

Statistics 215B: Lab 4

UC Berkeley, Spring 2011

Hoxie Ackerman

April 19, 2011

1 Introduction

The field of modern statistics owes a great deal to the scientific revolution that began with the publication of Darwin's 'On the Origin of Species' in 1859. Turning both the scientific and philosophical worlds on their respective heads, the idea of organisms evolving to adapt to changes in their environment was a radical one. Inspired by the theory of evolution and interested in both collecting data to support or refute this new hypothesis and studying patterns of genetic inheritance, Darwin's half-cousin and intellectual descendant Francis Galton founded the Anthropometric Laboratory in 1884 as well as the Eugenics Record Office and the Biometric Laboratory at University College London in 1904 [1].

One of the goals of these institutions was to collect evidence of evolution in action, i.e. to observe measurable physical change in organisms, and in an extreme case, to document the emergence of a new species in reaction to an environmental change. As a way to organize efforts and disseminate findings, Galton and his statistical disciple Karl Pearson founded the journal *Biometrika* in 1901. Though today, *Biometrika* is home to premiere theoretical statistics publications, 100 years ago, *Biometrika* (literally "the measure of life") concerned itself with the publication and analysis of datasets collected from various species around the world. Given these data, the then-new technology of 'parameter estimation' would be applied to various physical characteristics and simple statistical tests would be performed in an effort to find a significant change. For a more thorough account of this side of the birth of modern statistics, please see [2].

In the vein of measuring biological characteristics to examine changes over time, the data set under consideration (which perhaps not coincidentally comes to us from the Galton Laboratory) consists of four series of male Egyptian skulls, with each series coming from a different time period [3]. A number of physical distances on each skull were measured, and in her analysis, Barnard performs variable subset selection, tests for statistically significant differences between measurements from two series, and tries to use the significant variables to distinguish between the series. In this report, I attempt to recreate her data set by cleaning messier data files, explore the data visually from a classification perspective, and use a number of statistical machine learning algorithms to classify skull series based on physical measurements.

2 Recreating Data and EDA

2.1 Recreating Data

As described in the Introduction, the data under consideration are physical measurements of human skulls from different time periods. To begin her analysis, Barnard selects a subset of variables to consider. The seven variables (predictors / features) retained are Maximum breadth (B), Nasal width (NB), Glabello-occipital length (L), Basialveolar length (GL), Nasialveolar height (GH), Nasal height (NH), and Basibregmatic height (H) [3]. In this section, I clean the data based on these seven variables. In Section 3.1, I consider her variable selection choices more closely.

The data I had to analyze came in two files: `skulls1.txt`, which contained Series I, II, and IV, and `skulls2.txt`, which contained Series III. Though I initially believed that `skulls1.txt` contained these seven variables as described, Christine Ho pointed out that the columns of `skulls1.txt` were mislabeled. Fortunately, Google Books contains a digitized version of the source of the data, which I used to fix my columns [4]. (These column names were later fixed, but I was working with an earlier iteration of the data.) `skulls2.txt` contained exactly these seven variables, though separate left and right Nasal Height values were given. The left and right measurements were highly correlated ($r = 0.98$). To obtain a single Nasal Height value for comparison with the skulls in `skulls1.txt`, I averaged the left and right measurements. I also removed the Female skull measurements from `skulls2.txt`, since Barnard only considered males.

I next addressed missing values, considering only Barnard's seven variables for the time being. Barnard's goal in variable selection was to maximize the "number of skulls possessing all the measurements investigated," so I assumed that any skull missing a measurement on any of the seven variables should be removed from the data set. This filter reduced the total number of skulls from 517 to 394.

Some values in the remaining skulls, though not missing, were questionable: the values in the data files were literally followed by question marks. Whether these were added recently by 215B staff or years ago by some scientist unable to read someone's handwriting years ago, I didn't know. Since only about 8% of skulls had one or more questionable values, my first solution was to remove these skulls from the data

set. A more thorough solution would be to look at each questionable value individually, figure out where in the empirical distribution the questionable value lies, and informally assess the likelihood that the value is correct, but I wanted to see how close I was to Barnard’s values using my simpler approach first. Removing all skulls with questionable values reduced the total number of skulls from 394 to 363. (I found out later that these question marks were elements of the original data set, as published in [4], and that Barnard used the questionable values in her analysis. Background information is always key when considering missing or questionable values.)

Next, I considered bad data points. One distance had the value “8-May”. Dates-as-numbers errors usually come from Microsoft Excel helpfully converting decimal values to their date equivalents, but in this case, none of the values I could think of that would be converted to “8-May” fell in the range of X7:H. Thus, I removed this skull from Series II, leaving 362 skulls total.

As for bad data points with numeric values, I addressed those points which were clearly typos, as indicated by being on the wrong order of magnitude or negative (Figure 1a). Some of these values could be divided or multiplied by 10 (i.e. adjusting the decimal point by one position) or made positive, and they landed right in the middle of the distribution. Other typos were more subtle; Christine Ho pointed out that in the original data table, it was difficult to differentiate between 3s and 8s, and more than a few outliers could be made much more realistic by changing 8s to 3s. I also noticed one value that was made much more appropriate by changing a 7 to a 1. (For what I did exactly, please see attached code, lines 184–232.) Not all skulls could be fixed; after quality control, I had 361 skulls remaining. After fixing these typos, the data were much more reasonably distributed, as can be seen in Figure 1b.

In practice, I usually wouldn’t be so cavalier in adjusting observed values. However, I had two motives for doing so. First, the assignment said that we should only try to get our data close to Barnard’s, as measured by means of all series/variable combinations. Second, given that most male human skulls should be approximately the same size, it’s logical to seek corrections for observations with statistically outlying values. The fact that the original data are available make this correction process doable.

