A Pattern Recognition Approach to Identifying Different "Voices" in the Dao De Jing

Bruce R. Linnell, PhD

2012

Abstract

Analytical pattern recognition techniques are applied to the question of whether or not there are multiple "voices" in the Dao De Jing, and whether a 3-voice or 4-voice model fits better. The results are amazingly consistent independent of the chapters used as examples of their voice, or even the number of voices.

Introduction

While doing research on the Dao De Jing (DDJ), I discovered John Emerson's articles on the possibility of various "layers" in the DDJ, based on the presence or absence of certain symbols and "themes" in each chapter[2,3]. Of particular interest was his comment, "Many of the objections to my method seemed to be objections to the kinds of rough, empirical, non-algorithmic methods best used to disentangle historically confused material."[2] Having an extensive background in pattern recognition, I decided to find out whether or not each chapter could be classified as belonging to these layers based on purely analytical methods.

Pattern recognition classifies un-labeled examples ("unknowns") as belonging to one of a number of classes which are defined by some a-priori labeled examples (collectively called the "training set"). Every example (labeled or un-labeled) has a number of "features" that describe it (for example : height, weight, hair color, skin color, and eye color are features that describe a person's appearance). There are many kinds of classifiers available, but one of the best ones to use when the number of examples is small compared to the number of features[4] is the "Linear" classifier[1]. A Linear classifier draws a straight line between the classes, based on the "shape" of the classes, assuming a Gaussian distribution (the ovals) of the training sets (see Figure 1). Any unknowns on one side of the line are assigned to one class, unknowns on the other side are assigned to the other class. Another classifier that was investigated was the "K-nearest-neighbor" (KNN)[1]. This classifier looks at the K nearest labeled examples to the unknown, and assigns it the same class as the majority of its neighbors (see Figure 2). This classifier has the advantage of being able to separate non-Gaussian data fairly well.

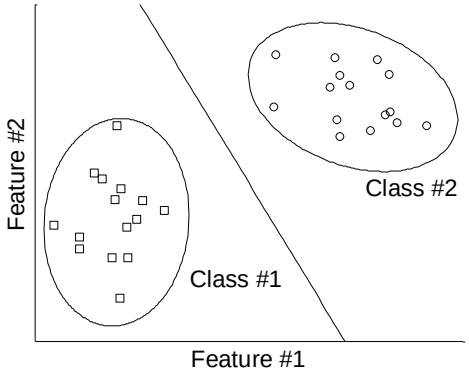

Figure 1 – Linear classifier

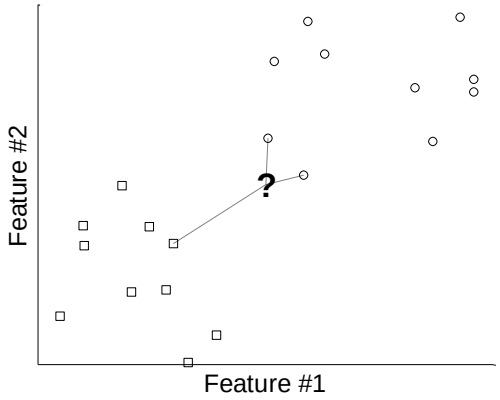

Figure 2 – KNN classifier (K=3)

The "leave-one-out" (LOO) algorithm[1] was used extensively in this research. When trying to estimate how well a classifier built from a particular training set will perform on future examples, the training set must be split into two non-overlapping parts – one set of examples to create a classifier, and one set of examples to run through it. Since the correct classes are known for all the examples run through the classifier, the performance of the training set on future examples can then be estimated. There are numerous estimation methods, but LOO is simple, reasonably accurate, relatively quick, and non-stochastic. Each example in the training set is removed one at a time, and all the remaining examples are used to create the classifier. The removed example is then run through the classifier. The estimated success rate of the entire training set is then the average number of examples that were correctly classified. At other points in this research described below, this leave-one-out concept is generalized to : given a set of examples, each example is removed from the set one at a time, and

then the rest of the set is processed in some way. This produces a list of results with one entry for each chapter in the set.

In his original analysis[2], Emerson identified Early, Middle, Late, and Added layers. After the Guodian texts were discovered, which do not have chapters 67-81, he added a Post-Guodian ("PostGuo") layer in a subsequent analysis[3]. The Added layer was very small, consisting of just one chapter and a handful of "sections" (described in the next paragraph) from other chapters, so it was ignored for this research. Emerson wasn't completely sure that the Middle layer was real, but found enough evidence to tentatively support its existence. So another goal of this research became to determine whether or not the Middle layer exists. To that end, classification was attempted both with and without the Middle layer, to see which performed better.

