Optimal Estimation Methods

(Lecture 8 – Maximum Likelihood & Maximum A Posteriori Estimation)

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

Maximum Likelihood (i)

- Maximum likelihood yields estimates for the unknown quantities which maximize the probability of obtaining the observed set of data
  - For motivational purposes, let $\tilde{\mathbf{y}}$ be a random sample from a simple Gaussian distribution, conditioned on some unknown parameter set denoted by $\mathbf{x}$
  - The density function is given by

$$p(\tilde{\mathbf{y}}|\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} e^{\left[-\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2/(2\sigma^2)\right]}$$

  - Clearly, the Gaussian distribution is a monotonic exponential function for the mean and variance
  - Due to the monotonic aspect of the function, this fit can be accomplished by also taking the natural logarithm, so that

$$\ln\left[p(\tilde{\mathbf{y}}|\mathbf{x})\right] = -\frac{m}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(\tilde{y}_i-\mu\right)^2$$

Maximum Likelihood (ii)

- Now the fit leads immediately to an equivalent quadratic optimization problem to maximize the function
- This leads to the concept of maximum likelihood estimation, which is stated as follows

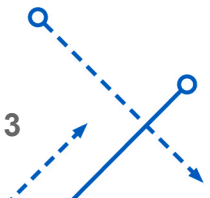
Given a measurement $\tilde{\mathbf{y}}$, the maximum-likelihood estimate $\hat{\mathbf{x}}$ is the value of $\mathbf{x}$ that maximizes $p(\tilde{\mathbf{y}}|\mathbf{x})$, which is the likelihood that $\mathbf{x}$ resulted in $\tilde{\mathbf{y}}$

- The likelihood function is also a probability density function

$$L(\tilde{\mathbf{y}}|\mathbf{x}) = \prod_{i=1}^{q} p(\tilde{\mathbf{y}}_i|\mathbf{x})$$

where q is the total number of density functions (a product of a number of density functions, known as a joint density, is also a density function in itself)

- Note that the distributions used in the likelihood function are the same, but the measurements belong to a different sample drawn from the conditional density



Maximum Likelihood (iii)

- Many likelihood functions contain exponential terms, which can complicate the mathematics involved in obtaining a solution
- However, since log is a monotonic function, finding $\mathbf{x}$ to maximize the log of the likelihood function is equivalent to maximizing the likelihood function itself

- Two required conditions for maximum likelihood
  - Necessary condition

↖ still a pdf

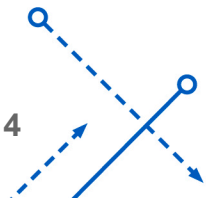
$$\left\{ \frac{\partial}{\partial \mathbf{x}} \ln \left[L(\tilde{\mathbf{y}} | \mathbf{x}) \right] \right\} \Bigg|_{\hat{\mathbf{x}}} = \mathbf{0}$$

  - Sufficient condition

maximization problem

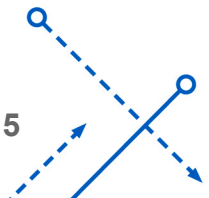
$$\frac{\partial^2}{\partial \mathbf{x} \, \partial \mathbf{x}^T} \ln \left[L(\tilde{\mathbf{y}} | \mathbf{x}) \right] \text{ must be negative definite}$$

(All λ < 0)



Maximum Likelihood (iv)

- Some properties of maximum likelihood estimation
  - Can yield biased estimates (be careful!). But it is asymptotically efficient, which means if the sample size is large, the maximum likelihood estimate is approximately unbiased and has a variance that approaches the smallest that can be achieved by any estimator
  - Obeys the *invariance principle*. The maximum likelihood estimate of any function $g(\mathbf{x})$ of these parameters is the function $g(\hat{\mathbf{x}})$ of the maximum likelihood estimate
    - Say we estimate the variance using maximum likelihood, then the estimate of the standard deviation is just the square root of the estimate of the variance
    - Very powerful!
  - The estimation errors in the maximum likelihood estimate can be shown to be *asymptotically Gaussian* no matter what density function is used in the likelihood function



Example I (i)

- Let $\tilde{\mathbf{y}}$ be a random sample from a Gaussian distribution
  - We desire to determine estimates for the mean and variance
  - The likelihood function is given by

$$L(\tilde{\mathbf{y}}|\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} e^{\left[-\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2/(2\sigma^2)\right]}$$

