

A Few Traditional (Non-Bayesian) Estimation Approaches

Outline:

- Linear-model examples, minimum-variance unbiased (MVU) estimator for the linear model,
- Best linear unbiased estimator (BLUE),
- General MVU estimation,
- Maximum-likelihood (ML) estimation.

Reading: Kay-I, Chapters 4, 6, 7, and 8.

Linear-model Examples

Polynomial Curve Fitting Example. Continuous signal $X(t)$ is modeled as a polynomial of degree $p - 1$ in additive noise:

$$X(t) = \theta_1 + \theta_2 t + \cdots + \theta_p t^{p-1} + W(t).$$

We observe $\{x[n]\} = \{x(t_n)\}$ $n = 0, 1, \dots, N - 1$. Define

$$\mathbf{X} = [X(t_0), \dots, X(t_{N-1})]^T$$

$$\mathbf{x} = [x(t_0), \dots, x(t_{N-1})]^T$$

$$\mathbf{W} = [W(t_0), \dots, W(t_{N-1})]^T$$

$$\boldsymbol{\theta} = [\theta_1, \dots, \theta_p]^T$$

$$H = \begin{bmatrix} 1 & t_0 & \cdots & t_0^{p-1} \\ 1 & t_1 & \cdots & t_1^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{N-1} & \cdots & t_{N-1}^{p-1} \end{bmatrix} \quad (\text{an } N \times p \text{ matrix}).$$

The data model is then

$$\mathbf{X} = H \boldsymbol{\theta} + \mathbf{W} \quad (1)$$

where H is known and $\boldsymbol{\theta}$ is the parameter vector to be estimated.

Sinusoidal Amplitude and Phase Estimation Estimation:

Measured signal $x(t)$ is modeled as a superposition of $p/2$ sinusoids (having known frequencies but unknown amplitudes and phases):

$$X(t) = \sum_{k=1}^{p/2} r_k \sin(\omega_k t + \phi_k) + W(t).$$

This model is linear in r_k but nonlinear in ϕ_k . However, we can rewrite it as

$$X(t) = \sum_{k=1}^{p/2} [A_k \cos(\omega_k t) + B_k \sin(\omega_k t)] + W(t)$$

and we get the model (1), with a different (trigonometric) H .

For $p/2 = 2$ sinusoids:

$$H = \begin{bmatrix} \cos(\omega_1 t_0) & \cos(\omega_2 t_0) & \sin(\omega_1 t_0) & \sin(\omega_2 t_0) \\ \cos(\omega_1 t_1) & \cos(\omega_2 t_1) & \sin(\omega_1 t_1) & \sin(\omega_2 t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(\omega_1 t_{N-1}) & \cos(\omega_2 t_{N-1}) & \sin(\omega_1 t_{N-1}) & \sin(\omega_2 t_{N-1}) \end{bmatrix}$$

and

$$\boldsymbol{\theta} = [A_1, \dots, A_{p/2}, B_1, \dots, B_{p/2}]^T.$$

Once we compute an estimate $\hat{\theta}$ of θ , \hat{r}_k and $\hat{\phi}_k$ are obtained using the simple conversion from rectangular to polar coordinates.

Note: Even if $\hat{\theta}$ is a minimum variance unbiased (MVU) estimator, $\{\hat{r}_k\}$ and $\{\hat{\phi}_k\}$ will only be *asymptotically* MVU (for large N), as we will see later.

A Problem Formulation for Linear Models

Consider the model (1) where H is a known *deterministic* $N \times p$ matrix, with $N \geq p$. We wish to estimate the *unknown* parameter vector $\boldsymbol{\theta}$.

Assume that the noise \mathbf{W} is white Gaussian:

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$$

where σ^2 is the noise variance. If σ^2 is known, the likelihood function of $\boldsymbol{\theta}$ for the measurement vector \mathbf{x} is

$$f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\sigma^2 I_N|}} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{x} - H\boldsymbol{\theta})^T (\mathbf{x} - H\boldsymbol{\theta})\right].$$

Recall the *identifiability condition* in (3) of handout # 0:

$$f_{\mathbf{X}|\boldsymbol{\theta}}(\cdot|\boldsymbol{\theta}_1) = f_{\mathbf{X}|\boldsymbol{\theta}}(\cdot|\boldsymbol{\theta}_2) \iff \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$

which, in this case, reduces to

$$H\boldsymbol{\theta}_1 = H\boldsymbol{\theta}_2 \iff \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

To satisfy this condition, we assume that H has full rank p .

Minimum Variance Unbiased (MVU) Estimator for the Linear Model

Theorem 1. *For the linear model*

$$\mathbf{X} = \mathbf{H} \boldsymbol{\theta} + \mathbf{W} \quad (2)$$

where

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$$

the MVU estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{X}. \quad (3)$$

The covariance matrix of $\hat{\boldsymbol{\theta}}$ attains the Cramér-Rao bound (CRB) for all $\boldsymbol{\theta} \in \mathbf{R}^p$ and is given by

$$\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \text{MSE}\{\hat{\boldsymbol{\theta}}\} = \mathbb{E}_{\mathbf{X}|\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] = \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1}.$$

Proof. Verifying the unbiasedness of $\hat{\boldsymbol{\theta}}$ and the covariance matrix expression for $\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$ proves the theorem. For the above model,

$$\text{CRB}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})$$

and the Fisher information matrix (FIM) $\mathcal{I}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is computed using the general Gaussian FIM expression (29) from handout # 2:

$$[\mathcal{I}(\boldsymbol{\theta})]_{i,k} = \frac{1}{\sigma^2} \cdot \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})^T}{\partial \theta_i} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_k}$$

where $\boldsymbol{\mu}(\boldsymbol{\theta}) = H \boldsymbol{\theta}$. Now,

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial (H \boldsymbol{\theta})}{\partial \theta_i} = \textit{ith column of } H$$

implying that

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2} H^T H \\ \text{CRB}(\boldsymbol{\theta}) &= \sigma^2 (H^T H)^{-1}. \end{aligned} \quad (4)$$

□

Comments:

- Since the joint FIM and CRB for

$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\theta} \\ \sigma^2 \end{bmatrix}$$

are block-diagonal matrices, $\boldsymbol{\theta}$ and σ^2 are decoupled:
CRB($\boldsymbol{\theta}$) is the same regardless of whether σ^2 is known

or not. To be more precise, $\text{CRB}(\boldsymbol{\theta})$ in (4) is *CRB for $\boldsymbol{\theta}$ assuming that σ^2 is known* and the full CRB for $\boldsymbol{\rho}$ when both $\boldsymbol{\theta}$ and σ^2 are *unknown* is

$$\text{CRB}_{\boldsymbol{\rho},\boldsymbol{\rho}}(\boldsymbol{\rho}) = \begin{bmatrix} \overbrace{\text{CRB}_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\rho})}^{\text{same as (4)}} & 0 \\ 0 & \text{CRB}_{\sigma^2,\sigma^2}(\boldsymbol{\rho}) \end{bmatrix}.$$

Therefore, $\hat{\boldsymbol{\theta}}$ in (3) is the MVU estimator of $\boldsymbol{\theta}$ regardless of whether or not σ^2 is known.

- $\hat{\boldsymbol{\theta}}$ in (3) coincides with the *least-squares (LS) estimator* of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{x} - H \boldsymbol{\theta}\|_{\ell_2}^2$$

which can be shown by differentiating $\|\mathbf{x} - H \boldsymbol{\theta}\|_{\ell_2}^2$ with respect to $\boldsymbol{\theta}$ and setting the result to zero or by completing the squares. Here,

$$\|\mathbf{a}\|_{\ell_2}^2 \triangleq \mathbf{a}^T \mathbf{a} \quad \ell_2 \text{ (Euclidean) norm}$$

for an arbitrary real-valued vector \mathbf{a} . Later in this handout, we will see a geometric interpretation of the LS approach.

Minimum Variance Unbiased Estimator for the Linear Model (cont.)

The solution from the above theorem is numerically not sound as given. It is better to use a QR factorization, briefly outlined below. Suppose that the $N \times p$ matrix H is factored as

$$H = QR = [Q_1 \ Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$$

where Q is orthonormal $N \times N$ and R_1 is upper-triangular $p \times p$ (**MATLAB**: `qr`). Then

$$(H^T H)^{-1} H^T = R_1^{-1} Q_1^T.$$

Thus, $\hat{\theta}$ in (3) can be obtained by solving the triangular system of equations

$$R_1 \hat{\theta} = Q_1^T x.$$

MATLAB has the backslash command for computing the LS solution:

$$\theta = H \backslash x;$$

Minimum Variance Unbiased Estimator for the Linear Model, Colored Noise

Suppose that the noise is colored, so that

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2 C)$$

where $C \neq I$ is a known positive-definite matrix. Here, σ^2 can be an unknown parameter or a known constant. We can use *prewhitening* to get back to the white-noise case. Compute the Cholesky factorization of C^{-1} :

$$C^{-1} = D^T D \quad \text{MATLAB: } D = \text{inv}(\text{chol}(C))';$$

(Any other square-root factorization could be used as well.)

Now, define the transformed measurement model:

$$\underbrace{D \mathbf{X}}_{\mathbf{X}^{\text{transf}}} = \underbrace{D H}_{H^{\text{transf}}} \boldsymbol{\theta} + \underbrace{D \mathbf{W}}_{\mathbf{W}^{\text{transf}}}.$$

Clearly,

$$\begin{aligned} \text{COV}_{\mathbf{W}^{\text{transf}}}(\mathbf{W}^{\text{transf}}) &= \text{COV}_{\mathbf{W}}(D \mathbf{W} D^T) = D (\sigma^2 C) D^T \\ &= \sigma^2 D (D^T D)^{-1} D^T = \sigma^2 I \end{aligned}$$

and

$$\mathbf{W}^{\text{transf}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

reducing the problem to the white-noise case.

