STOR 754 FINAL
DECEMBER 2–3 2008

This is a take-home exam. It is to be completed in your own time and returned to me no later than 4:00pm, Wednesday December 3.

Use of the computer and all course materials is allowed. Solutions may be handwritten or word-processed. Where you use the computer, you should hand in enough of the output that I can follow the steps you took (including any graphs etc. that you feel are appropriate to illustrate your answer — extra credit will be given for good illustrative answers) but I do not need or want you to document every step.

Any result derived in the course notes or texts may be quoted without proof or detailed citation, so long as you give enough information for me to follow your method. If you consult any other references (including Internet), you should give a full citation.

In a multi-part question, an error in one part of the question will not prevent you getting full marks in another part of the question, so long as your derivation is clear.

You are reminded that the Honor Code is in force during this exam. You are not allowed to consult with each other, or with any outside person, in any way at all. You are allowed to ask me questions in person or by phone or email.

1. (a) Suppose we have p observations from a stationary AR(1) process,

$$U_t = \phi U_{t-1} + \epsilon_t, \quad \epsilon_t \sim N[0, \tau].$$

Show that the covariance matrix of $U_1, ..., U_p$ is

$$V = \frac{\tau}{1-\phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{p-1} \\ \phi & 1 & \phi & \dots & \phi^{p-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{p-1} & \phi^{p-2} & \phi^{p-3} & \dots & 1 \end{bmatrix}. \quad (1)$$

(b) Show that V^{-1} is of the form

$$V^{-1} = \begin{bmatrix} a & b & 0 & 0 & \dots & 0 & 0 \\ b & c & b & 0 & \dots & 0 & 0 \\ 0 & b & c & b & \dots & 0 & 0 \\ 0 & 0 & b & c & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & c & b \\ 0 & 0 & 0 & 0 & \dots & b & a \end{bmatrix}$$

where a, b and c are constants (independent of p) to be determined.

You are also given (this doesn't have to be proved) that $|V| = \frac{\tau^p}{1-\phi^2}$.

(c) Consider the simple linear regression model without intercept,

$$Y_t = \beta x_t + U_t$$

where x_t is given (scalar), β is an unknown regression constant, and U_t is as in (a). Show that the generalized least squares estimator of β is

$$\hat{\beta} = \frac{\sum_{t=1}^{p} x_t y_t - \phi \sum_{t=1}^{p-1}(x_t y_{t+1} + y_t x_{t+1}) + \phi^2 \sum_{t=2}^{p-1} x_t y_t}{\sum_{t=1}^{p} x_t^2 - 2\phi \sum_{t=1}^{p-1} x_t x_{t+1} + \phi^2 \sum_{t=2}^{p-1} x_t^2}$$

and give an explicit expression for its variance.

Hint: Use the GLS regression formula: if $Y = X\beta + U$, where U has mean 0 and covariance matrix V, then $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ and has covariance matrix $(X^T V^{-1} X)^{-1}$.

(d) Now suppose $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are independent p-dimensional random vectors where each $\mathbf{Y}_i \sim MVN_p(\mu, V)$. Hypothesis H_0 is that V is of form (1) for some τ and ϕ. Hypothesis H_1 is that V is completely unrestricted. (In either case, μ is an arbitrary p-dimensional vector.) Let $S_0 = \{s_{ij},\ 1 \le i, j \le p\}$ be the maximum likelihood sample covariance matrix (with divisor n instead of $n-1$) defined from $\mathbf{Y}_1, ..., \mathbf{Y}_n$, and define $A = \sum_{i=1}^{p} s_{ii}$, $B = \sum_{i=1}^{p-1} s_{i,i+1}$, $C = \sum_{i=2}^{p-1} s_{ii}$. Find expressions that must be satisfied by the joint maximum likelihood estimators of τ and ϕ. (*Note:* You probably won't be able to find a closed-form expression for the estimators themselves, but you should still be able to find an explicit equation that the estimators must satisfy.) Hence show how to construct the likelihood ratio test of H_0 against H_1. What is the asymptotic distribution (as $n \to \infty$) of the test statistic?

