Exam 2 review. 10 sheets of notes and a calculator. Wednesday March 20.

Types of problems.

43) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$ where $0 < \alpha < 1$.

44) The classical $(1-\alpha)100\%$ large sample prediction region is $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\alpha}\}$ and works well if n is large and the data are iid MVN.

45) Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and $q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n)$, otherwise. If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$ is a large sample $(1-\alpha)100\%$ prediction regions if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i. The nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and the semiparametric prediction region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$. The parametric MVN prediction region
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq \chi^2_{p,q_n}\}$ also uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$.

46) These 3 regions can be displayed in an RMVN DD plot with cases in the nonparametric region corresponding to points to the left of the vertical line corresponding to $D_{(up)}(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Cases in the semiparametric region correspond to points below the horizontal line corresponding to $D_{(up)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ while cases in the parametric MVN region correspond to points below the horizontal line corresponding to $\sqrt{\chi^2_{p,q_n}}$. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_x$. The three prediction regions are asymptotically optimal if the data is MVN. The semiparametric and nonparametric prediction regions are asymptotically optimal on a large class of EC distributions and the nonparametric prediction region is a large sample $100(1-\alpha)\%$ prediction region, although large sample prediction regions with smaller volume may exist.

47) Suppose m independent large sample $100(1-\alpha)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\alpha_n)$ where $1 - \alpha_n$ is the true coverage and $1 - \alpha_n \to 1 - \alpha$ as $n \to \infty$. Simulation can be used to see if the true or actual coverage $1 - \alpha_n$ is close to the nominal coverage $1 - \alpha$. A prediction region with $1 - \alpha_n < 1 - \alpha$ is liberal and a region with $1 - \alpha_n > 1 - \alpha$ is conservative. It is better to be conservative by 5% than liberal by 5%. Parametric prediction regions tend to have large undercoverage and so are too liberal.

48) For prediction regions, want $n > 10p$ for the nonparametric prediction region and $n > 20p$ for the semiparametric prediction region.

49) Let $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ be a positive definite symmetric $p \times p$ dispersion matrix. A *generalized correlation matrix* $\boldsymbol{\rho} = ((\rho_{ij}))$ where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The generalized correlation matrix is the correlation matrix when second moments exist if $\boldsymbol{\Sigma} = c \operatorname{Cov}(\boldsymbol{x})$ for some constant $c > 0$.

50) Classical principal component analysis (PCA) gets the eigenvalues and eigenvectors $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ of the sample covariance matrix $\boldsymbol{S}$ or of the sample correlation matrix $\boldsymbol{R}$.

51) Let U be the subset of at least half of the cases from which the robust estimator is computed. Let $\boldsymbol{S}_U$ and $\boldsymbol{R}_U$ denote the sample covariance matrix and sample correlation matrix computed from the cases in $\boldsymbol{U}$. Then the robust estimator $\boldsymbol{C} = d\boldsymbol{S}_U$ for some constant $d > 0$ and $\boldsymbol{R}_U$ is the generalized correlation matrix corresponding to $\boldsymbol{C}$. The robust PCA uses U corresponding to the RMVN estimator.

52) Want $n > 10p$ for the classical PCA and $n > 20p$ for the robust PCA.

53) Both *R* and *SAS* output give the eigenvectors as shown in symbols for the following table.

PC1	PC2	$\cdots$	PCp
$\hat{\boldsymbol{e}}_1$	$\hat{\boldsymbol{e}}_2$	$\cdots$	$\hat{\boldsymbol{e}}_p$

R output shows the square roots of the eigenvalues

$$\sqrt{\hat{\lambda}_1}, \sqrt{\hat{\lambda}_2}, ..., \sqrt{\hat{\lambda}_p}$$

while *SAS* output gives the eigenvalues $\hat{\lambda}_i$.

54) Given the eigenvalues or square roots of the eigenvalues, be able to sketch a *scree plot* of i versus $\hat{\lambda}_i$.

55) The *trace explained* or *variance explained* by the first k principal components is $\frac{\sum_{i=1}^k \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i}$ where the denominator is equal to p if the correlation option $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, as recommended in point 58).

56) Use k principal components if the trace explained is bigger than some percentage like 90%, 80% or 70%. There is often a sharp bend in the scree plot when the components are no longer useful.