Emerson also discussed at length the concept that some chapters contained multiple "sections", with different sections within a chapter belonging to different layers. For example, he says that all sections starting with the phrase "Therefore the Sage…" are Late. While splitting chapters into smaller sections might make for a more accurate classification of each section compared to the whole chapter, this was not attempted here, for two reasons. First, the boundaries of sections are more subjective (such as sometimes having to decide when a "Thus the Sage…" or "chain" passage ends), whereas chapters are well-defined. Second, the resulting sections in many cases are very short, which would make it harder to extract features from them (described below).

While Emerson named the layers according to the order in which he thought they were added to the DDJ, this research is only interested in identifying how many different layers there are, and which chapters belong to them. So from this point on the term "voices" will be used instead of Emerson's term "layers", and the names (Early, Middle, Late, PostGuo) are kept only for convenience.

Methodology

Many factors influence the final voice assignments of the chapters. The first is which version of the DDJ to use. The Wang Bi version was selected mainly because its earliest manuscripts are complete, and the Chinese characters are "modern" enough that there is less doubt as to their meaning. However, some characters do differ between original manuscripts, so the results found here might vary a little depending on the specific Wang Bi source used.

The next choice is the selection of chapters that represent the Early, Late, PostGuo, and possibly Middle voices for the training set. An analysis of Emerson's analysis identified the following chapters as being "pure" representatives of their voice, without any sections that might contain other voices :

Early : [1 4 5 6 10 15 20 31 32 35 37 50 51 56]
Middle : [8 9 11 24 26 36 38-49]
Late : [3 12 17 18 19 53 57 58 60-66]
PostGuo : [67-81]

It should be noted that using a different initial training set may change the results below.

Next, there are the features extracted from the chapters. There are three features for the 3-voice classifier, and four features for the 4-voice classifier. Each feature represents how much in common the example (i.e., chapter) has with the representative chapters in the training set for each voice. For example, a 3-voice feature list of [0.05 0.11 0.84] would mean that this chapter has roughly 5% in common with Early, 11% in common with Late, and 84% in common with PostGuo. As one of Emerson's criteria for labeling chapters was whether or not certain symbols appeared in them, this feature list is built by combining three different metrics, each of which looks at the distribution of symbols within and between chapters in a different manner :

1. **Often used symbols that are in all voices ("InCommon")**
   This metric starts by finding all symbols that are common to all chapters in the training set, independent of voice. Then for each symbol in this list, it determines the number of chapters in each voice that use

the symbol. For example, a value of [10 3 2] means that this symbol appears 10 times in all the Early chapters of the training set, 3 times in all the Late chapters, and 2 times in the PostGuo chapters. One option that was explored here was whether to count each instance of the symbol in the chapter, or just count the symbol once for the chapter no matter how many times it actually appeared.

Then the ratio of the largest to the smallest value in the list is calculated (using the previous example : 10/2=5), so that larger ratios indicate a more non-uniform distribution of the use of this symbol between the voices. If this ratio is above a certain value ("THRESH1"), the counts (normalized by the number of symbols in each voice in the training set) are added to the running total for every chapter (not just training set chapters) that contains that symbol. The final totals for every chapter are then normalized so that they sum to unity.

2. **Rare symbols ("Rare")**
   This metric starts by finding all symbols that occur between two and some upper limit ("THRESH2") times in the entire document. As for InCommon, a count of the number of chapters that use the symbol is generated for each voice. However, since the minimum value in this list is often zero, a ratio is not calculated. Instead, if the maximum value in the list is above a certain threshold ("THRESH3"), the normalized counts are again added to the running total for each chapter that contains that symbol. The final totals are then normalized so that the sum for each chapter equals unity.

3. **Symbols common to all but one voice ("AllButOne")**
   For 3 voices, this metric starts by finding all symbols in the training set that are either (in Early & PostGuo but not in Late) or (in Early & Late but not in PostGuo) or (in Late & PostGuo but not in Early). A similar pattern is followed for 4 voices, using all combinations of 3 voices at a time. For each of these lists, processing proceeds exactly as for InCommon, using a different threshold for inclusion ("THRESH4").

