# Optimal Estimation Methods

# (Lecture 7 – Minimum Variance Estimation & Cramér-Rao Inequality)

Dr. John L. Crassidis

University at Buffalo – State University of New York
Department of Mechanical & Aerospace Engineering
Amherst, NY 14260-4400
johnc@buffalo.edu
http://www.buffalo.edu/~johnc

University at Buffalo The State University of New York

# Probability in Least Squares

- Previously stated that in many cases we wish to weight different measurements differently
  - We now derive the “optimal” weighting matrix based on probability
- Two main approaches shown here
  - Minimum Variance Estimation
  - Maximum Likelihood Estimation
- Note there are others, such as Minimum Risk
- Two main types of estimators (we’ll derive both)
  - Without *a priori* estimates
  - With *a priori* estimates
- Also, we’ll derive the covariance of the estimation errors
  - Discuss the Cramér-Rao lower bound too

Very Important

# Minimum Variance Estimation (i)

- Consider case without *a priori* estimates first  V- zero mean
  - Assume a linear observation model

Measurement Error with $E\{\mathbf{v}\} = \mathbf{0}$ and $E\{\mathbf{v}\mathbf{v}^T\} = R$

$$\overset{(m\times 1)}{\tilde{\mathbf{y}}} = \overset{(m\times n)}{H}\ \overset{(n\times 1)}{\mathbf{x}} + \overset{(m\times 1)}{\mathbf{v}}$$

  - We desire an estimate as a linear combination of the measurements

$$\overset{(n\times 1)}{\hat{\mathbf{x}}} = \overset{(n\times m)}{M}\ \overset{(m\times 1)}{\tilde{\mathbf{y}}} + \overset{(n\times 1)}{\mathbf{n}}$$

  - The minimum variance definition of "optimum" $M$ and $\mathbf{n}$ is that the variance of $n$ estimates from their respective "true" values is minimized

$E[(\hat{x}-x)^2]$ - Variance Definition

  - Leads to the following loss function

$$J_i = \frac{1}{2}E\left\{(\hat{x}_i - x_i)^2\right\}, \quad i = 1, 2, \ldots, n$$

    - This clearly requires $n$ minimizations depending upon the same $M$ and $\mathbf{n}$
  - Let's prove that the "uncoupled" loss function is valid

# Minimum Variance Estimation (ii)

- The linear model **must** also be true when no measurement errors exist, so in this case we have ($v=0$)

$$\tilde{\mathbf{y}} \equiv \mathbf{y} = H\mathbf{x}$$

  - An obvious requirement upon the desired estimator is that perfect measurements should result (if a solution is possible) when $\hat{\mathbf{x}} = \mathbf{x} =$ true state
  - This requirement can be written by substituting $\hat{\mathbf{x}} = \mathbf{x}$ and $\tilde{\mathbf{y}} = H\mathbf{x}$ into the linear measurement model, which gives

$$\mathbf{x} = MH\mathbf{x} + \mathbf{n}$$

  - Thus we conclude that

$$\mathbf{n} = \mathbf{0} \quad \text{and} \quad MH = I$$

  - Note that $MH = I$ will also be shown for unbiased estimates
  - The desired estimator then has the form

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}}$$

  - Need to now find $M$

# Minimum Variance Estimation (iii)

- The unknown $M$-matrix is partitioned by rows as

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix}, \quad M_i \equiv \{M_{i1}\ M_{i2}\ \cdots\ M_{im}\}$$

or

$$M^T = \begin{bmatrix} M_1^T & M_2^T & \cdots & M_n^T \end{bmatrix}$$

- The identity matrix can be partitioned by rows and columns as

$$I = \begin{bmatrix} I_1^r \\ I_2^r \\ \vdots \\ I_n^r \end{bmatrix} = \begin{bmatrix} I_1^c & I_2^c & \cdots & I_n^c \end{bmatrix}, \quad \text{note } I_i^r = (I_i^c)^T$$

# Minimum Variance Estimation (iv)

- The constraint $MH = I$ can now be written as

$$H^T M_i^T = I_i^c, \quad i = 1, 2, \ldots, n$$

or

$$M_i H = I_i^r, \quad i = 1, 2, \ldots, n$$

- The $i^{\text{th}}$ element of the estimate is given by

$$\hat{x}_i = M_i \tilde{\mathbf{y}}, \quad i = 1, 2, \ldots, n \tag{1}$$

- A glance at this equation reveals that the $i^{\text{th}}$ element of the estimate depends only upon the elements of $M$ contained in the $i^{\text{th}}$ row
- A similar statement holds for the constraint equations
  - The elements of the $i^{\text{th}}$ row are independently constrained
- This "uncoupled" nature is the key feature which allows one to carry out the $n$ separate minimizations of the loss function shown before
  - We will show another approach later that does not need this feature per se

# Minimum Variance Estimation (v)

- Substituting Eq. (1) into the loss function gives

$\tilde{y} = Hx + v$

$$J_i = \frac{1}{2} E\left\{ (M_i \tilde{\mathbf{y}} - x_i)^2 \right\}, \quad i = 1, 2, \ldots, n$$

- Substituting the measurement equation gives

$$J_i = \frac{1}{2} E\left\{ (M_i H \mathbf{x} + M_i \mathbf{v} - x_i)^2 \right\}, \quad i = 1, 2, \ldots, n$$

- Incorporating the constraint gives

$M_i H = I_i^r$

$$J_i = \frac{1}{2} E\left\{ (I_i^r \mathbf{x} + M_i \mathbf{v} - x_i)^2 \right\}, \quad i = 1, 2, \ldots, n$$

- Noting $I_i^r \mathbf{x} = x_i$ gives simply

$x_i$'s cancel

$$J_i = \frac{1}{2} E\left\{ (M_i \mathbf{v})^2 \right\}, \quad i = 1, 2, \ldots, n$$

$$= \frac{1}{2} E\left\{ M_i \left( \mathbf{v}\, \mathbf{v}^T \right) M_i^T \right\}, \quad i = 1, 2, \ldots, n$$

$E(vv^T) = R$

As stated previously, zero mean

$E[(v-0)(v-0)^T]$

- Note: the only random variable on the right-hand side is $\mathbf{v}$

$M_i$ can be pulled out.