  - The log likelihood function is given by

$$\ln\left[L(\tilde{\mathbf{y}}|\mathbf{x})\right] = -\frac{m}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2$$

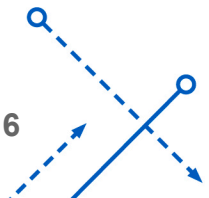
  - The maximizing mean is given by $\frac{\partial}{\partial\mu} = 0$

$$\left\{\frac{\partial}{\partial\mu}\ln\left[L(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right]\right\}\Big|_{\hat{\mu}} = \frac{1}{\sigma^2}\sum_{i=1}^{m}(\tilde{y}_i-\hat{\mu}) = 0$$

or

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\tilde{y}_i$$

This is the well known sample mean



Example I (ii)

- The maximizing variance is given by $\frac{\partial}{\partial \sigma^2} \ln [L(\tilde{\mathbf{y}}|\hat{\mathbf{x}})] \Big|_{\hat{\sigma}^2} = -\frac{m}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{m} (\tilde{y}_i - \mu)^2 = 0$

  or

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^{m} (\tilde{y}_i - \mu)^2$$

- From before we know that this is a biased estimate because we are dividing by m not $m-1$
- Thus, two different principles of estimation (unbiased estimator and maximum likelihood) give two different estimators
- Note that for large m the estimate becomes unbiased
  - It is asymptotically efficient as stated by one of the previously mentioned maximum likelihood properties

Example II (i)

- Estimate the covariance from a multivariate normal distribution given a set of observations

$$\{\tilde{\mathbf{y}}_1,\ \tilde{\mathbf{y}}_2,\ \ldots,\ \tilde{\mathbf{y}}_q\}$$

  - The likelihood function in this case is the joint density function

$$L(R) = \prod_{i=1}^{q} \frac{1}{(2\pi)^{m/2}\left[\det(R)\right]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T R^{-1}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]\right\}$$

  - The log likelihood function is given by

$$\ln[L(R)] = \sum_{i=1}^{q}\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T R^{-1}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right] - \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln\left[\det(R)\right]\right\}$$

  - Consider the following partials

$$\frac{\partial \ln\left[\det(R)\right]}{\partial R} = (R^T)^{-1}, \quad \frac{\partial \operatorname{Tr}(R^{-1}G)}{\partial R} = -(R^T)^{-1}G(R^T)^{-1}$$

Example II (ii)

- It can also be shown through simple matrix manipulations that

$$\sum_{i=1}^{q} [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}]^T R^{-1} [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}] = \mathrm{Tr}(R^{-1}G)$$

Important

where

$$G = \sum_{i=1}^{q} [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}] [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}]^T$$

- Since R is symmetric we have

$$\frac{\partial \ln[L(R)]}{\partial R} = -\frac{q}{2} R^{-1} + \frac{1}{2} R^{-1} G R^{-1}$$

- The maximum likelihood estimate for the covariance is given by

$$\hat{R} = \frac{1}{q} \sum_{i=1}^{q} [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}] [\tilde{\mathbf{y}}_i - \boldsymbol{\mu}]^T$$

  - It can also be shown that this estimate is biased

Example III (i)

- ML can work with non-Gaussian distributions as well
  - Suppose we wish to determine the probability of obtaining a certain number of heads in multiple flips of a coin
  - We are given $\tilde{y}$ "successes" in n trials, and wish to estimate the probability of success x of the binomial distribution

$$L(\tilde{y}|x) = \binom{n}{\tilde{y}} x^{\tilde{y}}(1-x)^{n-\tilde{y}}$$

  - The log likelihood function is given by

$$\ln\left[L(\tilde{y}|x)\right] = \ln\binom{n}{\tilde{y}} + \tilde{y}\ln(x) + (n-\tilde{y})\ln(1-x)$$

  - Sufficient condition gives

$$\left\{\frac{\partial}{\partial x}\ln\left[L(\tilde{y}|x)\right]\right\}\Big|_{\hat{x}} = \frac{\tilde{y}}{\hat{x}} - \frac{n-\tilde{y}}{1-\hat{x}} = 0$$

Example III (ii)

