LINEAR MODELS

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

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

Suppose that we are given $\{x(t_n)\}_{n=0}^{N-1}$. Define

$$\begin{aligned}
\boldsymbol{x} &= [x(t_0), \ldots, x(t_{N-1})]^T \\
\boldsymbol{w} &= [w(t_0), \ldots, w(t_{N-1})]^T \\
\boldsymbol{\theta} &= [\theta_1, \ldots, \theta_p]^T \\
\boldsymbol{H} &= \begin{bmatrix} 1 & t_0 & \cdots & t_0^{p-1} \\ 1 & t_1 & \cdots & t_1^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{N-1} & \cdots & t_{N-1}^{p-1} \end{bmatrix} \quad \text{(an } N \times p \text{ matrix).}
\end{aligned}$$

The data model is then

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{H}$ 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

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

Given $\boldsymbol{x} = [x(t_0), \ldots, x(t_{N-1})]^T$, we get the following model:

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}.$$

Linear Models (cont.)

For $p/2 = 2$ sinusoids:

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

and

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

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

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

General Problem Formulation

Consider the model

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \overbrace{\boldsymbol{w}}^{\text{noise}}$$

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

Assume that $\boldsymbol{w}$ is distributed as $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Recall the *identifiability condition:*

$$p(\boldsymbol{x}; \boldsymbol{\theta}_1) = p(\boldsymbol{x}; \boldsymbol{\theta}_2) \quad \Leftrightarrow \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$

which, in this case, reduces to

$$\boldsymbol{H}\boldsymbol{\theta}_1 = \boldsymbol{H}\boldsymbol{\theta}_2 \quad \Leftrightarrow \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

To satisfy this condition, we assume that $\boldsymbol{H}$ has full rank p.

Minimum Variance Unbiased Estimator for the Linear Model

Theorem 1. *For the model*

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$*, the MVU estimator of* $\boldsymbol{\theta}$ *is given by*

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

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

$$\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} = \mathrm{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \sigma^2 \left(\boldsymbol{H}^T \boldsymbol{H}\right)^{-1}.$$

Proof. Verifying the unbiasedness of $\widehat{\boldsymbol{\theta}}$ and the covariance matrix expression $\mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \mathrm{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \sigma^2 \left(\boldsymbol{H}^T \boldsymbol{H}\right)^{-1}$ proves the theorem.

For the above model,

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

and the Fisher information matrix (FIM) for $\boldsymbol{\theta}$, $\mathcal{I}(\boldsymbol{\theta})$, is computed using the general Gaussian FIM expression in 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 $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{\theta}$. Now

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial (\boldsymbol{H}\boldsymbol{\theta})}{\partial \theta_i} = i\text{th column of } \boldsymbol{H}$$

implying that

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \cdot \boldsymbol{H}^T\boldsymbol{H} \quad \Longrightarrow \quad \mathbf{CRB}(\boldsymbol{\theta}) = \sigma^2 \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}. \quad (2)$$

□

Comments:

- Since the joint FIM and CRB for $[\boldsymbol{\theta}^T, \sigma^2]^T$ are block-diagonal matrices, $\boldsymbol{\theta}$ and σ^2 are decoupled $\Longrightarrow$ **$\mathbf{CRB}(\boldsymbol{\theta})$ is the same regardless of whether σ^2 is known or not.** To be more precise, $\mathbf{CRB}(\boldsymbol{\theta})$ in (2) is the *CRB for $\boldsymbol{\theta}$ assuming that σ^2 is known* and here is the full CRB for $\underbrace{\boldsymbol{\theta} \text{ and } \sigma^2}_{\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\theta} \\ \sigma^2 \end{bmatrix}}$ for the case

where both $\boldsymbol{\theta}$ and σ^2 are unknown:

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

Therefore, $\widehat{\boldsymbol{\theta}}$ in (1) is the MVU estimator of $\boldsymbol{\theta}$ regardless of whether σ^2 is known or not.

- $\widehat{\boldsymbol{\theta}}$ in (1) coincides with the *least-squares (LS) estimate* of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|^2$$

which can be shown by differentiating $\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|^2$ with respect to $\boldsymbol{\theta}$ and setting the result to zero or by completing the squares. 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, say, briefly outlined below. Suppose that the $N \times p$ matrix $\boldsymbol{H}$ is factored as

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{R} = [\boldsymbol{Q}_1 \; \boldsymbol{Q}_2] \begin{bmatrix} \boldsymbol{R}_1 \\ 0 \end{bmatrix} = \boldsymbol{Q}_1 \boldsymbol{R}_1$$

where $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{R}_1$ is upper triangular $p \times p$ (MATLAB: `qr`). Then

$$(\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T = \boldsymbol{R}_1^{-1} \boldsymbol{Q}_1^T.$$

Thus, $\widehat{\boldsymbol{\theta}}$ can be obtained by solving the triangular system of equations

$$\boldsymbol{R}_1 \widehat{\boldsymbol{\theta}} = \boldsymbol{Q}_1^T \boldsymbol{x}.$$

MATLAB has the "backslash" command for computing the LS solution:

$$\boldsymbol{\theta} = \boldsymbol{H} \backslash \boldsymbol{x};$$

Minimum Variance Unbiased Estimator for the Linear Model, Colored Noise

Suppose that we have colored noise, so that $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{C})$, where $\boldsymbol{C} \neq \boldsymbol{I}$ is known and positive definite.

We can use *prewhitening* to get back to the old problem (i.e. the white-noise case). We compute the Cholesky factorization of $\boldsymbol{C}^{-1}$:

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

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

Now, define the transformed measurement model:

$$\underbrace{\boldsymbol{D}\,\boldsymbol{x}}_{\boldsymbol{x}^{\text{transf}}} = \underbrace{\boldsymbol{D}\,\boldsymbol{H}}_{\boldsymbol{H}^{\text{transf}}}\,\boldsymbol{\theta} + \underbrace{\boldsymbol{D}\,\boldsymbol{w}}_{\boldsymbol{w}^{\text{transf}}}\,.$$

Clearly, $\boldsymbol{w}^{\text{transf}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and the problem is reduced to the white-noise case.

MVU Estimation, Colored Noise (cont.)

Theorem 2. *For colored Gaussian noise with known covariance $\boldsymbol{C}$, the MVU estimate of $\boldsymbol{\theta}$ is*

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

The covariance matrix of $\widehat{\boldsymbol{\theta}}$ attains the CRB and is given by

$$\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} = (\boldsymbol{H}^T \boldsymbol{C}^{-1} \boldsymbol{H})^{-1}.$$

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

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|_{\boldsymbol{W}}^2 \\ &= \arg\min_{\boldsymbol{\theta}} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta})^T \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}). \end{aligned}$$

The "optimal weight matrix," $\boldsymbol{W} = \boldsymbol{C}^{-1}$, prewhitens the residuals.

Best Linear Unbiased Estimator

Given the model

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w} \tag{3}$$

where $\boldsymbol{w}$ has zero mean and covariance matrix $\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{C}$, we look for the *best linear unbiased estimator (BLUE)*. Hence, we restrict our estimator to be

- linear (i.e. of the form $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}^T\boldsymbol{x}$) and
- unbiased

and minimize its variance.

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

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

and its covariance matrix is

$$\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} = (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}.$$

The expression for $\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}}$ holds *independently* of the distribution of $\boldsymbol{w}$ — all we impose on $\boldsymbol{w}$ is that it has known mean vector and covariance matrix, equal to $\boldsymbol{0}$ and $\boldsymbol{C}$ (respectively).

The estimate $\widehat{\boldsymbol{\theta}}$ is (statistically) efficient if $\boldsymbol{w}$ is Gaussian (i.e. it attains the CRB), but it is not efficient in general. For non-Gaussian measurement models, there might be a better nonlinear estimate. (Most likely, there exists a better nonlinear estimate.)

Proof. (of Theorem 3). For simplicity, consider first the case where θ is scalar. Then, our measurement model is

$$x[n] = h[n]\,\theta + w[n] \quad \Longleftrightarrow \quad \boldsymbol{x} = \boldsymbol{h}\,\theta + \boldsymbol{w}.$$

The candidate linear estimates of θ have the following form:

$$\widehat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \boldsymbol{a}^T \boldsymbol{x}.$$

First, the bias is computed:

$$\mathrm{E}\,[\widehat{\theta}] = \boldsymbol{a}^T \mathrm{E}\,[\boldsymbol{x}] = \boldsymbol{a}^T \boldsymbol{h}\;\theta.$$

Thus, $\widehat{\theta}$ is unbiased if and only if $\boldsymbol{a}^T \boldsymbol{h} = 1$. Next, compute the variance of $\widehat{\theta}$. We have

$$\widehat{\theta} - \theta = \boldsymbol{a}^T (\underbrace{\boldsymbol{h}\,\theta + \boldsymbol{w}}_{\boldsymbol{x}}) - \theta = \boldsymbol{a}^T \boldsymbol{w}$$

where we have used the unbiasedness condition: $\boldsymbol{a}^T\boldsymbol{h} = 1$. Therefore, the variance is

$$\mathrm{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{E}\left[(\boldsymbol{a}^T\boldsymbol{w})^2\right] = \mathrm{E}\left[\boldsymbol{a}^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{a}\right] = \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}.$$

Note: The variance of $\widehat{\theta}$ depends only on the second-order properties of the noise. This result holds for any noise distribution that has second-order moments.

Thus, the BLUE problem is

$$\min_{\boldsymbol{a}} \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} \qquad \text{such that} \quad \boldsymbol{a}^T\boldsymbol{h} = 1.$$

Note the equivalence with MVDR beamforming. To read more about MVDR beamforming, see

H.L. Van Trees, *Detection, Estimation and Modulation Theory*, New York: Wiley, 2002, pt. IV.