# Minimum Variance Estimation (vi)

- Assuming that $\mathbf{v}$ has zero mean and $\text{cov}\{\mathbf{v}\} \equiv R = E\{\mathbf{v}\,\mathbf{v}^T\}$ gives

$$J_i = \frac{1}{2} M_i R M_i^T, \quad i = 1, 2, \ldots, n$$

- Need to also account for constraint equations
  - Use the Lagrange multiplier approach

$\frac{\partial}{\partial x}(x C x^T) = (C + C^T) x$

$C + C^T = 2C$ if $C$ is symmetric

$$J_i = \frac{1}{2} M_i \underset{\text{symmetric}}{R} M_i^T + \underbrace{\boldsymbol{\lambda}_i^T \left( I_i^c - H^T M_i^T \right)}_{\text{constraint}}, \quad i = 1, 2, \ldots, n$$

where

$$\boldsymbol{\lambda}_i^T = \{\lambda_{1_i}, \lambda_{2_i}, \ldots, \lambda_{n_i}\}$$

- The necessary conditions give

$$\nabla_{M_i^T} J_i = R M_i^T - H \boldsymbol{\lambda}_i = \mathbf{0}, \quad i = 1, 2, \ldots, n \tag{2}$$

$$\nabla_{\boldsymbol{\lambda}_i} J_i = I_i^c - H^T M_i^T = \mathbf{0}, \text{ or } M_i H = I_i^r, \quad i = 1, 2, \ldots, n \tag{3}$$

# Minimum Variance Estimation (vii)

- From Eq. (2) we have

Assume positive definite

$$M_i = \boldsymbol{\lambda}_i^T H^T R^{-1}, \quad i = 1, 2, \ldots, n$$

- Substituting this equation into Eq. (3) gives

$$\boldsymbol{\lambda}_i^T = I_i^r \left(H^T R^{-1} H\right)^{-1}$$

- Substituting this equation into $M_i$ gives

$$M_i = I_i^r \left(H^T R^{-1} H\right)^{-1} H^T R^{-1}, \quad i = 1, 2, \ldots, n$$

- It then follows that

$$M = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1}$$

# Minimum Variance Estimation (viii)

same as weighted least squares. Set $W = R^{-1}$

- Substituting $M$ into $\hat{\mathbf{x}} = M\tilde{\mathbf{y}}$ gives

$$\boxed{\hat{\mathbf{x}} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} \tilde{\mathbf{y}}}$$

- This is referred to as the *Gauss-Markov Theorem*

- Some observations
  - The minimal variance estimator is identical to the least squares estimator provided that the weight matrix is identified as the inverse of the observation error covariance optimal for this loss function
  - Also, the "sequential least squares estimation" results are seen to embody a special case "sequential minimal variance estimation"
    - It is simply necessary to employ $R^{-1}$ as $W$ in the sequential least squares formulation
    - But we still require $R^{-1}$ to have the block diagonal structure assumed for $W$

# Minimum Variance Estimation (ix)

- Another approach
  - Define the error covariance matrix for an unbiased estimator

$$P = E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\}$$

Variance of estimation errors

  - Minimum variance estimation is equivalent to minimizing the trace of $P$
  - Need to also satisfy constraint $MH = I$
  - Use method of Lagrange multipliers to append loss function

$$J = \frac{1}{2}\mathrm{Tr}\left[E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\}\right] + \mathrm{Tr}\left[\Lambda(I - MH)\right]$$

where $\Lambda$ is a matrix of Lagrange multipliers

  - Note: covariance can also be found using *Parallel Axis Theorem* for an unbiased estimate

x is not a random variable, pull out of integral, Truth

$$\begin{aligned} E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\} &= E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\} - E\left\{\mathbf{x}\right\}E\left\{\mathbf{x}\right\}^T \\ &= E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\} - \mathbf{x}\,\mathbf{x}^T \end{aligned}$$

# Minimum Variance Estimation (x)

- We have

$$
\begin{aligned}
P &= E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\} - \mathbf{x}\,\mathbf{x}^T \\
&= E\left\{M\tilde{\mathbf{y}}\,\tilde{\mathbf{y}}^T M^T\right\} - \mathbf{x}\,\mathbf{x}^T \\
&= E\left\{(MH\mathbf{x} + M\mathbf{v})(MH\mathbf{x} + M\mathbf{v})^T\right\} - \mathbf{x}\,\mathbf{x}^T
\end{aligned}
$$

Do this

  - Now use $E\{\mathbf{v}\} = \mathbf{0}$ and $E\{\mathbf{v}\,\mathbf{v}^T\} = R$

$$P = MRM^T + MH\mathbf{x}\,\mathbf{x}^T H^T M^T - \mathbf{x}\,\mathbf{x}^T$$

  - Noting that $MH = I$ leads to

$$P = MRM^T$$

- Therefore, the loss function becomes

$$J = \frac{1}{2}\mathrm{Tr}(MRM^T) + \mathrm{Tr}\left[\Lambda(I - MH)\right]$$

  - Again, the goal is to find $M$ that minimizes $J$

# Minimum Variance Estimation (xi)

- Consider the following useful identities

$$\frac{\partial}{\partial A}\mathrm{Tr}(BAC) = B^T C^T$$
$$\frac{\partial}{\partial A}\mathrm{Tr}(ABA^T) = A(B + B^T)$$

