Partitioning: sample covariance matrix has transpose of each other within the partitioned matrix. **Number of Elements:** $\frac{n(n+1)}{2}$ unique elements in a cov or corr matrix of n variables. **Generalized Sample Variance:** $|S|$ is product of eigenvalues of S, $|S| = 0$ if any eigenvalue is 0 (multi co-linearity). **Mahalanobis distance:** multivariate version of a z-score; $\delta^2 = (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)$ is in quadratic form and return s a scalar; **variables that have a high degree of variation will contribute less to the overall Mahalanobis distance**; how far a measurement is from the mean vector relative to what a typical deviation from the mean is. **Properties of Multivariate Normal:** y is $N_p(y_i, \Sigma)$; $z \sim N(a^T\mu, a^T\Sigma a)$ and individual y's must be normal too; rank(A) = q $\leq$ p then $z \sim N_q(A\mu, A^T\Sigma A)$; if $\mu = 0$ and $\Sigma = I$ then $Ay \sim N(0, I)$. **Independent Random Variables:** y_j and y_k are independent if and only if covariance $s_{jk} = 0$ or stated in terms of correlation of row jk; **Implication goes both ways since property of MV normal distribution.** *Handout 5:* **Disprove p vars are jointly MV normally distributed**: null hypothesis of these tests is that variables do follow a MV normal distribution. **Multivariate CLT:** $\bar{y} \sim N_p(\mu, \frac{\Sigma}{n})$ for a large enough sample size n. **Sample Variance and Cov Matrix Distribution:** univariate follows a chi-squared distribution; sample var/cov matrix follows a Wishart distribution: $(n-1)S \sim Wishart(n-1, \Sigma)$ if y follows a MV normal distribution. **Univariate 1-Sample T Test:** $t = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$. **Hotelling's 1-Sample T^2 Test:** $T^2 = n(\bar{y}-\mu_0)^T S^{-1}(\bar{y}-\mu_0)$; measures how far observed $\bar{y}$ is from its expected value, if the null were true, while taking into account the variance/cov of the sample mean vector. **Hotelling's T^2 Density:** no upper bound; reject when T^2 is large. **P-value accuracy:** data values must have been sampled from MV normal dist, S must be non-singular, and n > p (observations > variables). **Steps to Take After Rejecting Null:** check multivariate normality of data, conduct univariate tests on each variable. **Benefits of MV Tests:** using p univariate tests inflates the type I error rate (rejecting null when it is true); p = 4 and $\alpha = 0.05$ then $1 - (1 - 0.05)^4 = 0.19$ probability and quickly increases with p; does not ignore correlation structure between p vars; more power (prob of rejecting null when null is true) and high power is good; small deviations may be statistically significant when combined. **Limitations of MV Tests:** interpretations difficult without univariate tests when statistically significant MV test; **non-directional so null hypothesis for MV is 2-tailed.**— *Handout 6:* **Univariate 2-Sample t-Test:** H_0: $\mu_1 = \mu_2$; H_a: $\mu_1 \neq \mu_2, \mu_1 > \mu_2, \mu_1 < \mu_2$; $t = \frac{\bar{y}_1 - \bar{y}_2}{s_p l\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$; $s_{pl}^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$; b/c assuming equal variances; validity of p-val: sampled from normal dist. but diff means. **MV 2-Sample Hotelling's T Test:** H_0: $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$; H_a: $\boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$; $T^2 = (\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})^T[S_pl(\frac{1}{n_1}+\frac{1}{n_2})]^{-1}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})$; is the Maha. distance between mean vectors; $\boldsymbol{S_{pl}} = \frac{(n_1-1)\boldsymbol{S_1}+(n_2-1)\boldsymbol{S_2}}{n_1+n_2-2}$; validity of p-val is MV normal dist. with equal cov. matrices