7 Linear Least-Squares Estimation

In a given problem of interest, the Bayes' estimator for a parameter vector $\mathbf{x}$ based on data $\mathbf{y}$, even for the case of quadratic cost criterion

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

can be either difficult to determine or difficult to implement, or both. Indeed, determining

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] \tag{2}$$

requires that we have access to the posterior distribution $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$. To compute this posterior via

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{\int_{\mathcal{X}} p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}')\, p_{\mathbf{x}}(\mathbf{x}')\, \mathrm{d}\mathbf{x}'} \tag{3}$$

requires that we have access to a *complete* statistical characterization of the relationship between $\mathbf{x}$ and $\mathbf{y}$, i.e., full knowledge of $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and $p_{\mathbf{x}}(\mathbf{x})$. In practice, such information may be not available or difficult to obtain. Moreover, even when such information is available, evaluating the right-hand side of (3) may be computationally challenging or even intractable depending on the the form of the constituent distributions and/or the dimension of $\mathbf{x}$. Moreover, when it is possible to determine $\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y})$, the result is in general a nonlinear and often complicated function of the data $\mathbf{y}$ that may not be feasible to implement in some applications, such as when there are stringent real-time constraints and or very large data sets.

In these situations, we often must settle for a suboptimal estimator. One way to obtain such an estimator is to constrain the class of the estimators over which we optimize. Here we illustrate this approach with the practically very important example of *linear* estimators. In particular, we develop estimators that minimize the average cost (1), but subject to the additional constraint that the estimator be a linear[1] function of the data. Specifically, using $\hat{\mathbf{x}}_{\mathrm{LLS}}(\cdot)$ to denote this linear least-squares (LLS) estimator, we have:

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\cdot) = \operatorname*{arg\,min}_{\mathbf{f}(\cdot)\in\mathcal{B}} \mathbb{E}\left[\|\mathbf{x} - \mathbf{f}(\mathbf{y})\|^2\right] \tag{4a}$$

where

$$\mathcal{B} = \{\mathbf{f}(\cdot) \ : \ \mathbf{f}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{d} \text{ for some } \mathbf{A} \text{ and } \mathbf{d}\}\,. \tag{4b}$$

[1]Throughout, we follow convention, using the term "linear" to refer to estimators of the form $\mathbf{A}\mathbf{y} + \mathbf{d}$. More accurate terminology would refer to such estimators as "affine" and estimators of the form $\mathbf{A}\mathbf{y}$ as "linear."

As we'll see, this estimator is not only particularly efficient to implement, but it is straightforward to determine and conveniently requires access to only the joint second-order moment statistics of $\mathbf{x}$ and $\mathbf{y}$, which are often particularly easy to obtain.

7.1 Linear Algebra and Second-Order Moment Notation

Linear algebra plays a larger role in the development of linear estimation than in other topics in the course, so we summarize a little useful notation here. First, we use $\mathbf{0}$ to denote a matrix or vector of appropriate dimension whose entries are all zero. Likewise we use $\mathbf{1}$ to denote a matrix or vector whose entries are all unity. By contrast $\mathbf{I}$ denotes a square matrix whose entries on the main diagonal are all one, and whose remaining entries are all zero.

Additionally, since our solutions will be expressed in terms of (first- and) second-order moments, we summarize our notation for such moments. Specifically, for an arbitrary random vector $\mathbf{u}$, we use the mean and (auto-) covariance notation $\boldsymbol{\mu}_{\mathbf{u}} = \mathbb{E}[\mathbf{u}]$ and $\boldsymbol{\Lambda}_{\mathbf{u}} = \mathbb{E}\left[(\mathbf{u}-\boldsymbol{\mu}_{\mathbf{u}})(\mathbf{u}-\boldsymbol{\mu}_{\mathbf{u}})^{\mathrm{T}}\right]$, respectively. Moreover, for an arbitrary pair of random vectors $\mathbf{u}$ and $\mathbf{v}$, we use the (cross-) covariance notation $\boldsymbol{\Lambda}_{\mathbf{uv}} = \mathbb{E}\left[(\mathbf{u}-\boldsymbol{\mu}_{\mathbf{u}})(\mathbf{v}-\boldsymbol{\mu}_{\mathbf{v}})^{\mathrm{T}}\right]$.[2]

7.2 LLS Estimator Characterization

The most insightful derivation of the LLS estimator follows from re-interpreting the problem as an instance of approximation in abstract vector (Hilbert) space. Such a derivation provides a powerful geometric interpretation of the estimator and its properties. While our development will hint at such geometry, in this interests of brevity, we will generally take a shorter path if slightly less illuminating to our main results.

