\mathsf{T}}R_vW^o = (H^{\mathsf{T}}R_v^{-1}H)^{-1}$. Relation (29.127) expresses the cost $\mathcal{J}(K)$ as the sum of two nonnegative-definite terms: one is independent of W and is equal to $(W^o)^{\mathsf{T}}R_vW^o$, while the other is dependent on W. It is then clear, since $R_v > 0$, that the trace of the cost is minimized by choosing $W = W^o$. Note further that the matrix W^o satisfies the constraint $(W^o)^{\mathsf{T}}H = I_M$.

■

Example 29.4 (**Sample mean estimator**) Let us reconsider problem (29.110) where $\boldsymbol{x}$ is now modeled as an unknown *constant*, i.e., we now write

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\triangleq \boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\triangleq H} x + \underbrace{\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}}_{\triangleq \boldsymbol{v}} \tag{29.129}$$

where the boldface $\boldsymbol{x}$ is replaced by x. In this case, the value of x can be regarded as the mean value for each measurement, $\boldsymbol{y}_\ell$. Using expression (29.124) we find that the minimum-variance unbiased estimator for x is given by (compare with (29.112)):

$$\widehat{\boldsymbol{x}}_{\text{MVUE}} = \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{y}_\ell \tag{29.130}$$

which is simply the sample mean estimator that we are familiar with from introductory courses on statistics.

29.7 COMMENTARIES AND DISCUSSION

Linear estimation. In this chapter we covered the basics of linear mean-square error regression analysis and highlighted only concepts that are relevant to the subject matter of the book, motivated by the presentations in Sayed (2003,2008) and Kailath, Sayed, and Hassibi (2000). The pioneering work in this domain was done independently by the Russian mathematician **Andrey Kolmogorov (1903–1987)** in the works by Kolmogorov (1939,1941a,b) and the American mathematician **Norbert Wiener (1894–1964)** in the work by Wiener (1949); the latter reference was originally published in 1942 as a classified report during World War II. Kolmogorov was motivated by the work of Wold (1938) on stationary processes and solved a linear prediction problem for discrete-time stationary random processes. Wiener, on the other hand, solved a continuous-time prediction problem under causality constraints by means of an elegant technique now known as the Wiener-Hopf technique introduced in Wiener and Hopf (1931). Readers interested in more details about Wiener's contribution, and linear estimation theory in general, may consult the textbook by Kailath, Sayed, and Hassibi (2000).

Unbiased estimators. In the least-mean-square-error estimation problems studied here, the estimators were required to be unbiased. Sometimes, unbiasedness can be a hurdle to minimizing the mean-square error. This is because there are estimators that are biased but that can achieve smaller error variances than unbiased estimators — see, e.g., Rao (1973), Cox and Hinkley (1974), and Kendall and Stuart (1976–1979).

Two interesting examples to this effect are the following given in Sayed (2008): the first example is from Kay (1993, pp. 310–311) while the second example is from Rao (1973). In the derivation leading to (29.130) we studied the problem of estimating the mean value, x, of N measurements $\{\boldsymbol{y}_\ell\}$. The minimum-variance unbiased estimator for x was seen to be given by the sample mean estimator:

$$\widehat{\boldsymbol{x}}_{\text{MVUE}} \;=\; \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{y}_\ell \tag{29.131}$$

The value of x was not restricted in any way; it was only assumed to be an unknown constant and that it could assume any value in the interval $(-\infty,\infty)$. But what if we know beforehand that x is limited to some interval, say $[-\alpha,\alpha]$ for some finite $\alpha>0$? One way to incorporate this piece of information into the design of an estimator for x is to consider the following alternative construction:

$$\breve{\boldsymbol{x}} \;=\; \begin{cases} -\alpha, & \text{if } \widehat{x}_{\text{MVUE}} < -\alpha \\ \widehat{\boldsymbol{x}}_{\text{MVUE}}, & \text{if } -\alpha \le \widehat{x}_{\text{MVUE}} \le \alpha \\ \alpha, & \text{if } \widehat{x}_{\text{MVUE}} > \alpha \end{cases} \tag{29.132}$$

in terms of a realization for $\widehat{x}_{\text{MVUE}}$. In this way, $\breve{\boldsymbol{x}}$ will always assume values within $[-\alpha,\alpha]$. A calculation in Kay (1993) shows that although the above (truncated mean) estimator $\breve{\boldsymbol{x}}$ is biased, it nevertheless satisfies $\mathbb{E}\,(x-\breve{\boldsymbol{x}})^2 < \mathbb{E}\,(x-\widehat{\boldsymbol{x}}_{\text{MVUE}})^2$ — see Prob. 29.5. In other words, the truncated mean estimator results in a smaller mean-square error.

A second classical example from the realm of statistics is the variance estimator. In this case, the parameter to be estimated is the variance of a random variable $\boldsymbol{y}$ given access to several observations of it, say $\{\boldsymbol{y}_\ell\}$. Let σ_y^2 denote the variance of $\boldsymbol{y}$. Two well-known estimators for σ_y^2 are

$$\widehat{\boldsymbol{\sigma}_y^2} \;=\; \frac{1}{N-1}\sum_{\ell=1}^{N}(\boldsymbol{y}_\ell-\overline{y})^2 \quad \text{and} \quad \widetilde{\boldsymbol{\sigma}^2}_y \;=\; \frac{1}{N+1}\sum_{\ell=1}^{N}(\boldsymbol{y}_\ell-\overline{y})^2 \tag{29.133}$$

where $\bar{y} = \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{y}_\ell$ is the sample mean. The first one is unbiased while the second one is biased. However, it is shown in Rao (1973) that

$$\mathbb{E}\left(\sigma_y^2 - \widetilde{\boldsymbol{\sigma}^2}_y\right)^2 < \mathbb{E}\left(\sigma_y^2 - \widehat{\boldsymbol{\sigma}_y^2}\right)^2 \tag{29.134}$$

We therefore see that biased estimators can result in smaller mean-square errors. However, unbiasedness is often a desirable property in practice since it guarantees that, on average, the estimator agrees with the unknown quantity that we seek to estimate.