  - Thus, we have the following necessary conditions

$$\nabla_M J = MR - \Lambda^T H^T = 0$$
$$\nabla_\Lambda J = I - MH = 0$$

  - Two equations for $M$ and $\Lambda^T$
  - Solving the first equation for $M$ gives

$$M = \Lambda^T H^T R^{-1}$$

  - Substituting this into the second equation gives

$$\Lambda^T = (H^T R^{-1} H)^{-1}$$

# Minimum Variance Estimation (xii)

- Note that $\Lambda$ is a symmetric matrix
  - It also has a physical interpretation
  - This is equivalent to the error-covariance matrix, which tells us about the quality of the estimate
  - The "larger" its value, the worse the estimate will be
  - This will be discussed in detail later
- Substituting $\Lambda^T$ back into $M$ gives

$$M = (H^T R^{-1} H)^{-1} H^T R^{-1}$$

- This gives exactly the same solution as before
  - Note that the "decoupling" assumptions are actually in the loss function
  - We choose to minimize the trace of the covariance, which ignores the correlations (this is the decoupling)
  - Other possible forms for the loss function can be chosen, such as minimizing the infinity norm

# Unbiased Estimates (i)

$\hat{x}$ is random variable here because it depends on $\tilde{y}$ which depends on random variables

- An estimator $\hat{\mathbf{x}}(\tilde{\mathbf{y}})$ is said to be an "unbiased estimator" of $\mathbf{x}$ if $E\{\hat{\mathbf{x}}(\tilde{\mathbf{y}})\} = \mathbf{x}$ for every possible value of $\mathbf{x}$
  - If $\hat{\mathbf{x}}$ is biased, the difference $E\{\hat{\mathbf{x}}(\tilde{\mathbf{y}})\} - \mathbf{x}$ is called the "bias" of $\hat{\mathbf{x}}$
- Go back to the previous estimator form

$\tilde{y} = Hx + v$

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}}$$
$$= MH\mathbf{x} + M\mathbf{v}$$

truth

$MH\,E(\cancel{x}) = x$

$M\,E(v) = 0$

$E(\hat{x}) = x = E(MHx) + E(Mv) = MHx + 0$

  - Taking the expectation of both sides and assuming zero-mean measurement error, so that $E\{\mathbf{v}\} = \mathbf{0}$, gives

$$E\{\hat{\mathbf{x}}\} = MH\mathbf{x}$$

$x = MHx \Rightarrow MH = I$

  - Thus for an unbiased estimate we must have $MH = I$
  - Same result as before!

# Unbiased Estimates (ii)

- Sample Variance Example with Data $\{\tilde{y}(t_1), \tilde{y}(t_2), \ldots, \tilde{y}(t_m)\}$
  - Compute sample variance using
$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}[\tilde{y}(t_i) - \hat{\mu}]^2$$
  - Note, many calculators give the option of dividing by $m$ or $m-1$
  - Check to see if this estimate is unbiased using $m-1$
  - Defining $E\{\hat{\sigma}^2\} \equiv S^2$ with $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\tilde{y}(t_i)$ gives

$$S^2 = \frac{1}{m-1}E\left\{\left[\sum_{i=1}^{m}\tilde{y}^2(t_i) - 2\tilde{y}(t_i)\hat{\mu} + \hat{\mu}^2\right]\right\}$$
$$= \frac{1}{m-1}\left[E\left\{\sum_{i=1}^{m}\tilde{y}^2(t_i)\right\} - \frac{2}{m}E\left\{\sum_{i=1}^{m}\tilde{y}(t_i)\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]\right\} + \frac{1}{m^2}E\left\{\sum_{i=1}^{m}\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]^2\right\}\right]$$
$$= \frac{1}{m-1}\left[\sum_{i=1}^{m}E\left\{[\tilde{y}(t_i)]^2\right\} - \frac{2}{m}E\left\{\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]^2\right\} + \frac{m}{m^2}E\left\{\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]^2\right\}\right]$$
$$= \frac{1}{m-1}\left[\sum_{i=1}^{m}E\left\{[\tilde{y}(t_i)]^2\right\} - \frac{1}{m}E\left\{\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]^2\right\}\right]$$

# Unbiased Estimates (iii)

- For any random variable $z$ the variance is computed from (using the parallel axis theorem) $\mathrm{var}\{z\} = E\{z^2\} - E\{z\}^2$
  - Then applying the variance equation gives

$$
\begin{aligned}
S^2 &= \frac{1}{m-1}\left[\sum_{i=1}^{m}(\sigma^2+\mu^2) - \frac{1}{m}\left\{\mathrm{var}\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right] + \left[E\left\{\sum_{i=1}^{m}\tilde{y}(t_i)\right\}\right]^2\right\}\right] \\
&= \frac{1}{m-1}\left[m\sigma^2 + m\mu^2 - \frac{1}{m}m\sigma^2 - \frac{1}{m}m^2\mu^2\right] \\
&= \frac{1}{m-1}\left[m\sigma^2 - \sigma^2\right] \\
&= \sigma^2
\end{aligned}
$$

Divide by $\frac{1}{m}$ gives biased estimator

- Therefore, this estimator is unbiased
  - However, the sample variance shown in this example does not give an estimate with the smallest mean-square-error for Gaussian (normal) distributions

# Unbiased Estimates (iv)

- A more general definition for an unbiased estimator is

$$E\left\{\hat{\mathbf{x}}_k(\tilde{\mathbf{y}})\right\} = \mathbf{x} \quad \text{for all } k$$