Essentially all our results of interest are obtained rather immediately from the following characterization of the LLS estimator, which states that a defining property of this estimator is that its error is, in an appropriate sense, orthogonal to the data.

Theorem 1 (Orthogonality Principle)**.** *An estimator $\hat{\mathbf{x}}(\mathbf{y})$ for a parameter vector $\mathbf{x}$ based on data $\mathbf{y}$ is the LLS estimator $\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})$ if and only if*

$$\mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x}\right] = \mathbf{0} \tag{5}$$

and

$$\mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\mathbf{y}^{\mathrm{T}}\right] = \mathbf{0}. \tag{6}$$

Proof. First, note that if we establish the theorem for the case of a scalar parameter $\hat{x}(\mathbf{y})$, extending it to the case of a vector parameter

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_K \end{bmatrix}^{\mathrm{T}}$$

[2] Hence, $\boldsymbol{\Lambda}_{\mathbf{u}} = \boldsymbol{\Lambda}_{\mathbf{uu}}$.

is straightforward. Indeed, since the cost criterion (1) is additive, i.e., the errors from the estimation of each element of $\mathbf{x}$ add together, a linear estimator

$$\hat{\mathbf{x}}(\mathbf{y}) = \begin{bmatrix} \hat{x}_1(\mathbf{y}) & \hat{x}_2(\mathbf{y}) & \dots & \hat{x}_K(\mathbf{y}) \end{bmatrix}^{\mathrm{T}},$$

is the LLS estimator if and only if each of its elements $\hat{x}_i(\mathbf{y})$ is the LLS estimator for x_i, which is the case if and only if (5) and (6) are satisfied for each $\hat{x}_i(\mathbf{y})$, i.e.,

$$\mathbb{E}\left[\hat{x}_i(\mathbf{y}) - \mathbf{x}_i\right] = 0 \quad \text{and} \quad \mathbb{E}\left[(\hat{x}_i(\mathbf{y}) - \mathbf{x}_i)\mathbf{y}^{\mathrm{T}}\right] = \mathbf{0}, \qquad i = 1, \dots, K.$$

Stacking these constraints column-wise thus yields the necessary and sufficient conditions (5) and (6).

Thus it remains only to establish the theorem holds for a scalar parameter x. We begin by establishing the "only if" part. First, suppose (5) is not satisfied by the estimator $\hat{x}(\cdot)$, i.e.,

$$\mathbb{E}\left[\hat{x}(\mathbf{y}) - x\right] = b \neq 0. \tag{7}$$

Then we can construct an estimator with lower mean-square error. For example, the linear estimator $\hat{x}'(\mathbf{y}) = \hat{x}(\mathbf{y}) - b$ has a lower mean-square error. Specifically,

$$\begin{aligned} \mathbb{E}\left[(\hat{x}'(\mathbf{y}) - x)^2\right] &= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x - b)^2\right] && (8)\\ &= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right] - b^2 && (9)\\ &< \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right], && (10) \end{aligned}$$

where to obtain (9) and (10) we have used (7). Thus, $\hat{x}(\cdot)$ cannot be the LLS estimator.

Next, suppose that (6) is not satisfied by the estimator $\hat{x}(\cdot)$. Then we can also construct another linear estimator with lower mean-square error. In particular, the estimator

$$\hat{x}'(\mathbf{y}) = \hat{x}(\mathbf{y}) - \mathbf{\Lambda}_{e\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})$$

with

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

has mean-square estimator error

$$\begin{aligned} \mathbb{E}\left[(\hat{x}'(\mathbf{y}) - x)^2\right] &= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x - \mathbf{\Lambda}_{e\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}))^2\right]\\ &= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right] - \mathbf{\Lambda}_{e\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{e\mathbf{y}}^{\mathrm{T}}\\ &< \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right], \end{aligned}$$

so $\hat{x}(\cdot)$ cannot be the LLS estimator. Hence, both (5) and (6) must be satisfied by the LLS estimator.