Gauss-Markov theorem. Theorem 29.3 characterizes unbiased linear estimators of smallest error variance (or covariance), also known as BLUE estimators. Given an unknown $x \in \mathbb{R}^M$ and a random observation $\boldsymbol{y} \in \mathbb{R}^M$, the theorem was obtained by solving the constrained optimization problem (29.120), namely,

$$\min_{W} \mathbb{E}\,\|\widetilde{\boldsymbol{x}}\|^2, \;\; \text{subject to } \widehat{\boldsymbol{x}} = W^{\mathsf{T}}\boldsymbol{y}, \;\; \boldsymbol{y} = Hx + \boldsymbol{v}, \;\; W^{\mathsf{T}}H = I_M, \;\; \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} = R_v \tag{29.135}$$

The significance of the solution (29.124) is perhaps best understood if we recast it in the context of least-squares problems. We will study such problems in greater detail in Chapter 50; see also Sayed (2008). For now, let us assume that $R_v = \sigma_v^2 I_N$, so that the noise components are uncorrelated with each other and have equal variances. Then, expression (29.124) for the estimate of x reduces to

$$\widehat{x} = (H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}y \tag{29.136}$$

We are going to see later that this expression can be interpreted as the solution to the following least-squares problem. Given a noisy *deterministic* observation vector y satisfying

$$y = Hx + v \tag{29.137}$$

the unknown x can be estimated by solving

$$\widehat{x} = \underset{x\in\mathbb{R}^M}{\text{argmin}} \;\; \|y - Hx\|^2 \tag{29.138}$$

in terms of the squared Euclidean norm of the difference $y - Hx$. Indeed, if we expand the squared error we get

$$J(x) \;\triangleq\; \|y - Hx\|^2 \;=\; \|y\|^2 - 2y^{\mathsf{T}}Hx + x^{\mathsf{T}}H^{\mathsf{T}}Hx \tag{29.139}$$

Setting the gradient vector relative to x to zero at $\widehat{x}$ gives

$$\nabla_x J(x)\Big|_{x=\widehat{x}} = 2\widehat{x}^{\mathsf{T}}H^{\mathsf{T}}H - 2y^{\mathsf{T}}H = 0 \implies \widehat{x} = (H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}y \tag{29.140}$$

which is the same expression we had in (29.136). We therefore find that the standard least-squares estimate for x coincides with the BLUE estimate. More generally, consider a *weighted* least-squares problem of the form

$$\widehat{x} = \underset{x\in\mathbb{R}^M}{\text{argmin}} \;\; (y - Hx)^{\mathsf{T}}R(y - Hx) \tag{29.141}$$

where $R > 0$ denotes some weighting matrix. Differentiating again and solving for $\widehat{x}$ gives

$$\widehat{x} = (H^{\mathsf{T}}RH)^{-1}H^{\mathsf{T}}Ry \tag{29.142}$$

Comparing with (29.124) we find that the weighted least-squares solution would agree with the BLUE estimate if we select the weighting matrix as $R = R_v^{-1}$ (i.e., as the inverse of the noise covariance matrix). Therefore, the Gauss-Markov theorem is essentially stating that the least-squares solution leads to the best linear unbiased estimator (BLUE) if the weighting matrix is matched with R_v^{-1}.

The original version of the Gauss-Markov theorem with $R_v = \sigma_v^2 I_N$ is due to the German mathematician **Carl Friedrich Gauss (1777–1855)** and the Russian mathematician **Andrey Markov (1856–1922)**, who published versions of the result in Gauss (1821) and Markov (1912). The extension to the case of arbitrary covariance matrices R_v was given by Aitken (1935) — see also the overview by Plackett (1949,1950). The discussion in Sec. 29.6 on the Gauss Markov theorem, and the leading Secs. 29.4 and 29.5 on linear models and data fusion are adapted from the discussion in Kailath, Sayed, and Hassibi (2000).

PROBLEMS[1]

29.1 Show that the linear least-mean-square-error estimator defined by (29.70) also minimizes the determinant of the error covariance matrix, $\det(R_{\tilde{x}})$.

29.2 Show that the linear least-mean-square-error estimator defined by (29.70) also minimizes the $\mathbb{E}\,\widetilde{\boldsymbol{x}}^{\mathsf{T}} W \widetilde{\boldsymbol{x}}$ for any $W \geq 0$.

29.3 All variables are zero mean. Show that for any three random variables $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ it holds that

$$\widehat{\boldsymbol{x}}_{y,z} \;=\; \widehat{\boldsymbol{x}}_y \;+\; \widehat{(\widetilde{\boldsymbol{x}}_y)}_{\widetilde{z}_y}$$

where

$$\left\{\begin{array}{rcl} \widehat{\boldsymbol{x}}_{y,z} & = & \text{linear least-mean-squares estimator (l.l.m.s.e) of } \boldsymbol{x} \text{ given } \{\boldsymbol{z}, \boldsymbol{y}\}. \\ \widehat{\boldsymbol{x}}_y & = & \text{l.l.m.s.e of } \boldsymbol{x} \text{ given } \boldsymbol{y}. \\ \widehat{\boldsymbol{z}}_y & = & \text{l.l.m.s.e of } \boldsymbol{z} \text{ given } \boldsymbol{y}. \\ \widetilde{\boldsymbol{x}}_y & = & \boldsymbol{x} - \widehat{\boldsymbol{x}}_y \\ \widetilde{\boldsymbol{z}}_y & = & \boldsymbol{z} - \widehat{\boldsymbol{z}}_y \\ \widehat{(\widetilde{\boldsymbol{x}}_y)}_{\widetilde{z}_y} & = & \text{l.l.m.s.e of } \widetilde{\boldsymbol{x}}_y \text{ given } \widetilde{\boldsymbol{z}}_y. \end{array}\right.$$

What is the geometric interpretation of this result?

29.4 Verify that the mean-square-error values that correspond to the estimators $\{\widehat{\boldsymbol{x}}^o, \widehat{\boldsymbol{x}}^\bullet\}$ defined by (29.149a)–(29.149b) coincide.

29.5 Refer to the truncated mean estimator (29.132). Show that it results in a smaller mean-square error, namely, $\mathbb{E}\,(x - \breve{\boldsymbol{x}})^2 < \mathbb{E}\,(x - \widehat{\boldsymbol{x}}_{\text{MVUE}})^2$.

29.6 Let $\{\boldsymbol{x}, \boldsymbol{y}\}$ denote two zero-mean random variables with positive-definite covariance matrices $\{R_x, R_y\}$. Let $\widehat{\boldsymbol{x}}$ denote the linear least-mean-square-error estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$. Likewise, let $\widehat{\boldsymbol{y}}$ denote the linear least-mean-square-error estimator of $\boldsymbol{y}$ given $\boldsymbol{x}$. Introduce the estimation errors $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, and denote their covariance matrices by $R_{\tilde{x}}$ and $R_{\tilde{y}}$, respectively.

