Global Warming

Project 2

Jeremy Morris

March 24, 2006

1 A confidence set for the mean vector

1.1 t and T^2 Intervals

Johnson and Wichern outline a confidence set for a specific linear combination of the mean vector $\bar{\boldsymbol{x}}$, which is defined as follows. [2]

$$\left(\boldsymbol{a}'\bar{\boldsymbol{x}} - t_{n-1}(\alpha/2)\sqrt{\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}} \ , \quad \boldsymbol{a}'\bar{\boldsymbol{x}} + t_{n-1}(\alpha/2)\sqrt{\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}} \right) \tag{1}$$

Where $t_{n-1}(\alpha/2)$ is the upper $100(\alpha/2)^{th}$ percentile of a t-distribution with $n-1$ degrees of freedom. This confidence set holds for a specific choice of $\boldsymbol{a}$. The following confidence set can be used for $(1-\alpha)$ coverage for all choices of $\boldsymbol{a}$.

$$\left(\boldsymbol{a}'\bar{\boldsymbol{x}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} \ , \quad \boldsymbol{a}'\bar{\boldsymbol{x}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} \right) \tag{2}$$

Where $F_{p,n-p}(\alpha)$ is the upper $100(\alpha)$th percentile of the F distribution with p and $n-p$ degrees of freedom. This confidence interval is based on the T^2 statistic and should be used when $(1-\alpha)$ joint coverage is desired for several choices of $\boldsymbol{a}$. The consequence of having joint coverage is that the intervals will be larger than in (1).

1.2 The Bonferroni Method

For a small set of linear combinations of the mean vector, we can use the Bonferroni method to develop more precise confidence intervals. [2] The Bonferroni method is based on the following probabilistic argument. If we let C_i denote the confidence interval for $\boldsymbol{a}'\boldsymbol{\mu}$ with $P[C_i \text{ true}] = 1-\alpha_i$, where $i = 1, 2, \cdots, m$.

$$\begin{aligned} P[\text{all } C_i \text{ true}] &= 1 - P[\text{at least one } C_i \text{ false}] \\ &\geq 1 - \sum_{i=1}^{m} P[C_i \text{ false}] = 1 - \sum_{i=1}^{m}(1 - P[C_i \text{ true}]) \\ &= 1 - (\alpha_1 + \alpha_2 + \cdots + \alpha_m) \end{aligned} \tag{3}$$

This means that we can use the more precise confidence intervals in (1) and vary α based on the confidence set we desire. For example, if we would like to have a 0.95 confidence set for all means in the temperature data of this project, we would use the linear combinations $\boldsymbol{a}_i'\bar{\boldsymbol{x}}$, with $\boldsymbol{a}_i' = [0, 0, \ldots, a_i, \ldots, 0]'$, where $a_i = 1$, and let $\alpha_i = \alpha/12$ so that we have the following intervals.

$$\begin{array}{rcccl} \bar{x}_1 - t_{214}(\alpha/24)\sqrt{s_{1,1}/215} & \leq & \mu_1 & \leq & \bar{x}_1 + t_{214}(\alpha/24)\sqrt{s_{1,1}/215} \\ \bar{x}_2 - t_{214}(\alpha/24)\sqrt{s_{2,2}/215} & \leq & \mu_1 & \leq & \bar{x}_2 + t_{214}(\alpha/24)\sqrt{s_{2,2}/215} \\ & & \vdots & & \\ \bar{x}_{12} - t_{214}(\alpha/24)\sqrt{s_{12,12}/215} & \leq & \mu_{12} & \leq & \bar{x}_{12} + t_{214}(\alpha/24)\sqrt{s_{12,12}/215} \end{array} \tag{4}$$

1.3 Comparison of the T^2 and Bonferroni intervals

Notice that the only difference between the intervals in (1) and (2) is the multiplier corresponding to the critical value. For (1), we use the t distribution and the F distribution for (2). Table (1) shows the multipliers for all three types of intervals used. Notice that the

