6 Bayesian Parameter Estimation

The problem of hypothesis testing can be viewed as one of making decisions about the value of a binary variable taking values in the alphabet $\mathcal{H} = \{H_0, H_1\}$. In the Bayesian formulation, this variable H is modeled as random, while in the Neyman-Pearson formulation, H is modeled as deterministic, but unknown.

Hypothesis testing can be extended to $M > 2$ hypotheses in a fairly straightforward way, though we will not develop the details of such M-ary hypothesis testing in these notes. Instead we will proceed from a different perspective. In particular, an M-ary hypothesis test can be viewed as a problem of estimating a discrete-valued parameter of the model for the observations. Thus, we now turn our attention to problems of estimation, where the parameter(s) of interest are either discrete- or continuous-valued.

As with hypothesis testing, there are both Bayesian and nonBayesian approaches. In this section, we focus on the Bayesian formulation, where the (generally vector-valued) parameter of interest is modeled as a random variable $\mathbf{x}$ taking values in some alphabet $\mathcal{X}$. As before, our observations $\mathbf{y}$ are random and take values in some alphabet $\mathcal{Y}$.

Such problems arise in a wide range of applications. For example, in a face recognition setting, $\mathbf{x}$ might be a feature vector (consisting of, e.g., eye color, hair color, head proportions, etc), and $\mathbf{y}$ might be a digital image. As another example, in an air-traffic control setting, $\mathbf{x}$ might be a vector representing the position and velocity of an aircraft, and $\mathbf{y}$ might be a vector of radar return measurements from several sensors.

Our treatment generally encompasses both discrete- and continuous-valued parameters. However, we will typically focus on the continuous-valued case, which is most typically of interest. We will indicate when there are significant distinctions. Likewise, as in hypothesis testing, we accommodate either discrete- or continuous-valued observations. Note that when not otherwise specified, we will take $\mathcal{X}$ and $\mathcal{Y}$ to be all reals of appropriate dimensions.

In the Bayesian framework, there is an *a priori* distribution $p_{\mathbf{x}}(\cdot)$ for the unknown parameter. This represents our *belief* about $\mathbf{x}$ prior to any observation of the measurement $\mathbf{y}$.

In addition, the observation model takes the form of a conditional distribution $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\cdot)$, which fully specifies the way in which $\mathbf{y}$ contains information about $\mathbf{x}$.

Example 1. Suppose that $\mathbf{y}$ is a noise-corrupted measurement of some function of $\mathbf{x}$, viz.,

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \mathbf{w} \tag{1}$$

where $\mathbf{w}$ is a random noise vector that is independent of $\mathbf{x}$ and has density $p_{\mathbf{w}}(\mathbf{w})$. Then

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = p_{\mathbf{w}}(\mathbf{y} - \mathbf{h}(\mathbf{x})). \tag{2}$$

Suppose in addition, $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $\mathbf{w} \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda})$ where the matrix $\mathbf{A}$ and covariance matrix $\mathbf{\Lambda}$ are arbitrary. Then

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \mathtt{N}(\mathbf{y}; \mathbf{A}\mathbf{x}, \mathbf{\Lambda}).$$

The observation model $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and prior distribution $p_{\mathbf{x}}(\mathbf{x})$ together constitute a full statistical characterization of $\mathbf{x}$ and $\mathbf{y}$. In particular, the joint distribution is given by their product, i.e.,

$$p_{\mathbf{y},\mathbf{x}}(\mathbf{y}, \mathbf{x}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x}). \tag{3}$$

As in hypothesis testing, for a given observation $\mathbf{y}$, a complete characterization of our knowledge of the parameter $\mathbf{x}$ is given by the *posterior* distribution for $\mathbf{x}$, i.e.,

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{\int_{\mathcal{X}} p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}')\, p_{\mathbf{x}}(\mathbf{x}')\, \mathrm{d}\mathbf{x}'}. \tag{4}$$

This posterior, which we again observe is computed via Bayes' Rule, represents the revision of our belief based on the the availability of an observation.