(a) Show that $R_x R_{\tilde{x}}^{-1} \widehat{\boldsymbol{x}} \;=\; R_{xy} R_{\tilde{y}}^{-1} \boldsymbol{y}$.

(b) Assume $\{\boldsymbol{y}, \boldsymbol{x}\}$ are related via a linear model of the form $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{v}$, where H is a matrix of appropriate dimensions while $\boldsymbol{v}$ has zero-mean with covariance matrix R_v and is uncorrelated with $\boldsymbol{x}$. Verify that the identity of part (a) reduces to $R_{\tilde{x}}^{-1} \widehat{\boldsymbol{x}} = H^{\mathsf{T}} R_v^{-1} \boldsymbol{y}$.

29.7 Let $\boldsymbol{x}$ be a zero-mean random variable with an $M \times M$ positive-definite covariance matrix R_x. Let $\widehat{\boldsymbol{x}}_1$ denote the linear least-mean-square-error estimator of $\boldsymbol{x}$ given a zero-mean observation $\boldsymbol{y}_1$. Likewise, let $\widehat{\boldsymbol{x}}_2$ denote the linear least-mean-square-error estimator of the same variable $\boldsymbol{x}$ given a second zero-mean observation $\boldsymbol{y}_2$. That is, we have two separate estimators for $\boldsymbol{x}$ from two separate sources. Let P_1 and P_2 denote the corresponding error covariance matrices: $P_1 = \mathbb{E}\,\widetilde{\boldsymbol{x}}_1 \widetilde{\boldsymbol{x}}_1^{\mathsf{T}}$ and $P_2 = \mathbb{E}\,\widetilde{\boldsymbol{x}}_2 \widetilde{\boldsymbol{x}}_2^{\mathsf{T}}$ where

[1] Some problems in this section are adapted from exercises in Sayed (2003,2008) and Kailath, Sayed, and Hassibi (2000).

$\widetilde{\boldsymbol{x}}_j = \boldsymbol{x} - \widehat{\boldsymbol{x}}_j$, Assume $P_1 > 0$ and $P_2 > 0$ and that the cross-covariance matrix

$$\mathbb{E} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y}_2 \end{bmatrix}^{\mathsf{T}}$$

has rank M.

(a) Show that the linear least-mean-square-error estimator of $\boldsymbol{x}$ given *both* $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$, denoted by $\widehat{\boldsymbol{x}}$, satisfies $P^{-1}\widehat{\boldsymbol{x}} = P_1^{-1}\widehat{\boldsymbol{x}}_1 + P_2^{-1}\widehat{\boldsymbol{x}}_2$, where P denotes the resulting error covariance matrix and is given by $P^{-1} = P_1^{-1} + P_2^{-1} - R_x^{-1}$.

(b) Assume $\{\boldsymbol{y}_1, \boldsymbol{x}\}$ and $\{\boldsymbol{y}_2, \boldsymbol{x}\}$ are related via linear models of the form $\boldsymbol{y}_1 = H_1\boldsymbol{x} + \boldsymbol{v}_1$ and $\boldsymbol{y}_2 = H_2\boldsymbol{x} + \boldsymbol{v}_2$, where $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ have zero means with covariance matrices $\{R_{v_1}, R_{v_2}\}$ and are uncorrelated with each other and with $\boldsymbol{x}$. Verify that this situation satisfies the required rank-deficiency condition and conclude that the estimator of $\boldsymbol{x}$ given $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ is given by the expression in part (a).

29.8 Let $\boldsymbol{y}_1 = H_1\boldsymbol{x} + \boldsymbol{v}_1$ and $\boldsymbol{y}_2 = H_2\boldsymbol{x} + \boldsymbol{v}_2$ denote two linear observation models with the same unknown random vector $\boldsymbol{x}$. All random variables have zero-mean. The covariance and cross-covariance matrices of $\{\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2\}$ are denoted by

$$\mathbb{E} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} R_x & 0 & 0 \\ 0 & R_1 & C \\ 0 & C^{\mathsf{T}} & R_2 \end{bmatrix}$$

In particular, observe that we are assuming the noises to be correlated with $C = \mathbb{E}\,\boldsymbol{v}_1\boldsymbol{v}_2^{\mathsf{T}}$. All covariance matrices are assumed to be invertible whenever necessary.

(a) Show how you would replace the observation vectors $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ by two other observation vectors $\{\boldsymbol{z}_1, \boldsymbol{z}_2\}$ of similar dimensions such that they satisfy linear models of the form

$$\boldsymbol{z}_1 = G_1\boldsymbol{x} + \boldsymbol{w}_1, \quad \boldsymbol{z}_2 = G_2\boldsymbol{x} + \boldsymbol{w}_2$$

for some matrices G_1 and G_2 to be specified, and where the noises $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ are now uncorrelated. What are the covariance matrices of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ in terms of R_1 and R_2?

(b) Let $\widehat{\boldsymbol{x}}_1$ be the linear least-mean-square-error estimator (l.l.m.s.e.) of $\boldsymbol{x}$ given $\boldsymbol{z}_1$ with error covariance matrix P_1. Similarly, let $\widehat{\boldsymbol{x}}_2$ be the l.l.m.s.e. of $\boldsymbol{x}$ given $\boldsymbol{y}_2$ with error covariance matrix P_2. Let further $\widehat{\boldsymbol{x}}$ denote the l.l.m.s.e. of $\boldsymbol{x}$ given $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ with error covariance matrix P. Determine expressions for $\widehat{\boldsymbol{x}}$ and P in terms of $\{\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2, P_1, P_2, C, R_x, R_1, R_2\}$.

29.9 Let $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{v}$. All random variables have zero-mean. The covariance and cross-covariance matrices of $\{\boldsymbol{x}, \boldsymbol{v}\}$ are denoted by

$$\mathbb{E} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{v} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{v} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} R_x & C \\ C^{\mathsf{T}} & R_v \end{bmatrix}$$

with positive-definite R_x and R_v.

(a) What is the l.l.m.s.e. of $\boldsymbol{x}$ given $\boldsymbol{y}$? What is the corresponding m.m.s.e.?