but diff. means; y_ij is the jth obs. in group i and vectors with p entries; follow up with univariate t-tests to see where differences lie. **Paired MV Data:** Difference = Treatment - Control; MV and Paired (same subject); Apply 1-Sample Hotelling T Test; H_0: $\boldsymbol{\mu_d} = 0$; H_a: $\boldsymbol{\mu_d} \neq 0$; $T^2 = \bar{\boldsymbol{d}}^T[S_d(\frac{1}{n})]^{-1}\bar{\boldsymbol{d}}$; Maha. dist between $\bar{\boldsymbol{d}}$ and 0; validity of p-val is MV normal dist. with $N_z(\boldsymbol{\mu_d}, \Sigma_d)$; follow up with univariate t-tests to see which dist. from 0 is significant. *Handout 7:* **Uni 1-Way ANOVA:** $y_i.$ = sum across all measurements in sample i, $\bar{y}_i.$ = sample mean of all measurements in sample i, $\bar{y}_{..}$ = (grand) sample mean of all measurements across all samples; statistical model: $y_{ij} = \mu_i + \alpha_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}$, where (1) is overall pop mean response, (2) effect on mean resp. due to pop i, (3) population mean response for pop i, (4) random error assoc. with jth response in pop i; assume: (1) $\epsilon_{ij} \sim N(0, \sigma^2)$, (2) $E(y_ij) = \mu_i$, (3) $Var(y_ij) = \sigma^2$; var equal; H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$; H_a: at least 1 $\mu_i \neq \mu_j$; estimate of σ^2: **within** $MSE = \frac{\Sigma_{i=1}^k \Sigma_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2}{nk-k} = \frac{SSE}{nk-k}$, thought of as pooled estimate of σ^2; **between** $MSH = n\frac{1}{k-1}\Sigma_{i=1}^k(\bar{y}_i. - \bar{y}_{..})^2 = \frac{SSH}{k-1}$, thought of as estimate of σ^2; $F = \frac{MSH}{MSE}$, which is explained variation / unexp. variation (within is noise and between is signal); reject when F is large, look at graphs, sd of within group (noise). **1-Way MANOVA:** y_{ij} is a vector of $y_{ij1}, \ldots, y_{ijp}$; same stat model but with vectors and $y_i.$ estimate of μ; H_0: $\boldsymbol{\mu_1 = \mu_2 = \ldots = \mu_k}$; H_a: all pops. do not have the same mean vectors; **between** $H = n\Sigma_{i=1}^k(\bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{.} - \bar{\boldsymbol{y}}_{..})^T$; **within** $E = \Sigma_{i=1}^k\Sigma_{j=1}^{n_i}(y_{ij} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{.})(y_{ij} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{.})^T$; $E + H$ is total sample covariance; Det() of each give generalized within/total variation: $\Lambda = \frac{|E|}{|E+H|}$; Wilk's Λ stat which is noise / total; reject when Λ is small; Transform Λ to F and when large reject; `manova()` then `summary()` with "Wilks", $SS gives H and $Residuals gives E; follow up with univariate ANOVAs to see which var has diff means across groups; `TukeyHSD()` with Bonferroni correction: which groups are different; p-adj bumped up to protect from Type I error; compare p-vals to α/p to see which are significant, compare sample means this way. *Handout 8:* **MANOVA Test Statistics:** All functions of the eigenvalues of $E^{-1}H$, $s = min(k-1, p)$ is the rank of it; Wilk's $\Lambda = \prod_{i=1}^s \frac{\lambda_i}{1+\lambda_i}$, Roy's largest root $\frac{1}{1+\lambda_1}$, Pillai's $\sum_{i=1}^s \frac{1}{1+\lambda_i}$ (for hetero in cov matrix), Lawley-Hotelling $\sum_{i=1}^s \lambda_i$. **MANOVA Type 1 Error Rate:** As Corr between vars incr. type 1 error rate remains below 0.05. **Profile Analysis on MANOVA:** H_{01}: k profiles are parallel (slope from var to var is same across groups: $\mu_{12} - \mu_{11} = \mu_{22} - \mu_{21} = \ldots$), H_{02}: k profiles are at the same level (average of the mean elements are eq across groups (sums equal): $\sum \mu_{1i} = \sum \mu_{2i} = \ldots$) **offset in graph, sum of endpoints with 2 lines**, H_{03}: k profiles are flat (p vars avgs across groups are the same: $\mu_{11} + \mu_{21} + \cdots = \mu_{12} + \mu_{22} + \ldots$) **stacked points avgs are the same with 2 lines**; start with a MANOVA; Parallel Interp: moderate evidence that mean weights change at different rate depending on treatment (reject), Levels Interp: average of the mean weights may be the same across treatments (fail to reject), Flat Interp: mean weights do change regardless of group (reject); Repeated Measures Data: same experimental unit yields multiple observations (obsiv unit not same as exp unit), use profile analysis and MANOVA for this data. *Handout 9:* **Single Pop Cov Test:** H_0: $\Sigma = \Sigma_0$; H_a: $\Sigma \neq \Sigma_0$; **Test Stat:** $u = (n-1)[ln|\Sigma_0| - ln|S| - tr(S\Sigma_0^{-1}) - p]$; S close to Σ_0 then logs are similar, $S\Sigma_0^{-1} \sim I$, so trace is p, so $u \sim 0$; reject when stat is large; **DF for X^2 dist is the expected value**; Stat and P-val appropriate when n is relatively large and follows MV dist. (Mardia test). **Several Pop. Cov Test:** need for Hotelling's (2-Sample) and MANOVA; H_0: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$; H_a: all pop. do not have same covariance matrix; Box M Test Conditions are that k independent samples of data and each k comes from MV dist $N(\mu_i, \Sigma_i)$; Handout 6 pooled sample cov eq but now for several pops is the weighted avg of sample cov. matrices, weighting by sample size; Λ is ration of det of S_i to det of S_pl, when 1 then lambda is around 1; Transform to M-test and reject when M-stat is large; **DF is the expected value**. **Test for Indep. of P Vars:** Σ is diagonal and 0 elsewhere means p vars indep. of each other (can be Σ_0); P_ρ is pop. corr. matrix where diagonal are 1 and 0 elsewhere means p vars indep.; H_0: $P_\rho = I$; H_a: $P_\rho \neq I$; $|R|$ is det of sample corr. matrix and $|R| = 1$ if indep. and 0 if dep., 0-1 measures degree of evidence for/against null; $X^2 = -[(n-1) - \frac{2p+5}{6}]ln|R|$ with df being $\frac{p(p-1)}{2}$ which is the expected value. *Handout 10:* **Discriminant Analysis:** follow up to MANOVA; MANOVA and Hotelling's can distinguish linear combos of vars across groups; maximally separate z-bars of groups; $z = a'y = [S_{pl}^{-}1(\bar{y}_1 - \bar{y}_2)]^T y$, where a' is a scalar; t test on z scores reveals significant differences. **Disc. Analysis on Several Pops.:** $E^{-1}H$ eigenvector or largest eigenvalue is optimal disc. func. to maximally distinguish; handout 8 first line gives number of funcs; s disc. funcs. define s indep vars that are funcs. of original p vars.; **Relative Importance Formula:** $\frac{\lambda_i}{\Sigma_{i=1}^{s}\lambda_i}$, how much variability you retain between groups; $\bar{\boldsymbol{z}}$ of each group is pushed as far as possible from each other. *Handout 11:* **Stand. Coeffs. of Disc. Funcs.:** `scale()` to get z scores to compare relative importance of vars. **Stat. Signif. of Disc. Funcs.:** $\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}$, $V_m = -[N - \frac{p+k+1}{2}]ln(\Lambda_m)$, V_m transforms to F-stat with df $= (p-m+1)(k-m)$ as expected value; rejecting MANOVA tests whether at least first eigenvector provides significant dimension of separation; conditions: MV normal and equal cov. matrices across groups. **Stat. Sign. of Vars:** test if vars contribute sign. in group separation; $\Lambda = \frac{\Lambda_p}{\Lambda_{p-1}}$ where top is all vars and bottom is without var, useful if result is small (same if not useful); $F = \frac{1-\Lambda}{\Lambda}\frac{N-p-k+1}{p+k-1}$, large when var is useful, reject null that var does not contribute to group separation; partial F test: compare p-vals with Bonferroni correction of α/p. **Classification Trees:** want high node homo or low hetero; measure node hetero using misclassification rates; `rpart()` with output as `root, n, loss (misclassified), (Group1, ...)