MVU Estimation, Colored Noise (cont.)

Theorem 2. For colored Gaussian noise with covariance matrix $\sigma^2 C$ where C is a known positive-definite matrix, the MVU estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (H^T C^{-1} H)^{-1} H^T C^{-1} \mathbf{x}.$$

The covariance matrix of $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$ attains CRB for $\boldsymbol{\theta}$ and is given by

$$\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \text{MSE}\{\hat{\boldsymbol{\theta}}\} = \text{CRB}(\boldsymbol{\theta}) = (H^T C^{-1} H)^{-1}.$$

Note: $\hat{\boldsymbol{\theta}}$ is a weighted LS estimate,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta}} \|\mathbf{x} - H \boldsymbol{\theta}\|_Q^2 \\ &= \arg \min_{\boldsymbol{\theta}} (\mathbf{x} - H \boldsymbol{\theta})^T Q (\mathbf{x} - H \boldsymbol{\theta}). \end{aligned}$$

The optimal weight matrix

$$Q = C^{-1}$$

prewhitens the residuals.

Best Linear Unbiased Estimator (BLUE)

Consider the linear model (2) where \mathbf{W} has zero mean and a positive-definite covariance matrix

$$\text{cov}_{\mathbf{W}}(\mathbf{W}) = \mathbb{E}_{\mathbf{W}}(\mathbf{W} \mathbf{W}^T) = \mathbf{C}.$$

We look for the *best linear unbiased estimator (BLUE)* of $\boldsymbol{\theta}$. Hence, we restrict our estimators $\hat{\boldsymbol{\theta}}$ to be

- *linear*, i.e.

$$\hat{\boldsymbol{\theta}} = \mathbf{A}^T \mathbf{x}$$

and

- *unbiased*, i.e.

$$\mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}) = \boldsymbol{\theta}$$

and search for a $\hat{\boldsymbol{\theta}}$ that minimizes the estimator covariance matrix $\text{cov}_{\mathbf{x} | \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$.

Theorem 3. (Gauss-Markov) *The BLUE of $\boldsymbol{\theta}$ is*

$$\hat{\boldsymbol{\theta}}_{\text{BLUE}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{X} \quad (5)$$

and its covariance matrix is

$$\text{cov}_{\mathbf{x} | \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\text{BLUE}}) = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}. \quad (6)$$

The expression for $\text{COV}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}_{\text{BLUE}}|\boldsymbol{\theta})$ holds *independently* of the distribution of the noise \mathbf{W} : all we impose on \mathbf{W} is that it has zero mean and known positive-definite covariance matrix C .

Note:

- it is easy to generalize the above result and show that (5) is BLUE in the more general case where $\text{COV}_{\mathbf{W}}(\mathbf{W}) = \sigma^2 C$, C is known and positive definite, and σ^2 is *unknown*.
- The estimate $\hat{\boldsymbol{\theta}}$ is statistically efficient (attains CRB) if \mathbf{W} is Gaussian, but it is not efficient in general. For non-Gaussian measurement models, there might be a better nonlinear estimate.

Proof. (Theorem 3) Linear estimates of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = A^T \mathbf{X}. \quad (7)$$

The unbiasedness condition of $\hat{\boldsymbol{\theta}}$ in (7) implies:

$$\begin{aligned} \boldsymbol{\theta} &= \text{E}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}) \\ &= \text{E}_{\mathbf{X}|\theta}(A^T \mathbf{X}) \\ &\stackrel{\text{see (2)}}{=} \text{E}_{\mathbf{W}}[A^T (H \boldsymbol{\theta} + \mathbf{W})] = A^T H \boldsymbol{\theta} \end{aligned}$$

yielding

$$A^T H = I. \quad (8)$$

Now, use (7) and (8) to compute $\text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}})$ for an arbitrary linear unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$:

$$\begin{aligned} \text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}) &\stackrel{\text{see (2)}}{=} \text{cov}_{\mathbf{W}}[A^T (H \boldsymbol{\theta} + \mathbf{W})] \\ &= A^T \text{cov}_{\mathbf{W}}(\mathbf{W}) A \end{aligned}$$

and

$$\begin{aligned} \text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}_{\text{BLUE}}) &= (H^T C^{-1} H)^{-1} H^T C^{-1} C C^{-1} H (H^T C^{-1} H)^{-1} \\ &= (H^T C^{-1} H)^{-1}. \end{aligned}$$

To prove that $\hat{\boldsymbol{\theta}}_{\text{BLUE}}$ has the smallest variance [within the family of linear unbiased estimators $\hat{\boldsymbol{\theta}}$ in (7) satisfying (8)], we show that

$$\text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}_{\text{BLUE}}) \leq \text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}})$$

as follows:

$$\begin{aligned} \text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}) - \text{cov}_{\mathbf{X}|\theta}(\hat{\boldsymbol{\theta}}_{\text{BLUE}}) &= A^T C A - (H^T C^{-1} H)^{-1} \\ &\stackrel{A^T H=I}{=} A^T C A - A^T H (H^T C^{-1} H)^{-1} H^T A \\ &= A^T [C - H (H^T C^{-1} H)^{-1} H^T] A \\ &= A^T [C - H (H^T C^{-1} H)^{-1} H^T] C^{-1} [C - H (H^T C^{-1} H)^{-1} H^T] A \end{aligned}$$

which is always positive semidefinite. \square

Examples

Example 4.4 in Kay-I. Estimate DC level in colored noise:

$$X[n] = a + W[n]$$

for $n = 0, 1, \dots, N - 1$, where $\mathbf{W} = [W[0], W[1], \dots, W[N - 1]]^T$ is colored noise vector with zero mean and known covariance matrix

$$\text{cov}_{\mathbf{W}}(\mathbf{W}) = C.$$

Hence, substituting $H = \mathbf{h} = \mathbf{1} = [1, 1, \dots, 1]^T$ into (5) and (6) yields BLUE of a

$$\hat{a} = (\mathbf{h}^T C^{-1} \mathbf{h})^{-1} \mathbf{h}^T C^{-1} \mathbf{X} = \frac{\mathbf{1}^T C^{-1} \mathbf{X}}{\mathbf{1}^T C^{-1} \mathbf{1}}$$

and its variance given a

$$\text{var}_{\mathbf{X} | a}(\hat{a} | a) = \frac{1}{\mathbf{1}^T C^{-1} \mathbf{1}}.$$

Consider the Cholesky factorization

$$C^{-1} = D^T D.$$

Then, BLUE of a becomes

$$\hat{a} = \frac{\mathbf{1}^T D^T D \mathbf{X}}{\mathbf{1}^T D^T D \mathbf{1}} = \frac{(D \mathbf{1})^T \overbrace{D \mathbf{x}}^{\mathbf{x}^{\text{transf}}}}{\mathbf{1}^T D^T D \mathbf{1}} = \sum_{n=0}^{N-1} d_n x^{\text{transf}}[n]$$

where

$$d_n = \frac{[D \mathbf{1}]_n}{\mathbf{1}^T D^T D \mathbf{1}}.$$

Examples (cont.)

Sometimes, BLUE is completely wrong. For example, consider

$$X[n] \quad n = 0, 1, \dots, N - 1$$

white Gaussian noise with unknown variance σ^2 . The MVU estimator of σ^2 is

$$\hat{\sigma}^2 = \hat{\sigma}^2(\mathbf{X}) = (1/N) \cdot \sum_{n=0}^{N-1} X^2[n].$$

On the other hand,

$$\hat{\sigma}_{\text{BLUE}}^2 = \hat{\sigma}_{\text{BLUE}}^2(\mathbf{X}) = \sum_{n=0}^{N-1} a_n X[n].$$

For an estimator $\hat{\sigma}^2$ to be unbiased, we need

$$\mathbb{E}_{\mathbf{X} | \sigma^2}(\hat{\sigma}^2 | \sigma^2) = \sigma^2$$

but

$$\mathbb{E}_{\mathbf{X} | \sigma^2}(\hat{\sigma}_{\text{BLUE}}^2 | \sigma^2) = \sum_{n=0}^{N-1} a_n \mathbb{E}_{X | \sigma^2}(x[n] | \sigma^2) = 0.$$

It is impossible to find a_n s to make $\hat{\sigma}_{\text{BLUE}}^2$ unbiased.

Note: Although BLUE is not suitable for this problem, *transforming the data*

$$Y[n] = X^2[n]$$

and designing BLUE for $Y[n]$ would produce a viable estimator of σ^2 .

General MVU Estimation

What is MVU estimate in general?

Theorem 4. (Rao-Blackwell) *If $\tilde{\theta}(x)$ is any unbiased estimator and $T(x)$ is a sufficient statistic, then*

$$\hat{\theta}(\mathbf{X}) = E_{\mathbf{X} | T(\mathbf{X})} [\tilde{\theta}(\mathbf{X}) | T(\mathbf{X})] \quad (9)$$

is no worse than $\tilde{\theta}(\mathbf{X})$ in terms of MSE.

Problem: computing

$$E_{\mathbf{X} | T(\mathbf{X})} [\tilde{\theta}(\mathbf{X}) | T(\mathbf{X})]$$

may be difficult. Recall that this type of expectation occurred when proving sufficiency, but luckily, in the case of sufficiency, our efforts were greatly simplified by the factorization theorem.

Definition. $T(\mathbf{X})$ is *complete sufficient statistic* if only one estimator

$$\hat{\theta}(\mathbf{X}) = g(T(\mathbf{X}))$$

is unbiased.