- Estimate is then given by

$$\hat{x} = \frac{\tilde{y}}{n}$$

- This intuitively makes sense for our coin toss example
- Say we flip the coin 1,000 times, so that $n = 1000$
  - We expect to obtain about 500 heads, so that

$$\hat{x} = \frac{500}{1000} = \frac{1}{2}$$

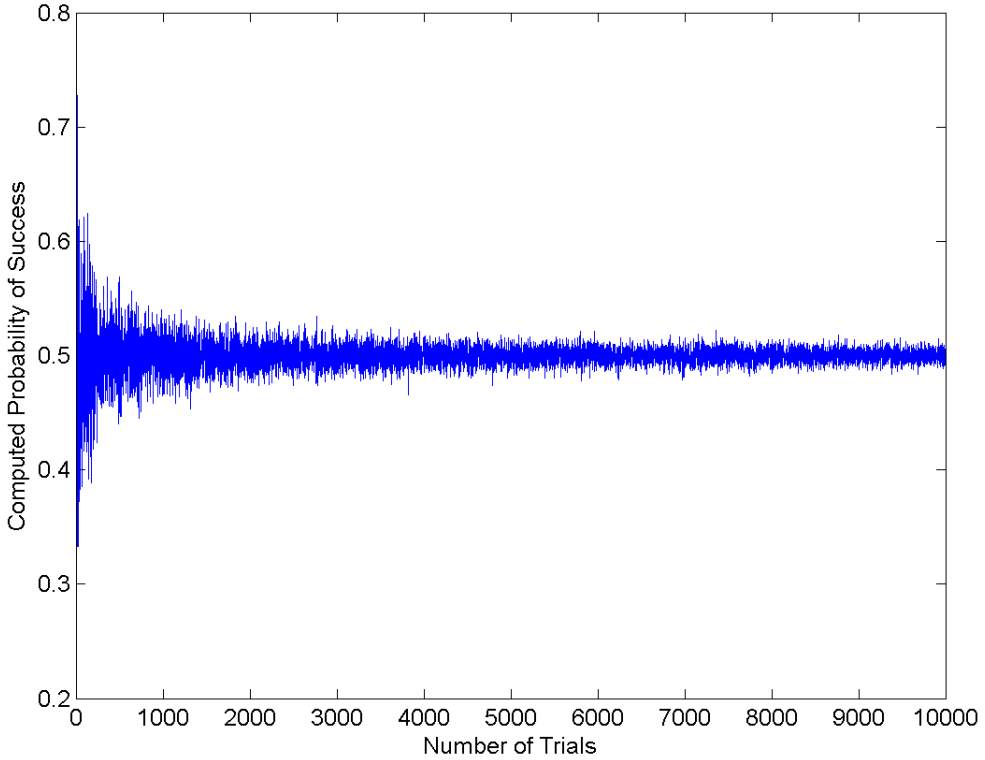
- We can write a simple program to test this concept
- Look at $n = 1$ to $m = 10000$ trials
  - Generate an integer random vector of zeros and ones
  - Compute the times a “heads” appears (i.e., the successes) by counting now many times a 1 appears in that random vector
  - Plot the number of successes divided by n (gives the computed probability) versus the number of trials
  - Should approach ½ as the number of trials increases

Example III (iii)

```
% Number of Trials
m=10000;
y_success=zeros(m,1);

% Main Trial Loop
for n=1:m,
% Generate Random Vector of 0's and 1's
% of Length n
 rand01=randi([0 1],n,1);
% Find Number of 1's and Divide by n
 y_success(n)=length(nonzeros(rand01))/n;
end

% Plot Results
plot(y_success)
set(gca,'fontsize',12)
ylabel('Computed Probability of Success')
xlabel('Number of Trials')
axis([0 m 0.2 0.8])
```



Relation to Least Squares (i)

- 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

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

  - The covariance is then given by

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

  - The likelihood function is then given by

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

    - Same as the conditional density from before

Relation to Least Squares (ii)

- The log likelihood function is given by

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

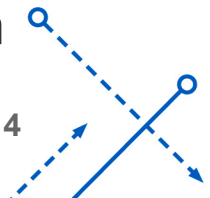
  - We can ignore the last two terms since they don't depend on $\mathbf{x}$
  - If we take the negative of the above equation, then maximizing the log likelihood function to determine the optimal estimate is equivalent to *minimizing*

Same as weighted least squares loss function

$$J(\hat{\mathbf{x}}) = \frac{1}{2}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right]$$

  - Leads to the same solution as given by minimum variance!
  - Therefore, for the case of Gaussian measurement errors the minimum variance and maximum likelihood estimates are identical to the least squares solution with the weight replaced with the inverse measurement-error covariance
  - The term ½ in the loss function comes directly from maximum likelihood (helps to simplify the math in the solution too)

