```
> X<-t(as.matrix(A[,1:p]) ## extract the data matrix X'
> one<-matrix(c(rep(1,n),n,1)
> nobs<-c(n1,n2,...,nk)   ## set up a vector of group sizes
> n<-sum(nobs)            ## total sample size
> G<-matrix(c(rep(1,nobs[1]),rep(0,n),rep(1,nobs[2]),
+ rep(0,n),...,rep(1,nk)),n,k) ## G is the group indicator
        ## matrix G_ij=1 if and only if case i is in group j
> F<-G%*%t(matrix(c(rep(1,nobs[1])/nobs[1],rep(0,n),
+ rep(1,nobs[2])/nobs[2],rep(0,n),...,rep(1,nk/nobs[k])),n,k))
        ## F is used in calculating group means in M
> M<-F%*%t(X)
> xbar<-t(one)%*%t(X)/n  ## overall mean vector
> W<-t(t(X)-G%*%M)%*%(t(X)-G%*%M)/(n-k) ##
> B<-t(G%*%M-one%*%xbar)%*%(G%*%M-one%*%xbar)/(k-1)
       ## W and B are within and between groups variances
```

A useful check on the calculations is to ensure that the analysis of variance is satisfied, i.e., that `(n-1)*var(t(X))=(n-k)*W+(k-1)*B` (up to rounding errors). Note that the line calculating $\boldsymbol{B}$ may appear to be different from the formula given in the previous section $\boldsymbol{B} = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})'$ but the **R** calculations are in terms of matrices with n_i copies of $(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})$.

If the number of groups k is large (i.e., more than 5 or 6, say) then calculation of the group indicator G is cumbersome and a quicker way of doing this ***if the group sizes are equal*** is to split the dataframe into separate group dataframes with the function `split(.)` (note that the argument of `split(.)` must be of class `"dataframe"` and not of class `"matrix"`) followed by calculation of variances of each group dataframe with the function `lapply(.)` as follows:

```
Xdat<-split(A[,1:p],A[,p+1])
     ## note A is a dataframe not a matrix
     ## creates a list of dataframes, one for each group
Xdat<-lapply(Xdat,as.matrix)
     ## convert group dataframes to matrices
Xvar<-lapply(Xdat,var)
     ## find variance of each group
W<-Reduce("+",Xvar)*(n/k-1)/(n-k)
     ## assumes groups are of equal sizes
B<-((n-1)*var(X)-(W*(n-k))/k-1
```

Having obtained $\boldsymbol{W}$ and $\boldsymbol{B}$, the within- and between-groups variances, it is then easy to calculate any of the statistics used for multivariate analysis of variance referred to in §9.2.5.2and §9.2.7.3. The crimcoords can be obtained by the first $k-1$ vectors given by `eigen(solve(W)%*%B)$vectors` but note that the scaling used for the eigenvectors ensures that $\boldsymbol{y}'\boldsymbol{y} = 1$ whereas the scaling used for the crimcoords produced by the function `lda(.)` ensures that $\boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} = 1$, i.e., multiplying the eigenvectors produced by `eigen(solve(W)%*%B)$vectors` by a

factor $(\boldsymbol{y}'\boldsymbol{y}/\boldsymbol{y}'\boldsymbol{W}\boldsymbol{y})^{1/2}$ will reproduce the crimcoords from `lda(.)` held in the matrix `lda(.)$scaling`.

The data transformed to crimcoords are given by $\boldsymbol{Y}' = (\boldsymbol{X} - \overline{\boldsymbol{X}})'\boldsymbol{V}$ where $\boldsymbol{V}$ is obtained either from `V=eigen(solve(W)%*%B)$vectors[,1:k-1]` or from `V=lda(.)$scaling`. The difference in scaling will not be apparent if scatterplots of the data referred to crimcoords are produced by the basic **R** command `plot(.)` but willl be noticeable if the `MASS` library routine `eqscplot(.)` is used.

