8 NonBayesian Parameter Estimation

In many estimation problems, it may be unnatural to assign a prior distribution to the latent variable of interest, and thus the Bayesian framework cannot be used. In such cases, an alternative is to treat the variable as deterministic, but unknown. In such cases, the observation model is not a distribution for $\mathbf{y}$ *conditioned* on the latent variable, but rather a distribution for $\mathbf{y}$ that is *parameterized* by this variable.

As an example, suppose we have a sequence of independent identically distributed Gaussian random variables $y_1, y_2, \ldots, y_N$, where the mean μ and variance σ^2 that parameterize the density are unknown. In this section we describe some classical approaches to the problem of developing good estimators for such nonrandom parameters.

In our treatment, we use $\mathbf{x}$ to denote the vector of parameters we seek to estimate, and write the density for the vector of observations $\mathbf{y}$ as $p_{\mathbf{y}}(\mathbf{y}; \mathbf{x})$ so as to make the parameterization explicit. In addition, we will use $\boldsymbol{\mu}_{\mathbf{y}}(\mathbf{x})$ and $\boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{x})$ to denote, respectively, the mean vector and covariance matrix of $\mathbf{y}$, again to make the parameterization explicit.

Throughout this section, we focus on mean-square estimation error as the performance measure of interest. We begin by noting that the Bayesian (least-squares) framework developed earlier cannot be immediately adapted to handle nonrandom parameter estimation. To see this, consider the case of a scalar parameter x. If we attempt to construct a minimum mean-square error estimate $\hat{x}(\mathbf{y})$ via

$$\hat{x}(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}\left[(x - f(\mathbf{y}))^2\right], \tag{1}$$

we encounter a difficulty. In particular, since the expectation in (1) is over $\mathbf{y}$ alone (since x is deterministic), we immediately obtain that the right-hand side of (1) is minimized by choosing $\hat{x}(\mathbf{y}) = x$, and hence the optimum estimator according to (1) depends on the very parameter we're trying to estimate!

This observation reveals an important insight on nonrandom parameter estimation: in any meaningful formulation of such problems, we need to explicitly restrict our search to estimators that don't depend explicitly on the parameters we're trying to estimate.

Definition 1. *An estimator is* valid *if it does not depend explicitly on the parameters being estimated.*

In the sequel, we describe some traditional approaches to finding valid estimators that yield good mean-square error performance.

8.1 Bias and Error Covariance

As in the case of random parameters, two important quantities that impact the mean-square error performance of an estimator for nonrandom parameters are the bias and error covariance. However, there are some significant distinctions between these quantities in the nonrandom case, which we emphasize in this section.

Using

$$\mathbf{e}(\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} = \hat{\mathbf{x}} - \mathbf{x} \tag{2}$$

as our notation for the error, we define the bias in an estimator $\hat{\mathbf{x}}(\cdot)$ as

$$\mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x}) = \mathbb{E}\left[\mathbf{e}(\mathbf{y})\right] = \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] = \left[\int_{-\infty}^{+\infty} \hat{\mathbf{x}}(\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\, \mathrm{d}\mathbf{y}\right] - \mathbf{x}. \tag{3}$$

Likewise, we express the error covariance as

$$\mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) = \mathbb{E}\left[\left[\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\right]\left[\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\right]^{\mathrm{T}}\right], \tag{4}$$

where, again, the expectation is with respect to $\mathbf{y}$. Not surprisingly, both the bias (3) and error covariance (4) are, in general, functions of the parameter $\mathbf{x}$.

The mean-square estimation error is the trace of the error correlation matrix $\mathbb{E}\left[\mathbf{e}(\mathbf{y})\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right]$, and this matrix in general depends on both bias and error covariance; specifically

$$\mathbb{E}\left[\mathbf{e}(\mathbf{y})\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) + \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\, \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})^{\mathrm{T}}. \tag{5}$$

In seeking to minimize mean-square estimation error, we have no particular preference for relative contributions from bias or error covariance terms. We simply want the combination to be as small as possible. However, in practice this has proven to be a difficult optimization. As a result, it has become common to resort to suboptimum estimators that are the solution to minimum mean-square error problems with additional (and somewhat arbitrary) constraints.

Perhaps the best known example of such a constrained optimization involves explicitly restricting the search for estimators to those that are valid and unbiased.

Definition 2. *An estimator* $\hat{\mathbf{x}}(\cdot)$ *for a nonrandom parameter* $\mathbf{x}$ *is unbiased if* $\mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x}) = 0$ *for all possible values of* $\mathbf{x}$.

This is the notion underlying well-known *minimum-variance unbiased estimators*, which we discuss next.

As one final comment before proceeding, note that in contrast to the case of random parameters, for nonrandom parameter estimators we have that the error covariance is the same as the covariance of the estimator itself, i.e.,

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

To see this, simply note that using (2)–(4), we have

$$\begin{aligned}
\mathbf{\Lambda_e}(\mathbf{x}) &= \mathbb{E}\left[\big(\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\big)\big(\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\big)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\big((\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}) - (\mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right] - \mathbf{x})\big)\big((\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}) - (\mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right] - \mathbf{x})\big)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\big(\hat{\mathbf{x}}(\mathbf{y}) - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right]\big)\big(\hat{\mathbf{x}}(\mathbf{y}) - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right]\big)^{\mathrm{T}}\right] \\
&= \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}).
\end{aligned}$$

In the remainder of our discussion, we restrict our attention to the case where the parameter to be estimated is a scalar x. As in the case of Bayesian estimation, vector parameter extensions with the mean-square error criterion can be constructed in a component-wise manner. At the end of the section we provide some additional insight into the vector parameter case.