For Gaussian Distribution



Bayesian Estimation

- Parameters that we have estimated using maximum likelihood have been assumed to be unknown constants
  - In Bayesian estimation, we consider that these parameters are random variables with some *a priori* distribution
  - Combines this *a priori* information with the measurements through a conditional density function
  - This conditional function is known as the *a posteriori* distribution of $\mathbf{x}$
  - Therefore, Bayesian estimation requires the probability density functions of both the measurement noise and unknown parameters
  - The posterior density function $p(\mathbf{x}|\tilde{\mathbf{y}})$ for $\mathbf{x}$ (taking the measurements into account) is given by *Bayes' rule*

$$p(\mathbf{x}|\tilde{\mathbf{y}}) = \frac{p(\tilde{\mathbf{y}}|\mathbf{x})p(\mathbf{x})}{p(\tilde{\mathbf{y}})}$$

MAP Estimation (i)

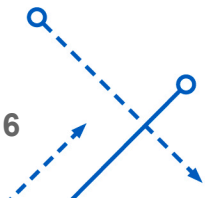
- Maximum *a posteriori* (MAP) estimation finds an estimate for $\mathbf{x}$ that maximizes $p(\mathbf{x}|\tilde{\mathbf{y}})$
  - Since $p(\tilde{\mathbf{y}})$ does not depend on $\mathbf{x}$, this is equivalent to maximizing $p(\tilde{\mathbf{y}}|\mathbf{x})p(\mathbf{x})$
  - We can again use the natural logarithm to simplify the problem by maximizing

$$J_{\text{MAP}}(\hat{\mathbf{x}}) = \ln\left[p(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right] + \ln\left[p(\hat{\mathbf{x}})\right]$$

    - The first term in the sum is actually the log-likelihood function
    - The second term gives the *a priori* information on the to-be-determined parameters
  - Therefore, the MAP estimator maximizes

$$\boxed{J_{\text{MAP}}(\hat{\mathbf{x}}) = \ln\left[L(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right] + \ln\left[p(\hat{\mathbf{x}})\right]}$$

    - Replaced first term with likelihood



MAP Estimation (ii)

- MAP Properties
  - If the *a priori* distribution $p(\hat{\mathbf{x}})$ is uniform, then MAP estimation is equivalent to maximum likelihood estimation
  - MAP estimation shares the asymptotic consistency and efficiency properties of maximum likelihood estimation
  - The MAP estimator converges to the maximum likelihood estimator for large samples
  - The MAP estimator also obeys the invariance principle

Example (i)

- Estimate the mean μ of a Gaussian variable from a sample of m independent measurements known to have a standard deviation of $\sigma_{\tilde{y}}$
  - We have been given that the *a priori* density function of μ is also Gaussian with zero mean and standard deviation σ_μ
  - The density functions are therefore given by

$$p(\tilde{y}_i|\mu) = \frac{1}{\sigma_{\tilde{y}}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\tilde{y}_i - \mu)^2}{\sigma_{\tilde{y}}^2}\right\}, \quad i = 1, 2, \ldots, m$$

$$p(\mu) = \frac{1}{\sigma_\mu\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma_\mu^2}\right\}$$

  - Since the measurements are independent we can write

$$p(\tilde{\mathbf{y}}|\mu) = \frac{1}{(\sigma_{\tilde{y}}\sqrt{2\pi})^m} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\frac{(\tilde{y}_i - \mu)^2}{\sigma_{\tilde{y}}^2}\right\}$$

Example (ii)

- Ignoring terms independent of μ we now seek to maximize

$$J_{\text{MAP}}(\hat{\mu}) = -\frac{1}{2}\left[\sum_{i=1}^{m}\frac{(\tilde{y}_i - \hat{\mu})^2}{\sigma_{\tilde{y}}^2} + \frac{\hat{\mu}^2}{\sigma_{\mu}^2}\right]$$

- Taking the partial w.r.t. to $\hat{\mu}$ and setting to zero gives

$$\sum_{i=1}^{m}\frac{(\tilde{y}_i - \hat{\mu})}{\sigma_{\tilde{y}}^2} - \frac{\hat{\mu}}{\sigma_{\mu}^2} = 0$$

