DATA FUSION III: Estimation Theory

Date: March 2, 2006 **Time:** 5:00 – 7:30 PM
Instructor: Dr. James K Beard
Credits: 1 **Course Code:** 0901-501-05 **Registration Number:** 13460

Today's topics:

1 Simple Example: Estimating Mean and Variance of a Distribution

This is an example that is often given as an application of the method of maximum likelihood to derive estimation of the mean and variance of a distribution, resulting in traditional estimators. Our model is a set of *M* samples of random variables that have a common mean *m* and variance σ^2. We write these measurements as a vector:

$$\underline{y} = m \cdot \underline{1} + \underline{v} \tag{1.1}$$

Our state vector is

$$\underline{x} = \begin{bmatrix} m \\ \sigma^2 \end{bmatrix} \tag{1.2}$$

and our likelihood function is

$$p\left(\underline{y} \middle| \underline{x}\right) = \frac{1}{\left(2\pi\right)^{M/2} \cdot \sigma^M} \cdot \exp\left(-\frac{1}{2} \cdot \left(\underline{y} - m \cdot \underline{1}\right) \cdot R^{-1} \cdot \left(y - m \cdot \underline{1}\right) \right) \tag{1.3}$$

where the measurement covariance matrix *R* is

$$R = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots \\ 0 & \sigma^2 & 0 & \cdots \\ 0 & 0 & \sigma^2 & \\ \vdots & \vdots & & \ddots \end{bmatrix} \tag{1.4}$$

because the variances of all the samples are the variance of the distribution and the measurements are uncorrelated. Our log likelihood function is

$$\begin{aligned} L\left(\underline{x}\right) = &-\frac{M}{2} \cdot \ln\left(2\pi\right) - \frac{M}{2} \cdot \ln\left(\sigma^2\right) \\ &-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{M} \left(y_i - m\right)^2 \end{aligned} \tag{1.5}$$

The likelihood equation for the estimate of the mean is

$$\frac{\partial L(\underline{x})}{\partial m} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{M} (y_i - \hat{m}) = 0 \tag{1.6}$$

so that our estimator for the mean is the sample mean

$$\hat{m} = \frac{1}{M} \cdot \sum_{i=1}^{M} y_i \tag{1.7}$$

The likelihood equation for the estimate of the variance is:

$$\frac{\partial L(\underline{x})}{\partial \sigma^2} = -\frac{M}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \cdot \sum_{i=1}^{M} (y_i - m)^2 = 0 \tag{1.8}$$

so that our estimator for the variance is the sample variance:

$$\hat{\sigma}^2 = \frac{1}{M} \cdot \sum_{i=1}^{M} (y_i - m)^2 \tag{1.9}$$

Using the estimator for *m* instead of the true value of *m* makes this estimator biased, so that we need to use *M-1* in place of *M* in the normalization constant.

$$\hat{\sigma}^2_{UNBIASED} = \frac{1}{M-1} \cdot \sum_{i=1}^{M} (y_i - \hat{m})^2 \tag{1.10}$$

The Cramer-Rao bound is a good estimator of variance because the estimator for the mean is efficient and the estimator for the variance is asymptotically efficient. For the Fisher information matrix, we need three more gradients:

$$\begin{aligned} \frac{\partial^2 L(\underline{x})}{\partial m^2} &= -\frac{M}{\hat{\sigma}^2} \\ \frac{\partial^2 L(\underline{x})}{\partial m \, \partial \sigma^2} &= -\frac{1}{\hat{\sigma}^4} \cdot \sum_{i=1}^{M} (y_i - \hat{m}) = 0 \\ \frac{\partial^2 L(\underline{x})}{(\partial \sigma^2)^2} &= \frac{M}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \cdot \sum_{i=1}^{M} (y_i - \hat{m})^2 = -\frac{M-2}{2\hat{\sigma}^4} \end{aligned} \tag{1.11}$$

where we have used the unbiased estimator for variance in the last term. Thus, the Fisher information matrix is

$$P^{-1} = \begin{bmatrix} \frac{M}{\hat{\sigma}^2} & 0 \\ 0 & \frac{M-2}{2\hat{\sigma}^4} \end{bmatrix} \tag{1.12}$$

and the Cramer-Rao bound is

$$P = \begin{bmatrix} \frac{\hat{\sigma}^2}{M} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{M-2} \end{bmatrix} \tag{1.13}$$

These results are consistent with our expectations for the distributions of the estimates. The estimate of the mean is Gaussian, and the variance is as expected from the sum of *M*

Gaussian variables. The estimate of the variance is chi-square distributed with *M-2* degrees of freedom, and again the variance is consistent with that distribution.