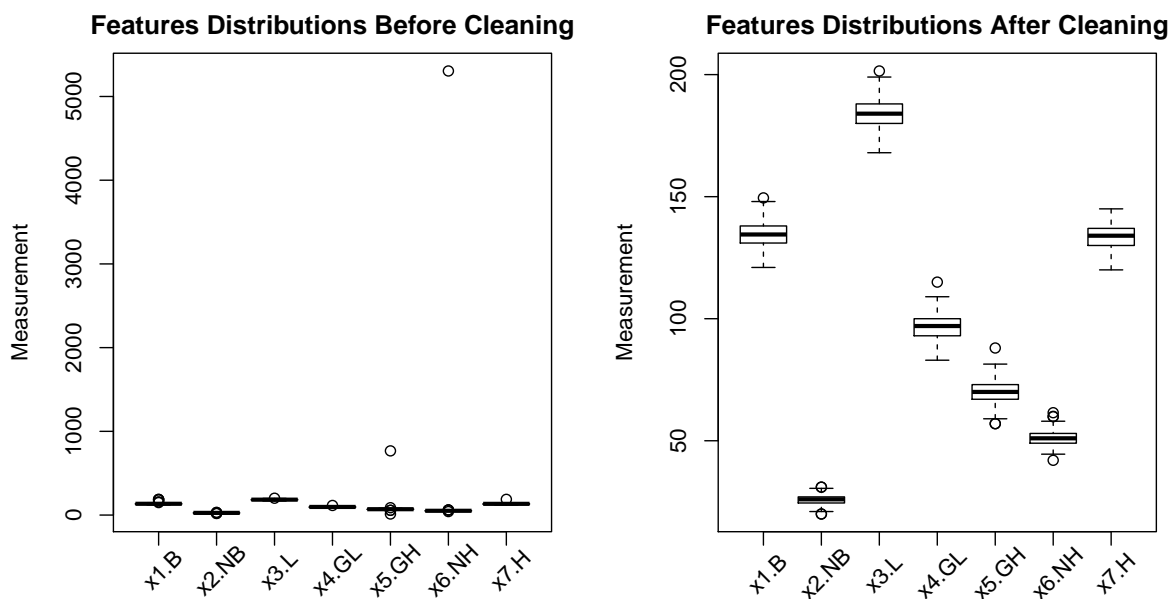


Figure 1: Univariate variable distribution before and after cleaning. (a) Values of the wrong order of magnitude were sometimes typos, sometimes errors. Distinguishing between the two was somewhat subjective. (b) After cleaning the data, univariate distributions were much more reasonable, containing only a few marginally outlying skulls.

To evaluate my data cleaning procedure, I computed summary statistics to compare with Barnard's Table 5 values. The values were, in my opinion, definitely close enough.

Distance	Series I: $n = 88$ (91)	Series II: $n = 139$ (162)	Series III: $n = 70$ (70)	Series IV: $n = 64$ (75)
X1: B	133.8 (133.6)	134.4 (134.3)	134.2 (134.4)	135.4 (135.3)
X2: NB	25.4 (25.5)	25.6 (25.8)	25.9 (26.0)	25.7 (25.7)
X3: L	185.2 (185.0)	182.6 (182.5)	184.5 (184.6)	183.5 (183.5)
X4: GL	98.3 (98.3)	96.9 (96.5)	95.8 (95.9)	95.0 (95.0)
X5: GH	70.4 (70.2)	70.4 (69.9)	69.2 (69.2)	71.2 (71.0)
X6: NH	50.9 (50.8)	51.2 (51.1)	50.0 (50.1)	51.7 (52.1)
X7: H	133.1 (133.0)	135.0 (134.9)	133.5 (133.6)	131.2 (131.5)

Table 1: Mean measurement values after data cleaning, with Barnard values in parentheses. The largest difference in means was 0.5 (Series II, X5:GH). Based on this table, I considered my data 'clean'.

2.2 EDA

Much of my own exploratory data analysis was performed during the data cleaning procedure explained in Section 2.1 and during the analyses given in Section 3. However, given the clean data set, I explored the data from a classification perspective by addressing the question, "How much separation between series is there at the bivariate level?" The answer to this question lies in Figure 7, which gives bivariate scatterplots for all series. The fact that all populations are essentially atop one another paints a relatively bleak picture from a classification perspective. Looking only at Series I and IV in Figure 8, which is another classification problem considered later in this report, we can see pockets of mild separation of a few points between the two groups, but most of the plots are again two populations transposed atop each other. Of course, these are only bivariate plots; it could be that when all dimensions are taken into account simultaneously, decent classification results emerge.

3 Further Data Considerations

3.1 Excluded Variables

As explained in her introduction, Barnard dropped four variables from consideration. These four variables and her explanations for doing so are summarized in Table 2. It was of interest to investigate her claims for dropping these variables. To do so, I started with the data in `skulls1.txt`, fixed the column labels, kept these four additional variables in the data set, and applied the same quality control procedures described in Section 2.1. Because there were more variables, there were more opportunities for skulls to be removed due to missing values, question marks, etc. My final sample sizes for Series I, II, and IV were (70, 120, 59) respectively, versus the (88, 139, 64) obtained using only the seven variables. Most additional removals were due to missing values. Series II was not considered here because we lack the four auxiliary variables for these skulls.

Measurement	Reason Given for Exclusion
Ophryo-occipital Length	"variation on the standard glabello-occipital length"
Biauricular Breadth	"variation on the standard maximum breadth"
Basinasal Length	"definitely connected with the glabello-occipital length"
Bizygomatic Breadth	"most frequently missing through fracture, number of skulls available would have been very much decreased"

Table 2: Four variables that Barnard dropped from the data set initially and her reasoning for doing so.