When, in an application of interest, one must go further than computing this belief and actually make a decision about (i.e., guess) the value of $\mathbf{x}$, then the tools of estimation theory are required, which we develop now. Before proceeding, we note that the revised belief is sometimes referred to as a "soft" decision, and an estimate is correspondingly referred to as a "hard" decision. However, such terminology is not universal.

We use $\hat{\mathbf{x}}(\mathbf{y})$ to denote our estimate of $\mathbf{x}$ based on observing that the measurement $\mathbf{y} = \mathbf{y}$. Note that what we are estimating is actually an entire vector-valued function $\hat{\mathbf{x}}(\cdot)$. In particular, for each possible observed value $\mathbf{y}$, the quantity $\hat{\mathbf{x}}(\mathbf{y})$ represents the estimate of the corresponding value of $\mathbf{x}$. This function is called the "estimator."

In general, to find a good estimator for $\mathbf{x}$, we need some measure of goodness of candidate estimators. In other words, we need a suitable performance criterion with respect to which we optimize our choice of estimator. In the Bayesian formulation, we begin by choosing a deterministic scalar-valued function $C(\mathbf{a}, \hat{\mathbf{a}})$ that specifies the cost of estimating an arbitrary vector $\mathbf{a}$ as $\hat{\mathbf{a}}$. Then, we choose our estimator $\hat{\mathbf{x}}(\cdot)$ as that function which minimizes the average cost, i.e.,

$$\hat{\mathbf{x}}(\cdot) = \arg\min_{\mathbf{f}(\cdot)} \mathbb{E}\left[C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\right]. \tag{5}$$

Note that the expectation in (5) is over $\mathbf{x}$ and $\mathbf{y}$ jointly, and hence $\hat{\mathbf{x}}(\cdot)$ is that function which minimizes the cost averaged over all possible $(\mathbf{x}, \mathbf{y})$ pairs.

As in Bayesian hypothesis testing, solving for the optimum function $\hat{\mathbf{x}}(\cdot)$ in (5) can, in fact, be accomplished on a *pointwise basis*, i.e., for each particular value $\mathbf{y}$ that is observed, we find the best possible choice (in the sense of (5)) for the corresponding estimate $\hat{\mathbf{x}}(\mathbf{y})$. Indeed, following the same approach used in hypothesis testing, we use (3) to first rewrite our objective function in (5) in the form

$$\begin{aligned}\mathbb{E}\left[C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\right] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\, p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}\right] p_{\mathbf{y}}(\mathbf{y})\, \mathrm{d}\mathbf{y}. \end{aligned} \tag{6}$$

Then, since $p_{\mathbf{y}}(\mathbf{y}) \geq 0$, we clearly will minimize (6) if we choose $\hat{\mathbf{x}}(\mathbf{y})$ to minimize the term in brackets for each individual value of $\mathbf{y}$, i.e.,

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}. \tag{7}$$

Note that when $\mathbf{x}$ is discrete-valued, the integral in (7) becomes a summation.

As we would expect, and as (7) indicates, the estimate depends on the model only through the belief $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$. That is, this belief summarizes everything we need to know about the $\mathbf{x}$ and $\mathbf{y}$ to construct the optimal Bayesian estimators for *any* given cost criterion.

In the remainder of this section, we focus on some examples of common cost criteria. For simplicity, we will generally restrict our attention to the case of estimating scalar variables x from (generally) vector observations $\mathbf{y}$. We note in advance, however, that for some classes of cost criteria—specifically, when the cost function is *additive*, i.e., of the form

$$C(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=1}^{N} C_i(x_i, \hat{x}_i) \tag{8}$$

for some $C_i(\cdot, \cdot)$, and with x_i and $\hat{x}_i$ denoting the ith elements of the vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively—the more general case of estimating vector variables $\mathbf{x}$ can be handled in a component-wise manner. We will see a specific instance of this.

6.1 Minimum Absolute-Error Estimation

One possible choice for the cost function is based on a minimum absolute-error (MAE) criterion. The cost function of interest in this case is

$$C(a, \hat{a}) = |a - \hat{a}|. \tag{9}$$

Claim 1. *The MAE estimate is the median of the belief* $p_{x|\mathbf{y}}(x|\mathbf{y})$.