These three separate "frequencies" are added together, and the result is again normalized so that the sum of all features is unity. While they are not probabilities in the strictest sense, they do generally represent the likelihoods that the chapter belongs to each voice. Using different metrics would most likely significantly change the results found below. Using all 9 or 12 features separately would drop the ratio of examples (labeled chapters) to features to an unacceptably low value. It is well-known in pattern recognition that many problems arise when this ratio becomes too small[4].

One perhaps obvious metric would be to use symbols that are unique to each voice as a measure for unknown chapters (indeed, this is one criteria Emerson used in his layer assignments). The problem with this metric is that it by definition always gives 100% correct predictions for the training set (which will be seen below to be a disadvantage), and testing showed that if an un-labeled chapter has even just a single symbol in it that is unique to only one voice, it would be strongly rated (100%) as belonging to that voice, which often conflicted with the other metrics. There are also many chapters (both training and unknown) which have none of the unique symbols from any voice, and so look like they belong to no voice because their features are [0 0 0] or [0 0 0 0]. For all these reasons, this metric was not used (but the concept will be revisited later).

Procedure

Each of the three metrics has one or more constants in them (the "thresholds" described above). The first step was to find the optimal values for the constants. "Optimal" in this sense means that the resulting features (for that metric only) for each chapter in the above training set had a maximum value in the feature list corresponding to its labeled voice. For example, a result of [0.15 0.20 0.65] would be considered better for a chapter labeled as voice #3 than would [0.65 0.15 0.20]. The constants were varied over a wide range of values, and those values that gave the best results determined the constant's final value.

For the *InCommon* metric with 3 voices, there was a range of optimality from THRESH1=3.5-3.7, so the value 3.6 was used. For 4 voices, there was a range of optimality from THRESH1=4.7-5.1, so the value 4.9 was used. For the *Rare* metric with either 3 or 4 voices, there was a wide plateau of optimality from THRESH2=9-13 and THRESH3=0.1-1.3, so the values THRESH2=11 and THRESH3=0.7 were used. For *AllButOne* with 3 or 4 voices, any value of THRESH4 ≥ 0 worked well, so a value of THRESH4=0 was used. In all cases, counting multiple occurrences of a symbol per chapter worked significantly better than counting the symbol only once. Changing any threshold by a large amount could significantly change the final results.

The next step was to reduce the number of chapters in the training set. This is because there is no guarantee that the chapters picked from Emerson's analysis would be the optimal examples of their voice as rated by the metrics described above, so inappropriate ones should be removed. The approach used was to find the largest subset of chapters for each voice that produced 1)the highest estimated classification success rate on the training set, 2)the most "consistent" results, and 3)the most "compact" results. Changing what constitutes a "good" training set may significantly change the final results. The estimated classification success rate of the training set ("Success") was calculated using LOO as described above.

"Consistent" here means that the classifications of chapters did not change as each example was left out of the training set. To calculate the consistency, each example is left out of the training set and then every chapter (labeled or un-labeled) is classified using that reduced training set. From that is calculated how often every chapter was classified as belonging to voice #1, #2, #3, or #4. For the training chapters, the consistency is the fraction of the time that chapter is classified as its labeled voice. For the unknown chapters, the consistency is the fraction of the time that chapter is classified as the voice it is assigned to by the Linear classifier using the entire training set. This gives two lists, a list of the consistencies of the training set, and a list of the consistencies of the unknowns. From each list, two values are calculated : the mean consistency ("mean") and the minimum consistency ("min"), and it is desirable that they both be *maximized*. To that end, two metrics are calculated to determine the quality of the training set :

$Consist_1 = (mean^2 + min^2)^{1/2}$ based on all labeled examples
$Consist_2 = (mean^2 + min^2)^{1/2}$ based on all un-labeled examples

Combining the values in this manner means that if one or both values increase then the overall metric increases, but if one value increases at the expense of another, the metric only increases if the increase of one value is greater than the decrease of the other. Including the minimum value in the formula emphasizes the significance of outliers and helps to reduce them.