The claim of extensive missingness in Bizygomatic Breadth was easy to investigate by counting the percentage of skulls missing this value in Series I, II, and IV. These percentages were 31%, 4%, and 10%, respectively. Though not as high as the value that Barnard suggests, removing 30% of the skulls in a series would significantly reduce the number of observations available. This decision therefore makes sense.

To investigate the relationships between Ophryo-occipital Length and Glabello-occipital Length, I made a simple jittered scatterplot (Figure 2a). With a strong visual relationship between the two variables and a correlation coefficient of $r = 0.97$, it does seem like including both variables in our analysis would be redundant and even generate multicollinearity issues.

Investigating the relationships between Basinasal Length and Glabello-occipital Length (Figure 2b) and Maximum Breadth and Glabello-occipital Length (Figure 2c), however, there were much weaker linear trends, with correlation coefficients of $r = 0.53$ and $r = 0.44$, respectively. Of course, r is only a measure of linear association, but there don't appear to be nonlinear relationships between these variables either. It's simply a noisy data cloud in both cases. Thus, while Barnard's claim that Ophryo-occipital Length is redundant after the inclusion of Glabello-occipital Length seems valid, making this claim for Basinasal Length and Biauricular Breadth seems unrealistic.

Looking at the correlation matrix for these 11 variables (calculated but not included), values of 0.53 and 0.44 are on the higher side, but there are certainly more correlated pairs of variables in this data set. Of course, Barnard didn't have the ability to instantly compute a correlation matrix for her skulls, but given that I do, the removal of these variables is difficult to support.

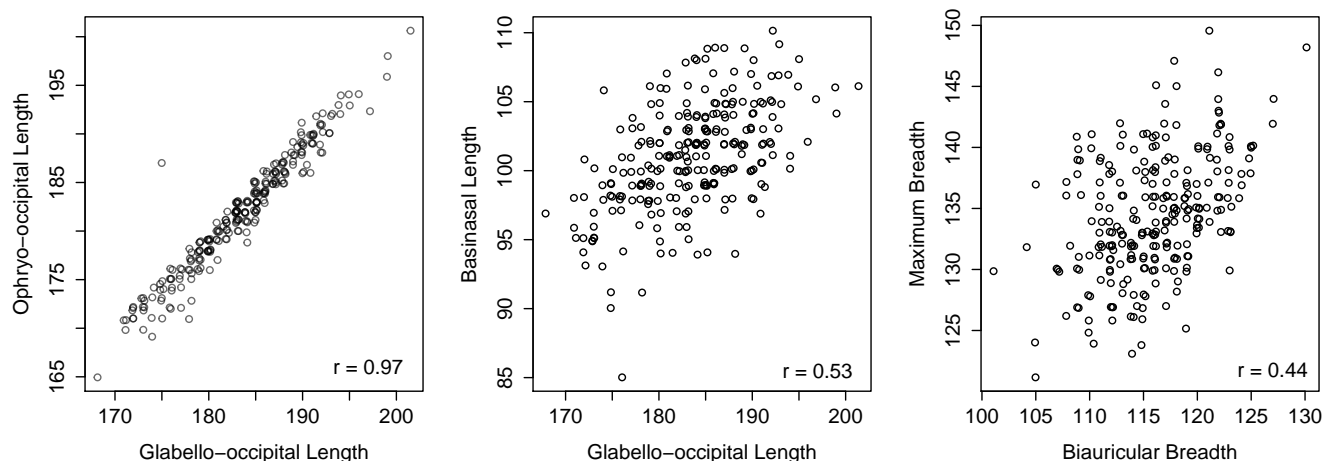


Figure 2: Investigating Barnard's removed variables. While Ophryo-occipital Length is strongly correlated with Glabello-occipital Length, there is no such strong relationship between the other two pairs of supposedly redundant variables.

3.2 Missing Data

As noted in Section 2, a fair number of skulls were missing one or more observations. Suffice to say, missing data is a tricky topic in statistical analysis. If data are missing completely at random, then removing observations with missingness shouldn't bias results. However, if the missing values are related to the value of the variable or to the value of other variables, then eliminating those observations can allow bias to creep into your analysis. There's an entire body of literature on missingness that will not be reviewed here.

For ease of analysis, I restricted my missing value investigation to the seven variables that Barnard used as features. I considered missingness from two different angles: the distribution of NAs across skulls within each series, and the distribution of NAs across all series within variables. Then I considered whether bias could stem from these removals.

The distribution of NA values for each series is given in Figure 3. There is an obvious discrepancy in the distribution of missingness the number of skulls with at least one missing value (and thus, the number eliminated from consideration) between Series. Series II and IV have less than 10% of skulls eliminated, while Series I and III have 29% and 50% of skulls eliminated due to missingness, respectively. Though there isn't much information about how these skulls were preserved or collected, I mentally hypothesized that more contemporary skulls would have fewer missing values. This doesn't seem to be the case.

More difficult to ascertain is whether the distribution of the number of missing values is approximately uniform or not. We really only have two series we can trust (Series I and III, due to their larger sample sizes), and while Series I suggests that a skulls with at least one missing value is more likely to have multiple missing values, Series III suggests that only having one or two missing values is more likely than have many missing values, given that a skull has at least one missing value.