(b) A new scalar observation, $\boldsymbol{\alpha}$, is added to $\boldsymbol{y}$ and a new row vector is added to H such that

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} H \\ h^{\mathsf{T}} \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{n} \end{bmatrix}$$

where h^{T} is a row vector, $\boldsymbol{n}$ is uncorrelated with all other variables and has variance σ^2. Let $\widehat{\boldsymbol{x}}_{\text{new}}$ denote the new estimator for $\boldsymbol{x}$ given $\{\boldsymbol{y}, \boldsymbol{\alpha}\}$. Relate $\widehat{\boldsymbol{x}}_{\text{new}}$ to $\widehat{\boldsymbol{x}}$ from part (a). Relate also their m.m.s.e. values.

29.10 All variables are zero-mean. Let

$$\begin{bmatrix} \boldsymbol{y}_a \\ \boldsymbol{y} \\ \boldsymbol{y}_b \end{bmatrix} = \begin{bmatrix} H_a \\ H \\ H_b \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} \boldsymbol{v}_a \\ \boldsymbol{v} \\ \boldsymbol{v}_b \end{bmatrix}$$

where $\{\boldsymbol{v}_a, \boldsymbol{v}, \boldsymbol{v}_b\}$, are uncorrelated with $\boldsymbol{x}$ and have zero mean and covariance matrices:

$$\mathbb{E} \begin{bmatrix} \boldsymbol{v}_a \\ \boldsymbol{v} \\ \boldsymbol{v}_b \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_a \\ \boldsymbol{v} \\ \boldsymbol{v}_b \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} R_a & S_a & 0 \\ S_a^{\mathsf{T}} & R & S_b \\ 0 & S_b^{\mathsf{T}} & R_b \end{bmatrix}$$

Let $\widehat{\boldsymbol{x}}_{\boldsymbol{y}_a,\boldsymbol{y}}$ denote the linear estimator of $\boldsymbol{x}$ given $\{\boldsymbol{y}_a, \boldsymbol{y}\}$. Let $\widehat{\boldsymbol{x}}_{\boldsymbol{y}_b,\boldsymbol{y}}$ denote the linear estimator of $\boldsymbol{x}$ given $\{\boldsymbol{y}_b, \boldsymbol{y}\}$. Can you relate these estimators, and their minimum-mean-square-error, to each other?

29.11 Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{x}$ and $\boldsymbol{v}$ are independent zero-mean Gaussian random variables with variances σ_x^2 and σ_v^2, respectively. Show that the linear least-mean-square-error estimator of $\boldsymbol{x}^2$ using $\{\boldsymbol{y}, \boldsymbol{y}^2\}$ is

$$\widehat{\boldsymbol{x}^2} \;=\; \sigma_x^2 \;+\; \frac{\sigma_x^4}{\sigma_x^4 + 2\sigma_x^2\sigma_v^2 + \sigma_v^4}\,(y^2 - \sigma_x^2 - \sigma_v^2)$$

29.12 A random variable $\boldsymbol{z}$ is defined as follows

$$\boldsymbol{z} = \begin{cases} -\boldsymbol{x}, & \text{with probability } p \\ H\boldsymbol{x} + \boldsymbol{v}, & \text{with probability } 1-p \end{cases}$$

where $\boldsymbol{x}$ and $\boldsymbol{v}$ are zero-mean uncorrelated random vectors. Assume we know the linear least-mean-square-error estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$, namely, $\widehat{\boldsymbol{x}}_{|\boldsymbol{y}}$, where $\boldsymbol{y}$ is a zero-mean random random variable that is also uncorrelated with $\boldsymbol{v}$.

(a) Find an expression for $\widehat{\boldsymbol{z}}_{|\boldsymbol{y}}$ in terms of $\widehat{\boldsymbol{x}}_{|\boldsymbol{y}}$.

(b) Find an expression for the linear least-mean-square-error estimator $\widehat{\boldsymbol{x}}_{|\boldsymbol{z}}$ and the corresponding m.m.s.e.

29.13 Consider the distributed network with m nodes, shown in Fig. 29.13. Each node k observes a zero-mean measurement $\boldsymbol{y}_k$ that is related to an unknown zero-mean variable $\boldsymbol{x}$ via a linear model of the form $\boldsymbol{y}_k = H_k\boldsymbol{x} + \boldsymbol{v}_k$, where the data matrix H_k is known, and the noise $\boldsymbol{v}_k$ is zero mean and uncorrelated with $\boldsymbol{x}$. The noises across all nodes are uncorrelated with each other. Let $\{R_x, R_k\}$ denote the positive-definite covariance matrices of $\{\boldsymbol{x}, \boldsymbol{v}_k\}$, respectively. Introduce the following notation:

- At each node k, the notation $\widehat{\boldsymbol{x}}_k$ denotes the linear least-mean-squares estimator of $\boldsymbol{x}$ that is based on the observation $\boldsymbol{y}_k$. Likewise, P_k denotes the resulting error covariance matrix, $P_k = \mathbb{E}\,\widetilde{\boldsymbol{x}}_k\widetilde{\boldsymbol{x}}_k^{\mathsf{T}}$.
- At each node k, the notation $\widehat{\boldsymbol{x}}_{1:k}$ denotes the linear least-mean-squares estimator of $\boldsymbol{x}$ that is based on the observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$, i.e., on the observations collected at nodes 1 through k. Likewise, $P_{1:k}$ denotes the resulting error covariance matrix, $P_{1:k} = \mathbb{E}\,\widetilde{\boldsymbol{x}}_{1:k}\widetilde{\boldsymbol{x}}_{1:k}^{\mathsf{T}}$.

The network functions as follows. Node 1 uses $\boldsymbol{y}_1$ to estimate $\boldsymbol{x}$. The resulting estimator, $\widehat{\boldsymbol{x}}_1$, and the corresponding error covariance matrix, $P_1 = \mathbb{E}\,\widetilde{\boldsymbol{x}}_1\widetilde{\boldsymbol{x}}_1^{\mathsf{T}}$, are transmitted to node 2. Node 2 in turn uses its measurement $\boldsymbol{y}_2$ and the data $\{\widehat{\boldsymbol{x}}_1, P_1\}$ received from node 1 to compute the estimator of $\boldsymbol{x}$ that is based on both observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$. Note that node 2 *does not* have access to $\boldsymbol{y}_1$ but only to $\boldsymbol{y}_2$ and the information received from node 1. The estimator computed by node 2, $\widehat{\boldsymbol{x}}_{1:2}$, and the corresponding error covariance matrix, $P_{1:2}$, are then transmitted to node 3. Node 3 evaluates $\{\widehat{\boldsymbol{x}}_{1:3}, P_{1:3}\}$ using $\{\boldsymbol{y}_3, \hat{\boldsymbol{x}}_{1:2}, P_{1:2}\}$ and transmits $\{\widehat{\boldsymbol{x}}_{1:3}, P_{1:3}\}$ to node 4 and so forth.