It was determined that it was not enough to only make the training set as consistent as possible by maximizing $Consist_1$, because that often resulted in highly inconsistent classifications of the unknowns (in other words, how they were labeled was very sensitive to the training set). Thus a final measure of consistency was calculated, using the same logic as above :

$Consist_3 = (Consist_1^2 + Consist_2^2)^{1/2}$

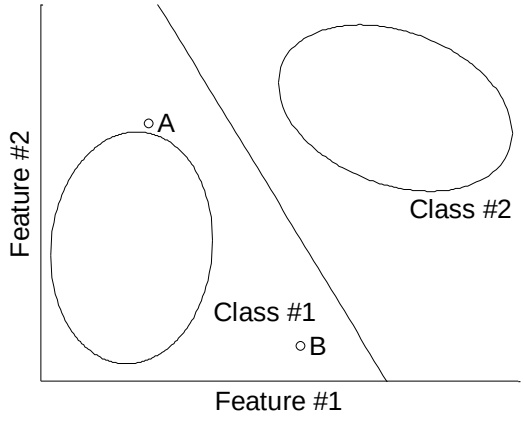

Figure 3

"Compactness" is measured as the average distance of the examples to their class mean. While the Linear classifier is robust and simple to implement, it does not take into account how far the unknown is from the examples in the training set (see Figure 3). While unknown "A" probably belongs to class #1, unknown "B" obviously does not, even though it is on the class #1 side of the line. To ensure that the unknown classes are also "close" to their assigned classes, the distance of each example to its class mean is calculated, using the Mahalanobis distance.

The Mahalanobis distance[1] differs from the Euclidean distance in that it takes into account the *shape* of the training set in its multi-dimensional space (see Figure 4). While examples A and B are the same Euclidean distance from the center of the class, points A and C are the same Mahalanobis distance from the center.

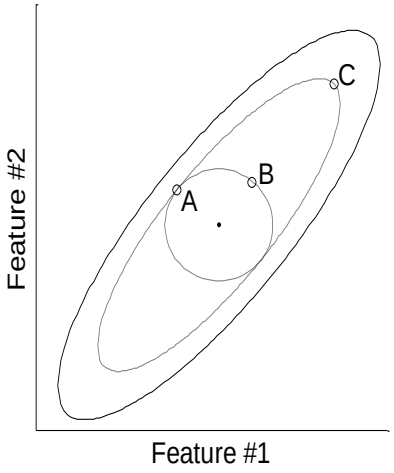

Figure 4

To calculate the compactness, the Mahalanobis distance from each example (labeled and unknown) to its class mean (as determined by the training set) is calculated. This also gives two lists, a list of the distances of the training set, and a list of the distances of the unknowns, both of which should be *minimized*.

As for consistency, two metrics are calculated to determine the quality of the training set :

$Maha_1 = (mean^2 + max^2)^{1/2}$ based on all labeled examples
$Maha_2 = (mean^2 + max^2)^{1/2}$ based on all un-labeled examples

And again a final measure of compactness was calculated :

$Maha_3 = (Maha_1^2 + Maha_2^2)^{1/2}$

Using the initial "pure" training set described above resulted in the following measures of quality :

3 voices : Success=95% $Consist_3$=1.43 $Maha_3$=20.1
4 voices : Success=92% $Consist_3$=1.37 $Maha_3$=19.2

Emerson classified chapter 8 as Middle but said, "There are many reasons to put Ch. 8 in the late layer: the water theme, non-contention, and the frequent use of the word *shan* 'good/expert'. But within my rules I could not justify doing so."[2] However, preliminary testing indicated that chapter 8 seemed to be more Late than Middle, and all of the metrics except $Maha_3$ are improved by putting it in the Late voice in the training set :

3 voices, 8=Late : Success=98% $Consist_3$=1.63 $Maha_3$=18.6
4 voices, 8=Late : Success=95% $Consist_3$=1.41 $Maha_3$=24.0

Thus 8 was moved from Middle to Late. This results in the final "pure" initial training set :

Early : [1 4 5 6 10 15 20 31 32 35 37 50 51 56]
Middle : [9 11 24 26 36 38-49]
Late : [3 8 12 17 18 19 53 57 58 60-66]
PostGuo : [67-81]

Leaving chapter 8 in the Middle voice for training (or leaving it out of the training set altogether, since there is some question as to where it belongs) might significantly change the final results.

The method used to reduce the training set also involved using a LOO approach. Each chapter in the current training set was left out to create a "reduced" training set, then the three measure-of-quality metrics described above were calculated. In general, the new reduced training set that had the largest Success, largest $Consist_3$, and smallest $Maha_3$ was chosen to be the new best training set. This process was repeated until removing every individual chapter resulted in a worse new training set than the current one.