Proof. Substituting (9) into (7) we obtain

$$\begin{aligned}\hat{x}_{\mathrm{MAE}}(\mathbf{y}) &= \arg\min_a \int_{-\infty}^{+\infty} |x-a|\, p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x \\ &= \arg\min_a \left\{ \int_{-\infty}^{a} (a-x) p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x + \int_{a}^{+\infty} (x-a) p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x \right\}. \end{aligned} \tag{10}$$

Differentiating the quantity inside braces in (10) with respect to a gives, via Leibnitz' rule, the condition

$$\left[\int_{-\infty}^{a} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x - \int_{a}^{+\infty} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x \right] \Bigg|_{a=\hat{x}_{\mathrm{MAE}}(\mathbf{y})} = 0. \tag{11}$$

Rewriting (11) we obtain

$$\int_{-\infty}^{\hat{x}_{\mathrm{MAE}}(\mathbf{y})} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x = \int_{\hat{x}_{\mathrm{MAE}}(\mathbf{y})}^{+\infty} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x = \frac{1}{2}. \tag{12}$$

From (12) we see that the $\hat{x}_{\mathrm{MAE}}(\mathbf{y})$ is the threshold in x of the belief $p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})$ for which half the probability is located above the threshold and, hence, half is also below the threshold. Hence, the MAE estimator for x given $\mathbf{y} = \mathbf{y}$ is the median of the belief. □

Example 2. Suppose we have the posterior density

$$p_{\mathsf{x}|y}(x|y) = \begin{cases} 1/(3y) & 0 < x < y \\ 2/(3y) & y < x < 2y \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\hat{x}_{\mathrm{MAE}}(y) = (1+\Delta)y$$

for an appropriate choice of $\Delta > 0$. To solve for Δ, we use (12) to obtain

$$\frac{1}{3y} \cdot y + \frac{2}{3y} \cdot y\Delta = 1/2$$

from which we deduce that $\Delta = 1/4$.

We also note that the median of a density is not necessarily unique, as the following example illustrates.

Example 3. Suppose

$$p_{\mathsf{x}|y}(x|y) = \begin{cases} 1/2y & 0 < x < y \text{ and } 2y < x < 3y \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Then the median of (13) is any number between y and $2y$; hence, the MAE estimators for x given $\mathsf{y} = y$ are all of the form

$$\hat{x}_{\mathrm{MAE}}(y) = \alpha$$

where α is any constant satisfying $y \leq \alpha \leq 2y$ (assuming $y \geq 0$) or $2y \leq \alpha \leq y$ (assuming $y < 0$).

6.2 Maximum A Posteriori Estimation

As an alternative to that considered in the previous section, consider the minimum uniform cost (MUC) cost criterion, whereby

$$C(a, \hat{a}) = \begin{cases} 1 & |a - \hat{a}| > \epsilon \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

which uniformly penalizes all estimation errors with magnitude bigger than ϵ.

Claim 2. *In the limit as $\epsilon \to 0$, the MUC estimate is the mode of the belief $p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})$, i.e., it is the* maximum a posteriori *(MAP) estimate*

$$\hat{x}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_a p_{\mathsf{x}|\mathbf{y}}(a|\mathbf{y}). \tag{15}$$

From this claim, we see that the MAP estimator can be viewed as resulting from a Bayes' cost formulation in which all errors are, in an appropriate sense, equally bad.

Proof. Substituting (14) into (7) we obtain that the MUC estimator satisfies

$$\begin{aligned} \hat{x}_{\mathrm{MUC}}(\mathbf{y}) &= \arg\min_a \left[1 - \int_{a-\epsilon}^{a+\epsilon} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y}) \, \mathrm{d}x \right] \\ &= \arg\max_a \int_{a-\epsilon}^{a+\epsilon} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y}) \, \mathrm{d}x. \end{aligned} \tag{16}$$

Note that via (16) we see that $\hat{x}_{\mathrm{MUC}}(\mathbf{y})$ corresponds to the value of a that makes $\mathbb{P}\left(|\mathsf{x} - \hat{x}_{\mathrm{MUC}}(\mathbf{y})| < \epsilon \mid \mathbf{y} = \mathbf{y}\right)$ as large as possible. This means finding the interval of length 2ϵ where the posterior density $p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})$ is most concentrated.