(a) Find an expression for $\widehat{\boldsymbol{x}}_{1:m}$ in terms of $\widehat{\boldsymbol{x}}_{1:m-1}$ and $\widehat{\boldsymbol{x}}_m$.

(b) Find an expression for $P_{1:m}^{-1}$ in terms of $\{P_{1:m-1}^{-1}, P_m^{-1}, R_x^{-1}\}$.

(c) Find a recursion relating $P_{1:m}$ to $P_{1:m-1}$.

(d) Show that $P_{1:m}$ is a non-increasing sequence as a function of m.

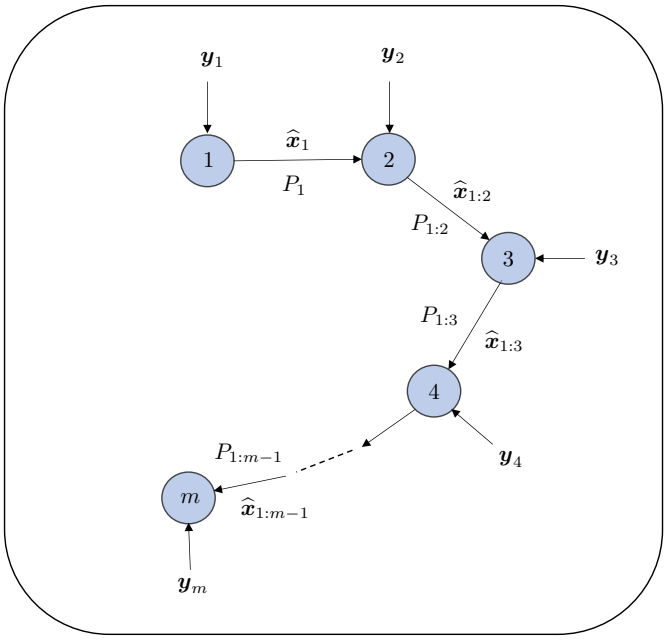

Figure 29.3 A distributed network with m nodes for Prob. 29.13.

(e) Assume $H_k = H$ for all k and $R_{v_k} = R_v > 0$. Assume further that H is tall and has full column rank. Find $\lim_{m\to\infty} P_{1:m}$.

29.14 Consider two sensors labeled $k = 1, 2$, and assume each sensor has an unbiased estimator, $\{\boldsymbol{w}_k, k = 1, 2\}$ for some $M \times 1$ column vector w^o. Let $\{P_k, k = 1, 2\}$ denote the error covariance matrix, $P_k = \mathbb{E}\,(w^o - \boldsymbol{w}_k)(w^o - \boldsymbol{w}_k)^{\mathsf{T}}$. Assume the errors of the two estimators are uncorrelated, i.e., $\mathbb{E}\,(w^o - \boldsymbol{w}_1)(w^o - \boldsymbol{w}_2)^{\mathsf{T}} = 0$. Consider a new aggregate estimator of the form $\widehat{\boldsymbol{w}} = \alpha \boldsymbol{w}_1 \ + \ (1 - \alpha)\boldsymbol{w}_2$.

(a) If α is nonnegative, determine the optimal scalar α that minimizes the mean-square-error, i.e., $\min_{\alpha \geq 0} \ \mathbb{E}\,\|w^o - \widehat{\boldsymbol{w}}\|^2$.

(b) Repeat part (a) when α is not restricted to being nonnegative. When would a negative α be advantageous?

(c) Let $P = \mathbb{E}\,(w^o - \widehat{\boldsymbol{w}})(w^o - \widehat{\boldsymbol{w}})^{\mathsf{T}}$. How does P compare to P_1 and P_2 in both cases (a) and (b)?

(d) Now assume the errors of the two estimators are *correlated* instead, i.e., $\mathbb{E}\,(w^o - \boldsymbol{w}_1)(w^o - \boldsymbol{w}_2)^{\mathsf{T}} = C$, for some matrix C. Repeat parts (a)–(c).

29.15 Consider N sensors labeled $k = 1, 2, \ldots, N$. Each node has an unbiased estimate of some unknown column vector $w^o \in \mathbb{R}^M$. We denote the individual estimator at node k by $\boldsymbol{w}_k$. We also denote the error covariance matrix of $\boldsymbol{w}_k$ by P_k and the cross-covariance matrix of $\boldsymbol{w}_k$ and $\boldsymbol{w}_\ell$ by $P_{k\ell}$. A sensor S wishes to combine the estimators $\{\boldsymbol{w}_k, k = 1, \ldots, N\}$ through $\widehat{\boldsymbol{w}}_S = \sum_{k=1}^{N} a_k \boldsymbol{w}_k$ in order to optimize the cost function:

$$\min_{\{a_k\}} \ \mathbb{E}\left\|w^o \ - \ \sum_{k=1}^{N} a_k \boldsymbol{w}_k\right\|^2$$

where the $\{a_k\}$ are real-valued scalars.

(a) Find a condition on the coefficients $\{a_k\}$ to ensure that the resulting $\widehat{\boldsymbol{w}}_S$ is an unbiased estimator for w^o.
(b) Under condition (a), find the optimal coefficients $\{a_k\}$. Your solution should not depend on w^o.
(c) Assume the reliability of each estimator $\boldsymbol{w}_k$ is measured by the scalar $\sigma_k^2 = \mathrm{Tr}(P_k)$. The smaller the σ_k^2 is, the more reliable the estimator will be. What is the relation between the optimal coefficients $\{a_k\}$ and the reliability factors $\{\sigma_k^2\}$?
(d) Evaluate the reliability of the estimator $\widehat{\boldsymbol{w}}_S$.
(e) Motivate and derive a stochastic gradient algorithm for updating the coefficients $\{a_k\}$ in part (b).
(f) How is the estimator of part (b) different from the unbiased linear least-mean-squares estimator of w^o based on the $\{\boldsymbol{w}_k\}$? Find the latter estimator.
(g) Find the minimum mean-square-error (m.m.s.e.) of the estimators in parts (b) and (f) for the case where $P_{k\ell} = 0$ when $\ell \neq k$. Specialize your result to the case $P_k = P$ for all k and compare the resulting mean-square-errors.