- Recall that the maximum likelihood estimate for the mean is given by

$$\hat{\mu}_{\text{ML}} = \frac{1}{m}\sum_{i=1}^{m}\tilde{y}_i$$

- Thus we have

$$\hat{\mu} = \frac{\sigma_{\mu}^2}{\frac{1}{m}\sigma_{\tilde{y}}^2 + \sigma_{\mu}^2}\hat{\mu}_{\text{ML}}$$

- Notice $\hat{\mu} \to \hat{\mu}_{\text{ML}}$ as either $\sigma_{\mu}^2 \to \infty$ or as $m \to \infty$
  - This is consistent with the properties discussed previously of a maximum *a posteriori* estimator

MAP Estimation (i)

- Consider case with *a priori* estimates now
  - 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 $E\{\mathbf{w}\,\mathbf{v}^T\} = E\{\mathbf{v}\,\mathbf{w}^T\} = 0$
  - To determine the mean of the *a priori* model, we take the expectation of both sides

$$\boldsymbol{\mu}_a \equiv E\{\hat{\mathbf{x}}_a\} = E\{\mathbf{x}\} + E\{\mathbf{w}\} = \mathbf{x}$$

  - The covariance is then given by

$$\begin{aligned}\text{cov}\{\hat{\mathbf{x}}_a\} &\equiv E\left\{(\hat{\mathbf{x}}_a - \boldsymbol{\mu}_a)(\hat{\mathbf{x}}_a - \boldsymbol{\mu}_a)^T\right\} \\ &= E\left\{\mathbf{w}\,\mathbf{w}^T\right\} = Q\end{aligned}$$

MAP Estimation (ii)

- The probability density functions are given by

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

$$p(\hat{\mathbf{x}}) = \frac{1}{(2\pi)^{n/2}\left[\det(Q)\right]^{1/2}} \exp\left\{-\frac{1}{2}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right]^T Q^{-1}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right]\right\}$$

- Taking the log of both gives

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

$$\ln\left[p(\hat{\mathbf{x}})\right] = -\frac{1}{2}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right]^T Q^{-1}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right] - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left[\det(Q)\right]$$

- Leads to minimizing (ignore terms independent of estimate)

$$J_{\text{MAP}}(\hat{\mathbf{x}}) = \frac{1}{2}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right] + \frac{1}{2}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right]^T Q^{-1}\left[\hat{\mathbf{x}}_a - \hat{\mathbf{x}}\right]$$

MAP Estimation (iii)

- Necessary condition gives

$$\frac{\partial J_{\text{MAP}}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}} = -H^T R^{-1} \tilde{\mathbf{y}} + H^T R^{-1} H \hat{\mathbf{x}} - Q^{-1} \hat{\mathbf{x}}_a + Q^{-1} \hat{\mathbf{x}} = \mathbf{0}$$

- Solving for the MAP estimate gives

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

- Same result obtained through minimum variance!
  - The solution using MAP estimation is much simpler since we do not need to solve a constrained minimization problem using Lagrange multipliers
- With poor *a priori* knowledge we have $Q \to \infty$ and $Q^{-1} \to 0$, which reduces down to the ML estimator, as expected
- With poor measurements we have $R \to \infty$ and $R^{-1} \to 0$, which gives the result $\hat{\mathbf{x}} = \hat{\mathbf{x}}_a$, as again expected

Cramér-Rao Inequality (i)

- Cramér-Rao inequality for a Bayesian estimator

$$P \equiv E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} \geq \left[F + E\left\{\left[\frac{\partial}{\partial \mathbf{x}}\ln p(\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}}\ln p(\mathbf{x})\right]^T\right\}\right]^{-1}$$

only difference

- The Fisher information matrix is given by

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

- Using the *a priori* density function from before leads to

$$E\left\{\left[\frac{\partial}{\partial \mathbf{x}}\ln p(\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}}\ln p(\mathbf{x})\right]^T\right\} = Q^{-1}E\left\{(\hat{\mathbf{x}}_a-\mathbf{x})(\hat{\mathbf{x}}_a-\mathbf{x})^T\right\}Q^{-1}$$

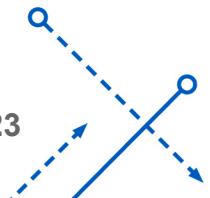
$$= Q^{-1}E\left\{\mathbf{w}\,\mathbf{w}^T\right\}Q^{-1} = Q^{-1}$$

- So the right hand side of the inequality is given by

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

p2

true only for unbiased estimator



Cramér-Rao Inequality (ii)