  - For the sequential estimator we wish to have the form

$$\hat{\mathbf{x}}_{k+1} = G_{k+1}\hat{\mathbf{x}}_k + K_{k+1}\tilde{\mathbf{y}}_{k+1}$$

    where $G_{k+1}$ and $K_{k+1}$ are deterministic matrices
  - Substituting the measurement equation at $k+1$ gives

$$\hat{\mathbf{x}}_{k+1} = G_{k+1}\hat{\mathbf{x}}_k + K_{k+1}H_{k+1}\mathbf{x}_{k+1} + K_{k+1}\mathbf{v}_{k+1}$$

  - Taking the expectation gives

$$E\{\hat{\mathbf{x}}_{k+1}\} = G_{k+1}E\{\hat{\mathbf{x}}_k\} + K_{k+1}H_{k+1}\mathbf{x}_{k+1}$$

  - Noting that the unbiased condition must be valid for all $k$ leads to

$$G_{k+1} = I - K_{k+1}H_{k+1}$$

    Then

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k + K_{k+1}(\tilde{\mathbf{y}}_{k+1} - H_{k+1}\hat{\mathbf{x}}_k)$$

  - This is exactly the sequential process
    - We have now shown that it produces unbiased estimates though

# *A Priori* Estimation (i)

- Consider case with *a priori* estimates now

v & w are both zero mean

  - Measurement model is same as before

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v}, \quad \text{with} \quad E\{\mathbf{v}\} = \mathbf{0} \quad \text{and} \quad E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R$$

  - Now consider an *a priori* estimate with model given by

$$\hat{\mathbf{x}}_a = \mathbf{x} + \mathbf{w}, \quad \text{with} \quad E\{\mathbf{w}\} = \mathbf{0} \quad \text{and} \quad E\left\{\mathbf{w}\,\mathbf{w}^T\right\} = Q$$

  - We also assume that the measurement errors and *a priori* errors are uncorrelated so that $E\{\mathbf{w}\,\mathbf{v}^T\} = E\{\mathbf{v}\,\mathbf{w}^T\} = 0$

- We desire to estimate $\mathbf{x}$ as a linear combination of the measurements and *a priori* estimates as

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a + \mathbf{n}$$

$E[Mv] = E[Nw] = 0$

  - For unbiased estimates we require

$$E\{\hat{\mathbf{x}}\} = E\{M(H\mathbf{x} + \mathbf{v})\} + E\{N(\mathbf{x} + \mathbf{w}) + \mathbf{n}\} = (MH + N)\mathbf{x} + \mathbf{n} = \mathbf{x}$$

  - Then $\mathbf{n} = \mathbf{0}$ and $MH + N = I$ is required for an unbiased estimate
  - So the actual form is given by

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a, \quad \text{subject to} \quad MH + N = I$$

# *A Priori* Estimation (ii)

- Loss function for this case becomes

$$J = \frac{1}{2}\mathrm{Tr}\left[E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\}\right] + \mathrm{Tr}\left[\Lambda(I - MH - N)\right]$$

  - Substituting the models into the estimate equation gives

$$\begin{aligned}\hat{\mathbf{x}} &= M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a \\ &= (MH+N)\mathbf{x} + M\mathbf{v} + N\mathbf{w} \\ &= \mathbf{x} + M\mathbf{v} + N\mathbf{w}\end{aligned}$$

    where the equality constraint $MH + N = I$ was used
  - Then we have

$$\begin{aligned}E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} = {} & ME\left\{\mathbf{v}\cancelto{R}{\mathbf{v}}^T\right\}M^T + NE\left\{\mathbf{w}\cancelto{Q}{\mathbf{w}}^T\right\}N^T \\ & + ME\left\{\mathbf{v}\cancelto{0}{\mathbf{w}}^T\right\}N^T + NE\left\{\mathbf{w}\cancelto{0}{\mathbf{v}}^T\right\}M^T\end{aligned}$$

  - Use $E\{\mathbf{v}\,\mathbf{v}^T\} = R$, $E\{\mathbf{w}\,\mathbf{w}^T\} = Q$ and $E\{\mathbf{w}\,\mathbf{v}^T\} = E\{\mathbf{v}\,\mathbf{w}^T\} = 0$

$$E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} = MRM^T + NQN^T$$

# *A Priori* Estimation (iii)

- So the loss function becomes

$$J = \frac{1}{2}\mathrm{Tr}(MRM^T + NQN^T) + \mathrm{Tr}\left[\Lambda(I - MH - N)\right]$$

- The necessary conditions are

$$\nabla_M J = MR - \Lambda^T H^T = 0 \qquad (1)$$

$$\nabla_N J = NQ - \Lambda^T = 0 \qquad (2)$$

$$\nabla_\Lambda J = I - MH - N = 0 \qquad (3)$$

- Solving Eq. (1) for $M$ gives

$$M = \Lambda^T H^T R^{-1}$$

- Solving Eq. (2) for $N$ gives

$$N = \Lambda^T Q^{-1}$$

- Substituting these into Eq. (3) and solving for $\Lambda^T$ gives

$$\Lambda^T = (H^T R^{-1} H + Q^{-1})^{-1}$$

  - This is the covariance for the *a priori* estimates

## *A Priori* Estimation (iv)

- Substituting $\Lambda^T$ into Eqs. (1) and (2) gives

$$M = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} H^T R^{-1}$$

$$N = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} Q^{-1}$$

- Therefore the *a priori* estimate equation is

$$\boxed{\hat{\mathbf{x}} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} \left(H^T R^{-1} \tilde{\mathbf{y}} + Q^{-1} \hat{\mathbf{x}}_a\right)}$$

- Some observations
  - With poor *a priori* knowledge we have $Q \to \infty$ and $Q^{-1} \to 0$, which reduces down to the minimum variance estimator! $\hat{x} = (H^T R^{-1} H)^{-1} (H^T R^{-1} \tilde{y})$
  - With poor measurements we have $R \to \infty$ and $R^{-1} \to 0$, which gives the result $\hat{\mathbf{x}} = \hat{\mathbf{x}}_a$, an intuitively pleasing result!