29.16 Consider a collection of N independent and identically-distributed random variables, $\{\boldsymbol{y}(n),\ n = 0, 1, \ldots, N-1\}$. Each $\boldsymbol{y}(n)$ has a Gaussian distribution with zero mean and variance σ_y^2. We want to use the observations $\{\boldsymbol{y}(n)\}$ to estimate the variance σ_y^2 in the following manner:

$$\widehat{\boldsymbol{\sigma}}_y^2 \ = \ \alpha \sum_{n=0}^{N-1} \boldsymbol{y}^2(n)$$

for some scalar parameter α to be determined.
(a) What is the mean of the estimator $\widehat{\boldsymbol{\sigma}}_y^2$ in terms of α and σ_y^2?
(b) Evaluate the mean-square-error below in terms of α and σ_y^2:

$$\text{m.s.e.} \ = \ \mathbb{E}\,(\widehat{\boldsymbol{\sigma}}_y^2 \ - \ \sigma_y^2)^2$$

(c) Determine the optimal scalar α that minimizes the m.s.e.. Is the corresponding estimator biased or unbiased?
(d) For what value of α would the estimator be unbiased? What is the m.s.e. of this estimator and how does it compare to the m.s.e. of the estimator from part (c)?

29.17 Consider noisy observations $\boldsymbol{y}(n) = \boldsymbol{x} + \boldsymbol{v}(n)$, where $\boldsymbol{x}$ and $\boldsymbol{v}(n)$ are independent random variables, $\boldsymbol{v}(n)$ is a white random process with zero mean and distributed as follows:

$$\begin{aligned} &\boldsymbol{v}(n) \text{ is Gaussian with variance } \sigma_v^2 \text{ with probability } q \\ &\boldsymbol{v}(n) \text{ is uniformly distributed over } [-a, a] \text{ with probability } 1-q \end{aligned}$$

Moreover, $\boldsymbol{x}$ assumes the values ± 1 with equal probability. The value of $\boldsymbol{x}$ is the same for all measurements $\{\boldsymbol{y}(n)\}$. All variables are real-valued.
(a) Find an expression for the linear least-mean-square-error estimator (l.l.m.s.e.) of $\boldsymbol{x}$ given the collection of N observations $\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(N-1)\}$.
(b) Find the l.l.m.s.e. of $\boldsymbol{x}$ given the observations $\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(N-1)\}$ and $\{\boldsymbol{y}^2(0), \boldsymbol{y}^2(1), \ldots, \boldsymbol{y}^2(N-1)\}$. How does the answer compare to part (a)?

29.18 This problem deals with constrained mean-square-error estimation. Let $\boldsymbol{d}$ denote a scalar zero-mean random variable with variance σ_d^2, and let $\boldsymbol{u}$ denote an $M \times 1$ zero-mean random vector with covariance matrix $R_u = \mathbb{E}\,\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}} > 0$. Consider the constrained optimization problem

$$\min_{w} \mathbb{E}\,(\boldsymbol{d} - \boldsymbol{u}^{\mathsf{T}} w)^2, \quad \text{subject to } \ c^{\mathsf{T}} w = \alpha$$

where c is a known $M \times 1$ vector and α is a known real scalar.
(a) Let $z = w - R_u^{-1} r_{ud}$ and $r_{du} = \mathbb{E}\,\boldsymbol{d}\boldsymbol{u}^{\mathsf{T}}$. Show that the above optimization problem is equivalent to the following:

$$\min_{z \in \mathbb{R}^M} \Big\{ \sigma_d^2 - r_{du} R_u^{-1} r_{ud} \ + \ z^{\mathsf{T}} R_u z \Big\}, \quad \text{subject to } \ c^{\mathsf{T}} z = \alpha - c^{\mathsf{T}} R_u^{-1} r_{ud}$$

(b) Show that the optimal solution, w^o, of the constrained optimization problem is given by

$$w^o = R_u^{-1} r_{ud} \; - \; \left(\frac{c^{\mathsf{T}} R_u^{-1} r_{ud} - \alpha}{c^{\mathsf{T}} R_u^{-1} c} \right) R_u^{-1} c$$

Verify that this solution satisfies the constraint $c^{\mathsf{T}} w^o = \alpha$.

29.19 Let $\boldsymbol{x}$ be a zero-mean random variable with an $M \times M$ positive-definite covariance matrix R_x. Let $\widehat{\boldsymbol{x}}_{\boldsymbol{y}_1}$ denote the linear least-mean-square-error estimator of $\boldsymbol{x}$ given a zero-mean observation $\boldsymbol{y}_1$ with covariance matrix R_{y_1}. Likewise, let $\widehat{\boldsymbol{x}}_{\boldsymbol{y}_2}$ denote the linear least-mean-square-error estimator of $\boldsymbol{x}$ given another zero-mean observation $\boldsymbol{y}_2$ with covariance matrix R_{y_2}. Let $R_{y_1,y_2} = \mathbb{E}\, \boldsymbol{y}_1 \boldsymbol{y}_2^{\mathsf{T}}$. We want to determine another estimator for $\boldsymbol{x}$ by combining $\widehat{\boldsymbol{x}}_{y_1}$ and $\widehat{\boldsymbol{x}}_{y_2}$ in a convex manner as follows:

$$\widehat{\boldsymbol{x}} = \lambda \widehat{\boldsymbol{x}}_{y_1} \; + \; (1-\lambda) \widehat{\boldsymbol{x}}_{y_2}$$

where λ is a real scalar lying inside the interval $0 \leq \lambda \leq 1$.

(a) Determine the value of λ that results in an estimator $\widehat{\boldsymbol{x}}$ with the smallest mean-square error.

(b) If λ is allowed to be any arbitrary real scalar (not necessarily limited to the range $[0, 1]$), how much smaller can the mean-square-error be?

29.20 Let $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{v}$ be a vector of measurements, where $\boldsymbol{v}$ is noise and $\boldsymbol{s}$ is the desired signal. Both $\boldsymbol{v}$ and $\boldsymbol{s}$ are zero-mean uncorrelated random vectors with covariance matrices $\{R_v, R_s\}$, respectively. We wish to determine a unit-norm column vector, w, such that the signal-to-noise ratio in the output signal, $\boldsymbol{y}^{\mathsf{T}} w$, is maximized.

(a) Verify that the covariance matrices of the signal and noise components in $\boldsymbol{y}^{\mathsf{T}} w$ are equal to $w^{\mathsf{T}} R_s w$ and $w^{\mathsf{T}} R_v w$, respectively.