8.2 Minimum-Variance Unbiased Estimators

We begin by defining the admissible set of estimators.

Definition 3. *An admissible estimator is one that is both valid (i.e., does not depend on x) and unbiased. We use*

$$\mathcal{A} = \{\hat{x}(\cdot) \,:\, \hat{x}(\cdot) \text{ is valid and } b_{\hat{x}}(x) = 0\}$$

to denote the set of admissible estimators.

In turn, when it exists, a minimum-variance unbiased (MVU) estimator for x is defined to be the admissible estimator with the smallest variance, i.e.,

$$\hat{x}_{\mathrm{MVU}}(\cdot) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}(x) \qquad \text{for all } x \tag{6}$$

Several observations regarding (6) are worth emphasizing. The first is that $\hat{x}_{\mathrm{MVU}}(\cdot)$ may not exist! For example, for some problems the set $\mathcal{A}$ is empty—there are no valid unbiased estimators. In other cases, $\mathcal{A}$ is not empty, but no estimator in $\mathcal{A}$ has a uniformly smaller variance than all the others, i.e., *for all values of the parameter* x. Suppose for example that $\mathcal{A}$ consists of three hypothetical estimators $\hat{x}_1(\cdot)$, $\hat{x}_2(\cdot)$, and $\hat{x}_3(\cdot)$, whose variances are plotted as a function of the unknown parameter x in Fig. 1. In this case, there is no estimator having a smaller variance than all the others for all values of x.

It should also be emphasized that even when $\hat{x}_{\mathrm{MVU}}(\cdot)$ does exist, it may be difficult to find. In fact in general there is no systematic procedure for either determining whether an MVU estimator exists, or for computing it when it is does exist. However, there are cases in which such estimators can be computed, as we'll discuss. To this end, it is sometimes useful to exploit a bound on $\lambda_{\hat{x}}(x)$ in the pursuit of MVU estimators. A celebrated bound for this purpose is the Cramér-Rao bound, as we develop next.

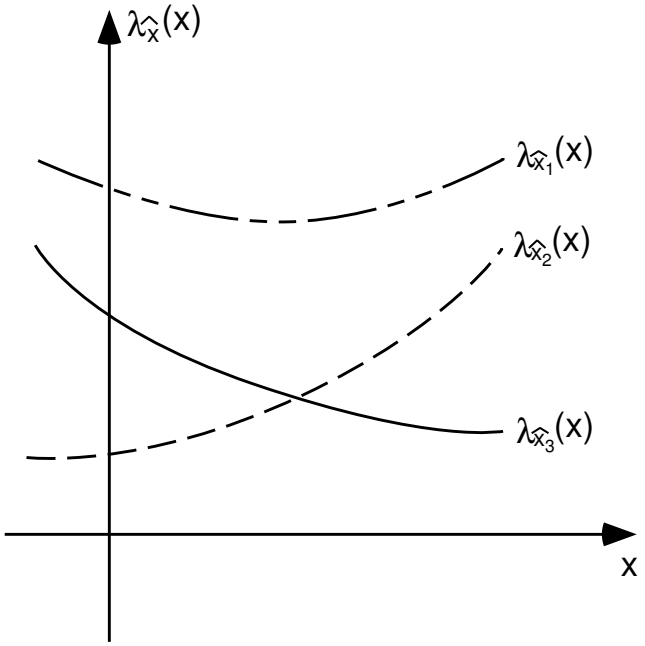

Figure 1: The variances of three hypothetical unbiased estimators.

8.3 The Cramér-Rao Bound

When it exists, the Cramér-Rao bound gives a lower bound on the variance of *any* admissible estimator $\hat{x}(\cdot)$ for x. In particular, we have the following.

Theorem 1 (Cramér-Rao bound, scalar version)**.** *Provided $p_{\mathbf{y}}(\mathbf{y}; x)$ satisfies the regularity condition*

$$\mathbb{E}\left[\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y}; x)\right] = 0 \qquad \textit{for all } x, \tag{7}$$

we have, for any $\hat{x}(\cdot)$ satisfying Definition 3,

$$\lambda_{\hat{x}}(x) \geq \frac{1}{J_{\mathbf{y}}(x)}, \tag{8}$$

where

$$J_{\mathbf{y}}(x) \triangleq \mathbb{E}\left[(S_{\mathbf{y}}(x))^2\right] \tag{9}$$

is referred to as the Fisher information *in* $\mathbf{y}$ *about x, and with*

$$S_{\mathbf{y}}(x) \triangleq \frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y}; x), \tag{10}$$

which is referred to as the score function *for x based on* $\mathbf{y}$.

Proof. To derive the Cramér-Rao bound (8), we begin by recalling that for unbiased estimators the error

$$e(\mathbf{y}) = \hat{x}(\mathbf{y}) - x \tag{11}$$

has zero mean, i.e.,

$$\mathbb{E}\left[e(\mathbf{y})\right] = 0, \tag{12}$$

and variance

$$\operatorname{var} e(\mathbf{y}) = \mathbb{E}\left[e^2(\mathbf{y})\right] = \lambda_{\hat{x}}(x). \tag{13}$$

Next we define

$$f(\mathbf{y}) = S_{\mathbf{y}}(x) \tag{14}$$

and note that using the identity

$$\frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x) = \frac{1}{p_{\mathbf{y}}(\mathbf{y}; x)} \frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x), \tag{15}$$

we get that $f(\mathbf{y})$ has zero mean:

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{y})\right] &= \mathbb{E}\left[\frac{1}{p_{\mathbf{y}}(\mathbf{y}; x)} \frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x)\right] \\ &= \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x)\, \mathrm{d}\mathbf{y} \\ &= \frac{\partial}{\partial x} \int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y}; x)\, \mathrm{d}\mathbf{y} = \frac{\partial}{\partial x} 1 = 0, \end{aligned} \tag{16}$$

and, in turn, variance

$$\operatorname{var} f(\mathbf{y}) = \mathbb{E}\left[f^2(\mathbf{y})\right] = J_{\mathbf{y}}(x). \tag{17}$$

Finally, again using the identity (15), the covariance between $e(\mathbf{y})$ and $f(\mathbf{y})$ is given by

$$\begin{aligned} \operatorname{cov}\left(e(\mathbf{y}), f(\mathbf{y})\right) &= \mathbb{E}\left[e(\mathbf{y}) f(\mathbf{y})\right] \\ &= \int_{-\infty}^{+\infty} (\hat{x}(\mathbf{y}) - x) \frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x)\, \mathrm{d}\mathbf{y} \\ &= \left[\frac{\partial}{\partial x} \int_{-\infty}^{+\infty} \hat{x}(\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y}; x)\, \mathrm{d}\mathbf{y}\right] - \left[x \frac{\partial}{\partial x} \int_{-\infty}^{+\infty} p_y(\mathbf{y}; x)\, \mathrm{d}\mathbf{y}\right] \\ &= 1 - 0 = +1. \end{aligned} \tag{18}$$

Now recall that correlation coefficients have at most unit magnitude; hence

$$\rho_{ef}^2 = \frac{\left[\operatorname{cov}\left(e(\mathbf{y}), f(\mathbf{y})\right)\right]^2}{\operatorname{var} e(\mathbf{y}) \operatorname{var} f(\mathbf{y})} \leq 1. \tag{19}$$

Finally, substituting (13), (17) and (18) into (19) we obtain (8). □

In addition, we note that Fisher information is often convenient to express as follows.

Corollary 1. *The Fisher information* (9) *can be equivalently expressed in the form*

$$J_{\mathbf{y}}(x) = -\,\mathbb{E}\left[\frac{\partial^2}{\partial x^2}\ln p_{\mathbf{y}}(\mathbf{y};x)\right]. \tag{20}$$

Proof. To verify (20), we begin by observing

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y} = 1. \tag{21}$$

Differentiating (21) with respect to x and using the identity (15) yields

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y} = 0. \tag{22}$$

Finally, differentiating (22) once more with respect to x and again using (15) we obtain

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\left[\frac{\partial^2}{\partial x^2}\ln p_{\mathbf{y}}(\mathbf{y};x)\right]\,\mathrm{d}\mathbf{y} + \int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\left[\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x)\right]^2\,\mathrm{d}\mathbf{y} = 0. \tag{23}$$

□

Some remarks on the Cramér-Rao bound:

1. Eq. (9) is the first of many interrelated information measures we will encounter in the subject. Some broader perspectives will come later.

2. We emphasize that the Fisher information cannot be computed in all problems, i.e., the regularity condition may not be satisfied, in which case no Cramér-Rao bound exists. For example, for densities such as

$$p_y(y;x) = \begin{cases} 1 & x < y < x+1 \\ 0 & \text{otherwise} \end{cases},$$

   which are not strictly positive for all x and y, the logarithm in (9) doesn't exist and hence $J_{\mathbf{y}}(x)$ cannot be calculated.

3. The Fisher information (9) can be interpreted as a measure of curvature: it measures, on average, how "peaky" $\ln p_{\mathbf{y}}(\mathbf{y};x)$ is as a function of x. As such, the larger $J_{\mathbf{y}}(x)$, the better we expect to be able to resolve the value of x from the observations, and hence the smaller we expect $\lambda_{\hat{x}}(x)$ to be. Our later development of information geometry will shed additional light on this behavior.

4. Note that in our derivation we have interchanged the order of integration and differentiation several times. As such we have implicitly made use of the regularity condition (7).

5. Any estimator that satisfies the Cramér-Rao bound with equality must be a MVU estimator. Note however, that the converse is not true: the Cramér-Rao bound may not be tight. Sometimes no estimator can meet the bound for all x, or even for any x!

6. The Cramér-Rao bound can be generalized in a variety of ways, as described in, e.g., Van Trees estimation text. For example, a Cramér-Rao bound can be constructed for *biased* estimates. However, in practice this bound is not particularly useful. Likewise, there is an analogous bound for random parameters, which can be more tractable than the tight bound on the error variance of random parameter estimates we have already, viz.,

$$\operatorname{var} e(\mathsf{x}, \mathbf{y}) = \operatorname{var}\left[\hat{x}(\mathbf{y}) - \mathsf{x}\right] \geq \mathbb{E}\left[\lambda_{\mathsf{x}|\mathbf{y}}(\mathbf{y})\right]$$

with equality if and only if $\hat{x}(\mathbf{y}) = \hat{x}_{\mathrm{BLS}}(\mathbf{y}) = \mathbb{E}\left[\mathsf{x}|\mathbf{y}\right]$.

We conclude the section with a simple example.

Example 1. Consider the scalar Gaussian problem

$$\mathsf{y} = x + \mathsf{w},$$

where $\mathsf{w} \sim \mathtt{N}(0, \sigma^2)$. To determine the Cramér-Rao bound, we first calculate

$$\ln p_{\mathsf{y}}(y; x) = -\frac{1}{2\sigma^2}(x - \mathsf{y})^2 - \frac{1}{2}\ln(2\pi\sigma^2), \tag{24}$$

from which we obtain

$$\frac{\partial}{\partial x}\ln p_{\mathsf{y}}(y; x) = -\frac{1}{\sigma^2}(x - \mathsf{y}) = \frac{1}{\sigma^2}\mathsf{w}.$$

Hence, the Fisher information is

$$J_{\mathsf{y}}(x) = \frac{1}{\sigma^4}\,\mathbb{E}\left[\mathsf{w}^2\right] = \frac{1}{\sigma^2}, \tag{25}$$

and, thus, the variance of any unbiased estimator satisfies

$$\lambda_{\hat{\mathsf{x}}}(x) \geq \sigma^2. \tag{26}$$

Moreover, we see that the smaller the variance σ^2 the sharper the peak of (24) is as a function of x.