`; large stretch in tree means helpful in split. *Handout 12:* **Fisher's Procedure:** after `lda() and predict()` compute $\bar{\boldsymbol{z}}$ of all groups, `predict()` on new y to get z score, find group using $D^2 = (z_1 - \bar{z_{i1}})^2 + (z_2 - \bar{z_{i2}})^2$. **Conf. Matrix:** accuracy: prop of obs. that classified correctly, misclass. rate is $1 - A$; sensitivity: prop of a given class that is classified as that class; specificity: prop of records not of a given class that are not classified as that class (ex 98/100). **Linear Class Func:** assume groups have same cov. matrix; minimize Maha. distance foild to maximize $L_i(y) = \bar{y_i}^T S_{pl}^{-1} y - (1/2)\bar{y_i}^T S_{pl}^{-1} y_i$. **Quad Class Func:** same minimize Maha dist. but with indiv cov matrix so less power (less data) b/c no assumption. **Steps on Analysis:** (1) Graphic and stats (2) mvn test (3) BoxM test for equal cov matrices across groups (4) standardize for disc analysis (5) lda() confusion matrix (6) qda() conf matrix. **LOO CV:** Fisher/QDA not good on new data; Training and Testing Set; CV by omit 1st and do lda on rest, but use model on all data and CV as honest estimator of how well model does; Interp: using CV estimates of accuracy/misc rates, sens, spec, one method does a bit better than the other; after CV use model for all predictions (not CV).

Handout 13: **Univariate Linear Regression:** setup is 1 quant resp var and q quant explanatory vars; random error allows random variation around trend. **Assumptions:** residuals are normal with mean 0, σ^2 constant; variance of resid is σ^2 and cov is 0; so cov matrix is $\sigma^2 I$; expected value of resid is $\mathbf{0}$, expected value of y is $\mathbf{X}\beta$. **Estimation:** $\hat{\beta} = (X^T X)^{-1} X^T y$; $\hat{y} = X\hat{\beta}$; minimize sum of squared residuals; sample variance of residuals is MSE $= s^2 = \frac{SSE}{n-q-1}$; df of SSE is $n - (q+1)$, where q+1 is the number of β's. **Coeff of Determ:** statistical significance of R^2 is tested with F-stat **MV Linear Regression:** modeled as $y = X\beta + \epsilon$. **Assumptions:** $E(\epsilon) = 0$, which means $E(y) = X\beta$; $cov(y_i) = \Sigma$, meaning cov between y's for same observation; $cov(y_i, y_j) = 0$ for $i \neq j$, meaning y's for diff observations is 0; both these imply that measurements of y on teh same observation can be correlated, but with the same cov structure for all obsv, measurements taken on diff obsv are uncorrelated; $\Sigma_i \sim N_p(0, \Sigma)$; **Estimation:** Dimension of β is $(q+1) \times p$. **Estimate Common Cov Matrix:** $S_e = \frac{1}{n-q-1}E$, where E is like SSE and is the MV version of unexplained variability in data (error SS); total $= E + H = (Y^T Y - \hat{\beta}^T X^T Y) + (\hat{\beta}^T X^T Y - n\bar{y}^T \bar{y})$. **Test of Overall Regression:** $\lambda = \frac{|E|}{|E+H|}$, which is unexplained over total; H_0 : no linear association between any of the x's (vector) and any of the y's (vector); H_0: $\beta_1 = 0$, where β_1 is the part of the matrix with the predictors (without intercept); reject when λ is small. **Conclusion:** lambda small: there is little unexplained variation. regression model explains a significant amount. At least one of the x's is linearly associated with at least one of y's; then look at univariate tests to see which x's are significant. *Handout 14:* **Canonical Correlation:** also the MV correlation coeff; partition data into μ and Σ, which has cov matrix of y and x; sample mean vector is $\bar{y}$ and $\bar{x}$; sample cov matrix is S_{yy}, 2 S_{yx}, and S_{xx}. **Intuition:** find 2 linear combs of y and x that will maximize Pearson's sample corr between observed transformed values; Find a_1 and b_1 such