Corollary 1. *If $T(\mathbf{X})$ is a complete sufficient statistic, then the unique unbiased estimate $\hat{\theta} = g(T(\mathbf{X}))$ is the MVU estimate.*

Comments:

- Conditioning always decreases the variance (does not increase, to be more precise).
- The definition of sufficient statistic $\mathbf{T}(\mathbf{X})$ implies that conditioning on it leads to a distribution that is not a function of the unknown parameters $\boldsymbol{\theta}$. Hence, (9) is a statistic and is therefore a proper (realizable) estimator.

Example: Suppose that

$$X[n] \quad n = 1, 2, \dots, N$$

are independent, identically distributed (i.i.d.) $\mathcal{N}(a, \sigma^2)$ and we wish to estimate the parameter vector

$$\boldsymbol{\theta} = [a, \sigma^2]^T.$$

Then,

$$\begin{aligned} f_{\mathbf{X} | \boldsymbol{\theta}}(\mathbf{x} | \boldsymbol{\theta}) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - a)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \\ &\cdot \exp \left\{ -\frac{1}{2\sigma^2} \left[(N-1) \frac{1}{N-1} \sum_{n=1}^N (x[n] - \bar{x})^2 + N(\bar{x} - a)^2 \right] \right\} \end{aligned}$$

where

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x[n].$$

Therefore, the jointly sufficient statistics for $\boldsymbol{\theta}$ are

$$T_1(\mathbf{x}) = \bar{x}, \quad T_2(\mathbf{x}) = \hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2.$$

It can be shown that $\hat{a} = \bar{x} = T_1(\mathbf{x})$ and $\hat{\sigma}^2 = T_2(\mathbf{x})$ are the only unbiased functions of

$$\mathbf{T}(\mathbf{x}) = [T_1(\mathbf{x}), T_2(\mathbf{x})]^T.$$

Hence, by Corollary 1, they make up the MVU estimate of $\boldsymbol{\theta}$ (although, in this case, the MVU estimate is *not efficient* and, therefore, could not have been found using Theorem 2 in handout # 2). Indeed, for $\hat{\boldsymbol{\theta}} = [\hat{a}, \hat{\sigma}^2]^T$,

$$\text{cov}_{\mathbf{X} | \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & 2(\sigma^2)^2/(N-1) \end{bmatrix}$$

but, CRB for this case [obtained by inverting the Fisher information matrix (23) in handout # 2] is:

$$\text{CRB}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & 2(\sigma^2)^2/N \end{bmatrix}.$$

Maximum-likelihood (ML) Estimation

$$\hat{\theta}(\mathbf{x}) = \arg \max_{\theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta).$$

Comments:

- Recall that $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$, viewed as function of θ , is the *likelihood function* of θ .
- For a given θ and discrete case, $p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ is the probability of observing the point \mathbf{x} . In the continuous case, $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ is approximately proportional to probability of observing a point in a small rectangle around \mathbf{x} . However, when we think of $p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ or $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ as functions of θ , they give, for a given observed \mathbf{x} , the *likelihood or plausibility* of various θ .
- ML estimate of θ is value of the parameter θ that *makes the probability of the data as great as it can be under the assumed model*.

ML Estimation (cont.)

Theorem 5. Assume that certain regularity conditions hold and denote the ML estimate of θ by $\hat{\theta} = \hat{\theta}(\mathbf{X})$. Then, as the number of measurements $N \nearrow +\infty$,

$$\hat{\theta} \rightarrow \theta \quad (\text{with probability 1}) \quad (\text{consistency}) \quad (10)$$

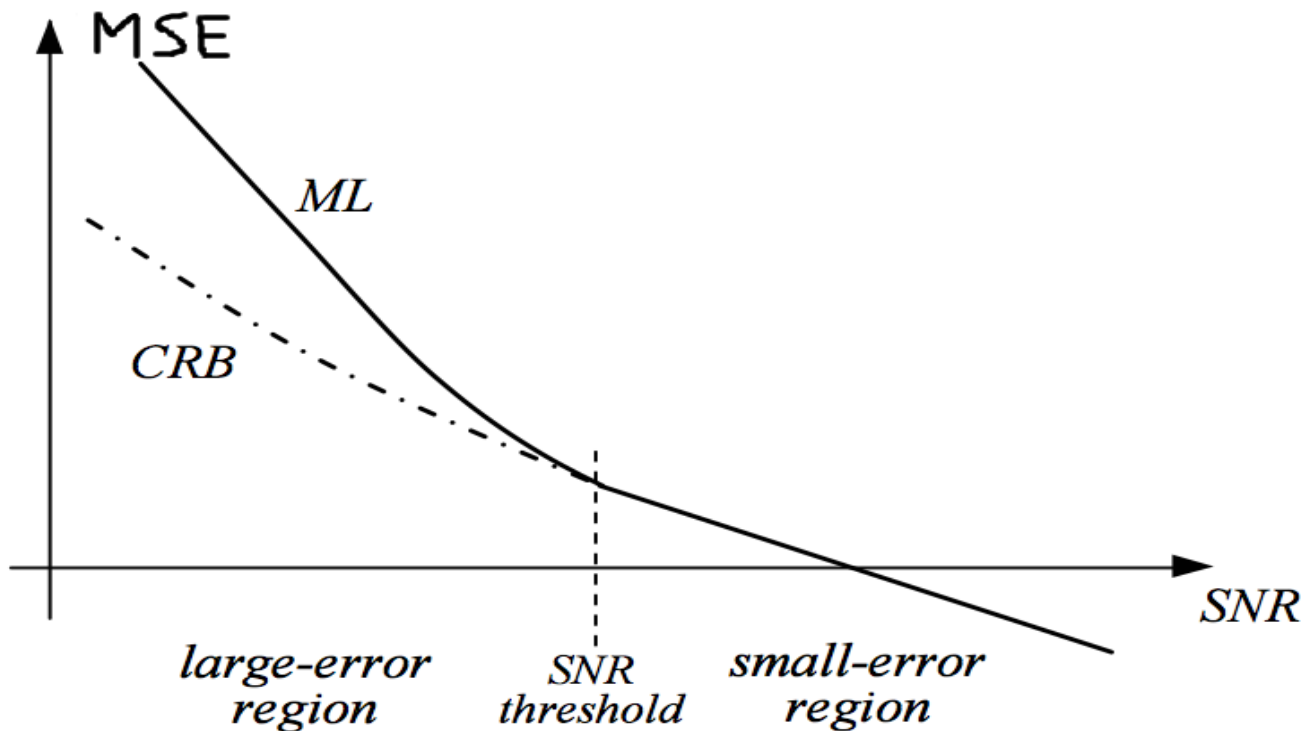
$$\sqrt{N} (\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, N \mathcal{I}^{-1}(\theta)) \quad (\text{asymptotic efficiency}) \quad (11)$$

where θ is the true value of the parameter and $\mathcal{I}(\theta_0)$ is the Fisher information [and $\mathcal{I}^{-1}(\theta)$ the CRB]. Moreover, if an efficient (for finite N) estimate of θ exists, it is the ML estimate.

Proof. See pp. 364–366, Chapter 5f.2 in

C.R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, 1973.

for the case of independent observations. \square



Note: At lower signal-to-noise ratios (SNRs), a threshold effect occurs: outliers give rise to increased MSE (more than predicted by the CRB). This behavior is characteristic of practically all (good) nonlinear estimators.

Example: $X[n]$ $n = 0, 1, \dots, N - 1$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$ given θ , where θ is the unknown parameter and σ^2 is a known constant. We wish to maximize $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ or, more conveniently and equivalently,

$$\ln f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$$

with respect to θ , where

$$\ln f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \text{const} - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \theta)^2$$

where const denotes terms that are not functions of θ . Thus, the ML estimate is the sample mean

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n] \quad (\text{sample mean})$$

and

$$\{\bar{X} \mid \theta\} \sim \mathcal{N}(\theta, \sigma^2/N).$$

In this example, ML estimator = MVU estimator = BLUE.

Note: When estimation error cannot be made small as $N \nearrow +\infty$, the asymptotic pdf in (11) is invalid. For asymptotics to work, there has to be an averaging effect.

Example 7.7 in Kay-I: Estimation of the DC level in *fully dependent non-Gaussian noise*:

$$X[n] = a + W[n].$$

We measure $X[0], X[1], \dots, X[N-1]$ but all noise samples are the same:

$$W[0] = W[1] = \dots = W[N-1].$$

Hence, we can discard $X[1], X[2], \dots, X[N-1]$; then,

$$\hat{a} = X[0]$$

say. Clearly, the pdf of this \hat{a} remains non-Gaussian as $N \nearrow +\infty$ and, therefore, (11) cannot hold. Furthermore, \hat{a} does not

satisfy the consistency result (10), since

$$\text{var}_{\mathbf{X}}(\hat{a}) = \text{var}_X(X[0]) \rightarrow 0 \quad \text{as } N \nearrow +\infty.$$

ML Estimation: Vector Parameters

Conceptually, nothing really changes: the ML estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} f_{\mathbf{X} | \boldsymbol{\theta}}(\mathbf{X} | \boldsymbol{\theta}).$$

Under appropriate regularity conditions, this estimate is consistent and

$$\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N \mathcal{I}^{-1}(\boldsymbol{\theta}))$$

where $\mathcal{I}(\boldsymbol{\theta})$ is now the *Fisher information matrix*.

ML Estimation: Properties

Theorem 6. (ML Invariance Principle) *The ML estimate of $\alpha = g(\theta)$ where the pdf/pmf $f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ is parametrized by θ , is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

where $\hat{\theta}$ is the ML estimate of θ [obtained by maximizing $f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ with respect to θ].