57) When $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, the correlation of the ith variable with the jth principal component is proportional to the ith entry of the jth eigenvector $\hat{\boldsymbol{e}}_j$. To try to explain the jth principal component, look at entries in $\hat{\boldsymbol{e}}_j$ that are large in magnitude and ignore entries close to zero. Sometimes only one entry is large. Sometimes all of the large entries have approximately the same size and sign, then the principal component is interpreted as an average of these entries. If exactly two entries are of similar large magnitude but of different sign, the principal component is interpreted as a difference of the two entries. If there are $j \geq 2$ large entries that differ in magnitude, then the principal component is interpreted as a linear combination of the corresponding variables.

58) PCA based on $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is easier to interpret than PCA based on $\boldsymbol{S}$ or $\boldsymbol{S}_U$.

i) If $\boldsymbol{S}$ is used, the variance explained by the first principal component could be large because one variable has much larger variance than the other variables.

ii) If $\boldsymbol{S}$ is used, the correlation of the ith variable with the jth principal component is proportional to the ith entry of the jth eigenvector $\hat{\boldsymbol{e}}_j$ divided by the standard deviation of ith variable: $e_{ij}/\sqrt{S_{ii}}$.

Hence PCA based on $\boldsymbol{S}$ is harder to interpret if p random variables do not have similar sample variances. The variances could differ if different units are used or if some variables are transformed while others are not. Hence PCA based on $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is recommended.

```
59) Typical Routput is shown. Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
Rotation: PC1         PC2         PC3         PC4         PC5
len       0.67271620 -0.21639022  0.05559575  0.15178244 -0.68883916
nasal    -0.22213361 -0.66957907  0.05173705 -0.68978370 -0.15440936
bigonal  -0.01373814  0.02995162  0.99668240  0.03545927  0.06542933
cephalic -0.67269993  0.21806615  0.02362841  0.16076405 -0.68812686
buxy     -0.21306252 -0.67556583 -0.01727087  0.68851877  0.15446292
```

60) Let $\hat{\boldsymbol{\Sigma}}$ be a consistent estimator of $\boldsymbol{\Sigma}$. The following theorems show that asymptotically, the eigenvalues and eigenvectors of $\hat{\boldsymbol{\Sigma}}$ act as those of $\boldsymbol{\Sigma}$ and vice verca. This result is useful since eigenvectors are not continuous functions of the dispersion matrix. The following theorem holds because eigenvalues and the generalized correlation matrix are continuous functions of the dispersion matrix.

i) **Theorem 6.1.** Suppose the dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Suppose $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let the eigenvalue eigenvector pairs of $\hat{\boldsymbol{\Sigma}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then $\hat{\lambda}_j(\hat{\boldsymbol{\Sigma}}) \xrightarrow{P} c\lambda_j(\boldsymbol{\Sigma}) = c\lambda_j$, $\hat{\boldsymbol{\rho}} \xrightarrow{P} \boldsymbol{\rho}$ and $\hat{\lambda}_j \left(\hat{\boldsymbol{\rho}}\right) \xrightarrow{P} \lambda_j \left(\boldsymbol{\rho}\right)$ where $\lambda_j(\boldsymbol{A})$ is the jth eigenvalue of $\boldsymbol{A}$ for $j = 1, ..., p$.

ii) **Theorem 6.2.** Assume the $p \times p$ symmetric dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. a) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i \boldsymbol{e}_i \xrightarrow{P} \boldsymbol{0}$.

b) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i \hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

If $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$, then

c) $\lambda_i \boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$, and

d) $\hat{\lambda}_i \hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.

e) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, and if the eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_j$ with $\boldsymbol{e}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$.

iii) **Theorem 6.3.** Under (E1), the correlation of the eigenvalues computed from the classical PCA and robust PCA converges to 1 in probability.

61) Centering uses $\boldsymbol{w}_i = \boldsymbol{x}_i - T$ where T is the sample mean or the sample mean of the standardized data for the full data set or for the set U used to compute the robust estimator. Centering does not change $\boldsymbol{S}, \boldsymbol{S}_U, \boldsymbol{R}$ or $\boldsymbol{R}_U$, but the jth principal component is $\hat{\boldsymbol{e}}_j^T \boldsymbol{w}_i = \hat{\boldsymbol{e}}_j^T(\boldsymbol{x}_i - T)$.