Lagrange-multiplier formulation:

$$L(\boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} + \lambda\cdot(\boldsymbol{a}^T\boldsymbol{h} - 1) \quad \overset{\text{differentiate}}{\Longrightarrow} \quad 2\,\boldsymbol{C}\boldsymbol{a} + \lambda\,\boldsymbol{h} = \boldsymbol{0}.$$

Hence

$$\boldsymbol{a} = -\frac{\lambda}{2}\cdot\boldsymbol{C}^{-1}\boldsymbol{h}$$

and then

$$\boldsymbol{a}^T\boldsymbol{h} = -\frac{\lambda}{2}\boldsymbol{h}^T\boldsymbol{C}^{-1}\boldsymbol{h} = 1 \quad \Rightarrow \quad \lambda = -\frac{2}{\boldsymbol{h}^T\boldsymbol{C}^{-1}\boldsymbol{h}}$$

and optimal $\boldsymbol{a}$ follows: $\boldsymbol{a} = (\boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{h})^{-1} \boldsymbol{C}^{-1} \boldsymbol{h}$. Returning to our estimator, we find the BLUE to be

$$\widehat{\theta} = (\boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{h})^{-1} \boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{x}.$$

and its variance is given by

$$\mathrm{E}\left[(\widehat{\theta} - \theta)^2\right] = (\boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{h})^{-1}.$$

□

Consider the vector case. Linear unbiased estimates of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{A}^T \boldsymbol{x}, \quad \text{where } \boldsymbol{A} \text{ is independent of } \boldsymbol{x}. \tag{4}$$

Remark: For LS estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$, $\boldsymbol{A}^T = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T$.

$$\boldsymbol{\theta} = \mathrm{E}\left[\widehat{\boldsymbol{\theta}}\right] = \mathrm{E}\left[\boldsymbol{A}^T \boldsymbol{x}\right] = \mathrm{E}\left[\boldsymbol{A}^T (\boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w})\right] = \boldsymbol{A}^T \boldsymbol{H} \boldsymbol{\theta}.$$

$$\Rightarrow \quad \boldsymbol{A}^T \boldsymbol{H} = \boldsymbol{I}.$$

Remark: For BLUE $\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}$, $\boldsymbol{A}^T_{\mathrm{BLUE}} = (\boldsymbol{H}^T \boldsymbol{C}^{-1} \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{C}^{-1}$. Since $\boldsymbol{A}^T_{\mathrm{BLUE}} \boldsymbol{H} = \boldsymbol{I} \Rightarrow \mathrm{E}\left[\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}\right] = \boldsymbol{\theta}$.

$$\mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \mathrm{E}\left\{[\boldsymbol{A}^T (\underbrace{\boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}}_{\boldsymbol{x}}) - \boldsymbol{\theta}][\boldsymbol{A}^T (\boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}) - \boldsymbol{\theta}]^T\right\} = \boldsymbol{A}^T \boldsymbol{C} \boldsymbol{A}.$$

and

$$\begin{aligned}\mathrm{cov}(\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}) &= (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \\ &= (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}.\end{aligned}$$

To prove that $\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}$ has the smallest variance [within the family of linear unbiased estimators $\widehat{\boldsymbol{\theta}}$ in (4)], we show that

$$\mathrm{cov}(\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}) \leq \mathrm{cov}(\widehat{\boldsymbol{\theta}})$$

as follows:

$$\begin{aligned}\mathrm{cov}(\widehat{\boldsymbol{\theta}}) - \mathrm{cov}(\widehat{\boldsymbol{\theta}}_{\mathrm{BLUE}}) &= \boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A} - (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \\ &\overset{\boldsymbol{A}^T\boldsymbol{H}=\boldsymbol{I}}{=} \boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A} - \boldsymbol{A}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{A} \\ &= \boldsymbol{A}^T[\boldsymbol{C} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T]\boldsymbol{A} \\ &= \boldsymbol{A}^T[\boldsymbol{C} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T]\boldsymbol{C}^{-1}[\boldsymbol{C} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T]\boldsymbol{A}\end{aligned}$$

which is always positive semidefinite.

Examples

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

$$x[n] = A + w[n]$$

for $n = 0, 1, \ldots, N-1$, where $\boldsymbol{w} = [w[0], w[1], \ldots, w[N-1]]^T$ is the colored noise with zero mean and covariance matrix $\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{C}$. Hence, $\boldsymbol{H} = \boldsymbol{h} = \boldsymbol{1} = [1, 1, \ldots, 1]^T$ in (3). The BLUE is

$$\widehat{A} = (\boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{h})^{-1} \boldsymbol{h}^T \boldsymbol{C}^{-1} \boldsymbol{x} = \frac{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{x}}{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{1}}$$

and its variance is

$$\mathrm{var}(\widehat{A}) = \frac{1}{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{1}}.$$

Consider the Cholesky factorization $(\boldsymbol{C})^{-1} = \boldsymbol{D}^T \boldsymbol{D}$; then the BLUE of A becomes

$$\widehat{A} = \frac{\boldsymbol{1}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{x}}{\boldsymbol{1}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{1}} = \frac{(\boldsymbol{D}\boldsymbol{1})^T \overbrace{\boldsymbol{D}\boldsymbol{x}}^{\boldsymbol{x}^{\mathrm{transf}}}}{\boldsymbol{1}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{1}} = \sum_{n=0}^{N-1} d_n x^{\mathrm{transf}}[n]$$

where

$$d_n = [\boldsymbol{D}\boldsymbol{1}]_n / \boldsymbol{1}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{1}.$$

Examples (cont.)

Sometimes, BLUE is completely wrong. For example, $x[n] = w[n]$, $n = 1, 2, \ldots, N$, white Gaussian noise with variance σ^2. The MVU estimator is $\widehat{\sigma}^2 = (1/N) \cdot \sum_{n=0}^{N-1} x^2[n]$. On the other hand,

$$\widehat{\sigma}^2_{\text{BLUE}} = \sum_{n=1}^{N} a_n \, x[n].$$

For an estimator $\widehat{\sigma}^2$ to be unbiased, we need $\mathrm{E}\left[\widehat{\sigma}^2\right] = \sigma^2$, but

$$\mathrm{E}\left[\widehat{\sigma}^2_{\text{BLUE}}\right] = \sum_{n=1}^{N} a_n \mathrm{E}\left(x[n]\right) = 0!$$

It is impossible to find a_ns to make $\widehat{\sigma}^2_{\text{BLUE}}$ unbiased.

Note: Although the BLUE is not suitable for this problem, utilizing the *transformed data* $y[n] = x^2[n]$ would produce a viable estimator.

General MVU Estimation

What is the MVU estimate in general?

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

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathrm{E}_{p(\boldsymbol{x}\,|\,T(\boldsymbol{x}))}\Big[\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})\,|\,T(\boldsymbol{X}) = T(\boldsymbol{x})\Big] \tag{5}$$

is no worse than $\widetilde{\boldsymbol{\theta}}(\boldsymbol{x})$ (in terms of MSE).

Problem: Computing $\mathrm{E}\,[\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})\,|\,T(\boldsymbol{X}) = T(\boldsymbol{x})]$ may be difficult! Recall that this type of expectation occurs when proving sufficiency, but luckily, in the case of sufficiency, our efforts were greatly simplified by the factorization theorem.

Definition. *$\boldsymbol{T}(\boldsymbol{x})$ is complete sufficient statistic if only one estimator $\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\boldsymbol{T}(\boldsymbol{x}))$ is unbiased.*

Corollary: If $\boldsymbol{T}(\boldsymbol{x})$ is a complete sufficient statistic, then the unique unbiased estimate $\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\boldsymbol{T}(\boldsymbol{x}))$ is the MVU estimate.

Comments:

- Conditioning always decreases the variance (does not increase, to be more precise).

- To get a realizable estimator, we need to condition on the sufficient statistics. The definition of sufficient statistic [denoted by $\boldsymbol{T}(\boldsymbol{x})$] implies that conditioning on it leads to a distribution that is not a function of the unknown parameters $\boldsymbol{\theta}$. Hence, (5) is a statistic, i.e. realizable.

Example: Suppose that $x[n]$, $n = 1, 2, \ldots, N$ are independent, identically distributed (i.i.d.) $\mathcal{N}(A, \sigma^2)$ with $\boldsymbol{\theta} = [A, \sigma^2]^T$. Then,

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x[n] - A)^2 \Big]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\Big\{ -\frac{1}{2\sigma^2} \Big[(N-1) \cdot \frac{1}{N-1} \cdot \sum_{n=1}^{N} (x[n] - \bar{x})^2 + N(\bar{x} - A)^2 \Big] \Big\}.$$

Therefore, the jointly sufficient statistics are

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

where $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x[n]$. It can be shown that $\widehat{A} = \bar{x}$ and $\widehat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x[n] - \bar{x})^2$ are the only unbiased functions of $\boldsymbol{T}(\boldsymbol{x}) = [T_1(\boldsymbol{x}), T_2(\boldsymbol{x})]^T$. Hence, the corollary at the previous page implies that they are the MVU estimates (although, in this case, the MVU estimates are *not efficient* and, therefore, could not have been found using the efficiency argument). Indeed, for $\widehat{\boldsymbol{\theta}} = [\widehat{A}, \widehat{\sigma}^2]^T$,

$$\mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & 2\sigma^4/(N-1) \end{bmatrix}$$

but, recall the CRB for this case (e.g. p. 24 of handout # 2):

$$\mathbf{CRB}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & 2\sigma^4/N \end{bmatrix}.$$

MAXIMUM LIKELIHOOD (ML) ESTIMATION

$$\widehat{\theta} = \arg\max_{\theta} p(\boldsymbol{x}; \theta).$$

The pdf $p(\boldsymbol{x}; \theta)$, viewed as function of θ, is the *likelihood function*.

Comments on the likelihood function: For given θ and discrete case, $p(\boldsymbol{x}; \theta)$ is the probability of observing the point $\boldsymbol{x}$. In the continuous case, it is approximately proportional to probability of observing a point in a small rectangle around $\boldsymbol{x}$. However, when we think of $p(\boldsymbol{x}; \theta)$ as a function of θ, it gives, for a given observed $\boldsymbol{x}$, the "likelihood" or "plausibility" of various θ.

ML estimate $\equiv$ 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 let $\widehat{\theta}$ be the ML estimate. Then, as $N \to \infty$,*

$$\widehat{\theta} \;\to\; \theta_0 \quad (\textit{with probability 1}) \qquad \textit{(consistency)} \tag{6}$$

$$\sqrt{N}\,(\widehat{\theta} - \theta_0) \;\xrightarrow{\mathrm{d}}\; \mathcal{N}\big(0, N\,\mathcal{I}^{-1}(\theta_0)\big) \qquad \textit{(asymptotic efficiency)} \tag{7}$$

where θ_0 is the true value of the parameter and $\mathcal{I}(\theta_0)$ is the Fisher information [and $\mathcal{I}^{-1}(\theta_0)$ the CRB]. Moreover, if an efficient (in finite samples) estimate exists, it is given by the ML estimate.