2 The Inverse Matrix Derivative Lemma

This is a trick that we will need to look at some maximum likelihood problems where we estimate variance of a distribution from samples of a random variable, but don't want to use the consistency property to estimate the Fisher information matrix instead.

We pose the problem in general terms for use in a variety of applications, beginning in the next section. We have a cost function

$$J = \mathrm{tr}\left(B \cdot A^{-1} \cdot C\right) \tag{2.1}$$

and we wish to find the matrix derivative or gradient

$$D = \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot C\right)}{\partial A} \tag{2.2}$$

We begin with the identity

$$B \cdot A^{-1} \cdot C = B \cdot A^{-1} \cdot A \cdot A^{-1} \cdot C \tag{2.3}$$

and we take the gradient of the trace of the matrix in this form. We use the product rule, with the matrix that is active in each derivative highlighted by enclosing it in square brackets:

$$\begin{aligned} \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot A \cdot A^{-1} \cdot C\right)}{\partial A} &= \frac{\partial \,\mathrm{tr}\left(B \cdot \left[A^{-1}\right] \cdot A \cdot A^{-1} \cdot C\right)}{\partial A} \\ &+ \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot \left[A\right] \cdot A^{-1} \cdot C\right)}{\partial A} \\ &+ \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot A \cdot \left[A^{-1}\right] \cdot C\right)}{\partial A} \end{aligned} \tag{2.4}$$

We note that

$$\frac{\partial \,\mathrm{tr}\left(B \cdot \left[A^{-1}\right] \cdot A \cdot A^{-1} \cdot C\right)}{\partial A} = \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot A \cdot \left[A^{-1}\right] \cdot C\right)}{\partial A} = D \tag{2.5}$$

This can be shown with the chain rule while we group $A \cdot A^{-1}$ or $A^{-1} \cdot A$ as a single matrix, or simply by inspection of that grouping and noting that this results in the original definition of *D*. The result is

$$D = D + \frac{\partial \,\mathrm{tr}\left(B \cdot A^{-1} \cdot \left[A\right] \cdot A^{-1} \cdot C\right)}{\partial A} + D \tag{2.6}$$

which, with Gelb's Equation (2.1-72) page 23 gets us to our final result:

$$D = -A^{-T} \cdot B^{T} \cdot C^{T} \cdot A^{-T} \tag{2.7}$$

3 Analysis of Questionnaire Returns

3.1 Problem Statement

We have a questionnaire with *N* questions and we have received *M* responses. The responses are interpreted as numerical values, such as one to five, or one to two (for true and false). The responses are numbered in order of receipt from one to *M*. The responses are not correlated with each other. We wish to know

- The mean value of each response.
- The variance of the estimate of the mean value of each response.
- Whether there is a trend in responses that depends on how quickly the questionnaire response was prepared or received.
- Whether there are correlations between responses to different questins.

3.2 Algebraic Interpretation of the Problem Statement

We have a set of measurements $\underline{yr}$ that are Gaussian-distributed about a mean that is a linear function of the response order *j*:

$$\underline{yr}_j = \underline{m} + \left(j - j_0\right) \cdot \underline{v} + \underline{n} \tag{3.1}$$

The parameter j_0 is an arbitrary point during the process of returning the questionnaires that we will use to look at the solution when we are done. We collect these in a measurement vector:

$$\underline{y} = \begin{bmatrix} \underline{yr}_1 \\ \underline{yr}_2 \\ \vdots \\ \underline{yr}_M \end{bmatrix} \tag{3.2}$$

At this point, we note that we can write the measurement vector in simpler form

$$\underline{y} = \underline{m} \cdot A + S \cdot \underline{v} \tag{3.3}$$

where the vector $\underline{v}$ represents a trend with the sequence in which the responses were returned, and the matrix A replicates the mean and the matrix *S* represents a weighting of the trends with sequence:

$$A = \begin{bmatrix} I_N \\ I_N \\ I_N \\ \vdots \\ I_N \end{bmatrix}, \; S = \begin{bmatrix} -j_0 \cdot I_N \\ \left(2 - j_0\right) \cdot I_N \\ \left(3 - j_0\right) \cdot I_N \\ \vdots \\ \left(M - j_0\right) \cdot I_N \end{bmatrix} \tag{3.4}$$

where I_N is an identity matrix of order *N*. Our state vector $\underline{x}$ is a combination of $\underline{m}$ and $\underline{v}$:

$$\underline{x} = \begin{bmatrix} \underline{m} \\ \underline{v} \end{bmatrix} \tag{3.5}$$