(e) Apply the method derived in (d) to the dataset `examdata1.txt` on the course webpage, where $n = 20$ and $p = 5$. Use an approximation to $\hat{\phi}$ if you cannot find the exact MLE. Do you accept or reject H_0?

(*Note:* You are not expected to use any time series or multivariate analysis software in solving (e). You may use the standard linear algebra commands that are available in R.)

2. The data file `raleighozone.txt` contains daily data, from 01/01/1987 through 12/31/2000, for daily maximum 8-hour ozone in Raleigh, NC. Units are parts per billion (ppb). This is ground-level data, not connected with the so-called ozone hole that arises from stratosphere data.

(a) Fit a suitable time series model to this dataset. Among the features you should take into account are:

- How to deal with missing values. Note that data are only available from the months April through October (which the EPA defines as the ozone period). Therefore, of necessity, your results will only cover those months. There are also isolated missing values during the ozone series. You will have to make your own decisions about how to deal with those, but whatever you do, make sure you explain it.
- Long-term trends
- Seasonality (even during the ozone season it is unlikely that the mean is constant across the whole season)
- Transformations of the data (e.g., to achieve homoscedasticity or normality)

(b) Ozone is a toxic pollutant that is regulated by the EPA. The current standard (since March, 2008) is 75 ppb. Suppose you want to institute a warning system that will sound

an alarm when the probability that the next day's ozone will exceed 75 ppb is at least 0.5. This alarm should be based on the current day's ozone and may (at your discretion) be also dependent on previous days. What is your proposed rule?

3. The data file `NCnormals.txt` contains summary meteorological data from 123 stations in North Carolina. The data are

Column	Description
1	Index number of station
2	Latitude
3	Longitude
4	Winter mean temperature
5	Spring mean temperature
6	Summer mean temperature
7	Fall mean temperature
8	Winter mean precipitation
9	Spring mean precipitation
10	Summer mean precipitation
11	Fall mean precipitation

   (a) Using just the meteorological data in the last eight columns, use PCA, factor analysis, or otherwise, to describe the main features of the data and (if appropriate) to propose a dimension reduction.

   (b) We can subdivide North Carolina into three regions as follows: "West" (west of longitude 80.7^{o}), "Central" (longitude between 78.2^{o} and 80.7^{o}), "East" (east of longitude 78.2^{o}). Is there a significant difference in the climate among these three regions?

   (c) Based on cluster analysis, is there any better way to subdivide the state than to use the method in (b)? [*Note:* For this part, you will have to decide whether to apply cluster analysis to the whole dataset or some reduction of it as determined in (a). In most cases, it would make sense to remove components that look like pure noise.]

Solutions and Comments

1. (a) {**3 Points**.} Immediate from ACVF of AR(1).

   (b) {**6 Points**.} $a = \frac{1}{\tau};\ b = -\frac{\phi}{\tau},\ c = \frac{1+\phi^2}{\tau}$; proof by multiplying out V and V^{-1} to verify that the product is I.

   (c) {**6 Points**.} We want to evaluate $X^TV^{-1}X$ and $X^TV^{-1}Y$ in the case when X is just a single-column matrix with entries $x_1, ..., x_p$. In that case, from the structure of V^{-1} we have

$$X^TV^{-1}Y \quad = \quad \frac{1}{\tau}\left\{x_1y_1 + x_py_p + (1+\phi^2)\sum_{t=2}^{p-1} x_ty_t - \phi\sum_{t=1}^{p-1} x_ty_{t+1} - \phi\sum_{t=1}^{p-1} x_{t+1}y_t\right\}.$$

   The expression for $X^TV^{-1}Y$ has the same structure (replace y_t by x_t) and $\hat{\beta}$ is then just $\frac{X^TV^{-1}Y}{X^TV^{-1}X}$, equivalent to the stated answer. Variance: $\frac{\tau}{\sum_{t=1}^{p} x_t^2 - 2\phi\sum_{t=1}^{p-1} x_tx_{t+1} + \phi^2\sum_{t=2}^{p-1} x_t^2}$.