Type	Source	Value
t interval	$t_{214}(0.05/2)$	1.971
Bonferroni t interval	$t_{214}(0.05/24)$	2.896
T^2 interval	$\sqrt{\frac{p(n-1)}{n-p}F_{12,203}(0.05)}$	4.772

Table 1: Critical Values

value for the t interval is the smallest, this will give the most precise interval if only one confidence interval is desired. However, this confidence interval will not give $(1-\alpha)$ confidence for multiple intervals. Notice also that the multiplier for the T^2 interval is roughly 60% larger than the multiplier using the t interval with the Bonferroni method. Any interval using the T^2 interval will be twice the length of an interval using the Bonferroni method. The conclusion reached here is that the most precise confidence intervals will use the t-statistic along with the Bonferroni method.

2 Changing Mean

In order to test if the mean vector is staying the same with in the data set, we will derive the generalized likelihood ratio test for the following.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k \neq \boldsymbol{\mu}_{k+1} = \cdots = \boldsymbol{\mu}_n \qquad H_a : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_n = \boldsymbol{\mu} \tag{5}$$

This will test for a single change point within the mean vector against the alternative that there is no change. In order to use the generalized likelihood ratio test, we will derive the likelihood functions for each hypothesis and determine the MLE's for the parameters. If we assume that our data comes from a multivariate normal distribution, then our data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ has the following density function.

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{6}$$

The derivation of the likelihood function for the null hypothesis follows.

$$L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, k, \boldsymbol{\Sigma}) = \prod_{i=1}^{k} f(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \prod_{j=k+1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \tag{7}$$

$$= L_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\ L_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \tag{8}$$

Where in (8), each likelihood function is defined so that $L_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ is taken over the first k observations and $L_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ is taken over the last $n-k$ observations. Then this problem reduces to exercise (6.11) from our text [2]. This means that we have the following maximum likelihood estimates.

$$\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{x}}_1 \tag{9}$$

$$\hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{x}}_2 \tag{10}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left[(k-1)\boldsymbol{S}_1 + (n-k-1)\boldsymbol{S}_2\right] \tag{11}$$

Where

$$\boldsymbol{S_1} = \sum_{i=1}^{k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)' \tag{12}$$

$$\boldsymbol{S}_2 = \sum_{j=k+1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_j - \bar{\boldsymbol{x}}_2)' \tag{13}$$

There is no closed form maximum likelihood estimate for the parameter k. Since we are looking for the maximum of the likelihood function, we can simply run through all choices of k in the data set for the function $L(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, k, \hat{\boldsymbol{\Sigma}})$ to determine where the maximum is reached, the particular value of k will be our MLE $\hat{k}$.

The likelihood function is more easily defined for the alternative hypothesis. Instead of (7), we have the following.

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{14}$$

With the the maximum reached at the MLE's that follow.

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} \tag{15}$$

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{n-1}{n}\right)\boldsymbol{S} \tag{16}$$

and

$$\boldsymbol{S} = \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' \tag{17}$$

Then, the generalized likelihood ratio (GLR) is defined by the following equation.

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{k}, \hat{\boldsymbol{\Sigma}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})} \tag{18}$$

Where $-2\log\lambda(\boldsymbol{x}) \sim \chi^2(2)$. The null hypothesis will be rejected if $-2\log\lambda(\boldsymbol{x}) \geq \chi^2_{1-\alpha}(2)$.

3 Constant Variance

Much like in the previous section, a generalized likelihood test can be used to test for a change in the variance. In this case we will be testing the following hypothesis.

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k \neq \boldsymbol{\Sigma}_{k+1} = \cdots = \boldsymbol{\Sigma}_n \qquad H_a : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma} \quad (19)$$

Likelihood functions for each of the hypotheses will have to be calculated. To do this, we refer again to equations (8) and (7). For the null hypothesis, we have the following likelihood function.

$$\begin{aligned} L(\boldsymbol{\mu}, k, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) &= \prod_{i=1}^{k} f(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}_1) \prod_{j=k+1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}_2) && (20) \\ &= L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)\, L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2) && (21) \end{aligned}$$

Which again reduces to a problem such as exercise (6.11) from our text. It should be fairly clear that the MLE for $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, due to the independence of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$. Then, using the argument in Result 4.10 from the text we can find the MLE's for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, which follow.