- Next, we need to compute the covariance matrix P
- Using $MH + N = I$, the estimate can be written as

$$\hat{\mathbf{x}} = \mathbf{x} + M\mathbf{v} + N\mathbf{w}$$

- Assuming $E\{\mathbf{v}\,\mathbf{w}^T\} = E\{\mathbf{w}\,\mathbf{v}^T\} = 0$ leads to

$$P = MRM^T + NQN^T$$

- Substituting the solutions for M and N from before we now have

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

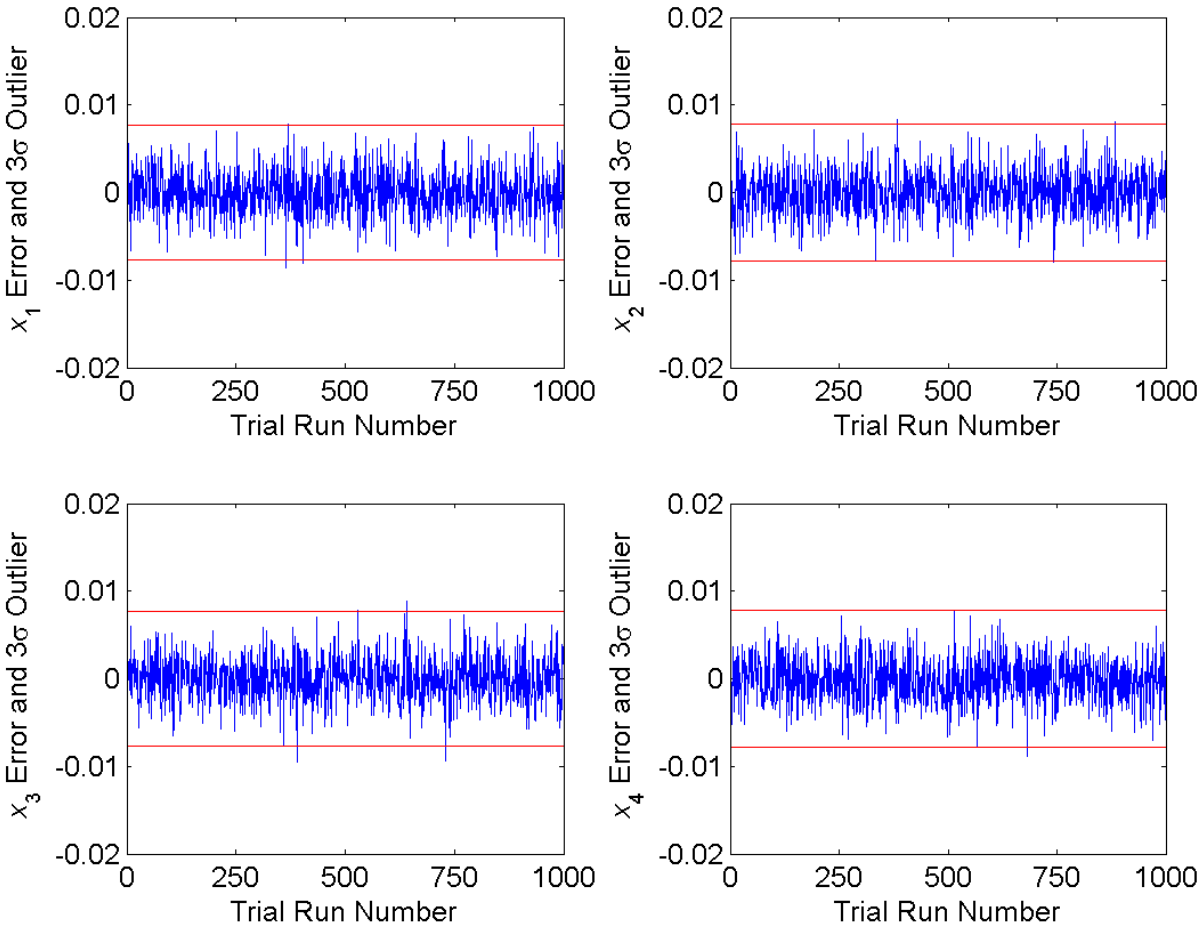
- Thus the Cramér-Rao lower bound is achieved, so the estimator is efficient
- Alternative form for P is given using the matrix inversion lemma

$$P = Q - QH^T \left(R + HQH^T\right)^{-1} HQ$$

  - May be preferred over other form if the dimension of R is less than the dimension of Q

Example (i)

$$\tilde{y}(t) = \cos(t) + 2\sin(t) + \cos(2t) + 2\sin(3t) + v(t), \quad R = 0.01I, \quad Q = 1 \times 10^{-5} I$$



Ran 1,000 Monte Carlo runs.