   [In econometrics this is known as the Cochrane-Orcutt estimator, based on a paper in *JASA* (1949). The case with an intercept is essentially the same if you first center x_t and y_t.]

   (d) {**10 Points**.} If we define the model as $\mathbf{Y}_i \sim MVN_p(\mu, V)$ then after maximizing the likelihood with respect to μ, ignoring some constants,

$$-2\log L(\hat{\mu}, V) \quad = \quad n\log|V| + n\ \text{tr}\ V^{-1}S_0. \tag{2}$$

   Under H_1, (2) is minimized by $V = S_0$, leading to $-2\log L_1 = n\log|S_0| + np$. These results so far follow from estimation results for the multivariate normal distribution that were in the class notes.

   Under H_0, we have $|V| = \tau^p/(1-\phi^2)$ as given in the question, while the trace of $V^{-1}S_0$ is of the form $aA + 2bB + (c-a)C$; then (2) becomes

$$-2\log L(\hat{\mu}, \tau, \phi) \quad = \quad n\left[p\log\tau - \log(1-\phi^2) + \frac{A - 2\phi B + \phi^2 C}{\tau}\right]. \tag{3}$$

   We minimize analytically with respect to τ; (3) becomes

$$-2\log L(\hat{\mu}, \hat{\tau}, \phi) \quad = \quad n\left[p\log(A - 2\phi B + \phi^2 C) - \log(1-\phi^2) + p - p\log p\right]. \tag{4}$$

   We wish to minimize (4) with respect to ϕ; differentiating, the equation to be satisfied is

$$\frac{p(\phi C - B)}{A - 2\phi B + \phi^2 C} - \frac{\phi}{1-\phi^2} \quad = \quad 0 \tag{5}$$

   and the resulting $\hat{\phi}$ leads to the likelihood ratio statistic

$$W \quad = \quad 2(\log L_1 - \log L_0) \tag{6}$$

$$\quad = \quad n\left[p\log(A - 2\hat{\phi}B + \hat{\phi}^2 C) - \log(1-\hat{\phi}^2) - p\log p - \log|S_0|\right]. \tag{7}$$

   Under H_0, W has an approximate χ^2_ν distribution, where $\nu = \frac{p(p+1)}{2} - 2$ (the difference in d.f. for the two models).

(e) {**5 Points**.} With the given data we find $\log|S_0| = 23.57233$, $A = 1075.966$, $B = 625.2579, C = 675.665$. The equation (5) is satisfied when $\hat{\phi} = 0.70524$, resulting in $\hat{\tau} = \frac{A-2\hat{\phi}B+\hat{\phi}^2C}{p} = 106.02$, $W = 8.67$. Under the null hypothesis this should have an approximate χ^2_{13} distribution; the p-value is 0.8. Accept H_0. [The actual dataset was a simulation from the AR(1) model with $\phi = 0.7$.] {**30 Points Total for Question 1**.}

2. {**12 Points for (a); 8 Points for (b)**.} There are various possible analyses but here is one that is fairly straightforward. I first removed all the Jan, Feb, Mar, Nov, Dec days from the ozone series. This reduced the series from length 5114 to length 2996. However, examination of the series shows that the first 445 values are missing, so we remove those. This leaves a series of length 2551 of which 85 are missing, and I decided just to remove those (slightly over 3%) with no further correction. Strictly speaking, we should allow for the missing values in fitting the time series model, but that is not easy to do, and would probably not change the following analysis much.

   Fitting this series by the arima command in R suggests that the best model according to AIC is the ARMA(1,3) model:

```
Call:
arima(x = x1, order = c(p, 0, q))

Coefficients:
         ar1      ma1      ma2      ma3  intercept
      0.9676  -0.3588  -0.2871  -0.1498    54.0837
s.e.  0.0082   0.0218   0.0214   0.0198     1.5844

sigma^2 estimated as 158.1:  log likelihood = -9742.21,  aic = 19496.41
```

   The standard diagnostic plots in R suggest that this model is perfectly satisfactory.