8.3.1 Efficiency

We next examine when the Cramér-Rao bound is satisfied with equality.

Definition 4. *An unbiased estimator is* efficient *if it satisfies the Cramér-Rao bound* (8) *with equality.*

We then have the following.

Corollary 2. *An estimator $\hat{x}(\cdot)$ is efficient if and only if it can be expressed in the form*

$$\hat{x}(\mathbf{y}) = x + \frac{1}{J_{\mathbf{y}}(x)} \frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x) \tag{27}$$

where the right-hand side must be independent of x for the estimator to be valid.

Proof. From our derivation of the Cramér-Rao bound (8) and in particular from (19), we note that the Cramér-Rao bound is satisfied with equality if and only if the functions $e(\mathbf{y})$ and $f(\mathbf{y})$ defined in (11) and (14), respectively, are perfectly positively correlated, i.e., if and only if there exists some constant $k(x) > 0$ (i.e., that can only depend on x) such that

$$e(\mathbf{y}) = k(x) f(\mathbf{y}) \qquad \text{for all } \mathbf{y}. \tag{28}$$

Rearranging (28) using (11) and (14), we obtain that an efficient estimator $\hat{x}(\cdot)$ must take the form

$$\hat{x}(\mathbf{y}) = x + k(x) \frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x). \tag{29}$$

Hence, an efficient estimator exists if and only if (29) is a valid estimator, i.e., if and only if the right-hand side of (29) is independent of x for some $k(x)$.

However, $k(x)$ cannot, in fact, be arbitrary. To see this, let us suppose that an efficient estimator exists, so that (8) is satisfied with equality. Then, via (13) we must have

$$\mathbb{E}\left[e^2(\mathbf{y})\right] = \lambda_{\hat{x}}(x) = \frac{1}{J_{\mathbf{y}}(x)}. \tag{30}$$

Next, using (28), (17), and (18) we obtain

$$\mathbb{E}\left[e^2(\mathbf{y})\right] = \mathbb{E}\left[e(\mathbf{y}) \cdot k(x) f(\mathbf{y})\right] = k(x)\ \mathbb{E}\left[e(\mathbf{y}) f(\mathbf{y})\right] = k(x). \tag{31}$$

Comparing (30) and (31), we can then conclude that

$$k(x) = \frac{1}{J_{\mathbf{y}}(x)}, \tag{32}$$

which when substituted into (29) yields our desired result.

As a final remark, we recall that the proof above is only valid for probability distributions that satisfy the regularity condition (7) in¡ Theorem 1. □

Three final remarks are important. First, note that an efficient estimator, i.e., a valid estimator satisfying (27) is guaranteed to be unbiased (and therefore admissible): taking the expectation of (27) we get, using (16),

$$\mathbb{E}\left[\hat{x}(\mathbf{y})\right] = x + \frac{1}{J_{\mathbf{y}}(x)}\,\mathbb{E}\left[f(\mathbf{y})\right] = x.$$

Second, we note that (27) implies that when it exists, an efficient estimator is also *unique*—clearly no two estimators can satisfy (27) and be distinct. Finally, since it meets a lower bound on the estimator variance, when it exists, an efficient estimator must be the unique MVU estimator for a problem.

Example 2. Let's continue Example 1, i.e.,

$$\mathsf{y} = x + \mathsf{w}, \tag{33}$$

where $\mathsf{w} \sim \mathtt{N}(0, \sigma^2)$. In this case, we have [cf. (25)]

$$J_{\mathsf{y}}(x) = \frac{1}{\sigma^2}. \tag{34}$$

Constructing the right-hand side of (27) using (24) and (34) we obtain

$$\hat{x}(y) = y \tag{35}$$

which we note is not a function of x and is therefore valid. Hence, we can immediately conclude that $\hat{\mathsf{x}} = \hat{x}(\mathsf{y})$ defined via (35) is unbiased and has a variance equal to the Cramér-Rao bound, i.e.,

$$\lambda_{\hat{\mathsf{x}}}(x) = \sigma^2. \tag{36}$$

Hence, we can conclude that (35) is an efficient estimator, and hence the unique MVU estimator for the problem.

8.4 Maximum Likelihood Estimation

Perhaps the most widely used estimators in practice are maximum likelihood estimators, defined as follows.

Definition 5. *The* maximum likelihood[1] *estimate* $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ *for parameter* x *based on observations* $\mathbf{y}$ *is defined via*

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{x \in \mathcal{X}} p_{\mathbf{y}}(\mathbf{y}; x). \tag{37}$$

[1]The function $p_{\mathbf{y}}(\cdot;\cdot)$ can be viewed two different ways. In particular, while we refer to $p_{\mathbf{y}}(\cdot; x)$ as a *model* for our data, we refer to $p_{\mathbf{y}}(\mathbf{y};\cdot)$ as the *likelihood* function for the possible parameters.

Maximum likelihood estimators have many desirable properties that underlie their popularity. In this section, we develop some initial perspectives based on their connection to efficient estimators.