## Cramér-Rao Inequality (i)

- One of the most useful and important concepts in estimation theory
  - The Cramér-Rao inequality can be used to give us a lower bound on the expected errors between the estimated quantities and the true values from the known statistical properties of the measurement errors
  - Consider the conditional density $p(\tilde{\mathbf{y}}|\mathbf{x})$
  - The Cramér-Rao inequality is given by

$$\boxed{P \equiv E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} \geq F^{-1}}$$

where the *Fisher information matrix*, $F$, is given by

$$F = E\left\{\left[\frac{\partial}{\partial \mathbf{x}} \ln p(\tilde{\mathbf{y}}|\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}} \ln p(\tilde{\mathbf{y}}|\mathbf{x})\right]^T\right\} = -E\left\{\frac{\partial^2}{\partial \mathbf{x}\,\partial \mathbf{x}^T} \ln p(\tilde{\mathbf{y}}|\mathbf{x})\right\}$$

  - Note, Cramér-Rao inequality is only valid for **unbiased estimates** $E(\hat{x}) = x$

# Cramér-Rao Inequality (ii)

- Proof begins by using

$$\int_{-\infty}^{\infty} p(\tilde{\mathbf{y}}|\mathbf{x})\, d\tilde{\mathbf{y}} = 1$$

- Taking the partial with respect to $\mathbf{x}$ gives

$$\frac{\partial}{\partial \mathbf{x}} \int_{-\infty}^{\infty} p(\tilde{\mathbf{y}}|\mathbf{x})\, d\tilde{\mathbf{y}} = \int_{-\infty}^{\infty} \left[\frac{\partial p(\tilde{\mathbf{y}}|\mathbf{x})}{\partial \mathbf{x}}\right] d\tilde{\mathbf{y}} = \mathbf{0}$$

- Since the estimate is assumed to be unbiased we have

$$E\left\{\hat{\mathbf{x}} - \mathbf{x}\right\} = \underbrace{\int_{-\infty}^{\infty} (\hat{\mathbf{x}} - \mathbf{x})\, p\left(\tilde{\mathbf{y}}|\mathbf{x}\right)}_{\text{definition}} d\tilde{\mathbf{y}} = \mathbf{0} \quad \leftarrow \text{unbiased}$$

- Differentiating both sides with respect to $\mathbf{x}$ gives

$$\int_{-\infty}^{\infty} (\hat{\mathbf{x}} - \mathbf{x}) \left[\frac{\partial p(\tilde{\mathbf{y}}|\mathbf{x})}{\partial \mathbf{x}}\right]^T d\tilde{\mathbf{y}} - I \cancelto{1}{\int_{-\infty}^{\infty} p\left(\tilde{\mathbf{y}}|\mathbf{x}\right) d\tilde{\mathbf{y}}} = 0$$

$$\int_{-\infty}^{\infty} (\hat{\mathbf{x}} - \mathbf{x}) \left[\frac{\partial p(\tilde{\mathbf{y}}|\mathbf{x})}{\partial \mathbf{x}}\right]^T d\tilde{\mathbf{y}} - I = 0$$

# Cramér-Rao Inequality (iii)

- Next, we use the following logarithmic differentiation rule

$$\frac{\partial p(\tilde{\mathbf{y}}|\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial}{\partial \mathbf{x}} \ln[p(\tilde{\mathbf{y}}|\mathbf{x})]\right] p(\tilde{\mathbf{y}}|\mathbf{x})$$

- Substitute this into the previous equation to give

$$I = \int_{-\infty}^{\infty} \left(\mathbf{a}\,\mathbf{b}^T\right) d\tilde{\mathbf{y}} \qquad (1)$$

where

$$\mathbf{a} \equiv p(\tilde{\mathbf{y}}|\mathbf{x})^{1/2}\,(\hat{\mathbf{x}} - \mathbf{x})$$

$$\mathbf{b} \equiv p(\tilde{\mathbf{y}}|\mathbf{x})^{1/2} \left[\frac{\partial}{\partial \mathbf{x}} \ln[p(\tilde{\mathbf{y}}|\mathbf{x})]\right]$$

- Note that $P$ and $F$ can be written now as

$$P = \int_{-\infty}^{\infty} \left(\mathbf{a}\,\mathbf{a}^T\right) d\tilde{\mathbf{y}}, \quad F = \int_{-\infty}^{\infty} \left(\mathbf{b}\,\mathbf{b}^T\right) d\tilde{\mathbf{y}}$$

# Cramér-Rao Inequality (iv)

- Multiply Eq. (1) on the left by an arbitrary row vector $\boldsymbol{\alpha}^T$ and on the right by an arbitrary column vector $\boldsymbol{\beta}$

$$\boldsymbol{\alpha}^T \boldsymbol{\beta} = \int_{-\infty}^{\infty} \boldsymbol{\alpha}^T \left(\mathbf{a}\,\mathbf{b}^T\right) \boldsymbol{\beta}\, d\tilde{\mathbf{y}}$$

- Next, we make use of the *Schwartz inequality*

$$\left[\int_{-\infty}^{\infty} g\left(\tilde{\mathbf{y}}|\mathbf{x}\right) h\left(\tilde{\mathbf{y}}|\mathbf{x}\right) d\tilde{\mathbf{y}}\right]^2 \leq \int_{-\infty}^{\infty} g^2\left(\tilde{\mathbf{y}}|\mathbf{x}\right) d\tilde{\mathbf{y}} \int_{-\infty}^{\infty} h^2\left(\tilde{\mathbf{y}}|\mathbf{x}\right) d\tilde{\mathbf{y}}$$

If $\int_{-\infty}^{\infty} a(\mathbf{x}) b(\mathbf{x})\, d\mathbf{x} = 1$ then $\int_{-\infty}^{\infty} a^2(\mathbf{x})\, d\mathbf{x} \int_{-\infty}^{\infty} b^2(\mathbf{x})\, d\mathbf{x} \geq 1$; the equality holds if $a(\mathbf{x}) = c\, b(\mathbf{x})$ where $c$ is not a function of $\mathbf{x}$.