   We can define the function $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{1-\phi_1 z}{1+\theta_1 z+\theta_2 z^2+\theta_3 z^3}$ and calculate the coefficients π_j recursively; the first five of these (after $\pi_0 = 1$) are $\pi_1 = -0.6089$, $\pi_2 = 0.0687$, $\pi_3 = -0.0004$, $\pi_4 = -0.0716$, $\pi_5 = -0.0155$; note that π_1 is by far the largest in magnitude. The optimal forecasting rule based on the infinite past would be $\hat{x}_{t+1} = \mu - \sum_{j=1}^{\infty} \pi_j (x_{t+1-j} - \mu)$; here $\mu = 54.0837$ and the forecast standard deviation is $\sqrt{158.1} = 12.6$. A possible rule for sounding the alarm is therefore that we would compute $p = 1 - \Phi\left(\frac{75-\hat{x}_{t+1}}{12.6}\right)$ (where Φ is the standard normal distribution function) and sound the alarm whenever $p > 0.5$.

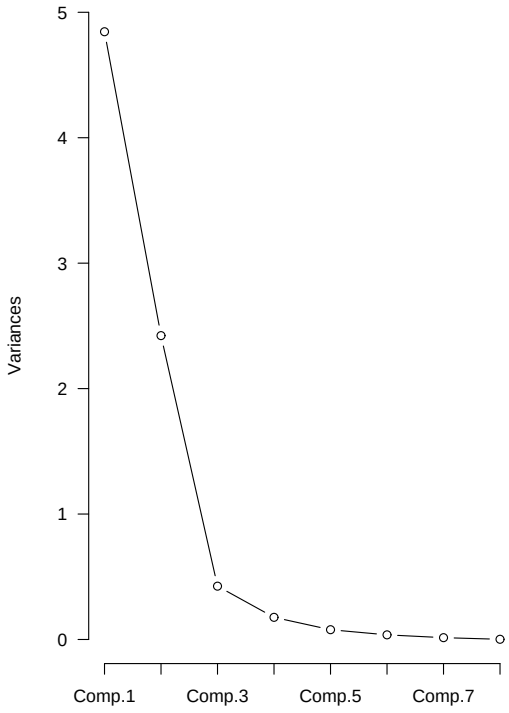
   The forecast standard deviation of 12.6 should be compared with the standard deviation of the series itself, which is 16.8; this reduction of standard deviation is a measure of the skill of the forecast (compared with an alarm rule that does not take past values of the series into account at all).

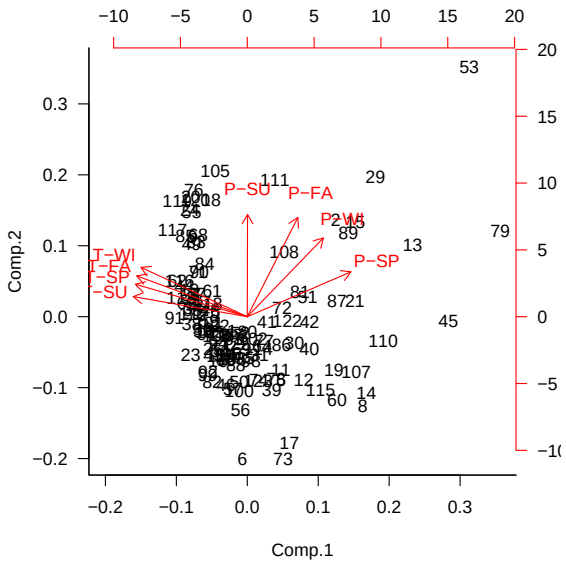
   However there's actually a simpler rule: the first partial autocorrelation coefficient is .644 and this suggests the predictor $\tilde{x}_{t+1} = \mu + .644(x_t - \mu)$ which has prediction standard deviation 12.8, almost as good as the above and far simpler to implement, especially in the presence of missing values.