Errors much smaller with *a priori* information.

Obviously depends on Q. Try $Q = 0.01$. Results are nearly identical to standard least squares.

Example (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);

% Measurement Covariance
r=0.01*eye(m);ri=inv(r);

% A Priori Information
xamean=[1;2;1;2];
q=1e-5*eye(4);qi=inv(q);

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

Example (ii)

```
% Monte Carlo Simulation
for i=1:1000,
 ym=y+sqrt(r(1,1))*randn(m,1);
 xa=xamean+sqrt(q(1,1))*randn(4,1);
 h=[cos(t) sin(t) cos(2*t) sin(3*t)];
 p=inv(h'*ri*h+qi);
 xe(i,:)=(p*(h'*ri*ym+qi*xa))';
 pcov(i,:)=diag(p)';
end
```

Example (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 (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')
```

Ridge Estimation (i)

- Used for ill-conditioned problems
  - If the matrix $H^T H$ is close to being ill-conditioned, then the model is known as weak *multicollinear*
  - We can clearly see that weak multicollinearity may produce a large covariance in the estimated parameters
  - A strong multicollinearity exists if there are exact linear relations among the observations so that the rank of H equals n
    - This corresponds to the case of having linearly dependent rows in H
  - Another situation for $H^T H$ ill-conditioning is due to H having linearly independent columns, which occurs when the basis functions themselves are not independent of each other
    - For example choosing t, t^2, and $at + bt^2$, where a and b are constants, as basis functions leads to an ill-conditioned H matrix
  - Ridge estimation can be used to overcome ill-conditioned cases
    - Useful, but we'll see that the estimates are biased

Ridge Estimation (ii)

- Involves adding a positive constant ϕ to $H^T H$

$$\boxed{\hat{\mathbf{x}} = (H^T H + \phi I)^{-1} H^T \tilde{\mathbf{y}}}$$

  - The positive constant ensures that the inverse is well-conditioned
- Substituting the measurement model gives

$$E\{\hat{\mathbf{x}}\} = (H^T H + \phi I)^{-1} H^T H \mathbf{x}$$

- Therefore, the bias is given by

$$\begin{aligned}
\mathbf{b} &\equiv E\{\hat{\mathbf{x}}\} - \mathbf{x} \\
&= \left[(H^T H + \phi I)^{-1} H^T H - I\right] \mathbf{x} \\
&= \left[(H^T H + \phi I)^{-1} H^T H - (H^T H + \phi I)^{-1}(H^T H + \phi I)\right] \mathbf{x} \\
&= (H^T H + \phi I)^{-1}(H^T H - H^T H - \phi I)\mathbf{x} \\
&= -\phi(H^T H + \phi I)^{-1}\mathbf{x}
\end{aligned}$$

  - We clearly see that the ridge estimates are unbiased only when $\phi = 0$, which reduces to the standard least squares estimator

Ridge Estimation (iii)

- Define the following and the measurement model

$$\Gamma \equiv (H^T H + \phi I)^{-1}$$

$$\hat{\mathbf{x}} = \Gamma H^T \tilde{\mathbf{y}} = \Gamma H^T H \mathbf{x} + \Gamma H^T \mathbf{v}$$

- The ridge covariance is then