62) Let $\boldsymbol{x}$ be the $p \times 1$ vector of predictors, and partition $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ where $\boldsymbol{w}$ is $m \times 1$ and $\boldsymbol{y}$ is $q \times 1$ where $m = p - q \leq q$ and $m, q \geq 2$. Canonical correlation analysis (CCA) seeks m pairs of linear combinations $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y}), ..., (\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ such that $\text{corr}(\boldsymbol{a}_i^T\boldsymbol{w}, \boldsymbol{b}_i^T\boldsymbol{y})$ is large under some constraints on the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ where $i = 1, ..., m$. The first pair $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y})$ has the largest correlation. The next pair $(\boldsymbol{a}_2^T\boldsymbol{w}, \boldsymbol{b}_2^T\boldsymbol{y})$ has the largest correlation among all pairs uncorrelated with the first pair and the process continues so that $(\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ is the pair with the largest correlation that is uncorrelated with the first $m - 1$ pairs. The correlations are called *canonical correlations* while the pairs of linear combinations are called *canonical variables.*

63) R output is shown in symbols for the following table.

corr					
$\hat{\rho}_1$	$\cdots$	$\hat{\rho}_m$			
wcoef					
$\boldsymbol{w}$	$\hat{\boldsymbol{a}}_1$	$\cdots$	$\hat{\boldsymbol{a}}_m$		
ycoef					
$\boldsymbol{y}$	$\hat{\boldsymbol{b}}_1$	$\cdots$	$\hat{\boldsymbol{b}}_m$	$\cdots$	$\hat{\boldsymbol{b}}_q$

```
64) $out$cor
[1] 0.98596703 0.06797587   $out$ycoef
$out$xcoef                          [,1]       [,2]       [,3]
        [,1]       [,2]   L 0.1625452  0.4237524 -2.8492678
S 0.14966183  0.6460117   W 0.2369692  1.5379681  0.9356495
M 0.03236328 -0.8543387   H 0.2530324 -2.6806462  1.7785931
```

65) Some notation is needed to explain CCA. Let the $p \times p$ positive definite symmetric dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Let $\boldsymbol{J} = \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_A = \boldsymbol{J}\boldsymbol{J}^T = \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1/2}$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_B = \boldsymbol{J}^T \boldsymbol{J} = \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{e}_i$ and $\boldsymbol{g}_i$ be sets of orthonormal eigenvectors, so $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$, $\boldsymbol{g}_i^T \boldsymbol{g}_i = 1$ and $\boldsymbol{g}_i^T \boldsymbol{g}_j = 0$ for $i \neq j$. Let the $\boldsymbol{e}_i$ be $m \times 1$ while the $\boldsymbol{g}_i$ are $q \times 1$.

Let $\boldsymbol{\Sigma}_a$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{a}_1), ..., (\lambda_m, \boldsymbol{a}_m)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Let $\boldsymbol{\Sigma}_A$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i = 1, ..., m$. Let $\boldsymbol{\Sigma}_b$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{b}_1), ..., (\lambda_q, \boldsymbol{b}_q)$. Let $\boldsymbol{\Sigma}_B$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{g}_i)$ for $i = 1, ..., q$. It can be shown that the m largest eigenvalues of the four matrices are the same. Hence $\lambda_i(\boldsymbol{\Sigma}_a) = \lambda_i(\boldsymbol{\Sigma}_A) = \lambda_i(\boldsymbol{\Sigma}_b) = \lambda_i(\boldsymbol{\Sigma}_B) \equiv \lambda_i$ for $i = 1, ..., m$. It can be shown that $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{e}_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{g}_i$. The eigenvectors $\boldsymbol{a}_i$ are not necessarily orthonormal and the eigenvectors $\boldsymbol{b}_i$ are not necessarily orthonormal.

Theorem 7.1. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. Assume $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_b$ are positive definite and that $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let $\boldsymbol{d}_i$ be an eigenvector of the corresponding matrix. Hence $\boldsymbol{d}_i = \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{e}_i$ or $\boldsymbol{g}_i$. Let $(\hat{\lambda}_i, \hat{\boldsymbol{d}}_i)$ be the ith eigenvalue eigenvector pair of $\hat{\boldsymbol{\Sigma}}_\gamma$.

a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ and $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_\gamma) = \lambda_i$ where $\gamma = A, a, B$ or b.

b) $\boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{d}}_i - \lambda_i \hat{\boldsymbol{d}}_i \xrightarrow{P} \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{0}$.

c) If the jth eigenvalue λ_j is unique where $j \leq m$, then the absolute value of the correlation of $\hat{\boldsymbol{d}}_j$ with $\boldsymbol{d}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{d}}_j, \boldsymbol{d}_j)| \xrightarrow{P} 1$.