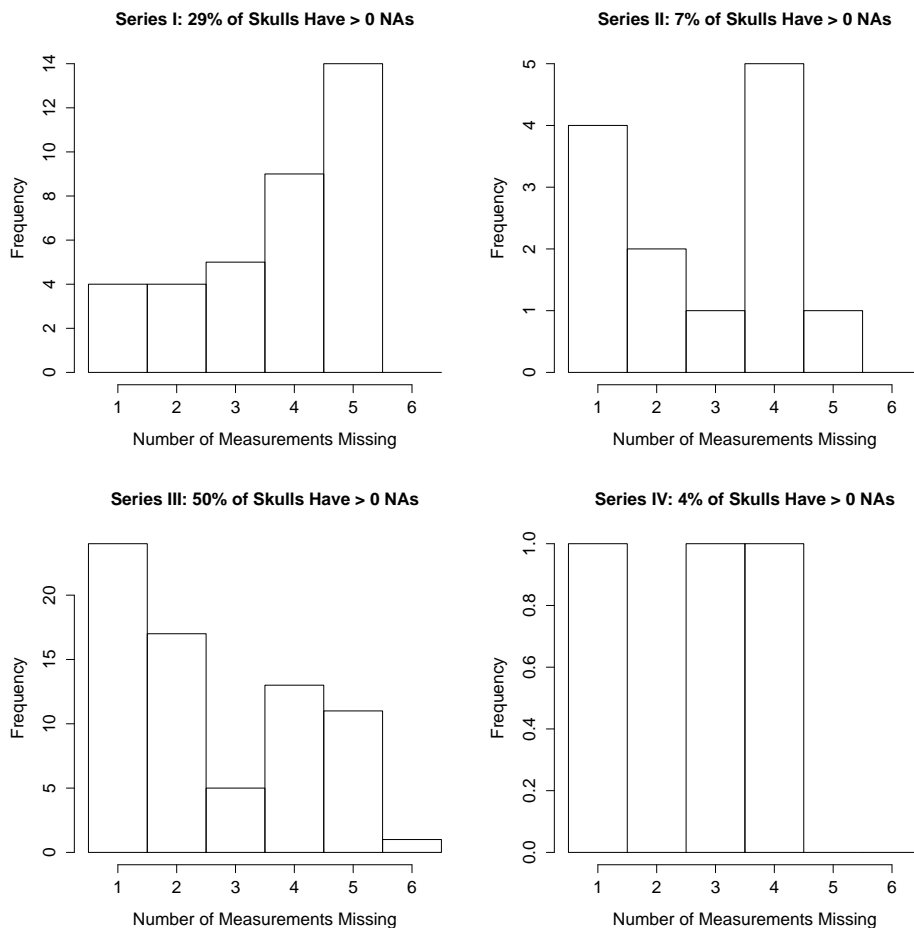


Figure 3: Distributions of the number of NA values for each series, with intact skulls excluded. Series I and III had far more missingness than Series II and IV, though their distributions ran counter to each other.

Considering the distribution of NAs across all series within variables, I combined all series and counted missing values for each variable. The resulting counts were surprisingly skewed (Table 3). At the extremes, only 4/517 skulls were missing a Glabella-occipital length measurement, while 85/517 were missing a Basia-aveolar length measurement. (A simple χ^2 test for equality across variables had a p-value less than $2.2e - 16$, but we don't even need a statistical test to notice the discrepancy.) Based on these data, I would guess that some portions of skulls are more fragile and prone to damage preventing measurement than others.

Variable	# Skulls Missing Value
Maximum breadth (B)	10
Nasal width (NB)	61
Glabello-occipital length (L)	4
Basialveolar length (GL)	85
Nasialveolar height (GH)	68
Nasal height (NH)	62
Basibregmatic height (H)	73

Table 3: Missingness counts across all series (517 skulls total) for each variable. These counts are highly skewed, with four variables in the 61–73 range but with extremes as low as 4 and as high as 85. I hypothesize that some portions of a human skull are more fragile and prone to damage than others.

To determine whether removing skulls with missing values could produce bias, I considered the distributions of variables in skulls with and without other variables. In light of Table 3.2, I examined the distributions of all variables in skulls with and without Basialveolar length. A few of these distributions can be seen in Figure 4. While the distributions between the two groups of skulls do appear to differ marginally, the bias introduced by removing skulls missing GL shouldn't be huge. Thus, while some series have much greater missingness than others, removing those observations with missing values is certainly the easiest solution, and it is probably a reasonable solution as well.

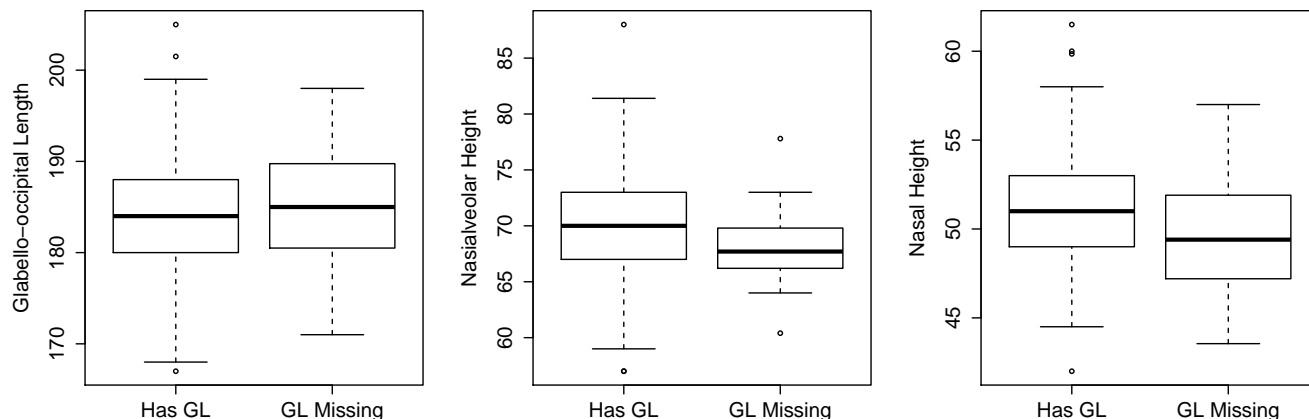


Figure 4: Variable distributions in skulls with ($n = 432$) and without ($n = 85$) Glabello-occipital Length (GL). Though distributions differ slightly between the two groups, removing skulls without GL measurements shouldn't greatly bias the analysis.