We now establish the "if" part. Suppose that (5) and (6) are satisfied by estimator $\hat{x}(\cdot)$. We show that any other estimator $\hat{x}'(\cdot)$ must have a larger mean-square error. Since our "only if" statement established that unbiased estimators are better than

biased estimators, it suffices to restrict our attention to unbiased $\hat{x}'(\cdot)$. In such cases, (6) implies

$$\mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)(\hat{x}'(\mathbf{y}) - \hat{x}(\mathbf{y}))\right] = 0. \tag{11}$$

since $\hat{x}'(\mathbf{y}) - \hat{x}(\mathbf{y})$ is a zero-mean linear function of $\mathbf{y}$. But then

$$\begin{aligned}
\mathbb{E}\left[(\hat{x}'(\mathbf{y}) - x)^2\right] &= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x) + (\hat{x}'(\mathbf{y}) - \hat{x}(\mathbf{y}))^2\right] && (12)\\
&= \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right] + \mathbb{E}\left[(\hat{x}'(\mathbf{y}) - \hat{x}(\mathbf{y}))^2\right] && (13)\\
&> \mathbb{E}\left[(\hat{x}(\mathbf{y}) - x)^2\right], && (14)
\end{aligned}$$

where to obtain (13) we have used (11). □

Geometrically, Theorem 1 can be interpreted as saying that an estimate is optimum if and only if it is a particular kind of projection of the parameter onto the space spanned by the data, in which case the error is orthogonal to the data in a particular sense. And while further discussion of such geometry is beyond the scope of our (abbreviated) treatment, our calculations in the sequel are guided by this geometry.

Proceeding, we next note that Theorem 1 can be equivalently expressed as follows.

Corollary 1. *A linear estimator* $\hat{\mathbf{x}}(\mathbf{y})$ *is the LLS estimator if and only if for all* $\mathbf{F}$ *and* $\mathbf{g}$ *of appropriate dimension*

$$\mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x})(\mathbf{F}\mathbf{y} + \mathbf{g})^{\mathrm{T}}\right] = \mathbf{0}. \tag{15}$$

Note that in this form we see that the LLS estimation error is orthogonal to all linear functions of the data, a collection that includes every linear estimator, a fact which is frequently useful.

Proof. For the "if" part, choosing $(\mathbf{F}, \mathbf{g}) = (\mathbf{0}, \mathbf{1})$ we obtain (5), and choosing $(\mathbf{F}, \mathbf{g}) = (\mathbf{I}, \mathbf{0})$ we obtain (6).

For the "only if" part, it suffices to rewrite the left-hand side of (15) as

$$\mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x})\mathbf{y}^{\mathrm{T}}\right]\mathbf{F}^{\mathrm{T}} + \mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x})\right]\mathbf{g}^{\mathrm{T}}$$

and note that each of the first and second expectations are zero via (6) and (5), respectively. □

Even more importantly, from the characterization Theorem 1, we immediately obtain the following explicit expression for the LLS estimator.

Corollary 2. *The LLS estimator can be expressed in the form*

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Lambda}_{\mathbf{xy}}\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}). \tag{16}$$

Note that, as we indicated would be the case, the LLS estimator depends only on the second-order moments of the joint distribution for x and y. Thus, the LLS estimator neither requires nor can exploit any further information about their statistical relationship. For example, note that if x and y are uncorrelated ($\mathbf{\Lambda}_{\mathsf{xy}} = \mathbf{0}$), the data is of no value. By contrast, the BLS would in general be able to make use of the data in such scenarios, as it can exploit higher-order statistical dependencies.

Proof. We start by substituting

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{d}$$

into (5), obtaining

$$\mathbf{A}\boldsymbol{\mu}_{\mathsf{y}} + \mathbf{d} - \boldsymbol{\mu}_{\mathsf{x}} = \mathbf{0},$$

so

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mathbf{A}(\mathbf{y} - \boldsymbol{\mu}_{\mathsf{y}}) + \boldsymbol{\mu}_{\mathsf{x}}, \tag{17}$$

from which we see

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x} = \mathbf{A}(\mathbf{y} - \boldsymbol{\mu}_{\mathsf{y}}) - (\mathbf{x} - \boldsymbol{\mu}_{\mathsf{x}}). \tag{18}$$

Next, due to (5), we can rewrite (6) as

$$\mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x})(\mathbf{y} - \boldsymbol{\mu}_{\mathsf{y}})^{\mathrm{T}}\right] = \mathbf{0}, \tag{19}$$

which when we substitute (18) into (19) yields

$$\mathbb{E}\left[(\mathbf{A}(\mathbf{y} - \boldsymbol{\mu}_{\mathsf{y}}) - (\mathbf{x} - \boldsymbol{\mu}_{\mathsf{x}}))(\mathbf{y} - \boldsymbol{\mu}_{\mathsf{y}})^{\mathrm{T}}\right] = \mathbf{A}\mathbf{\Lambda}_{\mathsf{y}} - \mathbf{\Lambda}_{\mathsf{xy}} = \mathbf{0}. \tag{20}$$

Solving (20) for $\mathbf{A}$ yields $\mathbf{A} = \mathbf{\Lambda}_{\mathsf{xy}}\mathbf{\Lambda}_{\mathsf{y}}^{-1}$, which when substituted into (17) yields (16). □

7.3 LLS Estimator Performance

The mean-square estimation error $\mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})\|^2\right]$ in the LLS estimator is the trace of the estimation error covariance

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

which is similarly straightforward to obtain from Theorem 1.