Comments:

- For a more precise formulation, see Theorems 7.2 and 7.4 in Kay-I.
- Invariance is often combined with the *delta method* which we introduce later in this handout.

More properties:

- If a given scalar parameter θ has a single sufficient statistic $T(\mathbf{x})$, say, then the ML estimate of θ must be a function of $T(\mathbf{x})$. Furthermore, if $T(\mathbf{x})$ is minimal and complete, then the ML estimate is unique.
- **(Connection between ML and MVU estimation)** If the ML estimate is unbiased, then it is MVU.

Statistical Motivation

ML has a nice intuitive interpretation, but is it justifiable statistically? We now answer this question for the case of i.i.d. observations.

Lehmann shows the following result in Theorem 2.1, Chapter 6.2 of

E.L. Lehmann, *Theory of Point Estimation*, New York: Wiley, 1983.

Theorem 7. *Suppose that the random observations $X[n]$ are i.i.d. with common pdf $f_{X|\theta}(x[n]|\theta)$ [or pmf $p_{X|\theta}(x[n]|\theta)$] where the true value of the parameter θ is in the interior of the parameter space. Then, as $N \nearrow +\infty$*

$$\Pr_{X|\theta} \left\{ \prod_{n=0}^{N-1} f_{X|\theta}(X[n]|\theta) > \prod_{n=0}^{N-1} f_{X|\theta}(X[n]|\theta') \right\} \rightarrow 1$$

or

$$\Pr_{X|\theta} \left\{ \prod_{n=0}^{N-1} p_{X|\theta}(X[n]|\theta) > \prod_{n=0}^{N-1} p_{X|\theta}(X[n]|\theta') \right\} \rightarrow 1$$

for any fixed $\theta' \neq \theta$.

This theorem states that, for large number of i.i.d. samples (i.e. large N), the joint pdf/pmf of $X[0], X[1], \dots, X[N-1]$ at the true parameter value

$$\prod_{n=0}^{N-1} f_{X| \theta}(X[n] | \theta) \quad \text{or} \quad \prod_{n=0}^{N-1} p_{X| \theta}(X[n] | \theta)$$

exceeds with probability one the joint pdf/pmf of $X[0], X[1], \dots, X[N-1]$ at any other parameter value. Consequently, as the number of observations increases, the parameter estimate that maximizes the joint pdf/pmf of the measurements (i.e. the ML estimate) must become close to the true value.

Regularity Conditions for I.I.D. Observations in Theorem 5

Not one set of regularity conditions applies to all scenarios.

Here are some typical regularity conditions for the i.i.d. case. Consider i.i.d. $X[0], X[1], \dots, X[N-1]$ (given θ) following

$$f_{X|\theta}(x[n]|\theta), \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}.$$

Regularity conditions:

(i) θ is identifiable under the model

$$f_{X|\theta}(x|\theta)$$

and the support of $f_{X|\theta}(x|\theta)$ is not a function of θ ;

(ii) The true value of the parameter θ lies in an open subset of the parameter space Θ ;

(iii) For almost all x , the pdf $f_{X|\theta}(x|\theta)$ has continuous derivatives to order three with respect to all elements of θ and all values in the open subset of (ii);

(iv) The following are satisfied:

$$\mathbb{E}_{X|\theta} \left[\frac{\partial}{\partial \theta_k} \ln f_{X|\theta}(X|\theta) \mid \theta \right] = 0 \quad k = 1, 2, \dots, d$$

and

$$\begin{aligned} \mathcal{I}_{i,k}(\theta) &= \mathbb{E}_{X|\theta} \left[\frac{\partial}{\partial \theta_i} \ln f_{X|\theta}(X|\theta) \frac{\partial}{\partial \theta_k} \ln f_{X|\theta}(X|\theta) \mid \theta \right] \\ &= -\mathbb{E}_{X|\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_k} \ln f_{X|\theta}(X|\theta) \mid \theta \right] \quad i, k = 1, 2, \dots, d. \end{aligned}$$

(v) FIM $\mathcal{I}(\theta) = [\mathcal{I}(\theta)]_{i,k}$ is positive definite;

(vi) Bounding functions $m_{i,k,l}(\cdot)$ exist such that

$$\left| \frac{\partial^3}{\partial \theta_i \partial \theta_k \partial \theta_l} \ln f_{X|\theta}(x|\theta) \right| \leq m_{i,k,l}(x)$$

for all θ in the open subset of (ii), and

$$\mathbb{E}_{X|\theta} [m_{i,k,l}(X) \mid \theta] < \infty.$$

Theorem 8. (\approx same as **Theorem 5**) *If*

$$X[0], X[1], \dots, X[N - 1]$$

are i.i.d. (conditional on $\boldsymbol{\theta}$) with pdf

$$f_{X|\boldsymbol{\theta}}(x[n] | \boldsymbol{\theta})$$

such that the conditions (i)–(vi) hold, then there exists a sequence of solutions $\{\hat{\boldsymbol{\theta}}_N\}$ to the likelihood equations such that

(1) $\hat{\boldsymbol{\theta}}_N$ *is consistent for $\boldsymbol{\theta}$:*

$$\hat{\boldsymbol{\theta}}_N \rightarrow \boldsymbol{\theta} \quad (\text{with probability 1}) \quad (\text{consistency});$$

(2)

$$\sqrt{N} (\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N \mathcal{I}^{-1}(\boldsymbol{\theta})) \quad (\text{asymptotic efficiency});$$

(3)

$$\sqrt{N} ([\hat{\boldsymbol{\theta}}_N]_i - \theta_i) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N [\mathcal{I}^{-1}(\boldsymbol{\theta})]_{i,i})$$

for $i = 1, 2, \dots, d$.

Comments. What we are *not* given:

- (a) uniqueness of $\hat{\theta}_N$;
- (b) existence for all $x[0], x[1], \dots, x[N - 1]$;
- (c) even if the solution exists and is unique, that we can find it.

An Array-processing Example

$$\mathbf{X}[n] = A(\boldsymbol{\phi}) \mathbf{s}[n] + \mathbf{W}[n] \quad n = 0, 1, \dots, N - 1$$

where

$$\boldsymbol{\theta} = [\boldsymbol{\phi}^T, \sigma^2, \mathbf{s}[0]^T, \mathbf{s}[1]^T, \dots, \mathbf{s}[N - 1]^T]^T$$

is the vector of unknown parameters and $\mathbf{W}[n]$ is WGN.

Note:

- $\mathbf{X}[n]$ are not *i.i.d.* and, therefore, the conditions on pp. 31–33 do not apply;
- the size of $\boldsymbol{\theta}$ grows with N (trouble).

It is well known that CRB cannot be attained asymptotically in this case, see

P. Stoica and A. Nehorai, “MUSIC, maximum likelihood and Cramér-Rao bound,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 720-741, May 1989.

What if

$$\mathbf{X}[n] = A(\boldsymbol{\phi}) \mathbf{S}[n] + \mathbf{W}[n] \quad n = 0, 1, \dots, N - 1$$

where $\mathbf{S}[n]$ are *random* $\sim \mathcal{N}(\mathbf{0}, \Gamma)$ *conditional on* Γ ? Then, $\mathbf{X}[n]$, $n = 0, 1, \dots, N - 1$ are *i.i.d.* given $\boldsymbol{\phi}, \Gamma$, and σ^2 :

$$\{\mathbf{X}[n] \mid \underbrace{\boldsymbol{\phi}, \Gamma, \sigma^2}_{\text{parameters}}\} \sim \mathcal{N}(\mathbf{0}, A(\boldsymbol{\theta})\Gamma A^T(\boldsymbol{\theta}) + \sigma^2 I).$$

Here, the number of parameters *does not grow with* N . If the regularity conditions that we stated for the i.i.d. case hold, CRB will be attained asymptotically. Furthermore, CRB for $\boldsymbol{\phi}$ will be different (smaller) than CRB for $\boldsymbol{\phi}$ when $s[n]$ is deterministic.

Digression: Delta Method

Theorem 9. (Gauss Approximation Formula, Delta Method)

Assume $\alpha = g(\theta)$ has bounded derivatives up to the second order. Then, if $\hat{\theta}$ is consistent, so is

$$\hat{\alpha} = g(\hat{\theta}).$$

Moreover, the asymptotic covariance matrices

$$\text{COV}_{\mathbf{X} | \theta}(\hat{\theta})$$

and

$$\text{COV}_{\mathbf{X} | \alpha}(\hat{\alpha})$$

are asymptotically equal to the corresponding MSE matrices (due to consistency) and are related as follows:

$$\text{COV}_{\mathbf{X} | \alpha}(\hat{\alpha}) = \frac{\partial g}{\partial \theta^T} \text{COV}_{\mathbf{X} | \theta}(\hat{\theta}) \frac{\partial g^T}{\partial \theta}.$$

Proof. Follows from the Taylor expansion around the true value $\alpha = g(\theta)$:

$$\hat{\alpha} = g(\theta) + \frac{\partial g(\theta)}{\partial \theta^T} (\hat{\theta} - \theta) + o(\|\hat{\theta} - \theta\|).$$

□

Example: Amplitude and Phase Estimation

Assume

$$X[n] = a \cos(\omega_0 n + \phi) + W[n] \quad n = 0, 1, \dots, N - 1$$

where ω_0 is a known constant and $W[n]$ is additive white Gaussian noise (AWGN). We wish to estimate the amplitude a and phase ϕ .