Sections covered: Olive (2012) 1.1, 1.2, 1.4, ch. 2, ch. 3 (skim ∮ 3.4) skim ch.4 with emphasis on p. 62, DGK, MG, FCH, RFCH and RMVN estimators, DD plot. From ∮ 5.1, Def. 5.1, Applications 5.1 and 5.2. Sections 6.1,6.2,7.1,7.2.

Johnson and Wichern (1988): 1.3, 1.4, ch. 2 is a review of linear algebra p. 45, 46, and sections 2.4, 2.5, 2.6 are important. Ch. 3: p. 89, 100, 103-4, ∮ 3.5 are important. Ch. 4: ∮ 4.2, 4.3 (omit proofs), p. 144-145, 155-157. Ch. 8, 10. Skip section 10.6.

66) In *supervised classification*, there are k known groups or populations and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., k$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the k groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups.

67) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., k$. This rule is robust to nonnormality and the assumption of equal population dispersion matrices, but $\hat{f}_j$ is hard to compute for $p > 1$.

68) Given the $\hat{f}_j(\boldsymbol{w})$ or a plot of the $\hat{f}_j(\boldsymbol{w})$, determine the maximum likelihood discriminant rule. See HW6 D, Q6.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\rho_a(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group a are unknown or equal: $\rho_a(\boldsymbol{w}_i) = 1/k$ for $a = 1, ..., k$. Often it is assumed that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-k} \sum_{j=1}^{k} (n_j - 1) \boldsymbol{S}_j$$

where $n = \sum_{j=1}^{k} n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, eg the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

69) Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \boldsymbol{w} - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., k$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$. LDA is robust to nonnormality and somewhat robust to the assumption of equal population covariance matrices.

70) The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2} \log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., k$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. QDA has some robustness to nonnormality.

71) The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$. This rule is robust to nonnormality and the assumption of equal $\boldsymbol{\Sigma}_j$, but needs $n_j > 10p$ for $j = 1, ..., k$.

72) Assume that $k = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in$ group 1). Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression (LR) estimate of $\rho(\boldsymbol{w})$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Equivalently, the LR rule allocates $\boldsymbol{w}$ to group 1 if $ESP > 0$ and allocates $\boldsymbol{w}$ to group 0 if $ESP < 0$.

73) Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i > 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$. If the LR response plot is good, $n_0 > 5p$ and $n_1 > 5p$, then the LR rule is robust to nonnormality and the assumption of equal population dispersion matrices. Know how to tell a good LR response plot from a bad one. See HW6 E, Q6.

74) Given LR output, as shown below in symbols and for a real data set, and given $\boldsymbol{x}$ to classify, be able to a) compute ESP, b) classify $\boldsymbol{x}$ in group 0 or group 1, c) compute $\hat{\rho}(\boldsymbol{x})$. See HW6 E, Q6.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p / se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
Binomial Regression Kernel mean function = Logistic
Response = Status  Terms = (Bottom Left) Trials = Ones
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -389.806         104.224        -3.740     0.0002
Bottom     2.26423         0.333233        6.795     0.0000
Left       2.83356         0.795601        3.562     0.0004
```

75) Suppose there is training data $\boldsymbol{x}_{ij}$ for $i = 1, ..., n_j$ for group j. Hence it is known that $\boldsymbol{x}_{ij}$ came from group j where there are $k \geq 2$ groups. Use the discriminant analysis method to classify the training data. If m_j of the n_j group j cases are correctly classified, then the *apparent error rate for group j* is $1 - m_j/n_j$. If $m_A = \sum_{j=1}^k m_j$ of the $n = \sum_{j=1}^k n_j$ cases were correctly classified. Then the *apparent error rate* AER $= 1 - m_A/n$.

76) For the `ddiscr` method, get the apparent error rate for each of the k groups with the following commands. Replace `ddiscr` by `ddiscr2` for the `ddiscr2` method.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$err
```

Get apparent error rates for `ddiscr,` `LDA` and `QDA` with the following commands.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$toterr

out2  <- lda(x,group)
1-mean(predict(out2,x)$class==group)

out3  <- qda(x,group)
1-mean(predict(out3,x)$class==group)
```

Get the AERs for the methods that use variables x_1, x_3 and x_7 with the following commands.