Corollary 3. *The estimation error covariance for the LLS estimator is*

$$\mathbf{\Lambda}_{\mathrm{LLS}} = \mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}))(\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}))^{\mathrm{T}}\right] = \mathbf{\Lambda}_{\mathsf{x}} - \mathbf{\Lambda}_{\mathsf{xy}}\mathbf{\Lambda}_{\mathsf{y}}^{-1}\mathbf{\Lambda}_{\mathsf{xy}}^{\mathrm{T}}. \tag{21}$$

Proof. We begin by writing

$$\begin{aligned}\boldsymbol{\Lambda}_{\mathrm{LLS}} &= \mathbb{E}\left[(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}))(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}))^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}))^{\mathrm{T}}\right] \qquad (22) \\ &= \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})((\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})-(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})-\boldsymbol{\mu}_{\mathbf{x}}))^{\mathrm{T}}\right] \\ &= \boldsymbol{\Lambda}_{\mathbf{x}} - \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})((\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})-\boldsymbol{\mu}_{\mathbf{x}}))^{\mathrm{T}}\right], \qquad (23)\end{aligned}$$

where to obtain (22) we have used both (5) and (6). Replacing the first factor of the second term in (23) with

$$\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}} = (\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})-\boldsymbol{\mu}_{\mathbf{x}}) - (\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})-\mathbf{x}),$$

and again applying (5) and (6) then yields

$$\boldsymbol{\Lambda}_{\mathrm{LLS}} = \boldsymbol{\Lambda}_{\mathbf{x}} - \boldsymbol{\Lambda}_{\hat{\mathbf{x}}(\mathbf{y})}. \qquad (24)$$

Finally, evaluating the second term in (24) via

$$\hat{\mathbf{x}}(\mathbf{y})-\boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\Lambda}_{\mathbf{xy}}\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}),$$

we obtain (21). □

We conclude the section with a simple example.

Example 1. Let's return to the random variables x and y from Example 4 in the last instalment of the notes (Bayesian Estimation), now finding the LLS estimator for x based on y. Recall that $y = \operatorname{sgn} x + w$, where x and w are both uniformly distributed on $[-1, 1]$ and independent.

First we note that by symmetry

$$\mu_x = \mu_y = 0. \qquad (25a)$$

Furthermore, since x and w are independent we have

$$\lambda_{xy} = \mathbb{E}[xy] = \mathbb{E}[x(\operatorname{sgn} x + w)] = \mathbb{E}[|x|] = \frac{1}{2}, \qquad (25b)$$

and

$$\lambda_y = \mathbb{E}\left[y^2\right] = \mathbb{E}\left[(\operatorname{sgn} x + w)^2\right] = \mathbb{E}\left[\operatorname{sgn}^2 x\right] + \mathbb{E}\left[w^2\right] = 1 + \frac{1}{3} = \frac{4}{3}. \qquad (25c)$$

Substituting (25) into (16) and (21) we obtain

$$\hat{x}_{\mathrm{LLS}}(y) = \frac{\lambda_{xy}}{\lambda_y} y = \frac{3}{8} y \qquad (26)$$

and

$$\lambda_{\mathrm{LLS}} = \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} = \frac{1}{3} - \frac{(1/2)^2}{(4/3)} = \frac{7}{48}. \qquad (27)$$

Comparing (27) with the performance of the BLS estimator, which we calculated to be $1/12$, we see that, as expected, constraining our estimator to be linear leads to a larger mean-square estimation error.

7.4 Jointly Gaussian Random Variables

The topic of linear least-squares estimation is intimately connected with that of jointly Gaussian distributions, as we will develop in the sequel. In preparation, we summarize the aspects of such distributions that will be required to appreciate our development. To keep our treatment concise, we omit proofs in this section.

We start with the familiar definition of a scalar Gaussian random variable.