Rewrite this model as a linear model:

$$\mathbf{X} = \begin{bmatrix} X[0] \\ \dots \\ X[N - 1] \end{bmatrix} = \mathbf{H} \boldsymbol{\theta} + \mathbf{W}$$

where $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ and

$$H_{i,1} = \cos[\omega_0 (i - 1)], \quad i = 1, 2, \dots, N$$

$$H_{i,2} = \sin[\omega_0 (i - 1)], \quad i = 1, 2, \dots, N$$

$$(A \cos \phi, -A \sin \phi) \leftrightarrow (\theta_1, \theta_2)$$

We have

$$\{\hat{\boldsymbol{\theta}} | \mathbf{X}\} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1}).$$

By the ML invariance principle, \hat{a} and $\hat{\phi}$ can be found from $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2]^T$ via rectangular-to-polar coordinate conversion:

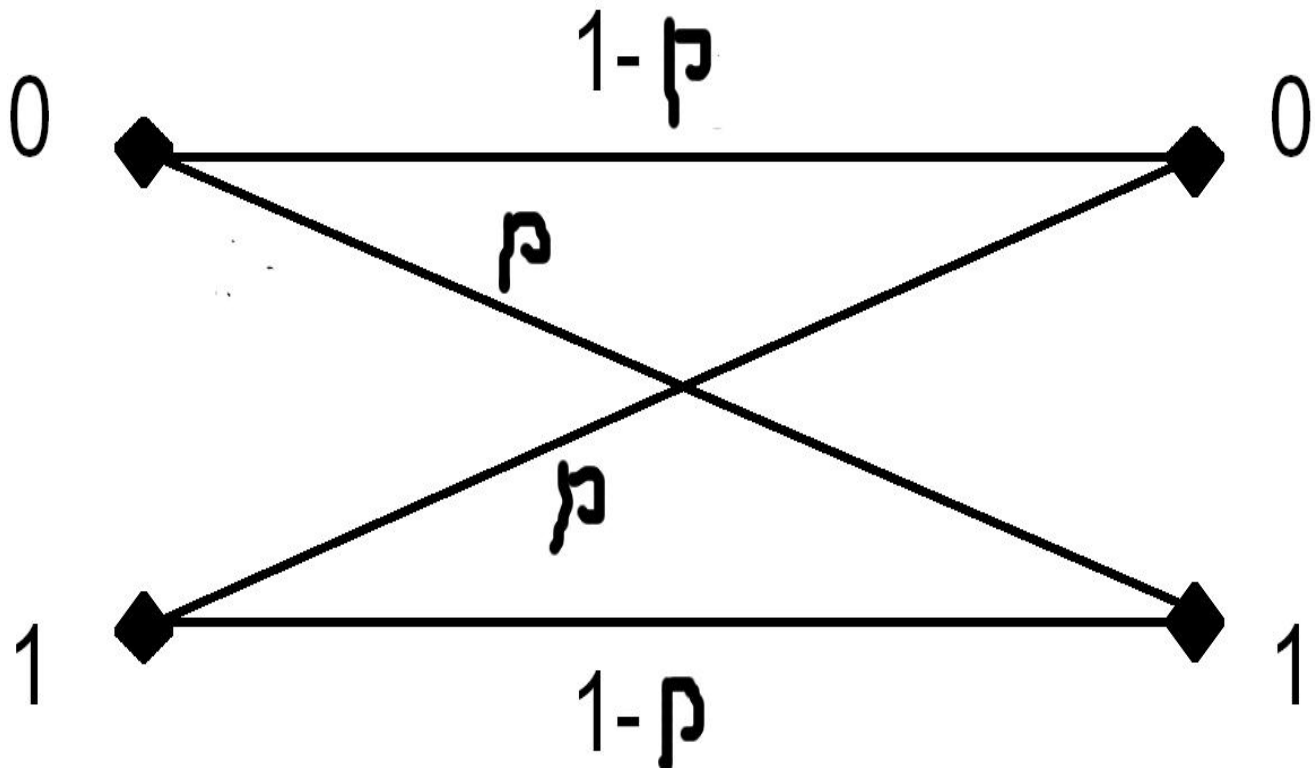
$$(\hat{\theta}_1, \hat{\theta}_2) \leftrightarrow (\hat{A} \cos \hat{\phi}, -\hat{A} \sin \hat{\phi}).$$

Define $\boldsymbol{\alpha} = [a, \phi]^T = \mathbf{g}(\boldsymbol{\theta})$. Then, the delta method yields

$$\text{cov}_{\mathbf{X}|\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}^T} \text{cov}_{\mathbf{X}|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \frac{\partial \mathbf{g}^T}{\partial \boldsymbol{\theta}}.$$

Here, $\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$ and $\text{cov}_{\mathbf{X}|\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}})$ are asymptotic covariance (and MSE) matrices of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$.

Example: ML Decoding



For a symmetric channel, the ML decoder is the minimum Hamming distance decoder.

Proof. \mathbf{x} and $\boldsymbol{\theta}$ are the received and transmitted vectors from a binary symmetric channel. The elements of \mathbf{x} and $\boldsymbol{\theta}$ are zeros and ones. Note that $\boldsymbol{\theta}$ belongs to a finite set of codewords, denoted by Θ . We wish to find which $\boldsymbol{\theta}$ was transmitted based on the received \mathbf{x} . We have

$$\{\mathbf{X} | \boldsymbol{\theta}\} = \boldsymbol{\theta} + \mathbf{W} \pmod{2} \triangleq \boldsymbol{\theta} \oplus \mathbf{W}$$

where $\mathbf{W} = [W_1, \dots, W_N]^T$ and W_i are i.i.d. Bernoulli random variables taking value 1 with probability p :

$$\Pr\{W_i = 1\} = p.$$

The likelihood function of $\boldsymbol{\theta}$ for data \boldsymbol{x} is given by

$$\begin{aligned} p_{\mathbf{X}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta}) &= \Pr\{\mathbf{X} = \boldsymbol{x}\} = \Pr\{\boldsymbol{\theta} \oplus \mathbf{W} = \boldsymbol{x}\} \\ &= \Pr\{\mathbf{W} = \boldsymbol{x} \oplus \boldsymbol{\theta}\} \\ &= p^{\sum_{i=1}^N x_i \oplus \theta_i} \cdot (1-p)^{N - \sum_{i=1}^N x_i \oplus \theta_i} \\ &= \left(\frac{p}{1-p}\right)^{d_H(\boldsymbol{x}, \boldsymbol{\theta})} (1-p)^N \end{aligned}$$

where

$$d_H(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^N x_i \oplus \theta_i$$

is the Hamming distance between \boldsymbol{x} and $\boldsymbol{\theta}$, i.e. the number of bits that are different between the two vectors. Hence, if $p < 0.5$, then

maximizing $p_{\mathbf{X}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$

is equivalent to

minimizing $d_H(\boldsymbol{x}, \boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$.

□

Computing ML Estimates

Finding the ML estimate typically requires a nonlinear d -dimensional optimization (for d -dimensional $\boldsymbol{\theta}$). More generally,

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$$

where, for ML estimation

$$V(\boldsymbol{\theta}) = -\ln f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}).$$

Newton-Raphson Iteration: Assume that a guess $\boldsymbol{\theta}^{(i)}$ is available. We wish to improve $\boldsymbol{\theta}^{(i)}$, yielding $\boldsymbol{\theta}^{(i+1)}$. Apply quadratic Taylor expansion:

$$V(\boldsymbol{\theta}) \approx V(\boldsymbol{\theta}^{(i)}) + \mathbf{g}_i^T \bar{\boldsymbol{\theta}}^{(i)} + \frac{1}{2} (\bar{\boldsymbol{\theta}}^{(i)})^T H_i \bar{\boldsymbol{\theta}}^{(i)}$$

where

$$\bar{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta} - \boldsymbol{\theta}^{(i)} \quad (12)$$

$$\mathbf{g}_i = \left. \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} \quad (13)$$

$$H_i = \left. \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} = - \left. \frac{\partial^2 \ln f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}. \quad (14)$$

Newton-Raphson Iteration (Ch. 7.7 in Kay-I)

Complete the squares:

$$V(\boldsymbol{\theta}) \approx (\bar{\boldsymbol{\theta}}^{(i)} + H_i^{-1} \mathbf{g}_i)^T \frac{1}{2} H_i (\bar{\boldsymbol{\theta}}^{(i)} + H_i^{-1} \mathbf{g}_i) + \text{const.}$$

We assume that $H_i > 0$

Hessian matrix of $V(\boldsymbol{\theta})$

(i.e. the second derivative of $V(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, computed at $\boldsymbol{\theta}_i$, is positive definite)

and thus choose

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - H_i^{-1} \mathbf{g}_i.$$

Newton-Raphson iteration achieves *quadratic convergence* near the optimum, i.e.

$$\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}\|_{\ell_2} \leq c \|\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}\|_{\ell_2}^2$$

where c is a positive constant. (This, of course, holds if the underlying Taylor approximation is good.) Therefore, we gain approximately one significant digit per iteration.

However, the algorithm can diverge if we start too far from the optimum. To facilitate convergence (to a *local optimum*, in

general), we can apply a damped Newton-Raphson algorithm. Here is one such damped algorithm:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i H_i^{-1} \mathbf{g}_i \quad (15)$$

where the step length μ_i is $\mu_i = 1, 1/2, 1/4, \dots$. In particular, in the i th iteration, start with the step length $\mu_i = 1$, compute $\boldsymbol{\theta}^{(i+1)}$ using (15), and check if

$$V(\boldsymbol{\theta}^{(i+1)}) < V(\boldsymbol{\theta}^{(i)})$$

holds; if yes, go to the $(i+1)$ st iteration. If no, keep halving μ_i (or we may apply some smarter update of μ_i) and recomputing $\boldsymbol{\theta}^{(i+1)}$ using (15) until

$$V(\boldsymbol{\theta}^{(i+1)}) < V(\boldsymbol{\theta}^{(i)})$$

is first satisfied; then, go to the $(i+1)$ st iteration. Once in the $(i+1)$ st iteration, reset $\mu^{(t+1)}$ to 1 and continue in the same manner.