   Further refinements: There is a day of week effect (ozone is lower at the weekend) but this doesn't seem to be very important. What is more significant is that if you look directly for a

seasonal effect, there is one: the mean ozone level drops off sharply in September and October compared with the main summer months. Therefore, a superior analysis would probably be to fit a separate time series model for each month, or in some other way to take account of this seasonal variation in the mean.

3. (a) {**7 Points**.} Apply correlation-based form of PCA (to put temperature and precipitation on comparable scale). The screeplot and biplot are as follows:



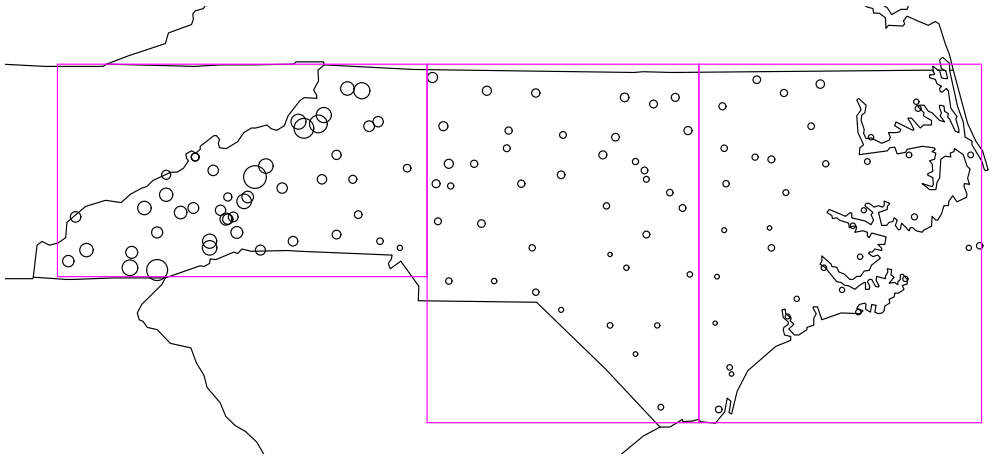


We can also plot the PC1 scores on a map (see next page).

The cumulative proportions of variance associated with the leading PCs are 0.61, 0.91, 0.962, 0.984,... and this as well as the screeplot suggests that at least 3 PCs are needed to capture most of the variability. The loadings on PC1 are –.407, –.429, –.437, –.423, .291, .387, 0, .194, so PC1 is high when temperatures are low and precipitation is high. The map shows a clear east to west drift (PC1 higher in the west), though there is also a north-south gradient. The other PCs did not have such an obvious spatial pattern, as far as I could tell.