- Define the following quantities

$$g\left(\tilde{\mathbf{y}}|\mathbf{x}\right) = \boldsymbol{\alpha}^T \mathbf{a}$$
$$h\left(\tilde{\mathbf{y}}|\mathbf{x}\right) = \mathbf{b}^T \boldsymbol{\beta}$$

# Cramér-Rao Inequality (v)

- Then the Schwartz inequality becomes

$$\left[\int_{-\infty}^{\infty} \boldsymbol{\alpha}^T(\mathbf{a}\mathbf{b}^T)\boldsymbol{\beta}\, d\tilde{\mathbf{y}}\right]^2 \leq \int_{-\infty}^{\infty} \boldsymbol{\alpha}^T(\mathbf{a}\,\mathbf{a}^T)\boldsymbol{\alpha}\, d\tilde{\mathbf{y}} \int_{-\infty}^{\infty} \boldsymbol{\beta}^T(\mathbf{b}\,\mathbf{b}^T)\boldsymbol{\beta}\, d\tilde{\mathbf{y}}$$

- Using the definitions of $P$ and $F$ and

$$\int_{-\infty}^{\infty} \mathbf{a}\mathbf{b}^T\, d\tilde{\mathbf{y}} = \int_{-\infty}^{\infty} (\hat{\mathbf{x}} - \mathbf{x}) \left[\frac{\partial p(\tilde{\mathbf{y}}|\mathbf{x})}{\partial \mathbf{x}}\right]^T d\tilde{\mathbf{y}} = I$$

gives

$$\left(\boldsymbol{\alpha}^T\boldsymbol{\beta}\right)^2 \leq \left(\boldsymbol{\alpha}^T P \boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}^T F \boldsymbol{\beta}\right)$$

- Finally, using the particular choice $\boldsymbol{\beta} = F^{-1}\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha}^T(P - F^{-1})\boldsymbol{\alpha} \geq 0$$

- Since is $\alpha$ arbitrary then $P \geq F^{-1}$ must be true, which proves the Cramér-Rao Inequality

- Consider the measurement model

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v}, \quad \text{with} \quad E\{\mathbf{v}\} = \mathbf{0} \quad \text{and} \quad E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R$$

  - To determine the mean of the observation model, we take the expectation of both sides

$Hx$

$$\boldsymbol{\mu} \equiv E\{\tilde{\mathbf{y}}\} = E\{H\mathbf{x}\} + E\{\mathbf{v}\} = H\mathbf{x}$$

$\tilde{y} - \mu = Hx + v - Hx = v$

  - The covariance is then given by

$$\begin{aligned} \operatorname{cov}\{\tilde{\mathbf{y}}\} &\equiv E\left\{(\tilde{\mathbf{y}} - \boldsymbol{\mu})(\tilde{\mathbf{y}} - \boldsymbol{\mu})^T\right\} \\ &= E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R \end{aligned}$$

  - The conditional density is then given by

$$p(\tilde{\mathbf{y}}|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}\left[\det(R)\right]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]\right\}$$

# Cramér-Rao Inequality (vii)

- Taking the natural log gives

$$\ln\left[p(\tilde{\mathbf{y}}|\mathbf{x})\right] = -\frac{1}{2}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right] - \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln\left[\det(R)\right]$$

- Carry out the computations for the Fisher Information Matrix

$$F = -E\left\{\frac{\partial^2}{\partial\mathbf{x}\,\partial\mathbf{x}^T}\ln p(\tilde{\mathbf{y}}|\mathbf{x})\right\} = (H^T R^{-1} H)$$

- Hence, the Cramér-Rao inequality is $P \geq (H^T R^{-1} H)^{-1}$
- Let us now find an expression for the estimate covariance $P$
- Estimate and measurement models

$$\hat{\mathbf{x}} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1}\tilde{\mathbf{y}}$$

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v}$$

- Substituting the measurement model into the estimate gives

$$\begin{aligned}\hat{\mathbf{x}} &= \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} H\mathbf{x} + \left(H^T R^{-1} H\right)^{-1} H^T R^{-1}\mathbf{v}\\ &= \mathbf{x} + \left(H^T R^{-1} H\right)^{-1} H^T R^{-1}\mathbf{v}\end{aligned}$$

# Cramér-Rao Inequality (viii)

- The expectation of the estimate is given by

$$E\{\hat{\mathbf{x}}\} = \mathbf{x} + \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} E\{\mathbf{v}\} = \mathbf{x}$$

since $E\{\mathbf{v}\} = \mathbf{0}$

- The covariance is

$$\begin{aligned} P &\equiv E\left\{\left(\hat{\mathbf{x}} - E\{\hat{\mathbf{x}}\}\right)\left(\hat{\mathbf{x}} - E\{\hat{\mathbf{x}}\}\right)^T\right\} \\ &= E\left\{\left(\hat{\mathbf{x}} - \mathbf{x}\right)\left(\hat{\mathbf{x}} - \mathbf{x}\right)^T\right\} \\ &= (H^T R^{-1} H)^{-1} H^T R^{-1} E\left\{\mathbf{v}\,\mathbf{v}^T\right\} R^{-1} H (H^T R^{-1} H)^{-1} \end{aligned}$$