Modification: Use an approximate form of the Hessian matrix of $V(\boldsymbol{\theta})$. In the case of ML estimation, the following algorithm is particularly popular:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i H_i^{-1} \mathbf{g}_i$$

where

$$H_i = \mathcal{I}(\boldsymbol{\theta}^{(i)})$$

i.e. we use FIM instead of the Hessian (14). The resulting algorithm is called *Fisher scoring*. This choice of H_i guarantees positive semidefiniteness of H_i , since $\mathcal{I}(\boldsymbol{\theta}^{(i)}) > 0$. It can be written as

$$\begin{aligned}\boldsymbol{\theta}^{(i+1)} &= \boldsymbol{\theta}^{(i)} - \mu_i \mathcal{I}^{-1}(\boldsymbol{\theta}^{(i)}) \mathbf{g}_i \\ &= \boldsymbol{\theta}^{(i)} + \mu_i \mathcal{I}^{-1}(\boldsymbol{\theta}^{(i)}) \left. \frac{\partial \ln f_{\mathbf{X} | \boldsymbol{\theta}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}.\end{aligned}$$

Note: The convergence point is a *local* minimum of $V(\boldsymbol{\theta})$. It is the global minimum if $V(\boldsymbol{\theta})$ is a unimodal function of $\boldsymbol{\theta}$ or if the initial estimate is sufficiently good. If (we suspect that) there are multiple local minima of $V(\boldsymbol{\theta})$ (i.e. multiple local maxima of the likelihood function), we should try many (wide-spread/different) starting values for our Newton-Raphson or Fisher-scoring iterations and select as our (best guess of the) ML estimate the convergence point that yields the largest log likelihood.

If the parameter space Θ is not \mathbf{R}^p , we should also examine the boundary of the parameter space to see if the global maximum of the likelihood function lies on this boundary. **Pay attention to this issue.**

Fisher Scoring: An Example

Consider the following model:

$$X[n] = s[n; \boldsymbol{\theta}] + W[n] \quad n = 0, 1, \dots, N-1$$

where $W[n]$ is white Gaussian noise with known variance σ^2 ,

$$s[n; \boldsymbol{\theta}] = \sin(\omega_1 n) + \sin(\omega_2 n)$$

and the unknown parameter vector is

$$\boldsymbol{\theta} = [\omega_1, \omega_2]^T.$$

The negative log-likelihood function of $\boldsymbol{\theta}$ is

$$V(\boldsymbol{\theta}) = -\ln f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}])^2 + \text{const}$$

where const denotes terms that do not depend on $\boldsymbol{\theta}$. Now, the negative score function of $\boldsymbol{\theta}$ is

$$V'(\boldsymbol{\theta}) = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n, \boldsymbol{\theta}]) \cdot \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}}$$

and FIM is

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \frac{\partial s[n; \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}^T}.$$

We can use $H = \mathcal{I}(\boldsymbol{\theta})$ and the damped Fisher-scoring iteration becomes

$$\begin{aligned} \boldsymbol{\theta}^{(i+1)} &= \boldsymbol{\theta}^{(i)} - \mu_i \mathcal{I}(\boldsymbol{\theta}^{(i)})^{-1} V'(\boldsymbol{\theta}^{(i)}) \\ &= \boldsymbol{\theta}^{(i)} \\ &+ \mu_i \left\{ \sum_{n=0}^{N-1} n^2 \begin{bmatrix} \cos^2(\omega_1^{(i)} n) & \cos(\omega_1^{(i)} n) \cos(\omega_2^{(i)} n) \\ \cos(\omega_1^{(i)} n) \cos(\omega_2^{(i)} n) & \cos^2(\omega_2^{(i)} n) \end{bmatrix} \right\}^{-1} \\ &\cdot \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}^{(i)}]) \begin{bmatrix} n \cos(\omega_1^{(i)} n) \\ n \cos(\omega_2^{(i)} n) \end{bmatrix}. \end{aligned}$$

Concentrated Likelihood: Example

Consider a situation in which a small-scale disease epidemic has been observed, with individuals exposed to the disease (e.g. virus) at a common place and time. Or, in a similar computer-analogous scenario, consider computers infected by a virus. We assume that a time interval is known for exposure, but not the exact time.

We collect times at which infection was detected at various computers ('incubation times'), say, with time 0 corresponding to the start of a known interval in which exposure occurred. The collected infection times after the exposure are

$$x[0], x[1], \dots, x[N - 1].$$

We model

$$X[0], X[1], \dots, X[N - 1]$$

as i.i.d. given $\boldsymbol{\theta}$, following

$$f_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta}) = \begin{cases} \frac{1}{(x-\alpha)\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}[\ln(x-\alpha) - \mu]^2\right\}, & x > \alpha \\ 0, & \text{otherwise} \end{cases}$$

with parameters

$$\boldsymbol{\theta} = [\alpha, \mu, \sigma]^T.$$

where the parameter $\alpha > 0$ represents the time at which the exposure took place. Since the support of the above distribution *depends on the parameter α* , regularity condition (i) on p. 31 does not hold.

Some references on the above model:

H.L. Harter and A.H. Moore, “Local-maximum-likelihood estimation of the parameters of three-parameter lognormal populations from complete and censored samples,” *J. Amer. Stat. Assoc.*, vol. 61, pp. 842–851, Sept. 1966.

B.M. Hill, “The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic,” *J. Amer. Stat. Assoc.*, vol. 68, pp. 72–84, Mar. 1963.

Note: this model is equivalent to $X[n]$ $n = 0, 1, \dots, N - 1$ satisfying

$$\{\ln(X[n] - \alpha) \mid \boldsymbol{\theta}\} \sim \mathcal{N}(\mu, \sigma^2) \quad (16)$$

which is useful for simulating the data from the above model, as well as for finding the ML estimates of μ and σ^2 (see the discussion below).

The log-likelihood function of $\boldsymbol{\theta}$ for the data $\boldsymbol{x} =$

$[x[0], x[1], \dots, x[N-1]]^T$ is

$$\begin{aligned}
 l(\boldsymbol{\theta}) &= \ln f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \ln f_{X|\boldsymbol{\theta}}(x[n]|\boldsymbol{\theta}) \\
 &= -\frac{N}{2} \ln(2\pi\sigma^2) - \sum_{n=0}^{N-1} \ln(x[n] - \alpha) \\
 &\quad - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \{ \ln(x[n] - \alpha) - \mu \}^2 \quad (17)
 \end{aligned}$$

where

$$x[n] > \alpha \quad \forall n = 0, 1, \dots, N-1.$$

For a fixed α , we can easily find μ and σ^2 that maximize (17):

$$\hat{\mu}(\alpha) = \frac{1}{N} \sum_{n=0}^{N-1} \ln(x[n] - \alpha) \quad (18)$$

$$\hat{\sigma}^2(\alpha) = \frac{1}{N} \sum_{n=0}^{N-1} \{ \ln(x[n] - \alpha) - \hat{\mu}(\alpha) \}^2 \quad (19)$$

which can be obtained directly by maximizing (17) or from (16), using our knowledge of the ML estimates of the mean and variance of i.i.d. Gaussian measurements. Now, we can *concentrate* the log-likelihood function of $\boldsymbol{\theta}$ with respect to μ

and σ^2 by substituting (18) and (19) into (17):

$$\underbrace{l([\alpha, \hat{\mu}(\alpha), \hat{\sigma}^2(\alpha)]^T)}_{\text{concentrated log-likelihood function of } \alpha}$$
$$= -\frac{N}{2} \ln[2\pi \hat{\sigma}^2(\alpha)] - \sum_{n=0}^{N-1} \ln(x[n] - \alpha) - \frac{N}{2}. \quad (20)$$

In statistics, concentrated likelihood function is sometimes referred to as *profile likelihood*.

Summary of Basic Classical Estimation

The CRB is a lower bound on the covariance of all unbiased estimators of an unknown parameter vector (information inequality). The information inequality holds only if the regularity conditions from p. 3 of handout # 2 hold; regularity condition (i) is easy to check.

The information-inequality theorem also gives us estimator that attains the bound, if it exists. Estimator that attains CRB is termed *efficient*. Efficient estimator \implies MVU estimator. If an efficient estimator exists, it coincides with the ML estimator.

Typically, MVU estimator does not exist. We have also seen that unbiasedness is often too restrictive and not needed.

CRB is generally used as a

- measure of the potential performance attainable from the system,
- benchmark for assessing algorithm performance,
- measure for system design,

- tool for constructing confidence regions for the unknown parameters,
- tool for constructing noninformative priors in Bayesian applications.

Under certain regularity conditions, CRB is attained *asymptotically* (for large numbers of measurements) by the ML estimator. Hence, ML method is *asymptotically efficient*.

Least-Squares (LS) Approach to Estimation

Consider the following signal model:

$$\mathbf{X} = H \boldsymbol{\theta} + \mathbf{W} \quad (21)$$

where $\mathbf{X} = [x[0], \dots, x[N-1]]^T$ is the vector of measurements, H is a known *regression vector matrix*, and \mathbf{W} is “error” vector.