Our likelihood function is

$$p\left(\underline{y}\middle|\underline{x}\right)=\frac{1}{\left(2\pi\right)^{M\cdot N/2}\cdot\left|R\right|^{M/2}}\cdot\exp\left(-\frac{1}{2}\cdot\left(\underline{y}-A\cdot\underline{m}-S\cdot\underline{v}\right)^{T}\cdot R_{M}^{-1}\cdot\left(\underline{y}-A\cdot\underline{m}-S\cdot\underline{v}\right)\right) \quad (3.6)$$

The matrix R_M is made up of submatrices that have the covariance of the unknown measurement noise R on the diagonal submatrices and zeros elsewhere. We will find maximum likelihood estimators for $\underline{m}$ and perform a significance test on values of $\underline{v}$ to determine whether trends exist. We will use part of the Fisher information matrix to find the accuracy of our estimate of $\underline{m}$ and to support our significance tests for the elements of $\underline{v}$.

We will quickly need nother form that collapses the quadratic form with the single-vector form for $\underline{y}$ to a sum on the measurements:

$$p\left(\underline{y}\middle|\underline{x}\right)=\frac{1}{\left(2\pi\right)^{M\cdot N/2}\cdot\left|R\right|^{M/2}}\cdot\exp\left(-\frac{1}{2}\cdot\sum_{j=1}^{M}\left(\underline{yr}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\underline{v}\right)^{T}\cdot R^{-1}\cdot\left(\underline{yr}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\underline{v}\right)\right) \quad (3.7)$$

The only loss of generality is that the measurements must be uncorrelated, which is consistent with our problem statement.

3.3 The Maximum Likelihood Estimators

The log likelihood function is

$$\begin{aligned}L\left(\underline{x}\right)=&-\frac{M\cdot N}{2}\cdot\ln\left(2\pi\right)-\frac{M}{2}\cdot\ln\left(\left|R\right|\right)\\&-\frac{1}{2}\cdot\sum_{j=1}^{M}\left(\underline{yr}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\underline{v}\right)^{T}\cdot R^{-1}\cdot\left(\underline{yr}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\underline{v}\right)\end{aligned} \quad (3.8)$$

Our likelihood equation for $\underline{m}$ is

$$\frac{\partial L\left(\underline{x}\right)}{\partial\underline{m}}=\sum_{j=1}^{M}R^{-1}\cdot\left(\underline{yr}_{j}-\hat{\underline{m}}-\left(j-j_{0}\right)\cdot\underline{v}\right)=\underline{0} \quad (3.9)$$

and our likelihood equation for $\underline{v}$ is almost identical

$$\frac{\partial L\left(\underline{x}\right)}{\partial\underline{v}}=\sum_{j=1}^{M}R^{-1}\cdot\left(j-j_{0}\right)\cdot\left(\underline{yr}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\hat{\underline{v}}\right)=\underline{0} \quad (3.10)$$

We can write the likelihood equation for $\underline{m}$ as

$$\hat{\underline{m}}+\left(\frac{1}{M}\sum_{j=1}^{M}\left(j-j_{0}\right)\right)\cdot\underline{v}=\frac{1}{M}\cdot\sum_{j=1}^{M}\underline{y}_{j} \quad (3.11)$$

The equation is simplified if sum on the left hand side is written in closed form

$$\hat{\underline{m}}+\left(\frac{M+1}{2}-j_{0}\right)\cdot\underline{v}=\frac{1}{M}\cdot\sum_{j=1}^{M}\underline{y}_{j} \quad (3.12)$$

Similarly, the likelihood equation for the trends $\underline{v}$ is

$$\sum_{j=1}^{M}\left(j-j_{0}\right)\cdot R^{-1}\cdot\left(\underline{y}_{j}-\underline{m}-\left(j-j_{0}\right)\cdot\hat{\underline{v}}\right)=\underline{0} \quad (3.13)$$