Claim 1. *When an efficient estimator $\hat{x}_{\text{eff}}(\cdot)$ exists, it is the maximum likelihood (ML) estimator, i.e., $\hat{x}_{\text{eff}}(\cdot) = \hat{x}_{\text{ML}}(\cdot)$.*

Proof. Suppose $\hat{x}_{\text{eff}}(\cdot)$ exists. Then, for any particular value of the data $\mathbf{y}$ we have, rewriting (27),

$$\hat{x}_{\text{eff}}(\mathbf{y}) = x + \frac{1}{J_{\mathbf{y}}(x)} \frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x). \tag{38}$$

which we can compute directly. Now since the right-hand side of (38) is independent of the value of x, we are free to choose *any* value of x in this expression,[2] so let us judiciously choose x to be the number $\hat{x}_{\text{ML}}(\mathbf{y})$ as defined in (37).

From (37) we see that provided the likelihood function is strictly positive and differentiable, the ML estimator satisfies

$$\left[\frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x) \right] \Bigg|_{x=\hat{x}_{\text{ML}}(\mathbf{y})} = 0. \tag{39}$$

Thus, since $J_{\mathbf{y}}(x) > 0$ for all x except in the trivial case, (38) becomes

$$\hat{x}_{\text{eff}}(\mathbf{y}) = \hat{x}_{\text{ML}}(\mathbf{y}). \tag{40}$$

□

Claim 1 establishes that *when it exists,* the (unique) efficient estimator is equivalent to the ML estimator for the problem. For future convenience, we'll use $\lambda_{\text{ML}}(x)$ to denote the variance (and hence error variance) of the estimator (37).

However, several points should be stressed. This does not mean the ML estimators are always efficient! When an efficient estimator doesn't exist for a problem, then the ML estimator need not have any special properties. This means, for example, that when an efficient estimator does not exist, the ML estimator may not have good variance properties or even be unbiased. Nevertheless, ML estimators often have desirable asymptotic properties, in the limit of a large number of (independent) observations, that make them attractive. We defer a discussion of asymptotics until later in the subject.

We conclude the section with some examples.

Example 3. Consider observations of the form

$$y = h(x) + w, \tag{41}$$

[2] In particular, we need not choose x to be its true value.

where $h(\cdot)$ is known and $\mathsf{w} \sim \mathtt{N}(0, \sigma^2)$. In this case,

$$p_{\mathsf{y}}(y; x) = \mathtt{N}(y; h(x), \sigma^2), \tag{42}$$

so that

$$\frac{\partial \ln p_{\mathsf{y}}(y; x)}{\partial x} = \left(\frac{y - h(x)}{\sigma^2} \right) \frac{\mathrm{d}h(x)}{\mathrm{d}x}. \tag{43}$$

Let's compute the Cramér-Rao bound on the performance of arbitrary unbiased estimates $\hat{x}(\cdot)$ for x. Using (43) we obtain that

$$J_{\mathsf{y}}(x) = \mathbb{E}\left[\left[\left(\frac{\mathsf{y} - h(x)}{\sigma^2} \right) \frac{\mathrm{d}h(x)}{\mathrm{d}x} \right]^2 \right] = \left[\frac{\mathrm{d}h(x)}{\mathrm{d}x} \right]^2 \mathbb{E}\left[\left(\frac{\mathsf{w}}{\sigma^2} \right)^2 \right] = \frac{1}{\sigma^2} \left[\frac{\mathrm{d}h(x)}{\mathrm{d}x} \right]^2, \tag{44}$$

so that

$$\lambda_{\hat{x}}(x) \geq \frac{\sigma^2}{(\mathrm{d}h(x)/\mathrm{d}x)^2} \tag{45}$$

for any unbiased estimate. Now an efficient estimate exists if and only if (27) is a valid estimator, i.e., if and only if

$$x + \frac{1}{J_{\mathsf{y}}(x)} \frac{\partial}{\partial x} \ln p_{\mathsf{y}}(y; x) = \left(x - \frac{h(x)}{\mathrm{d}h(x)/\mathrm{d}x} \right) + \frac{y}{\mathrm{d}h(x)/\mathrm{d}x} \tag{46}$$

is a function only of y. However, since the right-hand term in (46) is the only one that depends on y and since y can be arbitrary, we can conclude that no efficient estimate can exist unless $\mathrm{d}h(x)/\mathrm{d}x$ does not depend on x. However, this will only be the case when $h(\cdot)$ is a linear (affine) function. Hence, efficient estimates fail to exist in the strictly nonlinear case. Consider, for example, $h(x) = x^3$. In this case, (46) becomes

$$x + \frac{y - x^3}{3x^2} = \frac{2}{3}x + \frac{1}{3}\frac{y}{x^2}, \tag{47}$$

from which we see that there is no efficient estimate. Since an efficient estimate generally doesn't exist, the ML estimate, which we'll now compute, needn't have any special properties in the nonlinear case. When $h(\cdot)$ is invertible, as we'll assume in this example, we get immediately from (43) that

$$\hat{x}_{\mathrm{ML}}(y) = h^{-1}(y), \tag{48}$$

where $h^{-1}(\cdot)$ is the inverse function of $h(\cdot)$, i.e., $h^{-1}(h(x)) = x$. Calculating the bias $b_{\mathrm{ML}}(x)$ and variance $\lambda_{\mathrm{ML}}(x)$ of this estimate is difficult in general, though in general it will be biased. And when biased, this means we cannot conclude that its variance $\lambda_{\mathrm{ML}}(x)$ satisfies (45) for even one value of x.

Let's consider a couple of other examples of ML estimators that do happen to be efficient.