Proof. See e.g. Rao, Chapter 5f.2 at pp. 364–366 for the case of independent observations. □

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

Example: $x[n]$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, $n = 0, 1, \ldots, N-1$, for σ^2

known. Maximizing $p(\boldsymbol{x};\theta)$ is equivalent to

$$\max_{\theta} \log p(\boldsymbol{x};\theta) = \text{const} - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\theta)^2.$$

Thus, the ML estimate is the sample mean

$$\widehat{\theta} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \sim \mathcal{N}(\theta_0, \sigma^2/N).$$

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

Note: When estimation error cannot be made small as $N \to \infty$, the asymptotic pdf in (7) 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 observe $x[0], x[1], \ldots, x[N-1]$ but $w[0] = w[1] = \cdots = w[N-1]$, i.e. all noise samples are the same. Hence, we discard $x[1], x[2], \ldots, x[N-1]$. Then, $\widehat{A} = x[0]$, say. The pdf of $\widehat{A}$ remains non-Gaussian as $N \to \infty$. Also, $\widehat{A}$ is not consistent since $\text{var}(\widehat{A}) = \text{var}(x[0]) \not\to 0$ as $N \to \infty$.

ML Estimation: Vector Parameters

Nothing really changes. The ML estimate is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x}; \boldsymbol{\theta}).$$

Under appropriate regularity conditions, this estimate is consistent and

$$\sqrt{N}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(\mathbf{0}, N\,\mathcal{I}^{-1}(\boldsymbol{\theta}_0)\Big)$$

where $\mathcal{I}(\boldsymbol{\theta}_0)$ is the Fisher information *matrix* now.

ML Estimation: Properties

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

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{g}(\widehat{\boldsymbol{\theta}})$$

where $\widehat{\boldsymbol{\theta}}$ *is the ML estimate of* $\boldsymbol{\theta}$ *[obtained by maximizing* $p(\boldsymbol{x}; \boldsymbol{\theta})$ *with respect to* $\boldsymbol{\theta}$*].*

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 $T(\boldsymbol{x})$, say, then the ML estimate of θ must be a function of $T(\boldsymbol{x})$. Furthermore, if $T(\boldsymbol{x})$ 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? Now we try to add to the answer to this question, focusing on the case of i.i.d. observations.

In Ch. 6.2 of *Theory of Point Estimation,* Lehmann shows the following result.

Theorem 7. *Suppose that the random observations X_i are i.i.d. with common pdf/pmf $p(x_i;\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is in the interior of the parameter space. Then, as $N \to \infty$*

$$P\Big\{ \prod_{i=1}^{N} p(X_i;\boldsymbol{\theta}_0) > \prod_{i=1}^{N} p(X_i;\boldsymbol{\theta}) \Big\} \to 1$$

for any fixed $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Comment: This theorem states that, for large number of i.i.d. samples (i.e. large N), the joint pdf/pmf of $X_1, X_2, \ldots, X_N$ at the true parameter value

$$\prod_{i=1}^{N} p(X_i;\boldsymbol{\theta}_0)$$

exceeds the joint pdf/pmf of $X_1, X_2, \ldots, X_N$ at any other parameter value (with probability one). Consequently, as the

number of observations increases, the parameter estimate that maximizes the joint distribution of the measurements (i.e. the ML estimate) must become close to the true value.

Regularity Conditions for I.I.D. Observations

Not one set of regularity conditions applies to all scenarios.

Here are some typical regularity conditions for the i.i.d. case. Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf

$$p(x_i; \boldsymbol{\theta}), \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}.$$

Regularity conditions:

(i) $p(x; \boldsymbol{\theta})$ is identifiable for $\boldsymbol{\theta}$ and the support of $p(x_i; \boldsymbol{\theta})$ is not a function of $\boldsymbol{\theta}$;

(ii) The true value of the parameter, say $\boldsymbol{\theta}_0$, lies in an open subset of the parameter space $\boldsymbol{\Theta}$;

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

(iv) The following are satisfied:

$$\mathrm{E}_{\,p(x;\theta)} \left[\frac{\partial}{\partial \theta_k} \log p(X; \boldsymbol{\theta}) \right] = 0, \quad k = 1, 2, \ldots, p$$

and

$$\begin{aligned}\mathcal{I}_{i,k}(\boldsymbol{\theta}) &= \mathrm{E}_{\,p(x;\boldsymbol{\theta})}\left[\frac{\partial}{\partial\theta_i}\log p(X;\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta_k}\log p(X;\boldsymbol{\theta})\right] \\ &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_k}\log p(X;\boldsymbol{\theta})\right], \quad i,k=1,2,\ldots,p.\end{aligned}$$

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

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

$$\left|\frac{\partial^3}{\partial\theta_i\,\partial\theta_k\,\partial\theta_l}\log p(x;\boldsymbol{\theta})\right| \le m_{i,k,l}(x)$$

for all $\boldsymbol{\theta}$ in the open subset of (ii), and

$$\mathrm{E}_{\,p(x;\theta)}[m_{i,k,l}(X)] < \infty.$$

Theorem 8. **(≈ same as Theorem 5)** *If $X_1, X_2, \ldots, X_N$ are i.i.d. with pdf $p(x_i;\boldsymbol{\theta})$ such that the conditions* (i)–(vi) *hold, then there exists a sequence of solutions $\{\widehat{\boldsymbol{\theta}}_N\}$ to the likelihood equations such that*

(i) *$\widehat{\boldsymbol{\theta}}_N$ is consistent for $\boldsymbol{\theta}$;*

(ii) *$\sqrt{N}\,(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ is asymptotically Gaussian with mean $\mathbf{0}$ and covariance matrix $N\,\mathcal{I}^{-1}(\boldsymbol{\theta}) = N\,\mathbf{CRB}(\boldsymbol{\theta})$;*

(iii) $\sqrt{N}\left([\widehat{\theta}_N]_i - \theta_i\right)$ *is asymptotically Gaussian with mean* 0 *and variance* $N\left[\mathcal{I}^{-1}(\boldsymbol{\theta})\right]_{i,i}$, $i = 1, 2, \ldots, p$.

Comments: What we are *not* given:

(i) uniqueness of $\widehat{\boldsymbol{\theta}}_N$;

(ii) existence for all $x_1, \ldots, x_N$;

(iii) even if the solution exists and is unique, that we can find it.

An Array Processing Example

$$\boldsymbol{x}[n] = \boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{s}[n] + \boldsymbol{w}[n], \quad n = 0, 1, \ldots, N-1$$

where $\boldsymbol{\theta} = [\boldsymbol{\phi}^T, \boldsymbol{s}[0]^T, \boldsymbol{s}[1]^T, \ldots \boldsymbol{s}[N-1]^T]^T$ is the vector of unknown parameters and $\boldsymbol{w}[n]$ is complex WGN.

Note:

- $\boldsymbol{x}[n]$ are *not i.i.d.* (conditions that we stated are not enough);
- $\boldsymbol{\theta}$ grows with N.

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 $\boldsymbol{s}[n]$ are random $\sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{\Gamma})$? Then, $\boldsymbol{x}[n]$, $n = 0, 1, \ldots, N-1$ are i.i.d. with

$$\boldsymbol{x}[n] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{\Gamma}\boldsymbol{A}(\boldsymbol{\theta})^H + \sigma^2\boldsymbol{I}).$$

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

Digression: Delta Method

Theorem 9. *(Gauss Approximation Formula, Delta Method) Assume $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\theta})$ has bounded derivatives up to the 2nd order. Then, if $\widehat{\boldsymbol{\theta}}$ is consistent, so is $\widehat{\boldsymbol{\alpha}}$. Moreover, the asymptotic MSE matrices $\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}}$ and $\boldsymbol{C}_{\widehat{\boldsymbol{\alpha}}}$ are related by*

$$\boldsymbol{C}_{\widehat{\alpha}} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}^T} \boldsymbol{C}_{\widehat{\theta}} \frac{\partial \boldsymbol{g}^T}{\partial \boldsymbol{\theta}}.$$

Proof. Follows from Taylor expansion [around the true value $\boldsymbol{\alpha}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$]

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{g}(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|).$$

□

Example: Amplitude and Phase Estimation

Assume $x[n] = A\cos(\omega_0 n + \phi) + e[n]$, $n = 0, 1, \ldots, N-1$, where ω_0 is known and $e[n]$ is additive white Gaussian noise (AWGN). We wish to estimate A and ϕ.

We rewrite this model as a linear model:

$$\boldsymbol{x} = \begin{bmatrix} x[0] \\ \ldots \\ x[N-1] \end{bmatrix} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

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

$$\begin{aligned} H_{i,1} &= \cos[\omega_0(i-1)], \quad i = 1, 2, \ldots, N \\ H_{i,2} &= \sin[\omega_0(i-1)], \quad i = 1, 2, \ldots, N \\ (A\cos\phi, -A\sin\phi) &\leftrightarrow (\theta_1, \theta_2) \end{aligned}$$

We have

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x} \sim \mathcal{N}\Big(\boldsymbol{\theta}_0, \sigma^2 (\boldsymbol{H}^T\boldsymbol{H})^{-1}\Big).$$

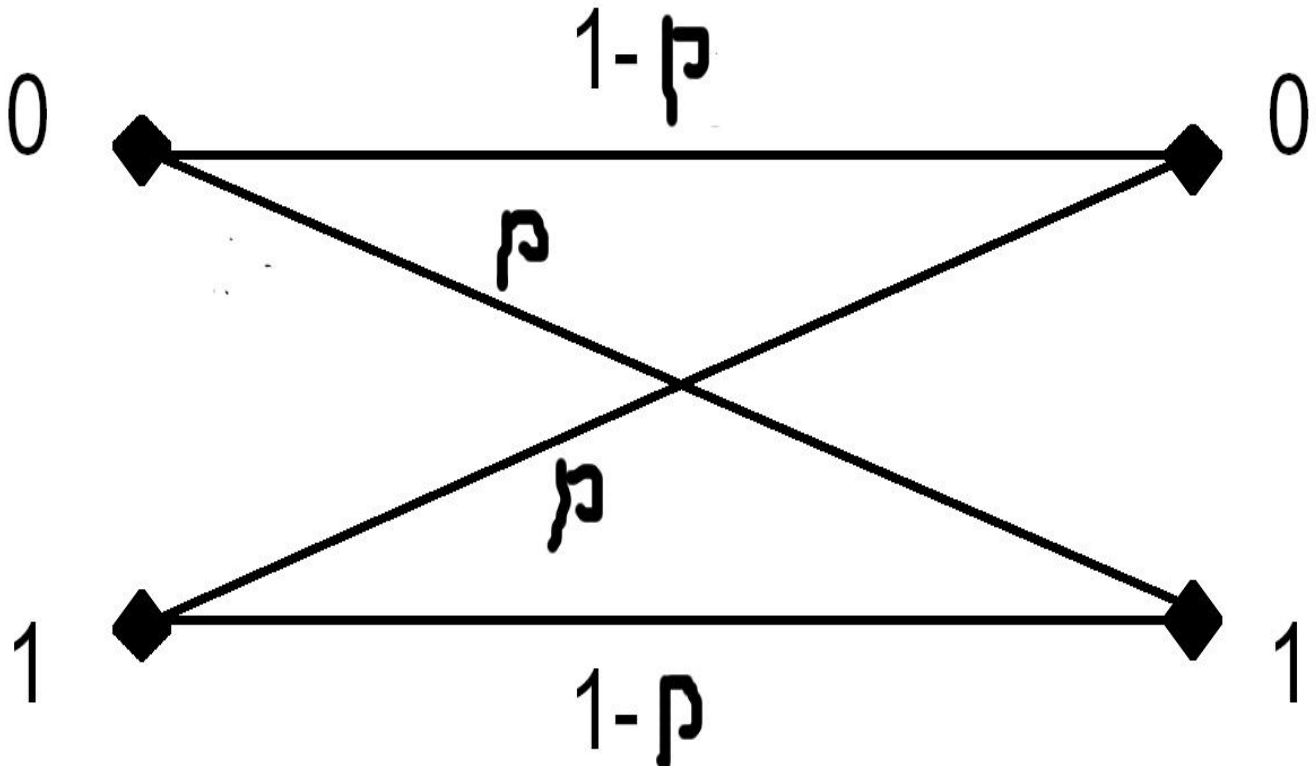
By the ML invariance principle, $\widehat{A}$ and $\widehat{\phi}$ can be found from $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1, \widehat{\theta}_2]^T$ via rectangular-to-polar coordinate conversion:

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

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

$$\boldsymbol{C}_{\widehat{\boldsymbol{\alpha}}} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}^T} \boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} \frac{\partial \boldsymbol{g}^T}{\partial \boldsymbol{\theta}}.$$

Example: ML Decoding



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

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{\theta}$ be the received and transmitted vectors from a binary symmetric channel (i.e. the elements of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are zeros or ones). Note that $\boldsymbol{\theta}$ belongs to a finite set of codewords. We wish to find which $\boldsymbol{\theta}$ was transmitted based on the received $\boldsymbol{x}$. We have

$$\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{w} \pmod 2 \stackrel{\triangle}{=} \boldsymbol{\theta} \oplus \boldsymbol{w}$$

where $\boldsymbol{w} = [w_1, \ldots, w_N]^T$ and w_i are i.i.d. $\mathrm{Bernoulli}(p)$. The

likelihood function is given by

$$\begin{aligned} p(\boldsymbol{x};\boldsymbol{\theta}) &= P\{\boldsymbol{X}=\boldsymbol{x}\} = P\{\boldsymbol{\theta} \oplus \underbrace{\boldsymbol{W}}_{\text{i.i.d. Bernoulli}} = \boldsymbol{x}\} \\ &= P\{\boldsymbol{W} = \underbrace{\boldsymbol{x} \oplus \boldsymbol{\theta}}_{\boldsymbol{w}=\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}\} \\ &= p^{\sum_{i=1}^N w_i} \cdot (1-p)^{N-\sum_{i=1}^N w_i} \\ &= \left(\frac{p}{1-p}\right)^{d_{\mathrm{H}}(\boldsymbol{x},\boldsymbol{\theta})} (1-p)^N \end{aligned}$$

where

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

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

$$\max_{\boldsymbol{x}} p(\boldsymbol{x};\boldsymbol{\theta}) \quad \Longleftrightarrow \quad \min_{\boldsymbol{\theta}} d_{\mathrm{H}}(\boldsymbol{x},\boldsymbol{\theta}).$$

□

Asymptotic ML for WSS Processes

Consider data $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}(\boldsymbol{\theta}))$. To find the ML estimate of $\boldsymbol{\theta}$, maximize

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{C}(\boldsymbol{\theta})|^{1/2}} \exp\left[-\tfrac{1}{2} \cdot \boldsymbol{x}^T \boldsymbol{C}(\boldsymbol{\theta})^{-1} \boldsymbol{x} \right]$$

over $\boldsymbol{\theta}$.

If $x[n]$ is WSS, then $\boldsymbol{C}(\boldsymbol{\theta})$ is Töplitz, so, as $N \to \infty$, we can approximate the log likelihood as:

$$\begin{aligned} \log p(\boldsymbol{x}; \theta) \quad &= \quad -\frac{N}{2} \log 2\pi \\ &\qquad -\frac{N}{2} \int_{-1/2}^{1/2} \left(\log P_{xx}(f; \boldsymbol{\theta}) + \frac{I_{\boldsymbol{x}}(f)}{P_{xx}(f; \boldsymbol{\theta})} \right) df \end{aligned}$$

where $I_{\boldsymbol{x}}(f)$ is the periodogram:

$$I_{\boldsymbol{x}}(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j2\pi f n} \right|^2$$

and $P_{xx}(f; \theta)$ is the PSD of $x[n]$.

This result is based on the Whittle approximation, see e.g.

P. Whittle, "The analysis of multiple stationary time series," *J. R. Stat. Soc., Ser. B* vol. 15, pp. 125–139, 1953.

Proof. See e.g. Ch. 7.9 in Kay-I. □

Note: Kay calls the Whittle approximation "asymptotic ML".

The discrete-frequency version of the above expression is also useful:

$$-\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\left\{\log[P_{yy}(f_k;\boldsymbol{\theta})] + \frac{I_{\boldsymbol{x}}(f_k)}{P_{yy}(f_k;\boldsymbol{\theta})}\right\} \quad (8)$$

where

$$f_k = k/N, \quad k = 0, 1, \ldots, N-1.$$

For example, (8) may exist even when the integral form does not. An example of such a case that we mentioned earlier is the Doppler PSD (which goes to infinity, causing an integrability problem), see

A. Dogandžić and B. Zhang, "Estimating Jakes' Doppler power spectrum parameters using the Whittle approximation," *IEEE Trans. Signal Processing*, vol. 53, pp. 987–1005, Mar. 2005.

Example. Autoregressive (AR) parameter estimation:

$$P_{xx}(f;\boldsymbol{\theta}) = \frac{\sigma_u^2}{|A(f;\boldsymbol{a})|^2}$$

where $\boldsymbol{\theta} = [\underbrace{a[1], a[2], \ldots, a[p]}_{\boldsymbol{a}^T}, \sigma_u^2]^T$, and

$$A(f; \boldsymbol{a}) = 1 + \sum_{m=1}^{p} a[m] \exp(-j2\pi f m).$$

So

$$\begin{aligned} \log p(\boldsymbol{x}; \boldsymbol{a}, \sigma_u^2) &= -\frac{N}{2} \log 2\pi \\ &\quad -\frac{N}{2} \int_{-1/2}^{1/2} \left(\log \frac{\sigma_u^2}{|A(f; \boldsymbol{a})|^2} + \frac{I_{\boldsymbol{x}}(f)}{\frac{\sigma_u^2}{|A(f;\boldsymbol{a})|^2}} \right) df. \end{aligned}$$

Assuming $\mathcal{A}(z) = 1 + \sum_{m=1}^{p} a[m] z^{-m}$ to be minimum-phase [typically required for stability of $1/\mathcal{A}(z)$], then

$$\int_{-1/2}^{1/2} \log |A(f; \boldsymbol{a})|^2 \, df = 0$$

see Problem 7.22 in Kay-I. Therefore

$$\begin{aligned} \log p(\boldsymbol{x}; \boldsymbol{a}, \sigma_u^2) &= -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma_u^2 \\ &\quad -\frac{N}{2\sigma_u^2} \int_{-1/2}^{1/2} |A(f; \boldsymbol{a})|^2 \, I_{\boldsymbol{x}}(f) \, df. \end{aligned}$$

Differentiating with respect to σ_u^2 and setting the result to zero, we obtain, for a fixed $\boldsymbol{a}$,

$$\widehat{\sigma}_u^2(\boldsymbol{a}) = \int_{-1/2}^{1/2} |A(f;\boldsymbol{a})|^2 \, I_{\boldsymbol{x}}(f) \, df.$$

Then, the concentrated Whittle log-likelihood function of $\boldsymbol{a}$ is obtained by substituting $\widehat{\sigma}_u^2(\boldsymbol{a})$ into the Whittle log-likelihood $\log p(\boldsymbol{x}; \boldsymbol{a}, \sigma_u^2)$:

$$\log p(\boldsymbol{x}; \boldsymbol{a}, \widehat{\sigma}_u^2(\boldsymbol{a})) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \widehat{\sigma}_u^2(\boldsymbol{a}) - N/2.$$

We will discuss the concentrated likelihood approach later, see p. 49. To find $\widehat{\boldsymbol{a}}$, we need to maximize the above concentrated likelihood or, equivalently, minimize $\widehat{\sigma}_u^2$:

$$\min_{\boldsymbol{a}} J(\boldsymbol{a}) = \min_{\boldsymbol{a}} \int_{-1/2}^{1/2} |A(f;\boldsymbol{a})|^2 \, I_{\boldsymbol{x}}(f) \, df.$$

The above function is quadratic in $\boldsymbol{a}$, resulting in the global minimum upon differentiation. For $k = 1, 2, \ldots, p$, we have

$$\frac{\partial J(\boldsymbol{a})}{\partial a[k]} = \int_{-1/2}^{1/2} \Big[A(f;\boldsymbol{a}) \underbrace{\frac{\partial A^\star(f;\boldsymbol{a})}{\partial a[k]}}_{\exp(j2\pi fk)} + \underbrace{\frac{\partial A(f;\boldsymbol{a})}{\partial a[k]}}_{\exp(-j2\pi fk)} \, A^\star(f;\boldsymbol{a}) \Big] I_{\boldsymbol{x}}(f) \, df.$$

Since $A(-f;\boldsymbol{a}) = A^{\star}(f;\boldsymbol{a})$ and $I_{\boldsymbol{x}}(-f) = I_{\boldsymbol{x}}(f)$, we have

$$\frac{\partial J(\boldsymbol{a})}{\partial a[k]} = 2\int_{-1/2}^{1/2} A(f;\boldsymbol{a}) I_{\boldsymbol{x}}(f)\exp(j2\pi fk)\,df.$$

Setting the above expression to zero, we get

$$\int_{-1/2}^{1/2}\left[1+\sum_{m=1}^{p} a[m]\exp(-j2\pi fm)\right] I_{\boldsymbol{x}}(f)\exp(j2\pi fk)\,df = 0$$

or

$$\begin{aligned}&\sum_{m=1}^{p} a[m]\int_{-1/2}^{1/2} I_{\boldsymbol{x}}(f)\exp[j2\pi f(k-m)]\,df\\ &= -\int_{-1/2}^{1/2} I_{\boldsymbol{x}}(f)\exp(j2\pi fk)\,df.\end{aligned}$$

But, $\int_{-1/2}^{1/2} I_{\boldsymbol{x}}(f)\exp(j2\pi fk)\,df$ is just the inverse discrete-time Fourier transform (DTFT) of the periodogram evaluated at k, which is equal to the biased sample estimate of the autocorrelation function:

$$\widehat{r}_{xx}[k] = \begin{cases} \frac{1}{N}\sum_{i=k}^{N-1} x[i]x[i-|k|], & |k| \le N-1,\\ 0, & k \ge N-1\end{cases}$$

Hence, the Whittle (asymptotic) ML estimate of the AR filter parameter vector $\boldsymbol{a}$ solves

$$\sum_{m=1}^{p} \widehat{a}[m]\, \widehat{r}_{xx}[k-m] = -\widehat{r}_{xx}[k], \quad k = 1, 2, \ldots, p$$

which are the *estimated Yule-Walker equations*.