We collect the states on the left hand side and drop R as before

$$\left(\frac{1}{M}\cdot\sum_{j=1}^{M}\left(j-j_0\right)\right)\cdot\underline{m}+\left(\frac{1}{M}\cdot\sum_{j=1}^{M}\left(j-j_0\right)^2\right)\cdot\underline{\hat{v}}=\frac{1}{M}\cdot\sum_{j=1}^{M}\left(j-j_0\right)\cdot\underline{y}_j \tag{3.14}$$

The sums can be written in closed form to simplify things again

$$\left(\frac{M+1}{2}-j_0\right)\cdot\underline{m}+f\left(M,j_0\right)\cdot\underline{\hat{v}}=\frac{1}{M}\cdot\sum_{j=1}^{M}\left(j-j_0\right)\cdot\underline{y}_j \tag{3.15}$$

where we have collected the closed form for the quadratic sum as

$$f\left(M,j_0\right)=\frac{1}{6}\cdot\left(2\cdot M^2+\left(3-6\cdot j_0\right)\cdot M+6\cdot j_0^2-6\cdot j_0+1\right) \tag{3.16}$$

We see immediately that we can set j_0 to the center point and simplify the equations,

$$j_0=\frac{M+1}{2} \tag{3.17}$$

and $f(M,j_0)$ becomes

$$f\left(M\right)=\frac{M^2-1}{12} \tag{3.18}$$

The maximum likelihood estimators are

$$\underline{\hat{m}}=\frac{1}{M}\cdot\sum_{j=1}^{M}\underline{y}_j \tag{3.19}$$

and

$$\underline{\hat{v}}=\frac{12}{M^2-1}\cdot\sum_{j=1}^{M}\left(j-\frac{M+1}{2}\right)\cdot\underline{y}_j \tag{3.20}$$

Our Fisher information matrix for the covariance of $\underline{\hat{m}}$ is

$$P_m^{-1}=-\frac{\partial^2 L\left(\underline{x}\right)}{\partial\underline{m}^2}=-A^T\cdot R_M^{-1}\cdot A=M\cdot R^{-1} \tag{3.21}$$

so that we have the covariance of $\underline{m}$ as

$$P_m=\frac{1}{M}\cdot R \tag{3.22}$$

The Fisher information matrix of $\underline{v}$ is

$$P_v^{-1}=S^T\cdot R_M^{-1}\cdot S=\frac{M^2-1}{12}\cdot R^{-1} \tag{3.23}$$

so that the covariance of $\underline{v}$ is

$$P_v=\frac{12}{M^2-1}\cdot R \tag{3.24}$$

We need to estimate R so we solve its' likelihood equation (using the lemma from the preceding section)

$$\begin{aligned}\frac{\partial L\left(\underline{x}\right)}{\partial R}&=-\frac{M}{2\cdot\left|R\right|}\cdot R^{-1}\\&+\frac{1}{2}R^{-1}\cdot\left(\sum_{j=1}^{M}\left(\underline{y}_j-\underline{m}-\left(j-\frac{M+1}{2}\right)\cdot\underline{v}\right)\cdot\left(\underline{y}_j-\underline{m}-\left(j-\frac{M+1}{2}\right)\cdot\underline{v}\right)^T\right)\cdot R^{-1}\end{aligned} \tag{3.25}$$

so that we have an estimate of R as

$$\hat{R} = \frac{1}{M} \cdot \sum_{j=1}^{M} \left(\underline{y}_j - \underline{\hat{m}} - \left(j - \frac{M+1}{2} \right) \cdot \underline{\hat{v}} \right) \cdot \left(\underline{y}_j - \underline{\hat{m}} - \left(j - \frac{M+1}{2} \right) \cdot \underline{\hat{v}} \right)^T \quad (3.26)$$

We use $\hat{R}$ and $\underline{v}$ to do statistical tests on the elements of $\underline{v}$ using the Student T test. We note as before that using estimates in the sums reduces the number of degrees of freedom in the summation so that a biased estimator results. The estimator with the bias removed is

$$\hat{R}_{UNBAISED} = \frac{1}{M-1} \cdot \sum_{j=1}^{M} \left(\underline{y}_j - \underline{\hat{m}} - \left(j - \frac{M+1}{2} \right) \cdot \underline{\hat{v}} \right) \cdot \left(\underline{y}_j - \underline{\hat{m}} - \left(j - \frac{M+1}{2} \right) \cdot \underline{\hat{v}} \right)^T \quad (3.27)$$