If we carry this perspective a little further, we see that if we let ϵ get sufficiently small then the $\hat{x}_{\mathrm{MUC}}(\mathbf{y})$ approaches the point corresponding to the peak of the posterior density, i.e., the MAP estimate (15):

$$\lim_{\epsilon \to 0} \hat{x}_{\mathrm{MUC}}(\mathbf{y}) = \arg\max_a p_{\mathsf{x}|\mathbf{y}}(a|\mathbf{y}). \tag{17}$$

□

6.3 Bayes' Least-Squares Estimation

Perhaps the most popular Bayesian estimator is based on a quadratic cost criterion, which we now develop. Specifically, we consider the mean-square error (MSE) cost criterion

$$C(\mathbf{a}, \hat{\mathbf{a}}) = \|\mathbf{a} - \hat{\mathbf{a}}\|^2 = (\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}(\mathbf{a} - \hat{\mathbf{a}}) = \sum_{i=1}^{N}(a_i - \hat{a}_i)^2, \tag{18}$$

from which we obtain what is termed the Bayes least-squares (BLS) estimator. Since this estimator minimizes the mean-square estimation error, it is often alternatively referred to as the minimum mean-square error (MMSE) estimator and denoted using $\hat{\mathbf{x}}_{\mathrm{MMSE}}(\cdot)$.

Claim 3. *The BLS estimate is the mean of the belief* $p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})$.

Proof. Substituting (18) into (7) yields

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{a})^{\mathrm{T}} (\mathbf{x} - \mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x} \tag{19}$$

Let us begin with the simpler case of scalar estimation, for which (19) becomes

$$\hat{x}_{\mathrm{BLS}}(\mathbf{y}) = \arg\min_{a} \int_{-\infty}^{+\infty} (x - a)^2\, p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x. \tag{20}$$

As we did in the case of MAE estimation, we can perform the minimization in (20) by differentiating with respect to a and setting the result to zero to find the local extrema.[1] Differentiating the integral in (20) we obtain

$$\begin{aligned} \frac{\partial}{\partial a}\left[\int_{-\infty}^{+\infty} (x-a)^2 p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x\right] &= \int_{-\infty}^{+\infty} \frac{\partial}{\partial a}(x-a)^2 p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x \\ &= -2\int_{-\infty}^{+\infty} (x-a) p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x. \end{aligned} \tag{21}$$

[1] Indeed the objective function is convex:

$$\frac{\partial^2}{\partial a^2}\left[\int_{-\infty}^{+\infty} (x-a)^2 p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x\right] = 2\int_{-\infty}^{+\infty} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\, \mathrm{d}x = 2 > 0.$$

Setting (21) to zero at $a = \hat{x}_{\mathrm{BLS}}(\mathbf{y})$ we see that

$$\begin{aligned}\left[\int_{-\infty}^{+\infty}(x-a)p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x\right]\Big|_{a=\hat{x}_{\mathrm{BLS}}(\mathbf{y})} &\\ &= \int_{-\infty}^{+\infty} x\,p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x - \int_{-\infty}^{+\infty}\hat{x}_{\mathrm{BLS}}(\mathbf{y})\,p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x \\ &= \mathbb{E}\left[\mathsf{x}|\mathbf{y}=\mathbf{y}\right] - \hat{x}_{\mathrm{BLS}}(\mathbf{y})\int_{-\infty}^{+\infty} p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})\,\mathrm{d}x \\ &= \mathbb{E}\left[\mathsf{x}|\mathbf{y}=\mathbf{y}\right] - \hat{x}_{\mathrm{BLS}}(\mathbf{y}) = 0. \end{aligned} \tag{22}$$

Hence,

$$\hat{x}_{\mathrm{BLS}}(\mathbf{y}) = \mathbb{E}\left[\mathsf{x}|\mathbf{y}=\mathbf{y}\right], \tag{23}$$

i.e., that the BLS or MMSE estimate of x given $\mathbf{y} = \mathbf{y}$ is the mean of the belief $p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})$.