Example 4. Suppose that the random variable y is exponentially-distributed with unknown mean $x \geq 0$, i.e.,

$$p_{\mathsf{y}}(y;x) = \frac{1}{x} e^{-y/x}, \quad y \geq 0. \tag{49}$$

Since $p_{\mathsf{y}}(y;x)$ and $\ln p_{\mathsf{y}}(y;x)$ have the same maximum, we obtain the ML estimate as the solution of

$$\begin{aligned} \frac{\partial}{\partial x} \ln p_{\mathsf{y}}(y;x) &= \frac{\partial}{\partial x} \left[-\ln x - \frac{y}{x} \right] \\ &= -\frac{1}{x} + \frac{y}{x^2} = 0. \end{aligned} \tag{50}$$

In particular, from (50) we get

$$\hat{x}_{\mathrm{ML}}(y) = y. \tag{51}$$

Since the mean of y is x, this estimate is unbiased. Furthermore, using the fact that

$$\lambda_{\mathrm{ML}}(x) = \operatorname{var} \mathsf{y} = x^2$$

we obtain

$$\begin{aligned} J_{\mathsf{y}}(x) &= \mathbb{E}\left[\left(\frac{\partial}{\partial x} \ln p_{\mathsf{y}}(y;x) \right)^2 \right] = \mathbb{E}\left[\frac{(y-x)^2}{x^4} \right] \\ &= \frac{1}{x^4} x^2 = \frac{1}{x^2} = \frac{1}{\lambda_{\mathrm{ML}}(x)}. \end{aligned} \tag{52}$$

Hence, the Cramér-Rao lower bound is tight and the ML estimate is efficient. Note that in this case the variance of the estimator and thus the Cramér-Rao bound are functions of x.

All of our results on nonrandom parameter estimation apply equally well to the case in which $\mathbf{y}$ is discrete-valued, as we illustrate with the following example.

Example 5. Suppose we observe a vector

$$\mathbf{y} = \begin{bmatrix} \mathsf{y}_1 & \mathsf{y}_2 & \cdots & \mathsf{y}_M \end{bmatrix}^{\mathrm{T}}$$

of independent Poisson random variables with unknown mean x, i.e., for $i = 1, 2, \ldots, M$ we have

$$p_{\mathsf{y}_i}[y_i;x] = \mathbb{P}\left(\mathsf{y}_i = y_i; x\right) = \frac{x^{y_i} e^{-x}}{y_i!}. \tag{53}$$

In this case,

$$\ln p_{\mathbf{y}}[\mathbf{y};x] = \sum_{i=1}^{M} \ln p_{\mathsf{y}_i}[y_i;x] = \sum_{i=1}^{M} (y_i \ln x - x) - \sum_{i=1}^{M} \ln(y_i!) \tag{54}$$

so that $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ is the unique solution to

$$\frac{\partial \ln p_{\mathbf{y}}[\mathbf{y}; x]}{\partial x} = \sum_{i=1}^{M} \left(\frac{y_i}{x} - 1 \right) = 0. \tag{55}$$

In particular, we obtain

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} y_i, \tag{56}$$

which again is then unbiased. Since the variance of a Poisson random variable equals its mean, we have

$$\lambda_{\mathrm{ML}} = \frac{1}{M^2} \sum_{i=1}^{M} x = \frac{x}{M}. \tag{57}$$

Using (20) with (54) we get that the Fisher information is

$$J_y(x) = \frac{1}{x^2} \, \mathbb{E} \left[\sum_{i=1}^{M} \mathsf{y}_i \right] = \frac{M}{x}, \tag{58}$$

so comparing (58) with (57) we get that the ML estimate is efficient.

It also worth noting that there is often a choice of parameterization in a problem of interest. In such settings it is worth noting that ML estimation commutes with invertible mappings. In particular, suppose that a parameter θ is related to x via

$$\theta = g(x),$$

where $g(\cdot)$ is an invertible transformation. Then it is a straightforward exercise to show that the ML estimates are also related by

$$\hat{\theta}_{\mathrm{ML}}(\mathbf{y}) = g(\hat{x}_{\mathrm{ML}}(\mathbf{y})). \tag{59}$$

Implicitly, we saw an instance of such behavior in Example 3.

It is worth observing that (59) is a property not shared by other estimators we've discussed. For example, Bayesian estimators almost never commute with nonlinear transformations, i.e., if x is a random parameter, then

$$\hat{\theta}_{\mathrm{B}}(\mathbf{y}) \neq g(\hat{x}_{\mathrm{B}}(\mathbf{y})),$$

for almost any nontrivial cost criterion.

We also remark that although (59) doesn't apply when $g(\cdot)$ is not invertible, straightforward extensions of this result can be developed to handle the non-invertible case. We leave such extensions as an exercise.

We remark, however, that properties possessed by $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ are often not preserved under the transformation (59). For example, since typically

$$\mathbb{E} \left[\hat{\theta}_{\mathrm{ML}}(\mathbf{y}) \right] = \mathbb{E} \left[g(\hat{x}_{\mathrm{ML}}(\mathbf{y})) \right] \neq g \left(\mathbb{E} \left[\hat{x}_{\mathrm{ML}}(\mathbf{y}) \right] \right),$$

we wouldn't expect $\hat{\theta}_{\mathrm{ML}}(\mathbf{y})$ to be unbiased even if $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ were.

8.5 Estimation of Nonrandom Vectors

In this section, we summarize some extensions of the preceding results to the problem of estimating a vector of nonrandom parameters $\mathbf{x}$. We begin with the extension of the Cramér-Rao bound to this case.