Computing the Estimates

Typically, finding the ML estimate requires a nonlinear $p-$dimensional optimization (for $\boldsymbol{\theta}$ of size p). More generally,

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

where, for ML estimation, we have $V(\boldsymbol{\theta}) = -\log p(\boldsymbol{x};\boldsymbol{\theta})$.

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

$$V(\boldsymbol{\theta}) \approx V(\boldsymbol{\theta}^{(i)}) + \boldsymbol{g}_i^T \overline{\boldsymbol{\theta}}^{(i)} + \tfrac{1}{2} \cdot (\overline{\boldsymbol{\theta}}^{(i)})^T \boldsymbol{H}_i \overline{\boldsymbol{\theta}}^{(i)}$$

where

$$\begin{aligned} \overline{\boldsymbol{\theta}}^{(i)} &= \boldsymbol{\theta} - \boldsymbol{\theta}^{(i)} \\ \boldsymbol{g}_i &= \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} \\ \boldsymbol{H}_i &= \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}. \end{aligned}$$

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

Complete the squares:

$$V(\boldsymbol{\theta}) \approx (\overline{\boldsymbol{\theta}}^{(i)} + \boldsymbol{H}_i^{-1}\boldsymbol{g}_i)^T \, \tfrac{1}{2} \, \boldsymbol{H}_i \, (\overline{\boldsymbol{\theta}}^{(i)} + \boldsymbol{H}_i^{-1}\boldsymbol{g}_i) + \text{const.}$$

We assume that $\boldsymbol{H}_i > 0$

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

and thus choose

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

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

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

where c is a positive constant. 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:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i \cdot \boldsymbol{H}_i^{-1} \boldsymbol{g}_i \tag{9}$$

where the step length μ_i is $\mu_i = 1, 1/2, 1/4, \ldots$. In particular, in the ith iteration, start with the step length $\mu_i = 1$, compute $\boldsymbol{\theta}^{(i+1)}$ using (9), 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 and recomputing $\boldsymbol{\theta}^{(i+1)}$ using (9) until $V(\boldsymbol{\theta}^{(i+1)}) < V(\boldsymbol{\theta}^{(i)})$ holds — 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 $V(\boldsymbol{\theta})$

$$\frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}.$$

In the case of ML estimation, use the FIM instead of this Hessian:

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

and the resulting algorithm is called *Fisher scoring*. This choice of $\boldsymbol{H}_i$ usually guarantees positive definiteness of the Hessian matrix: $\boldsymbol{H}_i = \mathcal{I}(\boldsymbol{\theta}^{(i)}) > 0$.

Note: The convergence point is a *local* minimum of $V(\boldsymbol{\theta})$. It is a 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.

If the parameter space $\boldsymbol{\Theta}$ is not $\mathbb{R}^p$, we should also examine the boundary of the parameter space to see if a global maximum of the likelihood function (i.e. a global minimum of $V(\boldsymbol{\theta})$) lies on this boundary. **Pay attention to this issue in your homework assignments and exams, as well as in general.**

Newton-Raphson Iteration: Example

Suppose $x[n] = s[n;\boldsymbol{\theta}] + e[n]$ where $e[n]$ is white Gaussian noise with known variance σ^2, $s[n;\boldsymbol{\theta}] = \sin(\omega_1 n) + \sin(\omega_2 n)$, $\boldsymbol{\theta} = [\omega_1, \omega_2]^T$, and $n = 0, 1, \ldots, N-1$.

Ignoring constants, we obtain the negative log likelihood:

$$V(\boldsymbol{\theta}) = -\log p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{2\,\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n;\boldsymbol{\theta}])^2 + \text{const}$$

its gradient:

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

and the FIM:

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

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

becomes

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i \cdot \mathcal{I}(\boldsymbol{\theta}^{(i)})^{-1} \cdot V'(\boldsymbol{\theta}^{(i)})$$

$$= \boldsymbol{\theta}^{(i)}$$

$$+\mu_i \cdot \left\{ \sum_{n=0}^{N-1} n^2 \begin{bmatrix} \cos^2(\omega_1^{(i)} n) & \cos(\omega_1^{(i)} n)\cos(\omega_2^{(i)} n) \\ \cos(\omega_1^{(i)} n)\cos(\omega_2^{(i)} n) & \cos^2(\omega_2^{(i)} n) \end{bmatrix} \right\}^{-1}$$

$$\cdot \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}^{(i)}]) \begin{bmatrix} n\cos(\omega_1^{(i)} n) \\ n\cos(\omega_2^{(i)} n) \end{bmatrix}.$$

Example: Concentrated Likelihood

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. Let $x_1, x_2, \ldots, x_n$ be the collected infection times after the exposure. Assume that $x_1, x_2, \ldots, x_n$ are i.i.d. following some distribution. Here, we adopt the following lognormal model:

$$p(x_i; \boldsymbol{\theta}) = \begin{cases} \frac{1}{(x_i-\alpha)\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}[\log(x_i-\alpha)-\mu]^2\}, & x_i > \alpha \\ 0, & \text{otherwise} \end{cases}$$

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, the regularity condition (i) on p. 29 does not hold.

Here are some references related to 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 having x_i, $i = 1, 2, \ldots, N$ such that $\log(x_i - \alpha) \sim \mathcal{N}(\mu, \sigma^2)$.

The log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(x_i; \boldsymbol{\theta}) \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \log(x_i - \alpha) \\ &\quad -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \big[\log(x_i - \alpha) - \mu\big]^2 \end{aligned}$$

where $x_i > \alpha, \ \forall i = 1, 2, \ldots, N$.

For a fixed α, we can easily find μ and σ^2 that maximize the

likelihood:

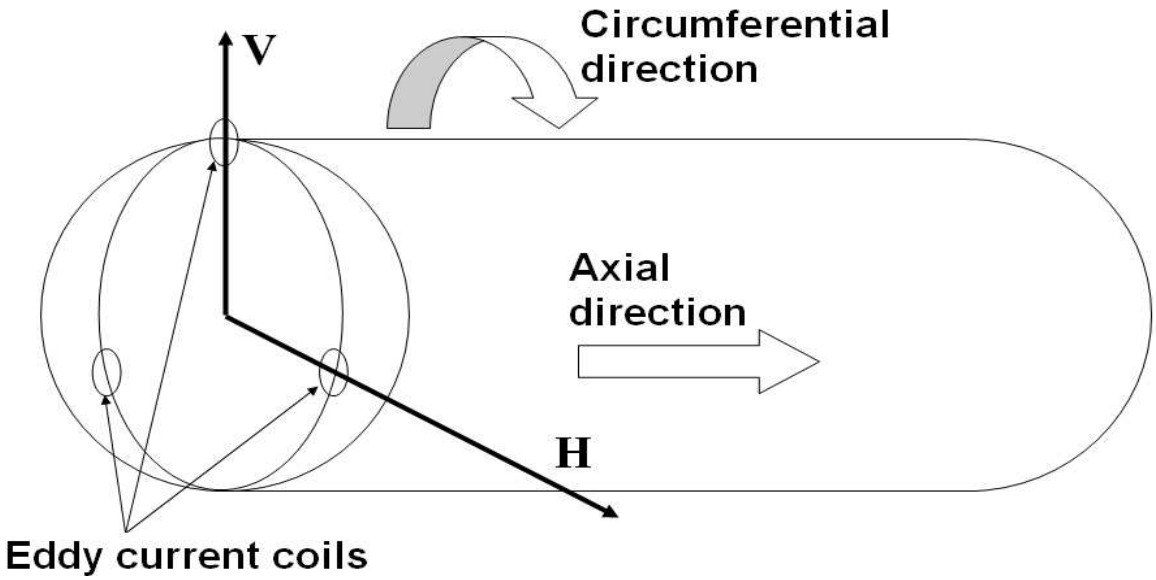
$$\begin{aligned}\widehat{\mu}(\alpha) &= \frac{1}{N}\sum_{t=1}^{N}\log(x_i-\alpha)\\ \widehat{\sigma}^2(\alpha) &= \frac{1}{N}\sum_{t=1}^{N}[\log(x_i-\alpha)-\widehat{\mu}(\alpha)]^2\end{aligned}$$

which follows from the above relationship with the normal pdf. Now, we can write the log-likelihood function as a function of α alone, by substituting $\widehat{\mu}(\alpha)$ and $\widehat{\sigma}^2(\alpha)$ into $l(\boldsymbol{\theta})$:

$$\begin{aligned}&l([\alpha,\widehat{\mu}(\alpha),\widehat{\sigma}^2(\alpha)]^T)\\ &\quad=-\frac{N}{2}\log[2\pi\widehat{\sigma}^2(\alpha)]-\sum_{i=1}^{N}\log(x_i-\alpha)-\frac{N}{2}.\end{aligned}$$

Example: ML Estimation for Eddy-Current Data

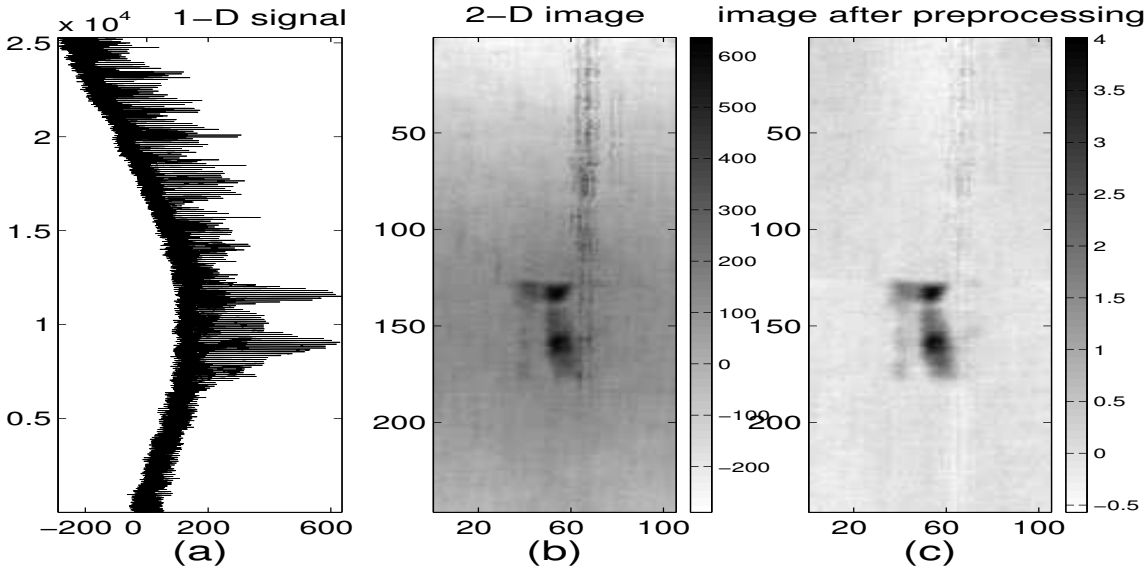
Data collected with a rotating probe, consisting of three coils spaced $2\pi/3$ rad (120°) apart. Each coil scans the inner surface of the tube by moving along a helical path.