When $\mathbf{x}$ is a vector, it suffices to note that since the cost criterion (18) is additive [cf.(8)], the minimum is achieved by minimizing the mean-square estimation error in each scalar component. Hence, we obtain

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right], \tag{24}$$

from which we see that the BLS estimate of $\mathbf{x}$ given $\mathbf{y} = \mathbf{y}$ is in general the mean of the belief $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$. □

6.3.1 Performance Characteristics of BLS Estimators

The error

$$\mathbf{e}(\mathbf{x},\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}, \tag{25}$$

of any estimator $\hat{\mathbf{x}}(\mathbf{y})$ can be expressed in the form

$$\mathbf{e}(\mathbf{x},\mathbf{y}) = \mathbf{b} + \left[\mathbf{e}(\mathbf{x},\mathbf{y}) - \mathbf{b}\right], \tag{26}$$

where

$$\mathbf{b} = \mathbb{E}\left[\mathbf{e}(\mathbf{x},\mathbf{y})\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}. \tag{27}$$

is referred to as the (global) bias in the estimator, and where the term in brackets has zero-mean.

Moreover, the mean-square estimation error in such an estimate is the trace of the error correlation matrix $\mathbb{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right]$, which can be expressed via (27) in the form

$$\mathbb{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right] = \mathbf{\Lambda}_{\mathbf{e}} + \mathbf{b}\mathbf{b}^{\mathrm{T}}, \tag{28}$$

where

$$\mathbf{\Lambda}_{\mathbf{e}} = \mathbb{E}\left[(\mathbf{e}(\mathbf{x},\mathbf{y}) - \mathbf{b})(\mathbf{e}(\mathbf{x},\mathbf{y}) - \mathbf{b})^{\mathrm{T}}\right] \tag{29}$$

is the error covariance matrix associated with the estimator. Hence we see that in general, both the bias and covariance contribute to the error correlation and, in turn, mean-square estimation error.

For the case of a Bayes' least-squares estimate specifically, it is straightforward to verify that there is no bias component to the error, as the following claim asserts.

Claim 4. *The BLS estimate is unbiased.*

Proof. Using (27) we have

$$\mathbf{b}_{\text{BLS}} = \mathbb{E}\left[\mathbf{e}(\mathbf{x},\mathbf{y})\right] = \mathbb{E}\left[\hat{\mathbf{x}}_{\text{BLS}}(\mathbf{y}) - \mathbf{x}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\right] - \mathbb{E}\left[\mathbf{x}\right] = \mathbf{0}, \tag{30}$$

where the last equality follows from a simple application of the law of iterated expectation. □

That the BLS estimate is unbiased means that the MSE performance of the estimator is simply the trace of the error covariance matrix $\mathbf{\Lambda}_{\text{BLS}}$. Hence, we have the following performance characterization.

Claim 5. *The error covariance matrix is the expected covariance of the belief, i.e.,*[2]

$$\mathbf{\Lambda}_{\text{BLS}} = \mathbb{E}\left[\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}(\mathbf{y})\right]. \tag{31}$$

Proof. Using (25), (29), and (30) we obtain that the associated error covariance is given by

$$\mathbf{\Lambda}_{\text{BLS}} \triangleq \mathbf{\Lambda}_{\mathbf{e}} = \mathbb{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right] = \mathbb{E}\left[\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\right)\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\right)^{\mathrm{T}}\right], \tag{32}$$

where we emphasize that the notation $\mathbf{\Lambda}_{\text{BLS}}$ is used to refer to the error covariance of the BLS estimator. Applying iterated expectation to (32) we see that the error covariance can be written as

$$\mathbf{\Lambda}_{\text{BLS}} = \mathbb{E}\left[\mathbb{E}\left[\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\right)\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\right)^{\mathrm{T}} \mid \mathbf{y}\right]\right]. \tag{33}$$

However, the inner expectation in (33) is simply the covariance of the belief, i.e., $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}$, whence (31). □

Example 4. Suppose x and w are independent random variables that are both uniformly distributed over the range $[-1, 1]$, and let

$$y = \operatorname{sgn} x + w.$$

Let's determine the BLS estimate of x given y. First we construct the joint density. Note that for $x > 0$, we have

$$p_{y|x}(y|x) = \begin{cases} 1/2 & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