```
out <- ddiscr(x[,c(1,3,7)],w=x[,c(1,3,7)],group,xwflag=T)
out$toterr

out <- lda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)

out <- qda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

Get the AERs for the methods that leave out variables x_1, x_4 and x_5 with the following commands.

```
out <- ddiscr(x[,-c(1,4,5)],w=x[,-c(1,4,5)],group,xwflag=T)
out$toterr

out <- lda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)

out <- qda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)
```

77) Expect the apparent error rate to be too low: the method works better on the training data than on the new data to be classified.

78) Cross validation (CV): for $i = 1, ..., n$ where the training data has n cases, compute the discriminant rule with case i left out and see if the rule correctly classifies case i. Let m_C be the number of cases correctly classified. Then the CV error rate is $1 - m_C/n$.

79) Suppose the training data has n cases. Randomly select a subset L of m cases to be left out when computing the discriminant rule. Hence $n - m$ cases are used to compute the discriminant rule. Let m_L be the number of cases from subset L that are correctly classified. Then the "leave a subset out" error rate is $1 - m_L/m$. Here m should be large enough to get a good rate. Often m uses between $0.1n$ and $0.5n$.

80) Variable selection is the search for a subset of variables that does a good job of classification.

81) Forward selection: suppose $X_1, ..., X_p$ are variables.

Step 1) Choose variable $W_1 = X_1$ that minimizes the AER.

Step 2) Keep W_1 in the model, and add variable W_2 that minimizes the AER. So W_1 and W_2 are in the model at the end of Step 2).

Step k) Have $W_1, ..., W_{k-1}$ in the model. Add variable W_k that minimizes the AER. So $W_1, ..., W_k$ are in the model at the end of Step k).

Step p) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

82) Backward elimination: suppose $X_1, ..., X_p$ are variables.

Step 1) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

Step 2) Delete variable $W_p = X_j$ such that the model with $p-1$ variables $W_1, ..., W_{p-1}$ minimizes the AER.

Step 3) Delete variable $W_{p-1} = X_j$ such that the model with $p-2$ variables $W_1, ..., W_{p-2}$ minimizes the AER.

Step k) $W_1, ..., W_{p-k+2}$ are in the model. Delete variable $W_{p-k+2} = X_j$ such that the model with $p - k + 1$ variables $W_1, ..., W_{p-k+1}$ minimizes the AER.

Step p) Have W_1 and W_2 in the model. Delete variable W_2 such that the model with 1 variable W_1 minimizes the AER.

83) Other criterion can be used and `proc stepdisc` in *SAS* does variable selection.

84) In *R*, using LDA, leave one variable out at a time as long as the AER does not increase much, to find a good subset quickly.

85) For PCA, the `summary(out)` statement shows

Importance of components:	PC1	PC2	$\cdots$	PCk	$\cdots$	PCp
Standard deviation	$\sqrt{\hat\lambda_1}$	$\sqrt{\hat\lambda_2}$	$\cdots$	$\sqrt{\hat\lambda_k}$	$\cdots$	$\sqrt{\hat\lambda_p}$
Proportion of variance	$\frac{\hat\lambda_1}{\sum_{i=1}^p \hat\lambda_i}$	$\frac{\hat\lambda_2}{\sum_{i=1}^p \hat\lambda_i}$	$\cdots$	$\frac{\hat\lambda_k}{\sum_{i=1}^p \hat\lambda_i}$	$\cdots$	$\frac{\hat\lambda_p}{\sum_{i=1}^p \hat\lambda_i}$
Cumulative Proportion	$\frac{\hat\lambda_1}{\sum_{i=1}^p \hat\lambda_i}$	$\frac{\sum_{j=1}^2 \hat\lambda_j}{\sum_{i=1}^p \hat\lambda_i}$	$\cdots$	$\frac{\sum_{j=1}^k \hat\lambda_j}{\sum_{i=1}^p \hat\lambda_i}$	$\cdots$	1

Recall that if $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, then $\sum_{i=1}^p \hat\lambda_i = p$. Typically want to keep the first m principal components where $\frac{\sum_{j=1}^m \hat\lambda_j}{\sum_{i=1}^p \hat\lambda_i} > a$ where the threshold a is a number like 0.9, 0.8 or 0.7.