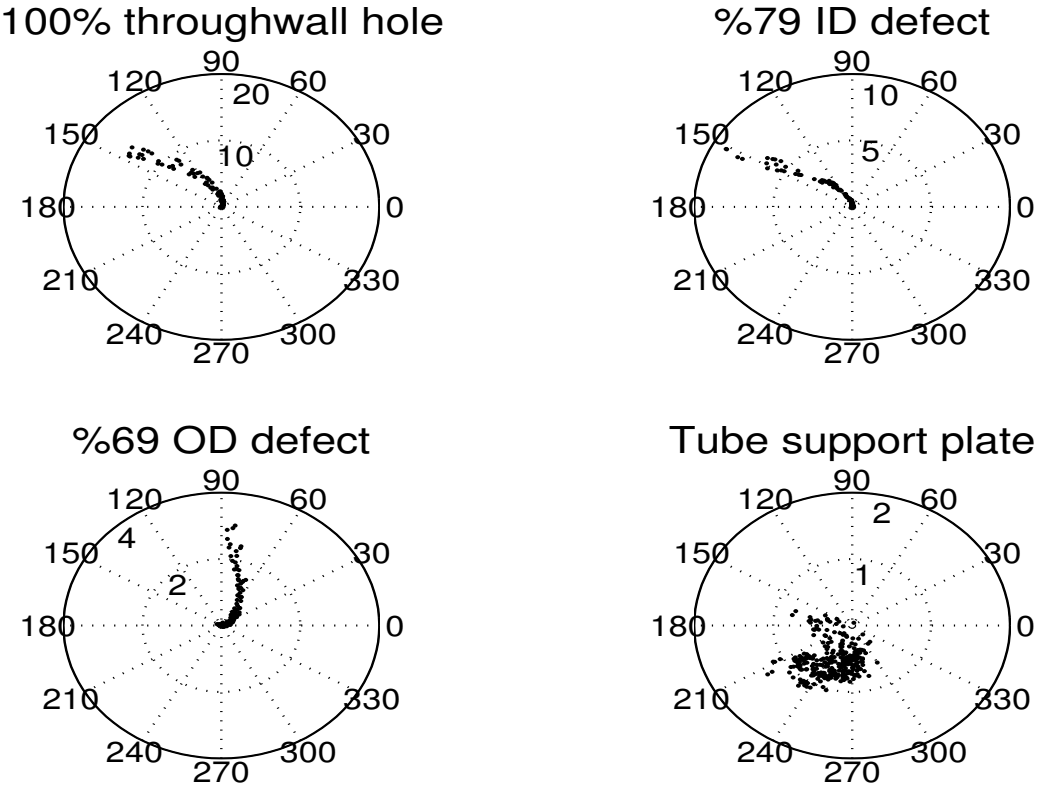
Data acquisition scheme.

A preprocessing step is performed first, with goal to extract meaningful information from the rotating-probe data, see

P. Xiang, S. Ramakrishnan, X. Cai, P. Ramuhalli, R. Polikar, S.S. Udpa, and L. Udpa, "Automated analysis of rotating probe multi-frequency eddy current data from steam generator tubes," *Intl. J. Applied Electrom. Mech.*, vol. 12, pp. 151–164, 2000.



Signal preprocessing: (a) 1-D raw data, (b) 2-D image, and (c) 2-D image after preprocessing.



Impedance-plane plots of typical signals measured by the rotating-probe eddy-current system.

Objective: Characterize the amplitude and phase probability distributions of the potential defects. The estimated distribution parameters can be used for:

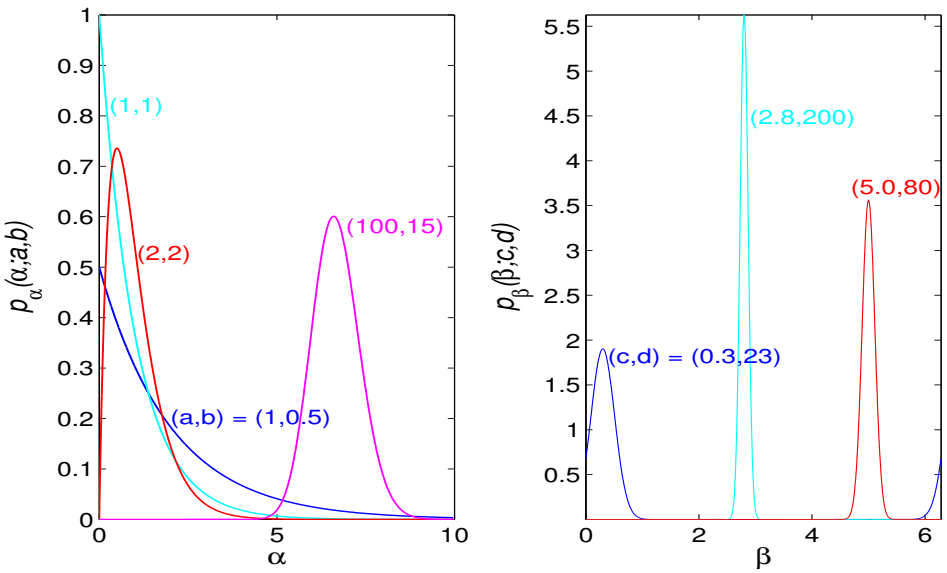
- defect detection,
- defect classification and characterization, e.g. discriminating between inner diameter (ID) and outer diameter (OD) defects,
- denoising.

Statistical Model: K complex measurements $x[k]$, $k = 0, 1, \ldots, K-1$ at the defect location modeled as

$$x[k] = \sqrt{\alpha_k} \cdot e^{j\beta_k} + e[k]$$

where

- $\alpha_k \equiv$ i.i.d. signal powers following a $\mathrm{Gamma}(a, b)$ distribution (Interestingly, in the special case where $a = 1$, the amplitudes $\sqrt{\alpha_k}$ follow a Rayleigh distribution.),
- $\beta_k \equiv$ i.i.d. signal phases independent of the amplitudes, where β_k follow a von Mises distribution (with parameters c and d),
- $e[k]$ i.i.d. zero-mean complex Gaussian noise independent of the signal amplitudes and phases, having known variance σ^2. [The noise variance σ^2 can be estimated from the neighboring pixels that contain only noise.]



Gamma distribution pdf (left)
and von Mises distribution pdf (right).

Pdfs of α_k and β_k:

$$p_\alpha(\alpha_k; a, b) = \frac{b^a}{\Gamma(a)} \alpha_k^{a-1} \exp(-b\alpha_k), \quad \alpha_k > 0, \quad a, b > 0$$

and

$$p_\beta(\beta_k; c, d) = \frac{1}{2\pi I_0(d)} \exp[d\cos(\beta_k - c)], \ 0 < \beta_k \leq 2\pi, \ d > 0$$

where $I_0(\cdot)$ denotes the modified Bessel function of the first kind and order zero. Define the unknown parameter vector

$$\boldsymbol{\lambda} = [a, b, c, d]^T$$

and the vectors of signal amplitudes and phases

$$\boldsymbol{\theta}_k = [\alpha_k, \beta_k]^T, \quad k = 0, 1, \ldots, K-1.$$

Marginal distribution of the kth observation:

$$p_x(x[k]; \boldsymbol{\lambda}) = \int_{\boldsymbol{\Theta}} p_{x|\theta}(x[k] \,|\, \boldsymbol{\theta})\, p_\theta(\boldsymbol{\theta}\,;\, \boldsymbol{\lambda})\, d\boldsymbol{\theta}, \;\; k = 0, 1, \ldots, K-1$$

where $\boldsymbol{\theta} = [\alpha, \beta]^T$, $\boldsymbol{\Theta} = \{(\alpha, \beta) : 0 < \alpha < \infty, 0 < \beta < 2\pi\}$, and

$$\begin{aligned} p_{x|\theta}(x[k] \,|\, \boldsymbol{\theta}) &= \frac{1}{\pi\sigma^2} \exp\Big[- \frac{|x[k] - \sqrt{\alpha} \cdot e^{j\beta}|^2}{\sigma^2} \Big] \\ p_\theta(\boldsymbol{\theta}; \boldsymbol{\lambda}) &= p_\alpha(\alpha; a, b)\; p_\beta(\beta; c, d). \end{aligned}$$

ML estimate of $\boldsymbol{\lambda}$ obtained by maximizing the log-likelihood of $\boldsymbol{\lambda}$ for all measurements $\boldsymbol{x} = [x[0], x[1], \ldots, x[K-1]]^T$:

$$L(\boldsymbol{\lambda}, \boldsymbol{y}) = \sum_{k=0}^{K-1} \log p_x(x[k]; \boldsymbol{\lambda}). \tag{10}$$

Newton-Raphson iteration for finding the ML estimates of $\boldsymbol{\lambda}$ [i.e. maximizing (10)]:

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - \delta^{(i)} \cdot \Big[\underbrace{\frac{\partial^2 L(\boldsymbol{\lambda}^{(i)})}{\partial \boldsymbol{\lambda}\, \partial \boldsymbol{\lambda}^T}}_{\text{Hessian of the log likelihood}} \Big]^{-1} \underbrace{\frac{\partial L(\boldsymbol{\lambda}^{(i)})}{\partial \boldsymbol{\lambda}}}_{\text{gradient}} \tag{11}$$

where the *damping factor* $0 < \delta^{(i)} \leq 1$ is chosen (at every step i) to ensure that the likelihood function (10) increases and the parameter estimates remain in the allowable parameter space $(a, b, d > 0)$.