4 Barnard vs. LDA/FDA

When linear discriminant analysis is approached from the perspective of trying to maximize the class posteriors under the assumption of a Normal distribution for each class with a common covariance matrix, it's difficult to see analogues between LDA and Barnard's approach. However, when the FDA framework is considered, under which Fisher tried to determine the linear transformation of the data that would push the class means as far away as possible while shrinking the variability within classes as much as possible, it's possible to draw parallels between these linear discriminant functions and Barnard's methods.

Much of Barnard's discriminatory work takes places in the later parts of her paper, but in Section 2, she uses t-tests and regression methods to select the two features with the greatest statistical difference between Series I and IV skulls. This is a simpler technique than LDA/FDA, which base decisions on linear combinations of the predictors, though computational restrictions make considering the predictors individually quite attractive. The work that she does in the second half of her paper is fairly similar to FDA, though FDA has built-in variable selection via weighting more important variables more heavily in the canonical variates. Barnard employed more of a two-step process, first selecting variables to be considered and then choosing appropriate weights for those variables.

5 Classification

5.1 Overview

Much of the work done at the turn of the century by groups such as the Galton Laboratory concerned inference: estimating parameter values for various groups and evaluating statistical significance. Barnard works in this vein for the first part of her paper, though she does consider a function that would discriminate between Series I and IV in the second part of her paper. Though prescient by the standards of her days, recent years have seen the field of statistics become increasingly focused on classification problems. With almost 100 years of technology, statistical advances, and machine learning algorithms on my side, I classified the period of skulls based on the seven variables retained by Barnard in her analysis using three classification methods: linear discriminant analysis (LDA), Fisher's discriminant analysis (FDA), and support vector machines (SVM).

5.2 Protocols

I used the `lda()` function from the R `MASS` library for LDA and FDA, and the `svm()` function from the R `e1071` library for SVM.

5.2.1 Cross-Validation

When performing classification analysis, classifiers are usually compared on the basis of their estimated prediction errors. The easiest way to estimate prediction error for a classifier is to use cross-validation. In K -fold cross-validation, the data are randomly partitioned into K approximately equally-sized groups. For each partition k ($k \in 1, \dots, K$), all data except partition k are used to train the classifier. Then the observations in partition k , which the classifier didn't get to see, are classified. This is performed for each partition, and the errors are combined. By separating the points used to build the classifier and the points being classified, more correct estimates of prediction error are obtained [5].

Cross-validation should actually be used twice in SVM: first, to choose the optimal tuning parameters, and second, to estimate the prediction error. The `tune.svm()` function, also in the `e1071` package, performs 10-fold cross-validated parameter selection. I used this function to perform three increasingly fine grid searches for each classification scenario, and used the best parameter values from the finest grid search as the final parameters used for classification. I then used cross-validation to estimate prediction error using the `cross` argument to the `svm` function. All SVMs performed used the radial kernel.

The `lda()` function has built-in leave-one-out cross-validation ($K = n$), but general K -fold cross-validation must be done manually. I wrote code that partitions the data into equally-sized groups and structures the partitions in a way that lends itself nicely to the `subset` argument of `lda()`. I used 10-fold

cross-validation for both LDA and FDA, so that the estimated prediction errors would be comparable to those obtained via SVM.

5.2.2 Multiclass Problems

LDA and FDA scale well to classification problems with more than two classes. SVM, however, is really a binary classifier. To get around this issue, various multiclass SVM strategies have been proposed. Many of them fall into either a One Versus All or One Versus One framework [6].

Given a classification problem with K classes, a One Versus All strategy constructs K SVMs, where classifier k ($k \in 1, \dots, K$) uses class k as the positive class and all other $K - 1$ classes are pooled into the negative class. A new observation is then assigned to the class whose classifier assigns it the largest discriminant value. In contrast, in a One Versus One framework, $\binom{K}{2}$ classifiers are built, one for each pair of classes. A new observation is analyzed by every classifier, and the results are used to ‘vote’ for that observation’s best class. Ties are often broken randomly.

By default, `svm()` uses One Versus One when faced with a multiclass problem. This is the option I therefore used when classifying skulls from all four series. For comparison purposes, I also performed One Versus One LDA in addition to standard multiclass LDA when classifying all series.

5.3 Series I and IV

Using the protocol described in Section 5.2, I classified the time periods of the 152 high-quality Series I and IV skulls using Barnard’s seven retained variables as features. The confusion matrices for LDA and SVM are given in Table 4. Looking across rows, LDA classified more than half of each series correct. SVM, on the other hand, was quite aggressive in classifying skulls as belonging to Series I, resulting in a much better Series I accuracy but a decreased Series IV accuracy. Overall, SVM had more correct classifications than LDA, but the computational costs of the three-resolution cross-validated parameter grid search exceed the meager costs of LDA by orders of magnitude. LDA is also more interpretable, understandable, and consistent than SVM.

LDA			SVM		
Truth \ Pred.	I	IV	Truth \ Pred.	I	IV
I	65	23	I	78	10
IV	29	35	IV	35	29

Table 4: Confusion matrices for Series I and IV 10-fold cross-validated classification using LDA and SVM. Though SVM made more correct classifications overall, LDA classified Series IV skulls more accurately.