Theorem 2 (Cramér-Rao bound, vector version)**.** *The covariance matrix* $\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})$ *of any unbiased estimator satisfies the matrix inequality*

$$\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) \geq \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x}), \tag{60}$$

i.e., $\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) - \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})$ *is positive semidefinite, where* $\mathbf{J}_{\mathbf{y}}(\mathbf{x})$ *is now the Fisher Information* matrix

$$\mathbf{J}_{\mathbf{y}}(\mathbf{x}) \triangleq \mathbb{E}\left[\mathbf{S}_{\mathbf{y}}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}(\mathbf{x})\right] \tag{61}$$

with

$$\mathbf{S}_{\mathbf{y}}(\mathbf{x}) \triangleq \frac{\partial \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}} \tag{62}$$

denoting the score (row) vector*.*

Note that from the diagonal elements of (60) we obtain a set of scalar Cramér-Rao bounds on the variances of individual components of $\mathbf{x}$. Note, too, that it is straightforward to verify that we can equivalently write the Fisher matrix as the expectation of a Hessian matrix, viz.,

$$\mathbf{J}_{\mathbf{y}}(\mathbf{x}) = -\,\mathbb{E}\left[\frac{\partial^2 \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^2}\right].$$

Proof. To derive the matrix Cramér-Rao bound (60), we follow an approach analogous to that used to obtain (8), but which requires some additional steps. In particular, we begin by recalling that for unbiased estimators the error

$$\mathbf{e}(\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} \tag{63}$$

has zero mean, i.e.,

$$\mathbb{E}\left[\mathbf{e}(\mathbf{y})\right] = 0 \tag{64}$$

and covariance

$$\mathbb{E}\left[\mathbf{e}(\mathbf{y})\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}). \tag{65}$$

Next we define

$$\mathbf{f}^{\mathrm{T}}(\mathbf{y}) = \mathbf{S}_{\mathbf{y}}(\mathbf{x}) \tag{66}$$

and note that using the identity

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = \frac{1}{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})} \frac{\partial}{\partial \mathbf{x}} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}), \tag{67}$$

we get that $\mathbf{f}(\mathbf{y})$ has zero mean:

$$\begin{aligned}\mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] &= \mathbb{E}\left[\frac{1}{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] \\ &= \int_{-\infty}^{+\infty}\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y} \\ &= \frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y} = \frac{\partial}{\partial \mathbf{x}}1 = \mathbf{0},\end{aligned} \tag{68}$$

and, in turn, covariance

$$\mathbf{\Lambda}_{\mathbf{f}}(\mathbf{x}) = \mathbb{E}\left[\mathbf{f}(\mathbf{y})\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{J}_{\mathbf{y}}(\mathbf{x}). \tag{69}$$

Finally, again using the identity (67), the covariance between $\mathbf{e}(\mathbf{y})$ and $\mathbf{f}(\mathbf{y})$ is given by

$$\begin{aligned}\operatorname{cov}\left(\mathbf{e}(\mathbf{y}),\mathbf{f}(\mathbf{y})\right) &= \mathbb{E}\left[\mathbf{e}(\mathbf{y})\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] \\ &= \int_{-\infty}^{+\infty}(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y} \\ &= \left[\frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}\hat{\mathbf{x}}(\mathbf{y})\,p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y}\right] - \left[\mathbf{x}\frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}p_{y}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y}\right] \\ &= \mathbf{I}-\mathbf{0} = \mathbf{I}.\end{aligned} \tag{70}$$

Next, for an arbitrary choice of $\mathbf{c}$ we let

$$\tilde{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{e}(\mathbf{y}) \tag{71}$$

and

$$\tilde{f}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{f}(\mathbf{y}) = \mathbf{f}^{\mathrm{T}}(\mathbf{y})\,\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\,\mathbf{c}. \tag{72}$$

Then both $\tilde{e}(\mathbf{y})$ and $\tilde{f}(\mathbf{y})$ have zero-mean and, using (65), (69) and (70), we have

$$\operatorname{var}\tilde{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})\mathbf{c} \tag{73a}$$

$$\operatorname{var}\tilde{f}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c} \tag{73b}$$

$$\operatorname{cov}\left(\tilde{e}(\mathbf{y}),\tilde{f}(\mathbf{y})\right) = \mathbf{c}^{\mathrm{T}}\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}. \tag{73c}$$

Now since the covariance between $\tilde{e}(\mathbf{y})$ and $\tilde{f}(\mathbf{y})$ satisfies the bound

$$\left[\operatorname{cov}\left(\tilde{e}(\mathbf{y}),\tilde{f}(\mathbf{y})\right)\right]^2 \leq \operatorname{var}\tilde{e}(\mathbf{y})\operatorname{var}\tilde{f}(\mathbf{y}) \tag{74}$$

we can substitute (73) into (74) to obtain, after some simple manipulation,

$$\mathbf{c}^{\mathrm{T}}\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}\left[\mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})\mathbf{c} - \mathbf{c}^{\mathrm{T}}\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}\right] \geq 0. \tag{75}$$

However, since $\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})$ is positive semidefinite, the term to the left of the brackets in (75) is non-negative. Hence, the term in brackets must be non-negative. But then since $\mathbf{c}$ is arbitrary this means $\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) - \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})$ must be positive semidefinite, which establishes (60) as desired.

It remains only to verify that the second form of the Fisher information in (61) is equivalent to the first. Analogous to our approach in the scalar case, we begin by observing

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,\mathrm{d}\mathbf{y} = 1. \tag{76}$$

Computing the gradient of (76) with respect to $\mathbf{x}$ and using the identity (67) yields

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) \left[\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}} \mathrm{d}\mathbf{y} = \mathbf{0}. \tag{77}$$

Finally, computing the Hessian of (76) with respect to $\mathbf{x}$ and again using (67) we obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) \left[\frac{\partial^2}{\partial \mathbf{x}^2} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] \mathrm{d}\mathbf{y} \\ + \int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) \left[\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}} \left[\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] \mathrm{d}\mathbf{y} = \mathbf{0},\end{aligned} \tag{78}$$

as desired. □

Next, we have the following.