Definition 1. *A random variable x is Gaussian if it has a distribution of the form*

$$p_{\mathsf{x}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

for some mean and variance parameters μ and $\sigma^2 > 0$, respectively. Additionally, we say that a (deterministic) constant is also, trivially, a Gaussian random variable.

Note that it is convenient to include the case of a constant in the definition, as this corresponds to the limit $\sigma \to 0$, i.e., constants are Gaussian random variables of zero variance.

We extend this definition to the multivariate case as follows.

Definition 2. *A random vector $\mathbf{x}$ is Gaussian if every linear combination of its elements, i.e., $\mathsf{z} = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ for every deterministic $\mathbf{a}$, is a Gaussian random variable.*

As additional terminology, we say the elements of a Gaussian random vector are *jointly Gaussian*.

It can be verified that Definition 2 implies that a Gaussian random vector must have a distribution of the following particular form. Although we will not discuss it further, a convenient proof can be constructed by working with characteristic functions.

Theorem 2. *A random vector $\mathbf{x}$ is Gaussian if and only if it has a distribution of the form*[3]

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{|2\pi\boldsymbol{\Lambda}|^{1/2}}. \tag{28}$$

for some parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda} \geq 0$.[4]

As one of many consequences of (28) in Theorem 2, note that a collection of jointly Gaussian random variables is independent if and only if they are uncorrelated (i.e., $\boldsymbol{\Lambda}$ is a diagonal matrix). Yet another consequence is that the elements of a Gaussian

[3] If $\boldsymbol{\Lambda}$ is singular (noninvertible) (28) must be interpreted more delicately. While we omit the details, such scenarios occur when some of the constituent random variables can be expressed as a deterministic linear combination of the others.

[4] The inequality inequality is to be interpreted in the sense of positive semidefiniteness. In essence, this condition ensures that the resulting distribution is nonnegative.

random vector are Gaussian. However, it is important to appreciate that Definition 2 imposes strong constraints on the relationship between variables in order that they be jointly Gaussian. In particular, two individually Gaussian random variables are, in general, not jointly Gaussian.

Finally, our definitions extend to multiple random vectors as follows.

Definition 3. *Random vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_N$ *are jointly Gaussian if*

$$\mathbf{z} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$

is a Gaussian random vector.

Using (3) the form of the distribution for $\mathbf{z}$ follows directly from Theorem 2.

7.5 Estimation in the Jointly Gaussian Case

In this section, we establish the following remarkable result.

Theorem 3. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are jointly Gaussian random vectors, then*

$$\hat{\mathbf{x}}_{\text{BLS}}(\mathbf{y}) = \hat{\mathbf{x}}_{\text{LLS}}(\mathbf{y}). \tag{29}$$

Proof. Let

$$\mathbf{e}_{\text{LLS}} = \hat{\mathbf{x}}_{\text{LLS}}(\mathbf{y}) - \mathbf{x}, \tag{30}$$

and note that by Theorem 1 we have that $\mathbf{e}_{\text{LLS}}$ must be orthogonal to every linear function of $\mathbf{y}$ and hence $\mathbf{y}$ itself. But since $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, this means that $\mathbf{e}$ is actually statistically independent of $\mathbf{y}$. This implies, for example, that

$$\mathbb{E}\left[\mathbf{e}_{\text{LLS}}|\mathbf{y}\right] = \mathbb{E}\left[\mathbf{e}_{\text{LLS}}\right] = 0 \tag{31}$$

where the last equality follows from the fact that the LLS estimate is unbiased. But we also have directly from (30) and from (2) that

$$\mathbb{E}\left[\mathbf{e}_{\text{LLS}}|\mathbf{y}\right] = \mathbb{E}\left[\hat{\mathbf{x}}_{\text{LLS}}(\mathbf{y})|\mathbf{y}\right] - \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] = \hat{\mathbf{x}}_{\text{LLS}}(\mathbf{y}) - \hat{\mathbf{x}}_{\text{BLS}}(\mathbf{y}). \tag{32}$$

Comparing (31) and (32) establishes our result. □

Note that one implication of this result is that there are no statistical dependencies between $\mathbf{x}$ and $\mathbf{y}$ beyond those described by second-order moments ones that the BLS can exploit in the jointly Gaussian case.

7.6 BLS Estimation with only Second-Order Moments

In the preceding development, we have seen that construction of the BLS estimator of a random variable x from a random vector $\mathbf{y}$ subject to the constraint that the estimator be linear requires only knowledge of the joint second-order moment properties of $(x, \mathbf{y})$.