LS problem formulation:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{x} - H \boldsymbol{\theta}\|_{\ell_2}^2.$$

Solution:

$$\hat{\boldsymbol{\theta}} = (H^T H)^{-1} H^T \mathbf{x}.$$

We can also use weighted least squares, which allows us to assign different weights to measurements. For example, if

$$E_{\mathbf{w}}(\mathbf{W} \mathbf{W}^T) = C$$

is known, we could use

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{x} - H \boldsymbol{\theta}\|_{C^{-1}}^2 = \arg \min_{\boldsymbol{\theta}} (\mathbf{x} - H \boldsymbol{\theta})^H C^{-1} (\mathbf{x} - H \boldsymbol{\theta})$$

Define $H = [\mathbf{h}_1 \cdots \mathbf{h}_p]$ and rewrite (21) as

$$\mathbf{X} = \sum_{k=1}^p \mathbf{h}_k \theta_k + \mathbf{W}.$$

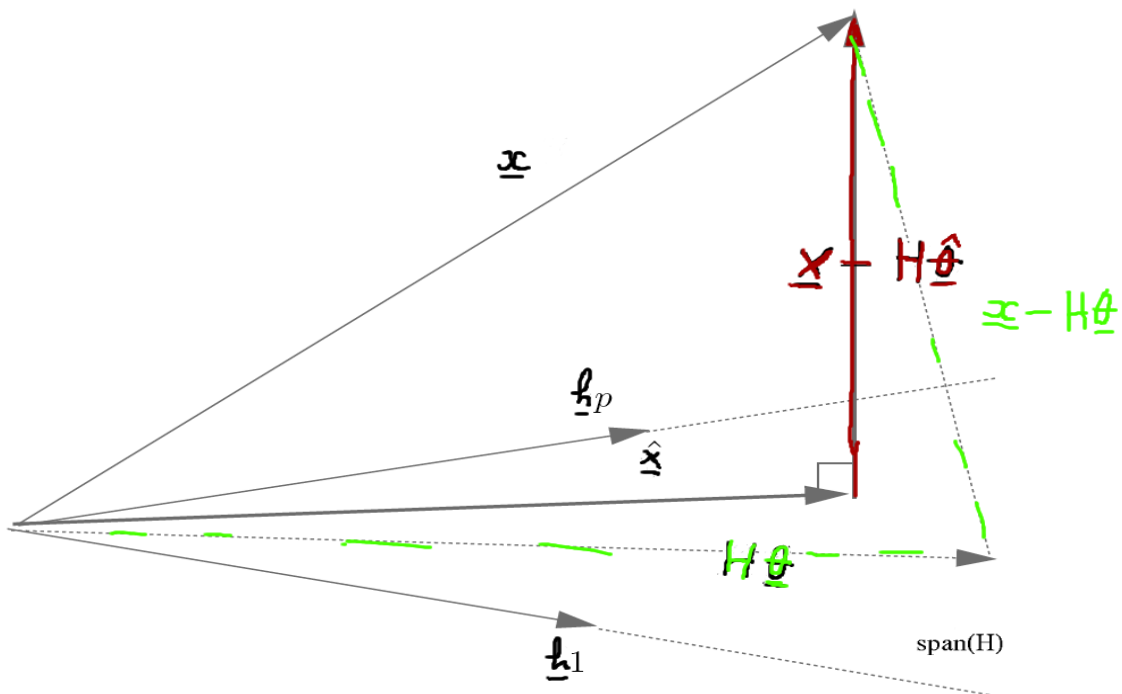
The “signal part” $H\theta$ of the N -vector \mathbf{X} is confined to the p -dimensional subspace spanned by $[\mathbf{h}_1 \cdots \mathbf{h}_p]$, the “signal subspace”. The signal estimate

$$\hat{\mathbf{x}} = H\hat{\theta} = \underbrace{H(H^T H)^{-1} H^T}_{P} \mathbf{x} = P \mathbf{x}$$

is the *orthogonal projection* of \mathbf{x} onto $\text{span}(H)$, the column space of H . The error

$$\hat{\mathbf{w}} = \mathbf{x} - H(H^T H)^{-1} H^T \mathbf{x}$$

is the *orthogonal projection* of \mathbf{x} onto the *orthogonal complement* of $\text{span}(H)$.



Note:

$$P = H (H^T H)^{-1} H^T$$

is the projection matrix onto the column space of H , and

$$P^\perp = I - P$$

is the complementary projection matrix.

Recall: Real-valued projection matrices are symmetric and idempotent:

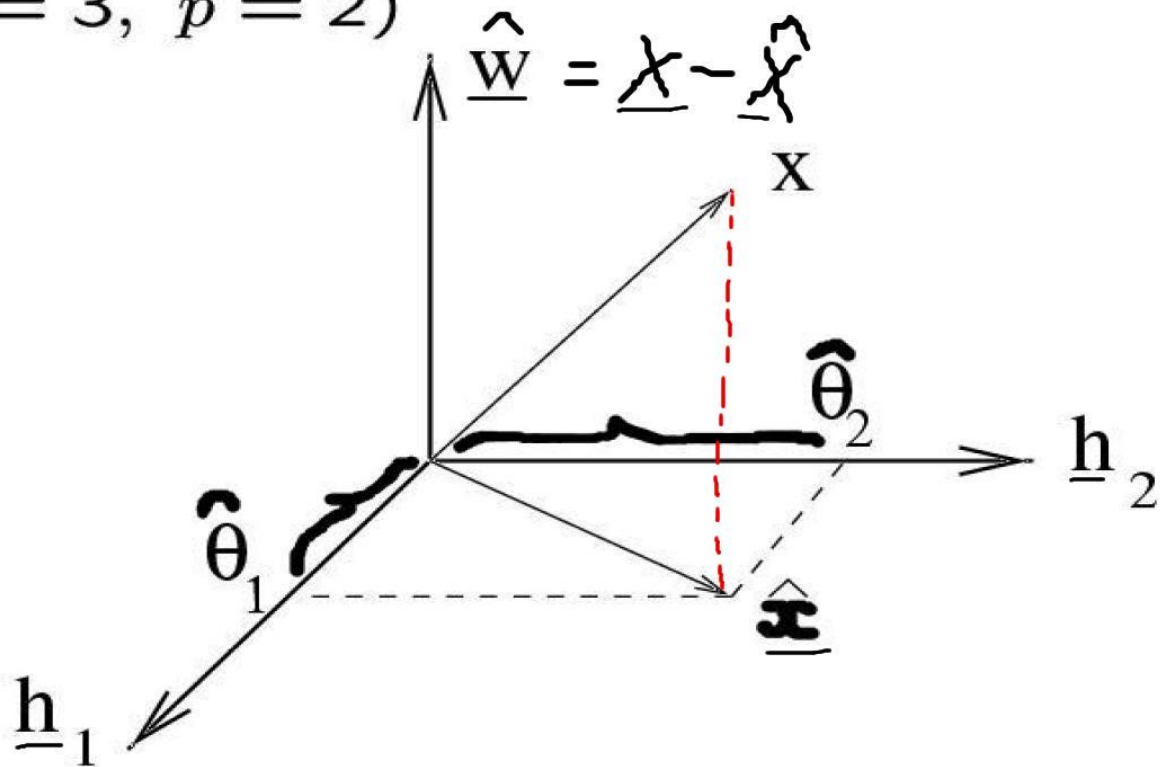
$$P = P^T = P^2.$$

Furthermore,

$$(\mathbf{x} - \underbrace{H \hat{\boldsymbol{\theta}}}_{\hat{\mathbf{x}}})^T H = \mathbf{0}.$$

Example:

$(N = 3, p = 2)$



Obviously, $(\mathbf{x} - \hat{\mathbf{x}}) \perp \{\mathbf{h}_1, \mathbf{h}_2\}$. This is an *orthogonality principle*: the minimum error is orthogonal to the columns of H (called regressors in statistics).

In general,

$$\mathbf{x} - \hat{\mathbf{x}} \perp \text{span}(H) \Leftrightarrow \mathbf{x} - \hat{\mathbf{x}} \perp \mathbf{h}_j, \forall \mathbf{h}_j \Leftrightarrow H^T(\mathbf{x} - H\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Computational Aspects of Least Squares

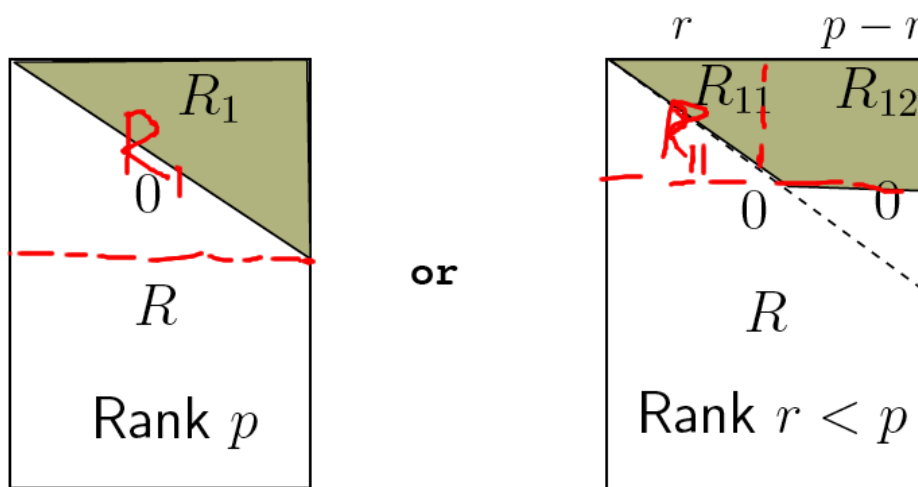
QR decomposition of H :

$$H_{N \times p} = Q_{N \times N} R_{N \times p}$$

$$= \begin{array}{|c|c|} \hline Q_1 & Q_2 \\ \hline \end{array} \begin{array}{|c|} \hline 0 \\ \hline R \\ \hline \end{array}$$

where Q has orthonormal columns: $Q^T Q = I$ (and rows, i.e. $Q Q^T = I$).