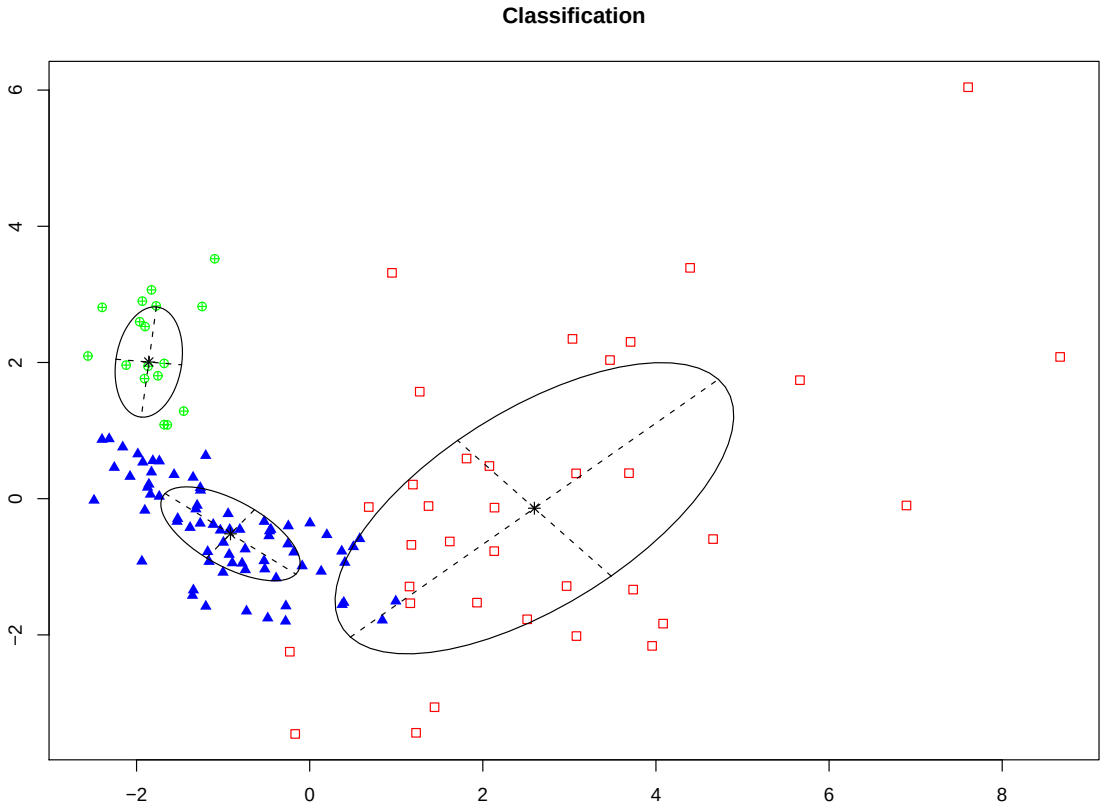
(b) {**6 Points**.} We can define a "region" variable according to longitude, with values 1,2,3, and then apply a MANOVA, as in the iris data example in class. The results for the four tests are

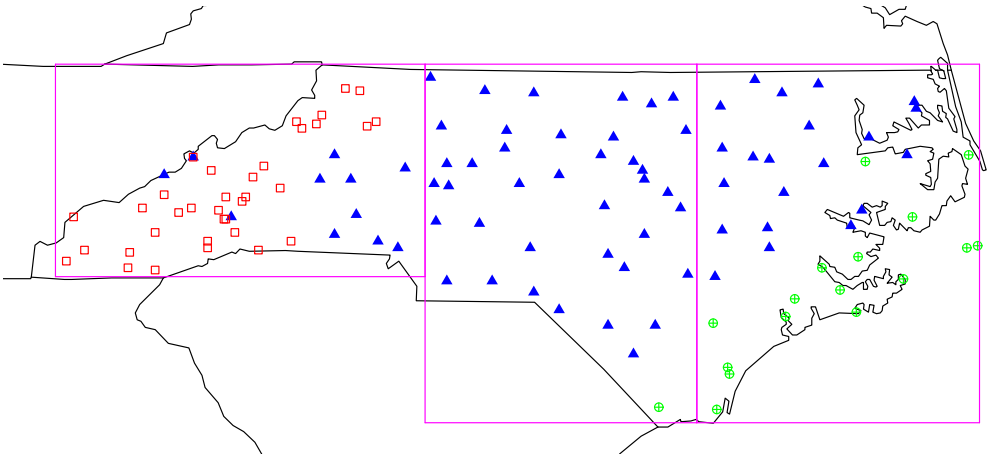
```
           Df  Pillai approx F num Df den Df    Pr(>F)
region      2  1.1477  19.1887     16    228 < 2.2e-16 ***
           Df   Wilks approx F num Df den Df    Pr(>F)
region      2  0.1553  21.7232     16    226 < 2.2e-16 ***
           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
region      2           3.4898  24.4284     16    224 < 2.2e-16 ***
           Df    Roy approx F num Df den Df    Pr(>F)
region      2  2.790   39.765      8    114 < 2.2e-16 ***
Residuals 120
```



By all four tests, the regional differences are very clearly significant.