[2]Note that given an observed value of $\mathbf{y}$, this posterior covariance $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}=\mathbf{y}}$ is in general a function of $\mathbf{y}$. As such, we'll sometimes use the alternative notation $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}(\mathbf{y})$ for this covariance.

while for $x < 0$, we have

$$p_{y|x}(y|x) = \begin{cases} 1/2 & -2 < y < 0 \\ 0 & \text{otherwise} \end{cases}.$$

Hence, the joint density is

$$p_{x,y}(x,y) = p_{y|x}(y|x)\, p_x(x) = \begin{cases} 1/4 & 0 < x < 1 \text{ and } 0 < y < 2 \\ 1/4 & -1 < x < 0 \text{ and } -2 < y < 0 \\ 0 & \text{otherwise} \end{cases}$$

and so for $y > 0$ we have

$$p_{x|y}(x|y) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{34a}$$

and for $y < 0$ we have

$$p_{x|y}(x|y) = \begin{cases} 1 & -1 < x < 0 \\ 0 & \text{otherwise} \end{cases}. \tag{34b}$$

Thus from (34) we conclude that

$$\hat{x}_{\text{BLS}}(y) = \mathbb{E}\left[\mathsf{x}|\mathsf{y} = y\right] = \frac{1}{2}\operatorname{sgn} y = \begin{cases} 1/2 & y > 0 \\ -1/2 & y < 0 \end{cases}. \tag{35}$$

Since

$$\lambda_{\mathsf{x}|\mathsf{y}}(y) = 1/12$$

is independent of y in this example, we have that the corresponding error variance is simply

$$\lambda_{\text{BLS}} = \mathbb{E}\left[\lambda_{\mathsf{x}|\mathsf{y}}(\mathsf{y})\right] = 1/12. \tag{36}$$

6.3.2 Orthogonality Characterization of BLS Estimators

Bayes' least-squares estimates are unique in having an important *orthogonality* property. Specifically, we have the following theorem.

Theorem 1. *An estimator $\hat{\mathbf{x}}(\cdot)$ is the Bayes' least-squares estimator, i.e., $\hat{\mathbf{x}}(\cdot) = \hat{\mathbf{x}}_{\text{BLS}}(\cdot)$, if and only if the associated estimation error $\mathbf{e}(\mathbf{x},\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$ is orthogonal to* any *(vector-valued) function $\mathbf{g}(\cdot)$ of the data, i.e.,*

$$\mathbb{E}\left[[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}]\, \mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{0}. \tag{37}$$

Proof. It is convenient to first rewrite (37) as

$$\mathbb{E}\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \tag{38}$$

and note, using the law of iterated expectation, that the left-hand side of (38) can in turn be expressed in the form

$$\mathbb{E}\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y}) \mid \mathbf{y}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \tag{39}$$

Then, to prove the "only if" statement, simply let $\hat{\mathbf{x}}(\cdot) = \hat{\mathbf{x}}_{\mathrm{BLS}}(\cdot)$ in (38), and note that in this case the right-hand expressions in both (39) and (38) are identical, verifying (37).

To prove the converse, let us rewrite (37) using (38) and (39) as

$$\begin{aligned}\mathbf{0} &= \mathbb{E}\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] \\ &= \mathbb{E}\left[\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \end{aligned} \tag{40}$$

Then, since (40) must hold for all $\mathbf{g}(\cdot)$, let us choose $\mathbf{g}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})$ where $\hat{\mathbf{x}}(\cdot)$ is our estimator. In this case (40) becomes

$$\mathbb{E}\left[\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]\left[\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]^{\mathrm{T}}\right] = \mathbf{0},$$

from which we can immediately conclude that $\hat{\mathbf{x}}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]$.[3] □

It is worth emphasizing that Theorem 1 ensures what we would expect of an estimator that yields the minimum mean-square error: that since the error $\mathbf{e}(\mathbf{x}, \mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$ is uncorrelated with *any* function of the data we might construct, there is no further processing that can be done on the data to further reduce the error covariance in the estimate.

[3]Here we are using a straightforward consequence of the Chebyshev inequality—that if $\mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] = 0$ then $\mathbf{z} = 0$, or more precisely, $\mathbb{P}\left(\mathbf{z} = 0\right) = 1$.