R is upper triangular, and may not have full rank:



Full-rank case,

$$\begin{aligned}\|\mathbf{x} - H\boldsymbol{\theta}\|_{\ell_2}^2 &= \|Q^T \mathbf{x} - R\boldsymbol{\theta}\|_{\ell_2}^2 \\ &= \|Q_1^T \mathbf{x} - R_1 \boldsymbol{\theta}\|_{\ell_2}^2 + \|Q_2^T \mathbf{x}\|_{\ell_2}^2 \\ \implies \hat{\boldsymbol{\theta}} &= R_1^{-1} Q_1^T \mathbf{x}.\end{aligned}$$

Comments:

- $Q^T \mathbf{x}$ yields coordinates of \mathbf{x} on columns of Q .
- $\hat{\mathbf{x}} = Q_1 Q_1^T \mathbf{x} = P \mathbf{x} = H (H^T H)^{-1} H^T \mathbf{x}$. Here, the projection matrix P is also known as the *hat matrix* (because it puts the hat on \mathbf{x}).
- Non-full-rank case: $\text{rank}(H) = r < p$. We need to solve

$$Q_1^T \mathbf{x} = R_{11} \boldsymbol{\theta}_1 + R_{12} \boldsymbol{\theta}_2$$

where Q_1 has r columns. There are infinitely many solutions: to get one, arbitrarily set $\boldsymbol{\theta}_2 = \mathbf{0}_{(p-r) \times 1}$ and solve for $\boldsymbol{\theta}_1$: $\boldsymbol{\theta}_1 = R_{11}^{-1} Q_1^T \mathbf{x}$. Here,

$$\hat{\mathbf{x}} = Q_1 Q_1^T \mathbf{x}$$

is still well defined and unique.

Nonlinear Least Squares (NLLS)

Often, the signal is *not* a linear function of $\boldsymbol{\theta}$, say $\mathbf{f}(\boldsymbol{\theta})$. Then, we obtain a NLLS estimate of $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) \\ V(\boldsymbol{\theta}) &= \|\mathbf{x} - \mathbf{f}(\boldsymbol{\theta})\|^2.\end{aligned}$$

Example:

$$s[n] = r \cos(\omega n + \phi) \quad n = 0, 1, 2, \dots, N - 1$$

gives

$$\mathbf{f}(r, \omega, \phi) = [r \cos(\phi), \dots, r \cos((N - 1)\omega + \phi)]^T.$$

This is a nonlinear problem and we need iterative optimization.

Recall the damped Newton-Raphson iteration:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} - \mu_k \cdot H_k^{-1} \mathbf{g}_k$$

where μ_k is the step length and H_k , \mathbf{g}_k are the Hessian and gradient of $V(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}^{(k)}$.

Nonlinear Least Squares Newton-Raphson Iteration

Define

$$\mathbf{f}_\theta^{(k)} = \left. \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}, \quad \mathbf{f}^{(k)} = \mathbf{f}(\boldsymbol{\theta}^{(k)}).$$

The needed partial derivatives are then

$$\mathbf{g}_k = \left. \frac{\partial (\mathbf{x} - \mathbf{f})^T (\mathbf{x} - \mathbf{f})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = -2(\mathbf{f}_\theta^{(k)})^T (\mathbf{x} - \mathbf{f}^{(k)})$$
$$H_k = \left. \frac{\partial^2 (\mathbf{x} - \mathbf{f})^T (\mathbf{x} - \mathbf{f})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = 2(\mathbf{f}_\theta^{(k)})^T \mathbf{f}_\theta^{(k)} - 2G^{(k)}$$

where

$$[G^{(k)}]_{i,l} = \left. \frac{\partial^2 \mathbf{f}^T}{\partial \theta_i \partial \theta_l} (\mathbf{x} - \mathbf{f}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}.$$

Assuming that we are close to the optimum and residuals are small:

$$\mathbf{x} - \mathbf{f}^{(k)} \approx \mathbf{0}$$

the Hessian is approximated by

$$H_k = 2 (\mathbf{f}_\theta^{(k)})^T \mathbf{f}_\theta^{(k)}.$$

Recall that

$$\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \mathbf{f}_\theta^T \mathbf{f}_\theta \quad \mathbf{f}_\theta \triangleq \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$$

is FIM for θ under the AWGN measurement model, hence this approach is equivalent to Fisher scoring when the noise is AWGN. It is also known as the *Gauss-Newton algorithm*.

Nonlinear Least Squares (cont.)

(Damped) Gauss-Newton:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} + \mu_k [(\mathbf{f}_{\boldsymbol{\theta}}^{(k)})^T (\mathbf{f}_{\boldsymbol{\theta}}^{(k)})]^{-1} (\mathbf{f}_{\boldsymbol{\theta}}^{(k)})^T (\mathbf{x} - \mathbf{f}^{(k)}).$$

The search direction

$$\boldsymbol{\gamma} = [(\mathbf{f}_{\boldsymbol{\theta}}^{(k)})^T (\mathbf{f}_{\boldsymbol{\theta}}^{(k)})]^{-1} (\mathbf{f}_{\boldsymbol{\theta}}^{(k)})^T (\mathbf{x} - \mathbf{f}^{(k)})$$

is LS solution to

$$\min_{\boldsymbol{\gamma}} \|(\mathbf{x} - \mathbf{f}^{(k)}) - \mathbf{f}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{\gamma}^{(k)}\|^2$$

which is efficiently computed in MATLAB using

$$\boldsymbol{\gamma} = \mathbf{f}_{\boldsymbol{\theta}}^{(k)} \backslash (\mathbf{x} - \mathbf{f}^{(k)}).$$

Note that the approximate Hessians $\mathbf{f}_{\boldsymbol{\theta}}^T \mathbf{f}_{\boldsymbol{\theta}}$ and $(\mathbf{f}_{\boldsymbol{\theta}}^{(k)})^T (\mathbf{f}_{\boldsymbol{\theta}}^{(k)})$ are always positive semidefinite, which is generally *not true for the exact Hessians*, and may cause trouble in an early iteration k where the parameter estimate is far from the optimum.

Separable NLLS

Consider the sinusoid example

$$s[n] = r \cos(\omega n + \phi) = A \sin(\omega n) + B \cos(\omega n).$$

A and B enter linearly in $s[n]$. We can write

$$\mathbf{f}(\boldsymbol{\theta}) = H(\underbrace{\boldsymbol{\alpha}}_{\omega}) \underbrace{\boldsymbol{\beta}}_{\begin{bmatrix} A \\ B \end{bmatrix}}$$

where $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}$.

Motivated by the above example, we now consider the *separable (partly linear) NLLS problem*:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \|\mathbf{x} - H(\boldsymbol{\alpha}) \boldsymbol{\beta}\|_{\ell_2}^2.$$

For a fixed $\boldsymbol{\alpha}$, the LS solution for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = [H^T(\boldsymbol{\alpha})H(\boldsymbol{\alpha})]^{-1}H^T(\boldsymbol{\alpha})\mathbf{x}.$$

Substituting $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ into $V(\boldsymbol{\theta})$ gives the concentrated criterion:

$$V_c(\boldsymbol{\alpha}) = \|\mathbf{x} - \underbrace{H(\boldsymbol{\alpha})[H^T(\boldsymbol{\alpha})H(\boldsymbol{\alpha})]^{-1}H^T(\boldsymbol{\alpha})}_{P(\boldsymbol{\alpha})}\mathbf{x}\|_{\ell_2}^2$$

where $P(\boldsymbol{\alpha})$ is the projection matrix onto the column space of $H(\boldsymbol{\alpha})$. Equivalently

$$\hat{\boldsymbol{\alpha}} = \arg \max_{\boldsymbol{\alpha}} \mathbf{x}^T \underbrace{H(\boldsymbol{\alpha})[H^T(\boldsymbol{\alpha})H(\boldsymbol{\alpha})]^{-1}H^T(\boldsymbol{\alpha})}_{P(\boldsymbol{\alpha})} \mathbf{x}.$$

Here, $\hat{\boldsymbol{\alpha}}$ maximizes the magnitude of the projection of \mathbf{x} onto the signal subspace.

We have used the fact that our cost function can be easily minimized with respect to a subset of parameters ($\boldsymbol{\beta}$, in our case) if the rest of the parameters $\boldsymbol{\alpha}$ are fixed. We have obtained a *concentrated cost function* to be maximized with respect to $\boldsymbol{\alpha}$ only.

Comments

- There is nothing fundamentally statistical about least squares: the least squares approach solves a minimization problem in vector spaces.
- In linear problems, least squares approach allows a closed-form solution.
- We need to replace T with H (the Hermitian transpose) to obtain the corresponding results for complex data:

$$\hat{\beta}(\alpha) = [H^H(\alpha)H(\alpha)]^{-1}H^H(\alpha)x$$

minimizes

$$\|x - H(\alpha)\beta\|_{\ell_2}^2 \triangleq [x - H(\alpha)\beta]^H [x - H(\alpha)\beta]$$

i.e.

$$\hat{\beta}(\alpha) = \arg \min_{\beta} \|x - H(\alpha)\beta\|_{\ell_2}^2$$

and α can be estimated using the concentrated criterion:

$$\hat{\alpha} = \arg \max_{\alpha} \underbrace{x^H H(\alpha) [H^H(\alpha)H(\alpha)]^{-1} H^H(\alpha) x}_{P(\alpha)}$$