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_1 &= \left(\frac{k-1}{k}\right) \sum_{j=1}^{k} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' && (22) \\ \hat{\boldsymbol{\Sigma}}_2 &= \left(\frac{n-k-2}{n-k}\right) \sum_{j=k+1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' && (23) \end{aligned}$$

Again, the parameter k cannot be found theoretically, so $\hat{k}$ will be calculated from the data set as the largest value of $L(\hat{\boldsymbol{\mu}}, k, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$.

The likelihood function for the alternative hypothesis is simply that found in the previous section defined by equations (14) - (16). The GLR, in this case is defined by the equation

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\boldsymbol{\mu}}, \hat{k}, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})} \quad (24)$$

Where $-2\log\lambda(\boldsymbol{x}) \sim \chi^2(2)$ and we reject for large values of $-2\log\lambda(\boldsymbol{x})$.

4 Correlation Between Monthly Averages

Sample correlations between components X_i and X_k will be calculated using the following formula.

$$r_{ik} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}} \quad (25)$$

The Pearson correlation test is to test the hypothesis

$$H_0 : \rho = 0 \qquad\qquad H_a : \rho \neq 0 \tag{26}$$

or that the correlation is significant. The Pearson test uses the following test statistic.

$$t \;=\; r\sqrt{\frac{n-2}{1-r^2}} \tag{27}$$

Where $t \sim t_{n-1}$. This test can be performed simply using `proc corr` in SAS. The null hypothesis will be rejected for large p-values.

5 The Normality Assumption

Johnson and Wichern suggest several ways to assess whether or not the data is actually normally distributed. The need to verify the normality of the data is not certain because the data is actually the mean vector of the daily temperatures and a central limit argument can be made that the monthly averages should be normally distributed.

Nevertheless, three methods will be used to verify the normality of the data set. One of the most effective ways is to look at Q-Q plots of each of the components. If the Q-Q plots turn out to be very close to linear, then normality can be assumed. Q-Q plots should not show any pattern when varying around the straight line and should not contain an excessive number of points at the tails that stray too far from the fitted line. Johnson and Wichern also suggest performing a test on the correlation between the data points and the quantile values, this amounts to performing a Shapiro-Wilk test on each of the components. These tests will be performed.

The text also suggests examining Q-Q plots of the first and last principal components, this seems like a good idea since most of the variability in the data will be contained in the first principal component and almost none in the last. The implication being that if the first and last principal components are normal, the rest should be well.

6 Grouping the Months

Principle component analysis can be used to group the months together. According to our text, we can determine the principle components by using the eigenvalues and eigenvectors of the sample covariance matrix. If we have the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ with corresponding eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$, then for a random sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$, transform $\boldsymbol{X}$ to get the new components that follow.

$$\begin{aligned} Y_1 &= \boldsymbol{e}_1'\boldsymbol{X} \\ Y_2 &= \boldsymbol{e}_2'\boldsymbol{X} \\ &\vdots \\ Y_p &= \boldsymbol{e}_p'\boldsymbol{X} \end{aligned}$$

Call the Y_i the principle components and calculate how much of the overall variance is due to each component with the following equation.

$$\begin{pmatrix} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } k\text{th principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots \lambda_p} \tag{28}$$

Then take the principal components that describe 85% - 90% of the variability and use them for all analysis because they are the most important. To determine this simply add (28) until the sum is larger than 0.85 or 0.90. It would not be necessary to normalize the data since each column is measured on the same scale.

This analysis can be done easily with `R` which has functions to calculate eigenvalues and eigenvectors. There are also several functions available to help in the analysis.

References

[1] Seber, George, and Alan Lee. Linear Regression Analysis. 2nd ed. Hoboken: John Wiley & Sons, 2003.

[2] Johnson, Richard, and Dean Wichern. Applied Multivariate Statistical Analysis. 5th ed. Upper Saddle River: Pearson Education, 2002.

[3] Everitt, Brian. An R and S-Plus© Companion to Multivariate Analysis. London: Springer, 2005.