For FDA, I used the procedure given in [5] to project the data onto the optimal subspaces for maximizing between-group variance and minimizing within-group variance. This projection onto the first two canonical variates is given in Figure 5, with the centroids for each group plotted as well. We can see that FDA has achieved exactly what it set out to achieve: the data were projected into the directions that maximize the signal to noise ratio for classification. In this case, only the first canonical variate is informative: the centroids are well-separated in terms of Coordinate 1 but almost completely aligned in terms of Coordinate 2. This can also be seen in the magnitude of the eigenvalues (computed but not given here): λ_1 was the only non-negligible eigenvalue. After finding the optimal separating line in Figure 5, a new point would be projected and classified based on which side of the line it fell on.

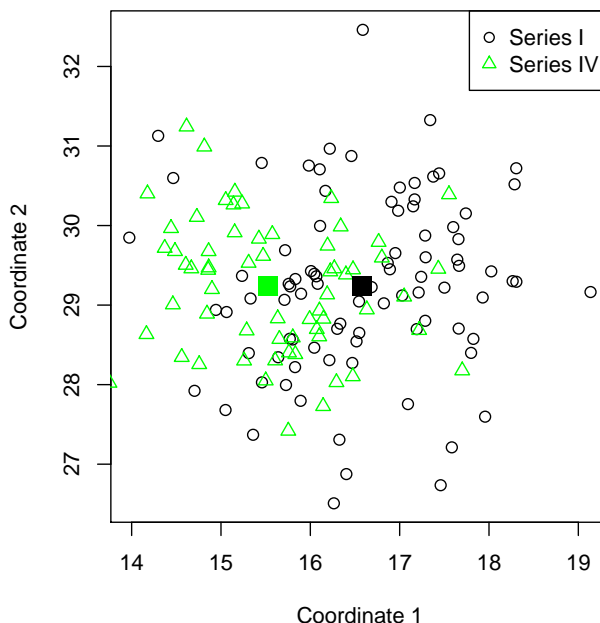


Figure 5: Projections of Series I and IV skulls onto canonical variates from FDA. Only Coordinate 1 produces meaningful separation in this dimension reduction, as seen in the centroid separation.

In Section 2 of her analysis, Barnard finds two variables that differ significantly between Series I and IV skulls and are uncorrelated with each other. A simple t test was used for the first variable; x_4 was chosen as most significant. Then a regression was performed to remove the effects of x_4 and another t test was performed; x_6 was selected in this iteration. To compare these results to my results from LDA/FDA, I ran LDA on all skulls in Series I and IV (no cross-validation). Though my data are slightly different than hers due to quality control choices, I still note that x_4 has the largest coefficient in magnitude of linear discriminant coefficients (Table 5).

Variable	x1:B	x2:NB	x3:L	x4:GL	x5:GH	x6:NH	x7:H
LDA Coefficient	0.06	0.141	-0.032	-0.16	0.05	0.098	-0.113

Table 5: Linear discriminant coefficients from a full LDA of Series I and IV skulls.

5.4 All Series

Using the protocols described previously, I also classified the time periods of the 361 skulls summarized in Table 1 based on Barnard’s seven variables. This is a multiclass problem, and I performed standard multiclass LDA, One Versus One LDA, multiclass FDA, and One Versus One SVM. The confusion matrices for both LDAs and SVM are given in Table 6.

Multiclass LDA did at least as well as One Versus LDA did for all four series. SVM outperformed both LDAs significantly, more than doubling the number of Series I and III skulls classified correctly. As in the two-class problem, the computational costs of SVM far exceeded those of LDA, but here in the harder four-class problem (recall Figure 7), those costs are probably worth paying for the exceptional improvements obtained. In terms of most common mistakes made, all classifiers mislabeled many Series I and III skulls as Series II. We have many more skulls in Series II than we do the others (88, 139, 70, and 64, respectively), so perhaps the algorithms are weighted to preferentially assign skulls to Series II.

Multiclass LDA					One Versus One LDA					One Versus One SVM				
Truth \ Pred.	I	II	III	IV	Truth \ Pred.	I	II	III	IV	Truth \ Pred.	I	II	III	IV
I	26	40	12	10	I	22	43	10	13	I	47	30	5	6
II	12	101	12	14	II	17	97	11	14	II	14	110	5	10
III	16	37	11	6	III	13	39	11	7	III	12	26	27	5
IV	15	23	5	21	IV	15	25	4	20	IV	13	16	3	32

Table 6: Confusion matrices for the classification of all skulls using two types of LDA and SVM. LDA results were comparable, with multiclass LDA slightly outperforming One Versus One. SVM blew both LDAs out of the water, vastly improving classification for all series.

FDA in a four-class problem is very similar to the two-class FDA described previously, although in this case, the top 3 eigenvalues were non-negligible, with the first about twice as large as the second and third (computed but not included here). Projections onto the two combinations of the first three canonical variates and centroids are given in Figure 6. Again, the first coordinate does the best job of separating the groups, though the separation is admittedly less than impressive. Coordinate 1 seems best able to separate Series III and IV (green and blue), while Coordinate 2 separates Series I and IV (black and blue) and Coordinate 3 separates Series II (red) from all other series.