We utilized the following formulas to compute the gradient vector and Hessian matrix in (11):

$$\frac{\partial}{\partial \lambda_i}\{\log p_x(x;\boldsymbol{\lambda})\} = \frac{1}{p_x(x;\boldsymbol{\lambda})}\int_{\boldsymbol{\Theta}} p_{x|\theta}(x|\boldsymbol{\theta})\frac{\partial p_\theta(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_i}\,d\boldsymbol{\theta}$$

$$\frac{\partial^2}{\partial \lambda_i \partial \lambda_m}\{\log p_x(x;\boldsymbol{\lambda})\} = \frac{1}{p_x(x;\boldsymbol{\lambda})}\int_{\boldsymbol{\Theta}} p_{x|\theta}(x|\boldsymbol{\theta})\frac{\partial^2 p_\theta(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_i \partial \lambda_m}\,d\boldsymbol{\theta}$$
$$-\frac{1}{[p_x(x;\boldsymbol{\lambda})]^2}$$
$$\cdot \int_{\boldsymbol{\Theta}} p_{x|\theta}(x|\boldsymbol{\theta})\frac{\partial p_\theta(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_i}\,d\boldsymbol{\theta} \cdot \int_{\boldsymbol{\Theta}} p_{x|\theta}(x|\boldsymbol{\theta})\frac{\partial p_\theta(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_m}\,d\boldsymbol{\theta}$$

for $i, m = 1, 2, 3, 4$.

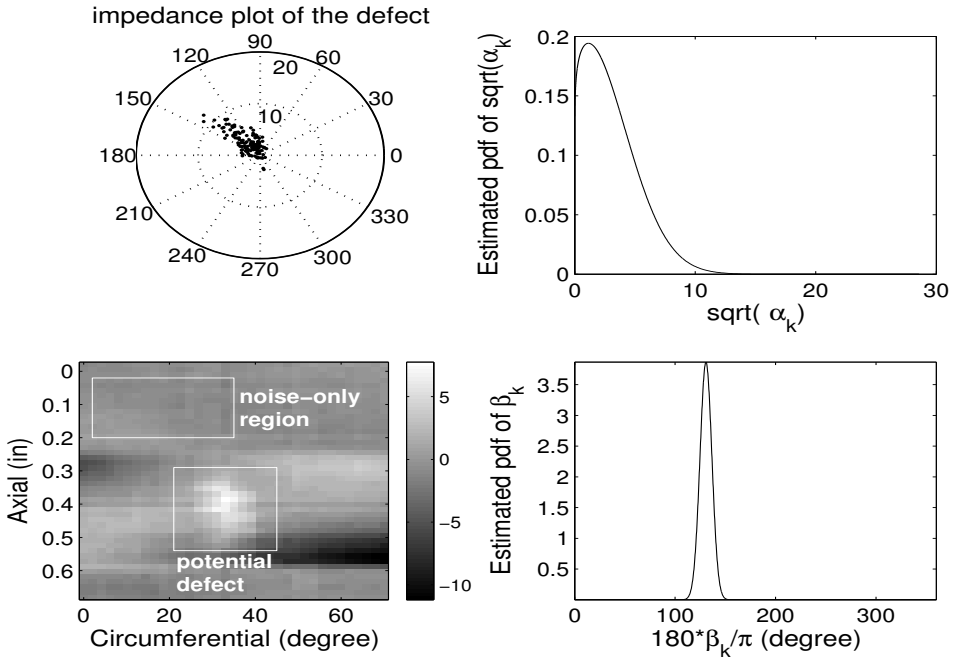
The above integrals with respect to $\boldsymbol{\theta}$ can be easily computed using Gauss quadratures.

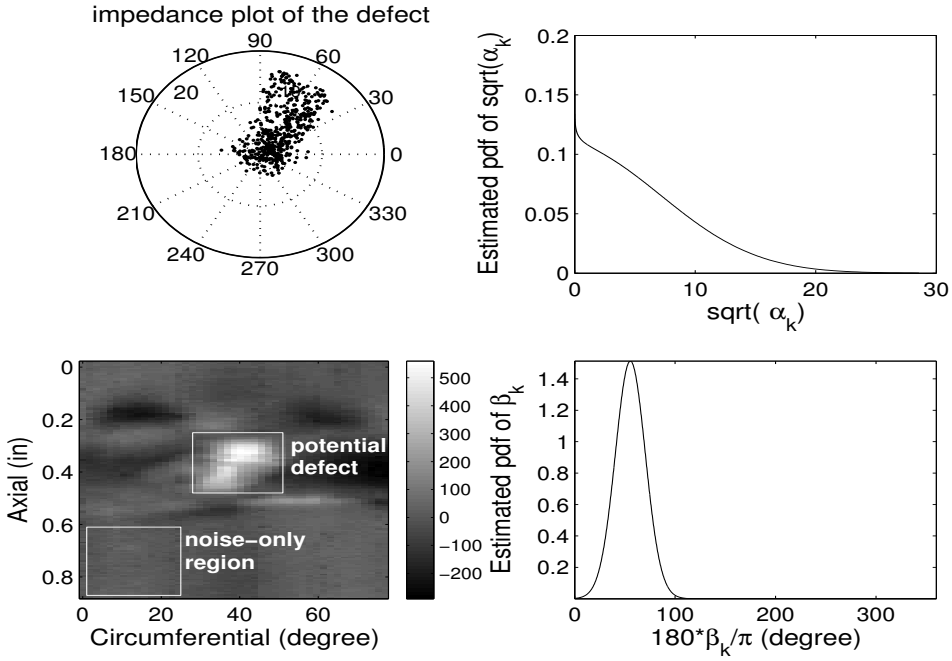
For more details, see

A. Dogandžić and P. Xiang, "A statistical model for eddy-current defect signals from steam generator tubes," in *Rev. Progress Quantitative Nondestructive Evaluation,* D.O.

Thompson and D.E. Chimenti (Eds.), Melville NY: Amer. Inst. Phys., vol. 23, 2004, pp. 605–612.

We now show the performance of the proposed ML estimation method using data containing two real defects, obtained by inspecting steam-generator tubes. The K measurements $x[k],\ k = 0, 1, \ldots, K-1$ were selected from potential defect regions, and the noise variance σ^2 was estimated from the neighboring regions that contain only noise.




impedance plot of the defect
90
120
60
150
20
30
180
0
210
330
240
300
270
Estimated pdf of sqrt(α_k)
sqrt(α_k)
potential defect
noise–only region
Axial (in)
Circumferential (degree)
Estimated pdf of β_k
180*β_k/π (degree)

Least-Squares Approach to Estimation

Suppose that we have a signal model

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{x} = [x[0], \ldots, x[N-1]]^T$ is the vector of observations, $\boldsymbol{H}$ is a known *regression vector matrix,* and $\boldsymbol{w}$ is "error" vector.

LS problem formulation:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|^2.$$

Solution:

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

We can also use weighted least squares, which allows us to assign different weights to measurements. For example, if $\mathrm{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right] = \boldsymbol{C}$ is known, we could use

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

Let $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_p]$:

$$\boldsymbol{x} = \sum_{k=1}^{p} \theta_k \, \boldsymbol{h}_k + \boldsymbol{w}.$$

The "signal part" $\boldsymbol{H\theta}$ of the N-vector $\boldsymbol{x}$ is confined to the p-dimensional subspace spanned by $[\boldsymbol{h}_1 \cdots \boldsymbol{h}_p]$ — the "signal subspace"!

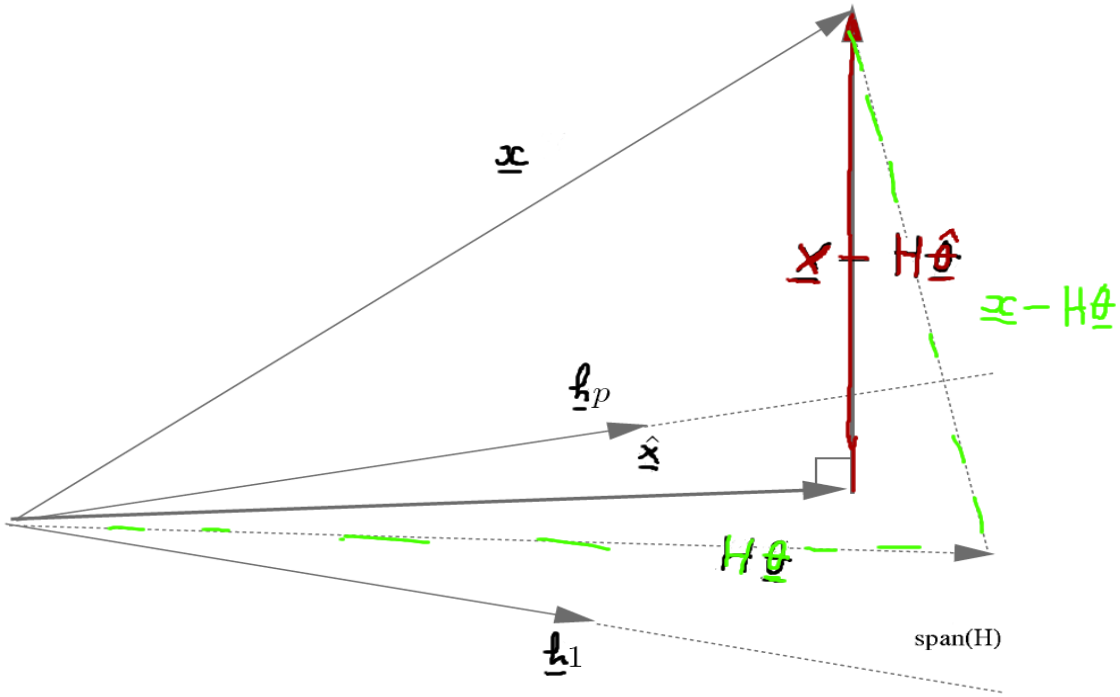
The signal estimate

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

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

$$\widehat{\boldsymbol{w}} = \boldsymbol{x} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}$$