It should be noted that this is a local search, not a global one. This means there could easily be local minima that prevent finding better solutions. For example, while removing any single chapter from the current training set might not result in any improvement in consistency, removing two chapters might. To help avoid local minima, another program was written which leaves out all possible combinations of two chapters at a time ("LOO^2"). However, as it took approximately 40 times longer to run, it was only used as needed. In addition, after every two chapters were removed, a program was run that added <u>all</u> the removed chapters back in to the

training set one at a time, to see if (for example) any chapter that had been removed earlier might improve the quality of the training set by being added back in now. Note that when using this program, a chapter would only be added back into a layer that Emerson thought it belonged to. In combination with the LOO and LOO^2 pruning, this is known as a "backward/forward" search, which also helps to avoid local minima.[4]

The pruning process for the 3-voice "pure" training set produced two "optimal" reduced training sets : removing [1 4 50 62 63] from the training set resulted in the best consistency of all training sets tested ($Consist_3$=1.87, $Maha_3$=12.2), while removing [1 4 10 50 58 60 63 65] resulted in a significantly smaller overall Mahalanobis distance ($Consist_3$=1.78, $Maha_3$=10.8). Note that the removed chapters have [1 4 50 63] in common, indicating these are "key" chapters that need to be removed in order to improve the training set. This does not necessarily mean that these chapters do not belong to the voice they were in, just that they may not be the best examples of their voice.

Two "optimal" training sets were also found for the 4-voice "pure" training set. Removing [9 36 38 45 49 50 57 58] from the original training set resulted in the best consistency ($Consist_3$=1.84, $Maha_3$=12.8), while removing [9 36 37 45 49 50 57 58 63] resulted in the best compactness ($Consist_3$=1.75, $Maha_3$=12.0). Both sets have [9 36 45 49 50 57 58] in common, indicating these are "key" chapters that need to be removed in order to improve the training set.

All four of these reduced training sets are the result of many runs of the pruning algorithm, using LOO, add-one-back, and LOO^2 to explore multiple possible best subsets. These were the best of all training sets discovered, but due to the local nature of the searches, it is not possible to guarantee that they are the very best ones. Note that chapter 8 was never once suggested for removal at any point during the chapter-pruning process, which would have indicated that it was ill-placed in the Late voice.

The chapters (training or unlabeled) that were consistently classified the same across all four reduced training sets and that agreed with Emerson's classification were then used to create another initial training set, called "consistent" :

Early : [1 4 5 6 9 10 15 16 20 21 23 28 31 32 35 37 51 52 56]
Middle : [11 24 25 26 39 40 41 42 43 44 45 47 48 50]
Late : [3 8 12 17 18 19 49 53 54 57 60 61 63 64 65 66]
PostGuo : [33 36 67:81]

This was then put through the same process to optimize the three metrics. This resulted in a single reduced 3-voice training set and two slightly different 4-voice training sets.

Another initial training set was created by placing every chapter in its *majority* voice according to Emerson, even if he thought it had sections from other voices. Only a few chapters which Emerson thought were equally mixed were left out of the initial training set, resulting in a "maximal" initial training set :

Early : [1 4 5 6 7 10 13 14 15 16 20 21 25 28 30 31 32 34 35 37 50 51 52 55 56]
Middle : [9 11 24 26 29 36 38 39 40 41 42 43 44 45 46 47 48 49]
Late : [3 12 17 18 19 27 53 57 58 59 60 61 62 63 64 65 66]
PostGuo : [67-81]

Starting from this initial training set resulted in five different 3-voice reduced training sets (three that were closely related to each other and another two that were closely related to each other) and three different 4-voice reduced training sets (two of which were closely related).

3-Voice Results

Both the Linear and KNN classifiers were tried (using 1, 3, and 5 neighbors for KNN). The Linear classifier always chose the voice with the highest consistency, whereas the KNN classifiers sometimes did not, and also sometimes disagreed with each other, so only the Linear classifier was used to classify the un-labeled chapters in the training sets.