- From $E\{\mathbf{v}\,\mathbf{v}^T\} = R$ we have

$$\begin{aligned} P &= (H^T R^{-1} H)^{-1} H^T R^{-1} R\, R^{-1} H (H^T R^{-1} H)^{-1} \\ &= (H^T R^{-1} H)^{-1} \end{aligned}$$

- Therefore, the equality is satisfied, so the least squares estimate from the Gauss-Markov Theorem is the most efficient possible estimate!
- Estimator is thus called ***efficient***

# Cramér-Rao Inequality (ix)

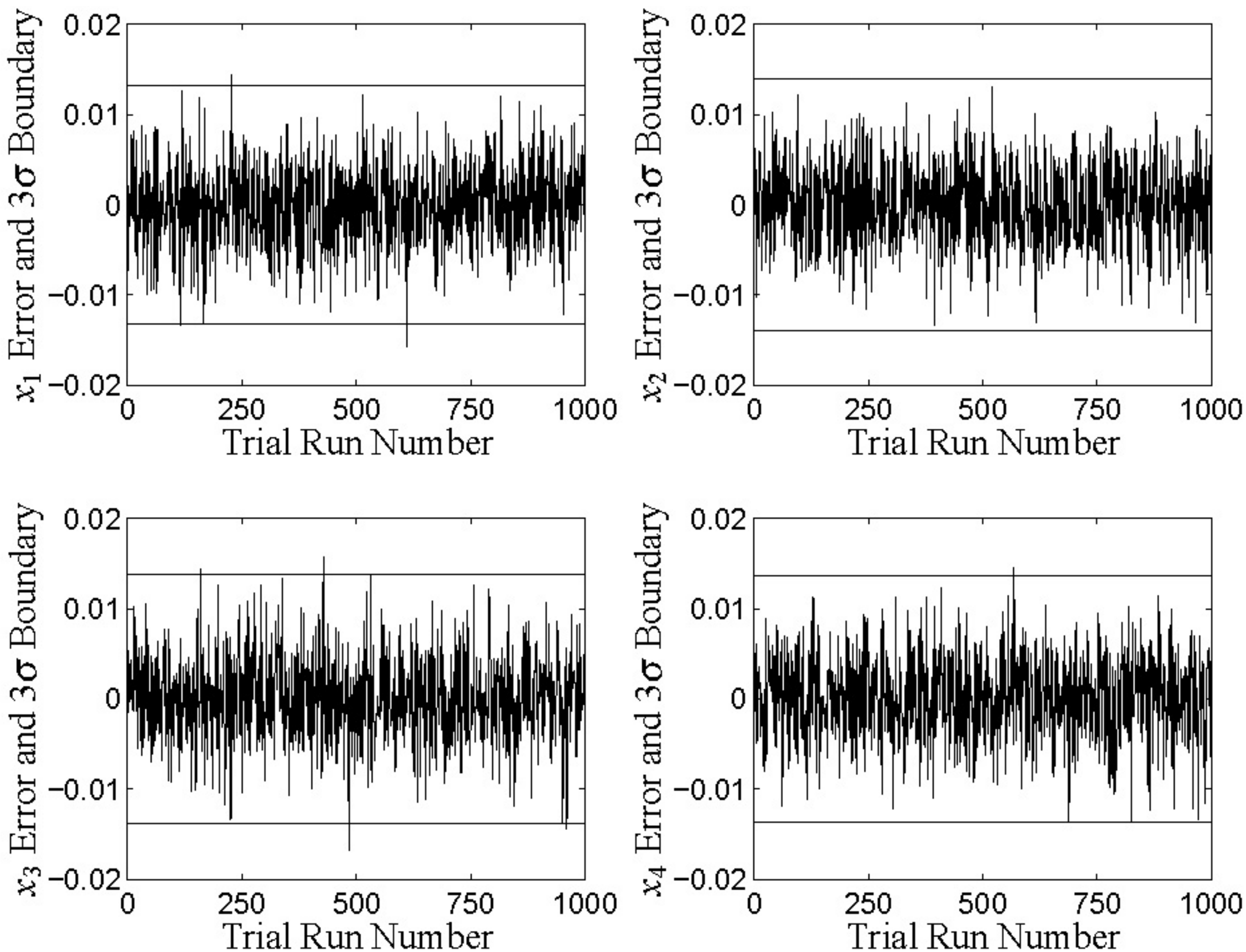
- Important result
  - Note again that the covariance of the estimation errors is given by

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

  - We never know the truth in the real world
  - But if we know the characteristics of the measurement errors (zero-mean with known covariance $R$) then we can determine a bound on the estimation errors from a statistical point of view
  - This is certainly useful information!
  - Note that the covariance of the estimation errors can be computed without ever computing the estimate
  - Helps to assess the performance of the estimator
    - For example, useful to develop an error budget for the total attitude errors in a spacecraft attitude control design

# Example I (i)

$$\tilde{y}(t) = \cos(t) + 2\sin(t) + \cos(2t) + 2\sin(3t) + v(t), \quad R = 0.01I$$




Ran 1,000 Monte Carlo runs

$3\sigma$ boundaries found from taking the square root of the diagonal elements of $P$ and multiplying the result by 3

Bounds actual errors well

# Example I (ii)

```
% True System
dt=0.01;tf=10;
t=[0:dt:tf]';
m=length(t);
y=cos(t)+2*sin(t)+cos(2*t)+2*sin(3*t);

% Pre-allocate Space
xe=zeros(1000,4);
pcov=zeros(1000,4);

% Monte Carlo Simulation
for i=1:1000,
 ym=y+0.1*randn(m,1);w=1/0.01;
 h=[cos(t) sin(t) cos(2*t) sin(3*t)];
 p=inv(h'*w*h);
 xe(i,:)=(p*h'*w*ym)';
 pcov(i,:)=diag(p)';
end
```