(b) Assume first that $R_v = \sigma_v^2 I$. Use the Rayleigh-Ritz characterization (1.16) to conclude that the solution of

$$\max_{\|w\|=1} \left(\frac{w^{\mathsf{T}} R_s w}{\sigma_v^2 \|w\|^2} \right)$$

is given by the unit-norm eigenvector that corresponds to the maximum eigenvalue of R_s, written as $w^o = q_{\max}$, where $R_s q_{\max} = \lambda_{\max} q_{\max}$. Verify further that the resulting maximum SNR is equal to $\lambda_{\max}/\sigma_v^2$.

(c) Assume now that $\boldsymbol{v}$ is colored noise so that its covariance matrix is not necessarily diagonal. Introduce the eigen-decomposition $R_v = U \Lambda U^{\mathsf{T}}$, where U is orthogonal and Λ is diagonal with positive entries. Let $L = U \Lambda^{1/2}$. Repeat the argument of part (b) to show that the solution of

$$\max_{\|w\|=1} \left(\frac{w^{\mathsf{T}} R_s w}{w^{\mathsf{T}} R_v w} \right)$$

is now related to the unit-norm eigenvector that corresponds to the maximum eigenvalue of $L^{-1} R_s \left(L^{\mathsf{T}}\right)^{-1}$.

29.21 Consider the optimization problem:

$$\min_{W} \; W^{\mathsf{T}} R_v W, \quad \text{subject to} \quad W^{\mathsf{T}} H = A, \;\; R_v > 0$$

where W^{T} is $M \times N$, H is $N \times P$, A is $M \times P$, $P < N$, $M < N$, and H has full rank. In the text we assumed A is square and equal to the identity matrix (see (29.120)). Show that the optimal solution is given by

$$(W^o)^{\mathsf{T}} = A (H^{\mathsf{T}} R_v^{-1} H)^{-1} H^{\mathsf{T}} R_v^{-1}$$

and that the resulting minimum cost is $A (H^{\mathsf{T}} R_v^{-1} H)^{-1} A^{\mathsf{T}}$.

29.22 Refer to (29.108a). Compare P to P_k, for each $k = 1, 2, \ldots, N$. Specifically, verify that the difference $P_k - P$ is non-negative definite.

29.23 Refer to (29.108a)–(29.108b). Let $\{\widehat{\boldsymbol{x}}, P\}$ be the estimator and the m.m.s.e. that result from estimating $\boldsymbol{x}$ from data across all N sensors. Let $\{\widehat{\boldsymbol{x}}', P'\}$ be the estimator and the m.m.s.e. that result from estimating $\boldsymbol{x}$ from data across the first $N-1$ sensors. Relate $\{\widehat{\boldsymbol{x}}, P\}$ to $\{\widehat{\boldsymbol{x}}', P'\}$.

29.24 All variables are zero-mean. Consider a complex-valued scalar random variable $\boldsymbol{d}$ and a complex-valued $M \times 1$ regression vector $\boldsymbol{u}$. Let

$$\widehat{\boldsymbol{d}} = (w^o)^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{u}^{\mathsf{T}}w^o$$

denote the linear least-mean-square-error error (l.l.m.s.e.) estimator of $\boldsymbol{d}$ given $\boldsymbol{u}$ for some $M \times 1$ vector w^o. Consider additionally the problem of estimating separately the real and imaginary parts of $\boldsymbol{d}$ using knowledge of the real and imaginary parts of $\boldsymbol{u}$, also in the linear least-mean-square-error sense, namely,

$$\widehat{\boldsymbol{d}}_{\text{real}} = (w^o_{\text{real}})^{\mathsf{T}}\begin{bmatrix}\text{Re}(\boldsymbol{u})\\ \text{Im}(\boldsymbol{u})\end{bmatrix}, \quad \widehat{\boldsymbol{d}}_{\text{imag}} = (w^o_{\text{imag}})^{\mathsf{T}}\begin{bmatrix}\text{Re}(\boldsymbol{u})\\ \text{Im}(\boldsymbol{u})\end{bmatrix}$$

for some $2M \times 1$ vectors w^o_{real} and w^o_{imag}.

(a) Argue that estimating the real and imaginary parts of $\boldsymbol{d}$ from the real and imaginary parts of $\boldsymbol{u}$ is equivalent to estimating the real and imaginary parts of $\boldsymbol{d}$ from $\{\boldsymbol{u}, \boldsymbol{u}^*\}$, where $\boldsymbol{u}^*$ is the complex conjugate transpose of u.

(b) What are the optimal choices for w^o, w^o_{real} and w^o_{imag}?

(c) Let $\widehat{\boldsymbol{d}}_2 = \widehat{\boldsymbol{d}}_{\text{real}} + j\widehat{\boldsymbol{d}}_{\text{imag}}$ denote the estimator that is obtained for $\boldsymbol{d}$ from this second construction. What is the corresponding m.m.s.e.? How does it compare to the m.m.s.e. obtained for $\widehat{\boldsymbol{d}} = (w^o)^{\mathsf{T}}\boldsymbol{u}$? Under what conditions will both constructions lead to the same m.m.s.e.?

29.A CONSISTENCY OF NORMAL EQUATIONS

In this appendix we verify that the normal equations (29.25) are always consistent, i.e., we establish that a solution w^o always exists. Moreover, the solution is either unique or there are infinitely many solutions. In the latter case, all solutions will differ by vectors in the nullspace of R_y and, moreover, all of them will lead to the *same* estimator for $\boldsymbol{x}$ and to the *same* mean-square error. Only these possibilities can occur.

Proof (Consistency of normal equations). We verify that at least one solution w^o exists to the normal equations $R_y w^o = r_{yx}$. For this purpose, we need to verify that r_{yx} belongs to the range space of R_y, i.e.,

$$r_{yx} \in \mathcal{R}(R_y) \tag{29.143}$$

We show this property by contradiction. Assume that (29.143) does not hold. Under this assumption, there should exist some nonzero vector $p \in \mathcal{N}(R_y)$ that is *not* orthogonal to r_{yx}, namely,

$$\exists\, p \text{ such that } R_y p = 0, \;\; r_{yx}^{\mathsf{T}} p \neq 0 \tag{29.144}$$

It follows from $R_y p = 0$ that $p^{\mathsf{T}} R_y p = 0$ so that

$$\begin{aligned} p^{\mathsf{T}} R_y p = 0 &\iff p^{\mathsf{T}}\left(\mathbb{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}}\right) p = 0 \\ &\iff \mathbb{E}\left((\boldsymbol{y}-\bar{y})^{\mathsf{T}} p\right)^2 = 0 \end{aligned} \tag{29.145}$$

from which we conclude that the zero-mean scalar random variable $(\boldsymbol{y}-\bar{y})^{\mathsf{T}}p$ has zero variance and, hence,

$$(\boldsymbol{y}-\bar{y})^{\mathsf{T}}p \;=\; 0, \quad \text{in probability} \tag{29.146}$$

This conclusion leads to a contradiction since it implies that

$$\begin{aligned} r_{yx}^{\mathsf{T}}p \quad &= \quad \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}}\,p \\ &\overset{(29.146)}{=} 0, \quad \text{in probability} \end{aligned} \tag{29.147}$$

which violates the assertion that $r_{yx}^{\mathsf{T}}p \neq 0$. We conclude that (29.143) holds and the normal equations (29.25) are consistent.