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



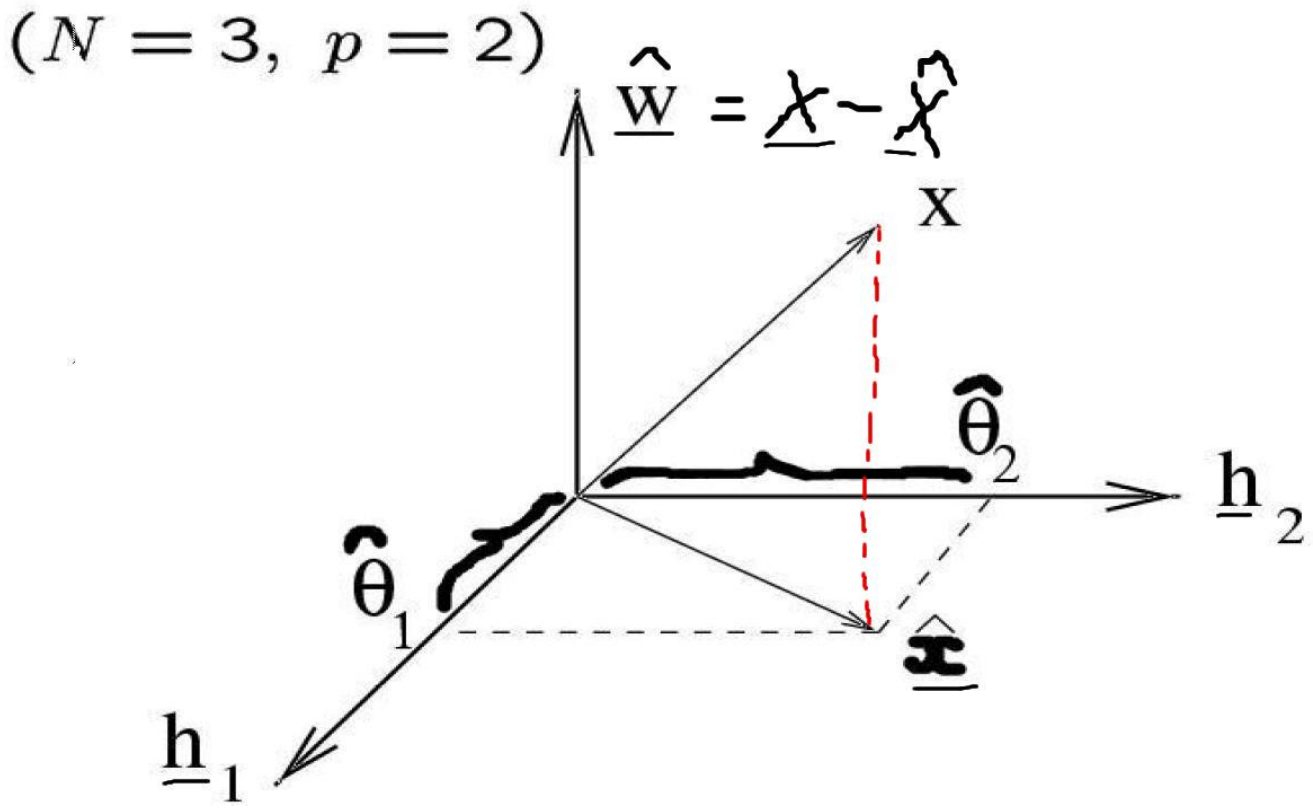
Note: $\boldsymbol{P} = \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$ is the projection matrix onto the column space of $\boldsymbol{H}$, and $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P}$ is the complementary projection matrix.

Recall: Projection matrices are symmetric and idempotent: $\boldsymbol{P} = \boldsymbol{P}^T = \boldsymbol{P}^2$.

Also

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

Example:



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

In general,

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

Computational Aspects of Least Squares

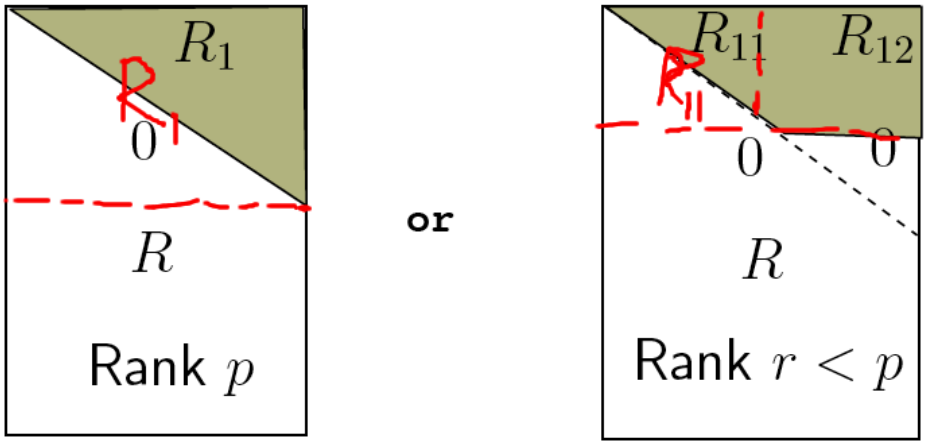
QR decomposition of $\boldsymbol{H}$:

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

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

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

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


R_1
0
R
Rank p
or
r
$p-r$
R_{11}
R_{12}
0
0
R
Rank $r < p$

For the full-rank case,

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|^2 &= \|\boldsymbol{Q}^T\boldsymbol{x} - \boldsymbol{R}\boldsymbol{\theta}\|^2 \\ &= \|\boldsymbol{Q}_1^T\boldsymbol{x} - \boldsymbol{R}_1\boldsymbol{\theta}\|^2 + \|\boldsymbol{Q}_2^T\boldsymbol{x}\|^2 \\ \Longrightarrow \quad \widehat{\boldsymbol{\theta}} &= \boldsymbol{R}_1^{-1}\boldsymbol{Q}_1^T\boldsymbol{x}. \end{aligned}$$

Comments:

- $\boldsymbol{Q}^T\boldsymbol{x}$ yields coordinates of $\boldsymbol{x}$ on columns of $\boldsymbol{Q}$.

- $\widehat{\boldsymbol{x}} = \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}$. Here, the projection matrix $\boldsymbol{P}$ is also known as the *hat matrix* (because it puts the hat on $\boldsymbol{x}$).

- Non full rank case: $\text{rank}(\boldsymbol{H}) = r < p$. We need to solve $\boldsymbol{Q}_1^T\boldsymbol{x} = \boldsymbol{R}_{11}\boldsymbol{\theta}_1 + \boldsymbol{R}_{12}\boldsymbol{\theta}_2$, where $\boldsymbol{Q}_1$ has r columns. There are infinitely many solutions — to get one, arbitrarily set $\boldsymbol{\theta}_2 = \boldsymbol{0}_{(p-r)\times 1}$ and solve for $\boldsymbol{\theta}_1$. Here, $\widehat{\boldsymbol{x}} = \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{x}$ is still well defined, and unique.

Nonlinear Least Squares (NLLS)

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

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

Example: $s[n] = r\cos(\omega n + \phi)$, $n = 0, 1, 2, \ldots, N-1$ gives

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

Nonlinear problem $\Longrightarrow$ we usually need iterative optimization.

Recall the damped Newton-Raphson's method:

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} - \mu_k \cdot \boldsymbol{H}_k^{-1} \boldsymbol{g}_k$$

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

Nonlinear Least Squares
Newton-Raphson Iteration

Define

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

The partial derivatives are then

$$\begin{aligned}
\boldsymbol{g}_k &= \frac{\partial (\boldsymbol{x}-\boldsymbol{f})^T(\boldsymbol{x}-\boldsymbol{f})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = -2(\boldsymbol{f}_\theta^{(k)})^T(\boldsymbol{x}-\boldsymbol{f}^{(k)}) \\
\boldsymbol{H}_k &= \frac{\partial^2 (\boldsymbol{x}-\boldsymbol{f})^T(\boldsymbol{x}-\boldsymbol{f})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = 2(\boldsymbol{f}_\theta^{(k)})^T\boldsymbol{f}_\theta^{(k)} - 2\boldsymbol{G}^{(k)}
\end{aligned}$$

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

Assuming that we have a "small residual" problem, such that $\boldsymbol{x} - \boldsymbol{f}^{(k)} \approx \boldsymbol{0}$ (close to the optimum), the Hessian is approximated by

$$\boldsymbol{H}_k = 2(\boldsymbol{f}_\theta^{(k)})^T\boldsymbol{f}_\theta^{(k)}.$$

Recall: $(\boldsymbol{f}_\theta^{(k)})^T\boldsymbol{f}_\theta^{(k)}$ is the FIM for $\boldsymbol{\theta}$ (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:

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

The search direction $\boldsymbol{\gamma} = [(\boldsymbol{f}_{\boldsymbol{\theta}}^{(k)})^T (\boldsymbol{f}_{\boldsymbol{\theta}}^{(k)})]^{-1} (\boldsymbol{f}_{\boldsymbol{\theta}}^{(k)})^T (\boldsymbol{x} - \boldsymbol{f}^{(k)})$ is the LS solution to

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

which is efficiently computed in MATLAB using

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

Note that the approximate Hessian $\boldsymbol{f}_{\boldsymbol{\theta}}^T \boldsymbol{f}_{\boldsymbol{\theta}}$ is always positive (semi)definite, which is generally *not true for the exact Hessian*!

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

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

where $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}$. For a fixed $\boldsymbol{\alpha}$, the LS solution for $\boldsymbol{\beta}$ is

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

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

$$V_{\mathrm{c}}(\boldsymbol{\alpha}) = \Big\|\boldsymbol{x} - \underbrace{\boldsymbol{H}(\boldsymbol{\alpha})[\boldsymbol{H}^T(\boldsymbol{\alpha})\boldsymbol{H}(\boldsymbol{\alpha})]^{-1}\boldsymbol{H}^T(\boldsymbol{\alpha})}_{\boldsymbol{P}(\boldsymbol{\alpha})}\,\boldsymbol{x}\Big\|^2$$

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

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

Here, $\widehat{\boldsymbol{\alpha}}$ maximizes the projection of $\boldsymbol{x}$ 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 LS: the least squares approach solves a minimization problem in vector spaces.
- In linear problems, LS allows a closed-form solution.
- We need to replace ^{T} with ^{H} (the Hermitian transpose) to obtain the corresponding results for complex data:

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

  minimizes $\|\boldsymbol{x} - \boldsymbol{H}(\boldsymbol{\alpha})\boldsymbol{\beta}\|^2 = [\boldsymbol{x} - \boldsymbol{H}(\boldsymbol{\alpha})\boldsymbol{\beta}]^H[\boldsymbol{x} - \boldsymbol{H}(\boldsymbol{\alpha})\boldsymbol{\beta}]$, i.e.

  $$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{x} - \boldsymbol{H}(\boldsymbol{\alpha})\boldsymbol{\beta}\|^2$$

  and $\boldsymbol{\alpha}$ can be estimated using the concentrated criterion:

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