9.5 Canonical Correlation Analysis

Canonical correlation analysis is concerned with investigating the relationship between two sets of variables measured on the same objects. In particular, the aim is to find which linear combination of variables of a $n \times p$ data matrix $\boldsymbol{X}'$ has the maximum correlation with which linear combination of variables of a $n \times q$ data matrix $\boldsymbol{Y}'$ amongst all such liner combinations. For example, in analyzing results of questionnaires eliciting subjects' opinions of a product, one set of variables may reflect the socio-economic aspects of the subjects and the other set may relate to their opinions on various properties of the product.

9.5.1 Derivation of canonical variates

If $\boldsymbol{S}_{xx}$, $\boldsymbol{S}_{yy}$ and $\boldsymbol{S}_{xy}$ are the sample variances and covariance matrices respectively of $\boldsymbol{X}'$, $\boldsymbol{Y}'$ and $(\boldsymbol{X}', \boldsymbol{Y}')$ and $\boldsymbol{x}$ and $\boldsymbol{y}$ are p-vectors then the correlations between $\boldsymbol{X}'\boldsymbol{x}$ and $\boldsymbol{Y}'\boldsymbol{y}$ is $\boldsymbol{x}'\boldsymbol{S}_{xy}\boldsymbol{y}/\sqrt{\boldsymbol{x}'\boldsymbol{S}_{xx}\boldsymbol{x}\boldsymbol{y}'\boldsymbol{S}_{yy}\boldsymbol{y}}$. It was shown in Example 6 (iv) that this is maximized with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ by taking $\boldsymbol{x}$ to be the eigenvector of $\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{xy}'$ corresponding to its largest eigenvalue and $\boldsymbol{y}$ to be the eigenvector of $\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{xy}'\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}$ corresponding to its largest eigenvalue. These eigenvectors are termed the first ***canonical variates*** of $\boldsymbol{X}'$ and $\boldsymbol{Y}'$.

If the complete set of eigenpairs of $\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{xy}'$ is $(\boldsymbol{u}_1, \lambda_1), \ldots, (\boldsymbol{u}_r, \lambda_r)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ and those of $\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{xy}'\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}$ are $(\boldsymbol{v}_1, \lambda_1), \ldots, (\boldsymbol{v}_r, \lambda_r)$ (noting the eigenvalues are identical) where $r = \min(p, q)$ then it can be shown that the linear combinations of $\boldsymbol{X}'$ with $\boldsymbol{Y}'$ given by $(\boldsymbol{u}_2, \boldsymbol{v}_2) \ldots, (\boldsymbol{u}_r, \boldsymbol{v}_r)$ maximise the correlation between linear functions of the $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ variables subject to the constraints of orthogonality with earlier ones and thus are termed the canonical variates of $\boldsymbol{X}'$ and $\boldsymbol{Y}'$. Plots of the [mean corrected] data referred to canonical variates may provide insight into the structure of a relationship between the data sets, i.e., plots of $(\boldsymbol{X} - \overline{\boldsymbol{X}})'\boldsymbol{u}_i$ against $(\boldsymbol{X} - \overline{\boldsymbol{X}})'\boldsymbol{u}_j$ (typically with $j = i + 1$) or $(\boldsymbol{X} - \overline{\boldsymbol{X}})'\boldsymbol{u}_i$ against $(\boldsymbol{Y} - \overline{\boldsymbol{Y}})'\boldsymbol{v}_i$ and $(\boldsymbol{Y} - \overline{\boldsymbol{Y}})'\boldsymbol{v}_i$ against $(\boldsymbol{Y} - \overline{\boldsymbol{Y}})'\boldsymbol{v}_j$ can all be useful in informal investigation the structure of the data.

It can be shown that if the $\boldsymbol{Y}'$ variables are group indicators or a set of binary dummy variables then the canonical variates of the $\boldsymbol{X}'$ variables are precisely the