Corollary 3. *An unbiased efficient estimate* $\hat{\mathbf{x}}(\mathbf{y})$ *exists if and only if*

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x}) \left[\frac{\partial \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}}\right]^{\mathrm{T}} \tag{79}$$

is a valid estimator, i.e., if and only if the right-hand side of (79) *does not depend on* $\mathbf{x}$*.*

Proof. Equality is satisfied in (74) (and therefore (75)) if and only if $\tilde{e}(\mathbf{y}) = k(\mathbf{x})\tilde{f}(\mathbf{y})$ for some function $k(\mathbf{x})$ that doesn't depend on $\mathbf{y}$, i.e., if and only if,

$$\mathbf{c}^{\mathrm{T}}\mathbf{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}} k(\mathbf{x})\, \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\, \mathbf{f}(\mathbf{y}). \tag{80}$$

However, since (80) holds for any choice of $\mathbf{c}$ we must have

$$\mathbf{e}(\mathbf{y}) = k(\mathbf{x})\, \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\, \mathbf{f}(\mathbf{y}). \tag{81}$$

Again $k(\mathbf{x})$ can't be arbitrary. In particular, when the bound (60) is satisfied with equality we have

$$\mathbb{E}\left[\mathbf{e}(\mathbf{y})\, \mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) = \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x}). \tag{82}$$

However, using (81), (69), and (70) we have

$$\mathbb{E}\left[\mathbf{e}(\mathbf{y})\,\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbb{E}\left[\mathbf{e}(\mathbf{y})\,\mathbf{f}^{\mathrm{T}}(\mathbf{y})\,\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\,k(\mathbf{x})\right] = \mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\,k(\mathbf{x}). \tag{83}$$

Comparing (82) with (83) we obtain

$$k(\mathbf{x}) = 1, \tag{84}$$

which when substituted into (81) yields our desired result. □

Finally, we have the following.

Corollary 4. *If an efficient unbiased estimate exists, it is the ML estimate.*

Proof. Since (79) must not be a function of $\mathbf{x}$ when an efficient estimator exists, we can then freely choose any value of $\mathbf{x}$ in this expression without effect. If we choose the value $\mathbf{x} = \hat{\mathbf{x}}_{\text{eff}}(\mathbf{y})$, we obtain

$$\mathbf{J}_{\mathbf{y}}^{-1}(\mathbf{x})\left[\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}}\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_{\text{eff}}(\mathbf{y})} = \mathbf{0}. \tag{85}$$

But since $\mathbf{J}_{\mathbf{y}}(\mathbf{x})$ is nonsingular except in the trivial case, we have that the term in brackets in (85) must be zero, i.e.,

$$\hat{\mathbf{x}}_{\text{eff}}(\mathbf{y}) = \hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}) = \arg\max_{\mathbf{x}} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}). \tag{86}$$

□

Again we stress that one should not infer from these results that the ML estimator is always efficient. When no efficient estimator exists, the ML estimate can still be computed; however it need not have any special properties. As in the scalar case, though, even when an efficient estimator doesn't exist, the ML estimator often has good asymptotic properties.

We conclude this section with an example in which ML estimation is efficient in the vector case.

Example 6. Suppose we that our observed data $\mathbf{y}$ depends on our parameter vector $\mathbf{x}$ through the model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{87}$$

where $\mathbf{w} \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_{\mathbf{w}})$. In this case

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = \mathtt{N}(\mathbf{y};\mathbf{H}\mathbf{x},\mathbf{\Lambda}_{\mathbf{w}}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right] \tag{88}$$

so that maximizing $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ with respect to $\mathbf{x}$ is equivalent to *minimizing*

$$\varphi(\mathbf{x}) = \frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) \tag{89}$$

with respect to $\mathbf{x}$. Since (89) is a non-negative function, its unique stationary point, which we obtain by setting the gradient of (89) to zero, is its global minimum and thus gives the ML estimate

$$\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = (\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{y} \tag{90}$$

This estimate is unbiased, since

$$\mathbb{E}\left[\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})\right] = (\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{H}\mathbf{x}+\mathbb{E}\left[\mathbf{w}\right]) = \mathbf{x} \tag{91}$$

and its error covariance is

$$\begin{aligned}\boldsymbol{\Lambda}_{\mathrm{ML}} &= \mathbb{E}\left[\left[(\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{w}\right]\left[(\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{w}\right]^{\mathrm{T}}\right] \\ &= (\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\boldsymbol{\Lambda}_{\mathbf{w}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1} \\ &= (\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}.\end{aligned} \tag{92}$$

Note that for this estimate to make sense, $\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}$ must be invertible, and this in turn requires that the dimension of $\mathbf{y}$ (or, more precisely, the rank of $\boldsymbol{\Lambda}_{\mathbf{w}}$) be at least as large as the dimension of $\mathbf{x}$. Phrased differently, the number of degrees of freedom in the measurements must equal or exceed the number of parameters to be estimated.

The Fisher information matrix for this problem is obtained using the second form of (61) and yields

$$J_{\mathbf{y}}(\mathbf{x}) = -\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{x}^2}\varphi(\mathbf{x}) = \mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H} \tag{93}$$

which by comparison to (92) allows us to conclude that the ML estimate is, in fact, efficient. Note as well that the estimator covariance (and thus the Fisher matrix) is independent of $\mathbf{x}$ in this example. Since our ML estimate is efficient, it is the MVU estimator for the nonrandom parameter estimation problem. This result is referred to as the Gauss-Markov theorem.