This observation raises a natural question. Suppose we only have knowledge of the joint second-order moment properties of a pair $(x, \mathbf{y})$, then what is the best possible estimator $\hat{x}(\mathbf{y})$ (in a BLS sense) we can construct, and how does it perform? In particular, we might reasonably ask whether the LLS estimator is also the solution to this problem.

To answer this question requires a minimax formulation, i.e., posing the problem as a game between two adversaries: the system designer tries to find the best estimator, and nature tries to find the model that makes the performance of the chosen estimator as bad as possible subject to the constraint that the $(x, \mathbf{y})$ statistics match the prescribed second-order moment information.

For this game, we now determine the best estimator choice for the system designer, the worst joint distribution we can encounter, and the resulting mean-square estimator error. With the given moments being μ_x, $\boldsymbol{\mu}_\mathbf{y}$, σ_x^2, σ_y^2, and λ_{xy}, we seek to evaluate

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right], \tag{33}$$

where

$$\mathcal{M} = \left\{ p_{x\mathbf{y}} : \quad \mathbb{E}\left[\begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}\right] = \begin{bmatrix} \mu_x \\ \boldsymbol{\mu}_\mathbf{y} \end{bmatrix}, \quad \operatorname{cov}\left(\begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}\right) = \begin{bmatrix} \sigma_x^2 & \boldsymbol{\Lambda}_{x\mathbf{y}} \\ \boldsymbol{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} & \boldsymbol{\Lambda}_{\mathbf{y}\mathbf{y}} \end{bmatrix} \right\}. \tag{34}$$

We first recall from our development of minimax hypothesis testing that, in general, a max-min can be upper bounded by a min-max. For our problem, this means that

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right] \leq \min_{f(\cdot)} \max_{p_{x\mathbf{y}} \in \mathcal{M}} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right]. \tag{35}$$

Intuitively, the minimizer is more powerful on the left side of (35) while the maximizer is more powerful on the right side. Indeed, the minimizer on the left side of (35) gets to choose an estimating function f that depends on the distribution chosen by the maximizer, while the opposite is true for the right side.

We can further upper bound the right side of (35) by substituting any function f_0. That is,

$$\min_{f(\cdot)} \max_{p_{x\mathbf{y}} \in \mathcal{M}} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right] \leq \max_{p_{x\mathbf{y}} \in \mathcal{M}} \mathbb{E}\left[(x - f_0(\mathbf{y}))^2\right]. \tag{36}$$

For any linear function $f_0(y) = d + \mathbf{a}^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})$ the right side of (36) only depends on the second-order moment statistics, which are fixed. Let us further choose $d = \mu_x$ and $\mathbf{a}^{\mathrm{T}} = \boldsymbol{\Lambda}_{x\mathbf{y}}\boldsymbol{\Lambda}_\mathbf{y}^{-1}$, which results in

$$\mathbb{E}\left[(x - f_0(\mathbf{y}))^2\right] = \mathbb{E}\left[\left(x - \mu_x - \boldsymbol{\Lambda}_{x\mathbf{y}}\boldsymbol{\Lambda}_\mathbf{y}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})\right)^2\right] = \sigma_x^2 - \boldsymbol{\Lambda}_{x\mathbf{y}}\boldsymbol{\Lambda}_\mathbf{y}^{-1}\boldsymbol{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} \tag{37}$$

for any $p_{x\mathbf{y}} \in \mathcal{M}$. Thus, the right side of (37) is an upper bound to (33).

We now proceed to show that the right side of (37) is also a lower bound to (33). Similarly to our upper bound in (36), we can lower bound (33) by choosing any distribution on x and y, i.e.,

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right] \geq \min_{f(\cdot)} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right] \tag{38}$$

where the expectation on the righthand side of (38) is with respect to an arbitrary distribution $p^*_{x\mathbf{y}} \in \mathcal{M}$. Let us choose $p^*_{x\mathbf{y}}$ as that corresponding to x and y being jointly Gaussian with the specified second-order moment statistics. For jointly Gaussian random variables, the BLS estimate is the linear estimate $\hat{x} = \mu_x + \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})$. The resulting mean square error is that given on the right side of (37).

Since (33) is both upper and lower bounded by the right-hand side of (37), we conclude that 1) the system designer should choose the LLS estimator, 2) nature should choose the jointly Gaussian model matching the second-order moment statistics, and 3) the resulting mean-square error performance will be $\sigma_x^2 - \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}}$.