# Example I (iii)

```
% Plot Results
subplot(221)
plot([1:1000],xe(:,1)-1,[1:1000],pcov(:,1).^(0.5)*3,[1:1000],-pcov(:,1).^(0.5)*3);
axis([0 1000 -0.02 0.02]);
set(gca,'fontsize',12);
set(gca,'xtick',[0 250 500 750 1000]);
set(gca,'ytick',[-0.02 -0.01 0 0.01 0.02]);
xlabel('Trial Run Number')
ylabel('{\it x}_1 Error and 3{\sigma} Outlier')

subplot(222)
plot([1:1000],xe(:,2)-2,[1:1000],pcov(:,2).^(0.5)*3,[1:1000],-pcov(:,2).^(0.5)*3);
axis([0 1000 -0.02 0.02]);
set(gca,'fontsize',12);
set(gca,'xtick',[0 250 500 750 1000]);
set(gca,'ytick',[-0.02 -0.01 0 0.01 0.02]);
xlabel('Trial Run Number')
ylabel('{\it x}_2 Error and 3{\sigma} Outlier')
```

# Example I (iv)

```
subplot(223)
plot([1:1000],xe(:,3)-1,[1:1000],pcov(:,3).^(0.5)*3,[1:1000],-pcov(:,3).^(0.5)*3);
axis([0 1000 -0.02 0.02]);
set(gca,'fontsize',12);
set(gca,'xtick',[0 250 500 750 1000]);
set(gca,'ytick',[-0.02 -0.01 0 0.01 0.02]);
xlabel('Trial Run Number')
ylabel('{\it x}_3 Error and 3{\sigma} Outlier')

subplot(224)
plot([1:1000],xe(:,4)-2,[1:1000],pcov(:,4).^(0.5)*3,[1:1000],-pcov(:,4).^(0.5)*3);
axis([0 1000 -0.02 0.02]);
set(gca,'fontsize',12);
set(gca,'xtick',[0 250 500 750 1000]);
set(gca,'ytick',[-0.02 -0.01 0 0.01 0.02]);
xlabel('Trial Run Number')
ylabel('{\it x}_4 Error and 3{\sigma} Outlier')
```

# Example II (i)

- Suppose we wish to estimate a nonlinear appearing parameter, $a > 0$, of the following exponential model

$$\tilde{y}_k = B\, e^{a\, t_k} + v_k, \quad k = 1, 2\, \ldots, m$$

  where $v_k$ is a zero-mean Gaussian white-noise process with variance given by $\sigma^2$

  - We can choose to employ nonlinear least squares to iteratively determine the parameter $a$, given the measurements and a known $B > 0$ coefficient
  - The covariance of the estimate error is given by $P = \sigma^2 (H^T H)^{-1}$ (this is the Cramér-Rao bound too) with

$$H = \begin{bmatrix} B\, t_1\, e^{a\, t_1} & B\, t_2\, e^{a\, t_2} & \cdots & B\, t_m\, e^{a\, t_m} \end{bmatrix}^T$$

  - Note that $H$ is a function of the true parameter $a$ now
  - This can be replaced by the final estimate after the nonlinear least squares iteration is complete (errors are second-order in nature)

# Example II (ii)

- Let's instead employ linear least squares by using a change of variables, as shown before, with $\tilde{z}_k \equiv \ln \tilde{y_k}$
  - Question: How optimal is this approach?
  - Expanding $\tilde{z}_k$ in a first-order series gives

$$\ln \tilde{y}_k - \ln B \approx a\, t_k + \frac{2\, v_k}{2\, B\, e^{a\, t_k} + v_k}$$

  - The least squares "$H$ matrix" is now simply given by

$$\mathcal{H} = \begin{bmatrix} t_1 & t_2 & \cdots & t_m \end{bmatrix}^T$$

  - A first-order expansion using the binomial series of the new measurement noise is given by

$$\varepsilon_k \equiv 2\, v_k (2\, B\, e^{a\, t_k} + v_k)^{-1} \approx \frac{v_k}{B\, e^{a\, t_k}} \left(1 - \frac{v_k}{2\, B\, e^{a\, t_k}}\right)$$

  - The variance can be shown to be given by

$$\varsigma_k^2 = E\{\varepsilon_k^2\} - E\{\varepsilon_k\}^2 = E\left\{\left(\frac{v_k}{B\, e^{a\, t_k}} - \frac{v_k^2}{2\, B^2 e^{2\, a\, t_k}}\right)^2\right\} - \frac{\sigma^4}{4\, B^2 e^{4\, a\, t_k}}$$

# Example II (iii)

- This leads to

$$\varsigma_k^2 = \frac{\sigma^2}{B^2 e^{2\,a\,t_k}} + \frac{\sigma^4}{2\,B^4 e^{4\,a\,t_k}}$$

- Contains both Gaussian and $\chi^2$ components
- The covariance of the linear approach is given by

$$\mathcal{P} = \left(\mathcal{H}^T \mathrm{diag}\left[\varsigma_1^{-2} \quad \varsigma_2^{-2} \quad \cdots \quad \varsigma_m^{-2}\right] \mathcal{H}\right)^{-1}$$

- Both covariances are equivalent if $\sigma^4/(2\,B^4 e^{4\,a\,t_k})$ is negligible
- If this is not the case, then the Cramér-Rao lower bound is not achieved and the linear approach does not lead to an efficient estimator
- This clearly shows how the Cramér-Rao inequality can be particularly useful to help quantify the errors introduced by using an approximate solution instead of the optimal approach