that $\mu_1 = a_1^T y$ and $\nu_1 = b_1^T x$ have max corr. **Solution:** a_1 is first eigenvector of $S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{yx}$ and b_1 is first eigenvector of $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{xy}$; max corr coeff is first eigenvalue of any of the above eqs; **μ and ν are the canonical variates**; 2 matrices should have same eigenvalues but different vectors. **Bivariate Graph:** first canonical corr should be at least max corr between any pair of y and x in graph. **Correlation:** squaring first eigenvalue gives correlation coeff; in R `corr(all_u, all_v)`; largest λ_1 is the largest possible squared corr of any 2 linear comb of x and y vars; $\lambda_1 = (corr(\mu_1, \nu_1))^2$. **Definition:** First canonical corr r_1 is $\sqrt{\lambda_1} = |corr(\mu_1, \nu_1)|$, magnitude of largest possible corr coeff; s = min(p, q) eigenvalues so s canonical corr coeff. **Test of Significance of Canonical Corr:** (Test of Overall Regression); Wilk's $\Lambda = \prod_{i=1}^{s}(1 - r_i^2)$, **can test sign of next canonical corr coef by starting product at $i+1$**; in R `linearHypothesis()`; reversing the order of x and y will give same test of sign but diff H and E. *Handout 15:* **Intuition Comp Analysis:** goal is dimension reduction; SE are large and lack of significance b/c of multicoll; reduce vars to increase precision (lower SEs) estimates and predictions; can be not significant b/c of covering 0 in CI interval; benefit from having exp vars with larger variance. **Principal Comp Analysis:** vars will be uncorrelated and max variance; **Derivation:** $z = Ay$ and $\Sigma = A\Sigma_y A^T$, where cov matrix of z is diagonal: 0s elsewhere, cov(zi, zj) = 0 and var(zi) = $\sigma_{z_i^2}$. **Spectral Decomp:** $\Sigma_y = CDC^T$ where C is the normalized eigenvectors and D contains eigenvalues of Σ_y, so $D = C^T \Sigma_y C$, which is a diagonalized sigma matrix; $z = C^T y$ so $\Sigma_z = C^T \Sigma_y C = D$; tr($\Sigma_y$) is the total population variance; total variance of 1st k princ comps / total var of all y's $= \frac{\Sigma_{i=1}^{k}\lambda_i}{\Sigma_{i=1}^{p}\lambda_i}$, bottom is tr($\Sigma_y$); z's have all of the var/cov of the y's in terms of generalized and total variance. **Implementation:** use sample corr matrix R_y when (1) some of vars have much larger variance than others (2) y-vars on much diff scales; use as many components to retain at least 80% of og vars total variance; Hotelling's test to see if they differ by class. **Handout 16: Objective of Cluster Analysis:** only 1 way to put all obsv into 1 cluster, only 1 way to put data into 16 clusters; 10 billion ways to cluster these 16 observations example. **Hierarchical Agglomerative:** start with m any small then join most similar, g suitable degree of homo, stop until 1 cluster will all obsv; use euclidean distance; single linkage or nearest neighbor: uses distance between 2 nearest obsv or vectors as distance between clusters; complete linkage or farthest: distance between 2 clusters is max possible dist between obsv (poor); average: average distance between obsv in clusters; centroid: distance between sample mean vector of obsv in clusters; Ward's: measure within cluster variability by sums of squares, combine 2 clusters with smallest increase in within-cl variability, minimize $SSE_{AB} - (SSE_A + SSE_B)$; dendrogram for linkage. **Hierarchical Divisive:** few large and successively divide into more and more similar to make homo; **Partitioning Methods:** approx of all possible partitions that can be considered and assign obsv in way that makes clusters as homo as possible; K-Means: get seeds of clusters, measure distance of each osbv to centroid, closest are assigned, repeat until no longer assigned to new clusters.