$$\begin{aligned} P_{\text{ridge}} &\equiv E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\} - E\left\{\hat{\mathbf{x}}\right\} E\left\{\hat{\mathbf{x}}\right\}^T \\ &= E\left\{(\Gamma H^T H \mathbf{x} + \Gamma H^T \mathbf{v})(\Gamma H^T H \mathbf{x} + \Gamma H^T \mathbf{v})^T\right\} - \Gamma H^T H \mathbf{x}\,\mathbf{x}^T H^T H\,\Gamma \\ &= E\left\{\Gamma H^T \mathbf{v}\,\mathbf{v}^T H\,\Gamma\right\} + E\left\{\Gamma H^T \mathbf{v}\,\mathbf{x}^T H^T H\,\Gamma\right\} + E\left\{\Gamma H^T H \mathbf{x}\,\mathbf{v}^T H\,\Gamma\right\} \\ &= \Gamma H^T R H\,\Gamma \\ &= (H^T H + \phi I)^{-1} H^T R\,H (H^T H + \phi I)^{-1} \end{aligned}$$

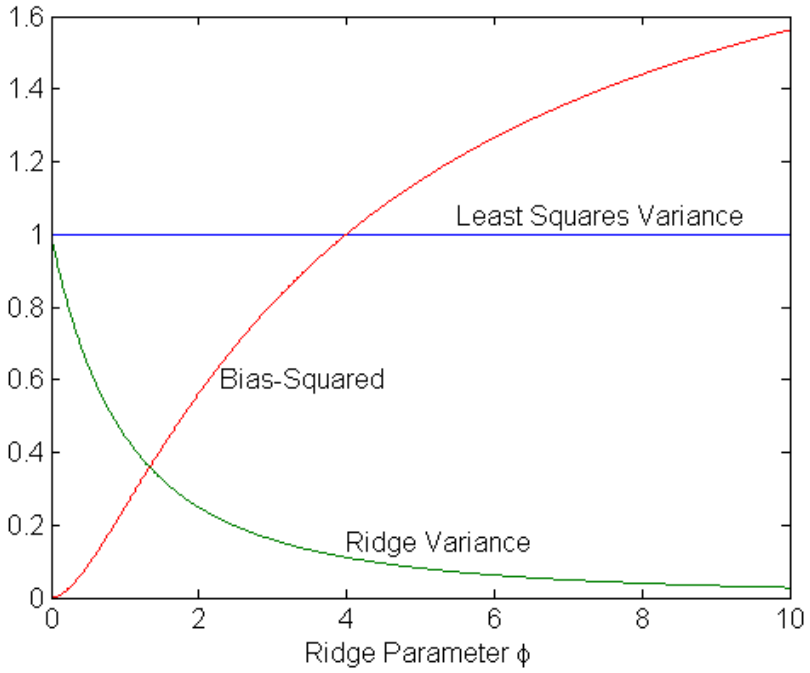
($E\left\{\Gamma H^T \mathbf{v}\,\mathbf{x}^T H^T H\,\Gamma\right\} \to 0$, $E\left\{\Gamma H^T H \mathbf{x}\,\mathbf{v}^T H\,\Gamma\right\} \to 0$)

- As ϕ increases the ridge covariance decreases, but at a price!
  - The estimate becomes more biased; again the bias is given by

$$\mathbf{b} = -\phi(H^T H + \phi I)^{-1}\mathbf{x}$$

Ridge Estimation (iv)

- Scalar example



- Ridge variance is always less than least squares variance
  - Not true in general case but a ϕ can be found to make it true
- But the bias increases, as expected
- Generally do not want to make this tradeoff!
  - Choose ϕ to minimize $E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\}$

Covariance Analysis (i)

- Assume that the assumed covariance is given by $\tilde{R}$
  - Least squares estimate

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

  - Substituting the measurement model gives

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

  - Using $E\{\mathbf{v}\} = \mathbf{0}$ and $E\{\mathbf{v}\,\mathbf{v}^T\} = R$ gives

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

  - Note that this expression reduces to $P = (H^T R^{-1} H)^{-1}$ when either of the following is true
    - The assumed covariance is equal to the true covariance, so that

$$\tilde{R} = R$$

    - The matrix H is square

Covariance Analysis (ii)

- Define the following relative inefficiency parameter

$$e = \frac{\det\left[(H^T\tilde{R}^{-1}H)^{-1}H^T\tilde{R}^{-1}R\,\tilde{R}^{-1}H(H^T\tilde{R}^{-1}H)^{-1}\right]}{\det\left[(H^TR^{-1}H)^{-1}\right]}$$

- Can prove that $e \geq 1$
  - It's equal to 1 only when there is no errors in the assumed covariance $(R = \tilde{R})$
- The specific value of e gives an indication of the inefficiency of the estimator
- Can be used to perform a sensitivity analysis given bounds on matrix R
- A larger value for e means that the estimates are further (in a statistical sense) from their true values

Covariance Analysis (iii)

- Two measurement case with true covariance set to I
  - Assumed covariance and H matrix

$$\tilde{R} = \begin{bmatrix} 1+\alpha & 0 \\ 0 & 1+\beta \end{bmatrix}, \quad H = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

  - Plot the efficiency for varying α and β
  - The minimum value, 1, is given when $\alpha = \beta = 0$ as expected
  - Note the steep increase along the edges
  - Does not increase as much when is α near β
  - Useful to assess the overall sensitivity to errors