(c) {**7 Points**.} This could be answered my either hierarchical or model-based clustering. I applied this using the scores from the first two PCs as the raw data for the cluster algorithm. With model-based clustering, the best model by BIC is "VEV with 3 components", producing the scatterplot shown below. Then we display the same clusters on a map.





The map shows a more refined spatial grouping than the original three boxes, with red dots in the mountains to the west, green dots along the coast, and blue dots for everything else. So it probably was worth doing the clustering: there is a clear geographical interpretation to the result.

Comments on Student Solutions

Question 1 was generally well answered though some students got lost in the detailed algebraic calculations of part (d). The part that seemed to cause most difficulty was actually (e); only a handful of students got the correct answer for the likelihood ratio test. For the calculation of $\hat{\phi}$, I was perfectly happy with a curve-drawing or trial-and-error solution, which is actually how I did it myself, but several students used optimization or equation-solving programs in R, which is of course a superior method in principle, though not if you got the wrong answer; in cases where that happened, however, I think it was because of putting the wrong function into the program and not misuse of the program itself.

Question 2 was the most open-ended and I tried therefore to be open-ended also in my grading (not requiring that you do any specific steps, but trying to evaluate what you did do to see how completely you answered the question). Common errors were (i) failing to describe sufficiently clearly what you were doing, (ii) failing to give an explicit answer to part (b) (I really did want to see an explicit prediction rule here, rather than just general description of the steps to be followed — some of you didn't even define the fitted model explicitly). Here is a summary of some of the steps that could be taken — I didn't expect every answer to contain all of these steps, but at some level a complete solution would have to consider them.

- Treatment of missing values — probably best to omit the November–March values completely since there are too few of them to perform a complete analysis (also, it's very unlikely that there would ever be an ozone alert during these months). Also, omit the first two years for which there are no data anyway. However, for the isolated missing values that arise during the summer series, there are essentially two things you can do, (i) just leave out these values and analyze the time series without them, (ii) fill in these values by interpolating in some way. One student downloaded some multiple imputation software form Gary King's webpage at Harvard, which showed commendable initiative, though I was perfectly happy if students used simpler interpolation rules such as linear interpolation between the nearest neighbors.

- Transformations — some used a square root transformations; other used likelihood methods to select a Box-Cox transformation, coming up with a λ of around 0.8. Either of these is probably an improvement on applying time series analysis to the untransformed data, though it's not a clear-cut decision.
- Seasonal variation — two reasonable ways of doing this, (i) fit a seasonal trend term (e.g. by polynomial trend estimation) and subtract from the ozone data to produce an approximately mean 0 series, (ii) use differencing across a time lag of 1 year. Either produces a reasonable solution but it's important to take account of this part of the model when answering part (b).
- Trends — various among you fitted a linear, quadratic or in one case cubic trend to the data. All I can say is I looked for a trend myself, both by simple linear regression at the beginning and by including time as a covariate in the eventual ARMA model, and I didn't see any evidence of a trend. So I don't know why some of you decided there was.

Question 3 was really less open-ended because I did ask you for a specific sequence of steps, though there is a lot of latitude in how those individual steps are performed.

In (a), I think it really is necessary to use a correlation-based analysis or else account in some other way for the downweighting of the precipitation values if you used covariance-based PCA. Some students used covariance-based PCA and even commented on the fact that the precipitation values received less weight, but didn't draw the obvious conclusion (i.e. either rescale the precipitation values or use correlation-based PCA). Most students preferred 2 for the number of significant PCAs, but I did want some form of interpretation of the PCAs themselves for full credit.

Not much to say about (b); I was looking for a MANOVA test here. Some students did that but did not obtain the numerical results given above.

In (c), I was willing to accept either hierarchical or model-based clustering (or some combination of both, as many did) but the main weakness in the solutions was failing to explain the result, especially how the division of stations by cluster analysis was related to the geographical division in (b). In hierarchical clustering, the software produces the dendogram automatically so I didn't give too much credit just for that; the key issue is whether you know how to interpret it.