Next we verify that the solution w^o is either unique or there are infinitely many solutions. To being with, it is clear from (29.25) that the solution is unique whenever R_y is invertible, in which case $w^o = R_y^{-1}r_{yx}$. On the other hand, when R_y is singular, then infinitely many solutions exist. This is because if we let p denote any nontrivial vector in the nullspace of R_y, i.e., $R_y p = 0$, then the vector $w^o + p$ will also satisfy the normal equations (29.25).

The next property we verify is that when infinitely many solutions exist, any solution will continue to lead to the *same* estimator for $\boldsymbol{x}$ and to the *same* mean-square-error value. Let w^o and $w^\bullet$ denote any two solutions to the normal equations, i.e.,

$$R_y w^o = r_{yx}, \qquad R_y w^\bullet = r_{yx} \tag{29.148}$$

The corresponding estimators for $\boldsymbol{x}$ are denoted by

$$\widehat{\boldsymbol{x}}^o = \bar{x} + (\boldsymbol{y}-\bar{y})^{\mathsf{T}}w^o \tag{29.149a}$$

$$\widehat{\boldsymbol{x}}^\bullet = \bar{x} + (\boldsymbol{y}-\bar{y})^{\mathsf{T}}w^\bullet \tag{29.149b}$$

Subtracting both equalities in (29.148) gives

$$R_y(w^o - w^\bullet) = 0 \tag{29.150}$$

so that any two solution vectors differ by vectors in the nullspace of R_y, namely,

$$w^\bullet = w^o + p, \quad \text{for some } p \in \mathcal{N}(R_y) \tag{29.151}$$

Moreover, we obtain from (29.150) that

$$\begin{aligned} R_y(w^o - w^\bullet) = 0 &\Longrightarrow (w^o - w^\bullet)^{\mathsf{T}} R_y (w^o - w^\bullet) = 0 \\ &\Longrightarrow (w^o - w^\bullet)^{\mathsf{T}} \left(\mathbb{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}}\right) (w^o - w^\bullet) = 0 \\ &\Longrightarrow \mathbb{E}\,\Big((\boldsymbol{y}-\bar{y})^{\mathsf{T}}(w^o - w^\bullet)\Big)^2 = 0 \end{aligned} \tag{29.152}$$

which implies that the following zero-mean scalar random variable is equal to zero in probability:

$$\boldsymbol{\alpha} \;\overset{\Delta}{=}\; (\boldsymbol{y}-\bar{y})^{\mathsf{T}}(w^o - w^\bullet) = 0, \quad \text{in probability} \tag{29.153}$$

That is, for any $\epsilon > 0$,

$$\mathbb{P}(|\boldsymbol{\alpha}| \geq \epsilon) = 0 \tag{29.154}$$

Subtracting expressions (29.149a)–(29.149b) gives

$$\begin{aligned} \widehat{\boldsymbol{x}}^o - \widehat{\boldsymbol{x}}^\bullet \quad &= \quad (\boldsymbol{y}-\bar{y})^{\mathsf{T}}\,(w^o - w^\bullet) \\ &\overset{(29.153)}{=} 0, \quad \text{in probability} \end{aligned} \tag{29.155}$$

which confirms our claim that different solutions to the normal equations continue to

lead to the same estimator. It is left as an exercise to check that the mean-square-errors corresponding to $\widehat{\boldsymbol{x}}^o$ and $\widehat{\boldsymbol{x}}^\bullet$ also agree with each other — see Prob. 29.4.

■

REFERENCES

Aitken, A. C. (1935), "On least squares and linear combinations of observations," *Proceedings of the Royal Society of Edinburgh*, vol. 55, pp. 42–48.

Cox, D. R. and D. V. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall, NY.

Gauss, C. F. (1821), *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae,* H. Dieterich, vol. 2, pp. 1–58.

Kailath, T., A. H. Sayed, and B. Hassibi (2000), *Linear Estimation*, Prentice Hall, NJ.

Kay, S. (1993), *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, NJ.

Kendall, M. and A. Stuart (1976–1979), *The Advanced Theory of Statistics*, vols. 1–3, Macmillan, NY.

Kolmogorov, A. N. (1939), "Sur l'interpolation et extrapolation des suites stationnaires," *C. R. Acad. Sci.*, vol. 208, p. 2043.

Kolmogorov, A. N. (1941a), "Stationary sequences in Hilbert space (in Russian)," *Bull. Math. Univ. Moscow*, vol. 2.

Kolmogorov, A. N. (1941b), "Interpolation and extrapolation of stationary random processes," *Bull. Acad. Sci. USSR,* vol. 5. [A translation has been published by the RAND Corp., Santa Monica, CA, as Memo. RM-3090-PR, Apr. 1962.]

Markov, A. A. (1912), *Wahrscheinlichkeitsrechnung,* B. G. Teubner, Leipzig, Berlin.

Plackett, R. L. (1949), "A historical note on the method of least squares," *Biometrika*, vol. 36, no. 3–4, pp. 458–4601.

Plackett, R. L. (1950), "Some theorems in least-squares," *Biometrika*, vol. 37, no. 1–2, pp. 149–157.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, Wiley, NY.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley, NJ.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley, NJ.

Wiener, N. (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* Technology Press and Wiley, NY. [Originally published in 1942 as a classified National Defense Research Council Report. Also published under the title *Time Series Analysis* by MIT Press, Cambridge, MA.]

Wiener, N. and E. Hopf (1931), "On a class of singular integral equations," *Proc. Prussian Acad. Math. – Phys. Ser.*, p. 696, 1931.

Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Upssala, Sweden.