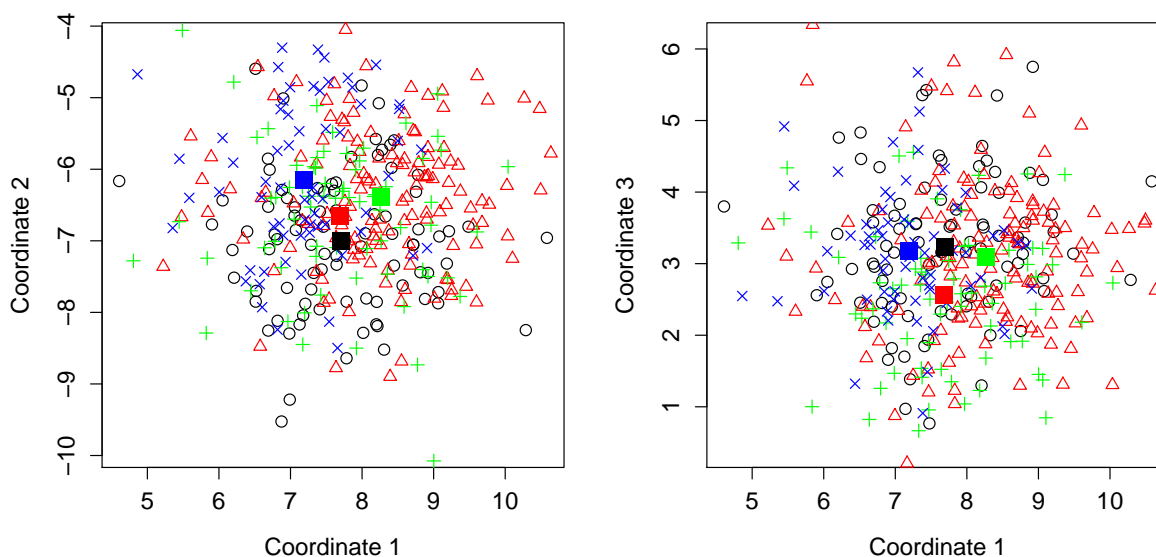


Figure 6: FDA projections for all series. The top three coordinates contribute to the separation of skulls from different series, though their contributions decrease with increasing coordinate number.

In light of the four-group classifications described here, Barnard’s methods don’t seem very robust, in the sense that performing her variable selection steps wouldn’t scale well to more than two classes. In a two-class case, performing t-tests is logical, and because there are only two groups being compared, the number of tests is $O(p)$, i.e. linear in the number of variables under consideration. In a four-class case, however, the number of pairwise t-tests would grow as $O(p^2)$, which could definitely require multiple testing considerations. (For this problem, there would be $\binom{4}{2} * 7 = 42$ tests to conduct.) It’s also not clear how the top variables would even be chosen, given pairwise t-tests; some variables might be significant for some pairs and insignificant for others.

Furthermore, by focusing only on Series I and IV, Barnard overlooked Series II, which was both the largest and the series responsible for many of the misclassifications that occurred in the multiclass classifications (Table 6). I understand that, from an evolutionary perspective, she maximized her chances of seeing

significant variation between series by considering the two series most separated in time, but it still overlooks the majority of points in the dataset. Then again, the lack of computational power available in there 1930s makes the fact that she could do any of this impressive in its own right.

References

- [1] J. S. Jones. The Galton Laboratory Today. http://www.galtoninstitute.org.uk/Newsletters/GINL9112/Galton_Laboratory_Today.htm. [Online; accessed Apr 8 2011].
- [2] D. Salsburg. *The Lady Tasting Tea*. Henry Holt and Company, 2001.
- [3] M. M. Barnard. The secular variations of skull characters in four series of Egyptian skulls. *Annals of Eugenics*, 6(4), 1935.
- [4] A. Thomas and R. Randall-Maciver. *Ancient Races of the Thebaid*. Oxford, 1905.
- [5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.
- [6] K. Duan and S. S. Keerthi. Which is the best multiclass SVM method? An empirical study. Technical report, Proceedings of the Sixth International Workshop on Multiple Classifier Systems, 2005.

A Scatterplot Matrices

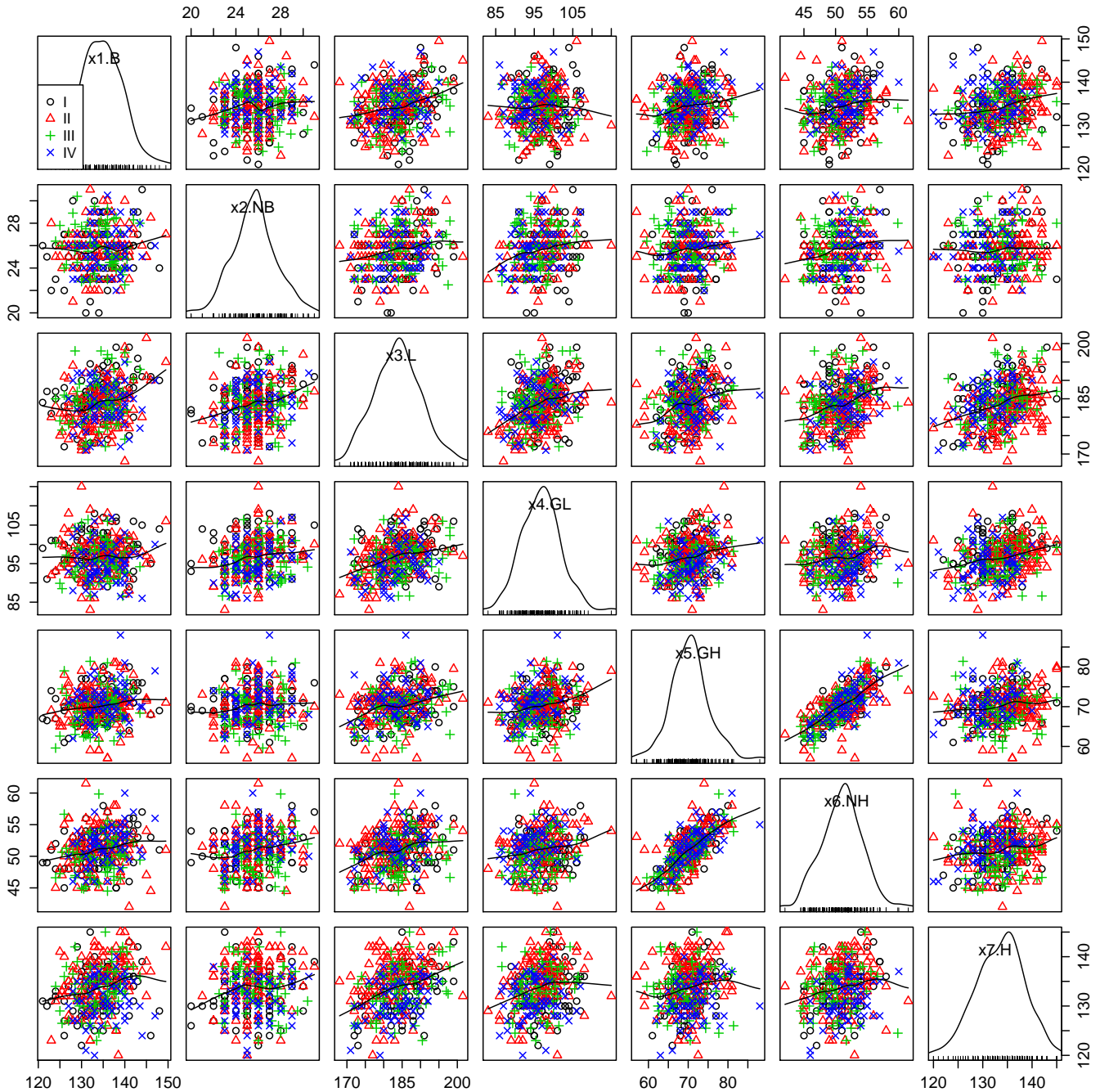


Figure 7: Bivariate scatterplot matrix for all skull series. Though this is only the bivariate perspective, it suggests that distinguishing between the four series could be a difficult classification problem.

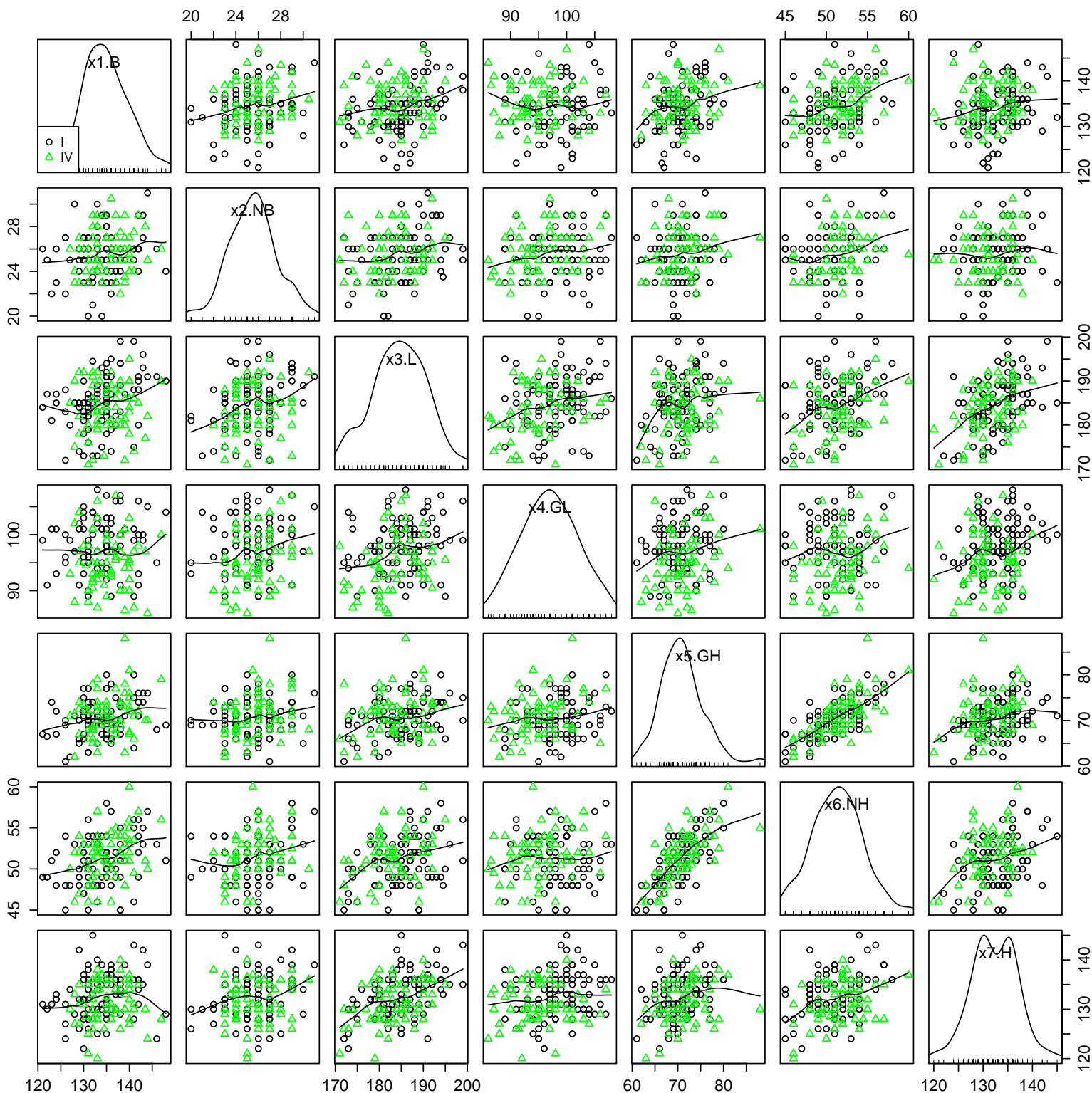


Figure 8: Bivariate scatterplot matrix for Series I and IV only. Though there are pockets of separation between the two groups (including the nice bimodal peak and left/right separation for X7:H), most bivariate relationships are cluttered and suggest a challenging classification situation lies ahead.