A “voice” is defined by the symbols in its chapters in the training set. As the training sets vary, the symbols in each voice change and so the classification of the unlabeled chapters may change as well. Due to the large number of reduced training sets and the variability of how they classified the unlabeled chapters, they were all compared to each other and the majority result was taken. The results for 3 voices are shown in Table 1, where in the “Final” column :

[X] = always in training set
X = all or all-but-one agree
(X) = all but two agree
{X} = less agreement, but some basis for decision
* = classified as PostGuo, but exists in the Guodian manuscript

	Pure Consist			Maximal						
Ch	#1	#2	#1	1a	1b	1c	2a	2b	Emerson	Final
1	E	E	E	E	E	E	E	E	E	E
2	P*	P*	E	P*	P*	P*	L	L	E/M/L	???
3	L	L	L	L	L	L	L	L	L	[L]
4	E	E	E	E	E	E	E	E	E	E
5	E	E	E	E	E	E	E	E	E	E
6	E	E	E	E	E	E	E	E	E	[E]
7	P	P	E	E	E	E	E	E	E/L	(E)
8	L	L	L	E	E	E	L	L	M/L	{L}
9	E	E	E	E	E	E	E	E	M	E
10	E	E	E	E	E	E	E	E	E	E
11	P	P	P	P	P	P	P	P	M	P
12	L	L	L	L	L	L	L	L	L	[L]
13	E	P*	E	E	E	E	E	E	E	E
14	E	P	E	E	E	E	E	E	E	E
15	E	E	E	E	E	E	E	E	E	E
16	E	E	E	E	E	E	E	E	E	E
17	L	L	L	L	L	L	L	L	L	[L]
18	L	L	L	L	L	L	L	L	L	[L]
19	L	L	L	L	L	L	L	L	L	[L]
20	E	E	E	E	E	E	E	E	E	[E]
21	E	E	E	E	E	E	E	E	E	E
22	L	L	L	E	E	E	L	L	M/L	{L}
23	E	E	E	E	E	E	E	E	E/M	E
24	P	P	E	E	E	E	P	L	M	???
25	E	E	E	E	E	E	E	E	E	E
26	P	P	P	P	P	P	L	P	M	P
27	P	L	P	L	L	L	L	L	M/L	(L)
28	E	E	E	E	E	E	E	E	E	E
29	L	L	L	L	L	L	L	L	M/L	L
30	P*	P*	P*	E	E	E	E	E	E	{E}
31	E	E	E	E	E	E	P*	E	E	E
32	E	E	E	E	E	E	E	E	E	[E]
33	P	P	P	P	P	P	P	P	M	P
34	E	E	E	E	E	E	E	E	E/L	E
35	E	E	E	E	E	E	P*	P*	E	(E)
36	P	P	P	E	P	E	P	P	M	(P)
37	E	E	E	E	E	E	E	E	E	[E]
38	L	L	L	L	L	L	L	L	M	L
39	E	E	E	L	L	L	E	E	M	{E}
40	P*	P*	P*	P*	P*	P*	P*	P*	M	P*
41	E	E	E	E	E	E	E	E	M	E
42	P	P	P	P	P	P	P	P	M	P
43	P	P	P	P	P	P	P	P	M	P
44	P*	P*	P*	E	E	E	E	E	M	{E}
45	L	L	L	E	E	E	E	L	M	???
46	L	P*	P*	L	L	L	P*	L	M	???

47	L	L	L	L	P	P	L	L	M	(L)
48	L	L	L	L	L	L	L	L	M	L
49	L	L	L	E	E	E	L	L	M	{L}
50	P	P	P	P	P	P	P	P	E	P
51	E	E	E	E	E	E	E	E	E	[E]
52	E	E	E	E	E	E	E	E	E	E
53	L	L	P	L	L	L	L	L	L	L
54	L	L	L	L	L	L	E	E	Added	(L)
55	P*	P*	E	E	E	E	E	E	E	(E)
56	E	E	E	E	E	E	E	E	E	E
57	L	L	L	L	L	L	L	L	L	[L]
58	L	L	E	L	L	L	L	L	L	L
59	P*	P*	E	L	L	L	L	L	L	{L}
60	L	L	L	L	L	L	L	L	L	L
61	L	L	L	L	L	L	L	L	L	L
62	P	L	E	E	E	E	L	L	L	???
63	L	L	L	L	L	L	L	L	L	L
64	L	L	L	L	L	L	L	L	L	[L]
65	L	L	L	L	L	L	L	L	L	L
66	L	L	L	L	L	L	L	L	L	[L]
67	P	P	P	P	P	P	P	P	P	[P]
68	P	P	P	P	P	P	P	P	P	[P]
69	P	P	P	P	P	P	P	P	P	[P]
70	P	P	P	P	P	P	P	P	P	[P]
71	P	P	P	P	P	P	P	P	P	[P]
72	P	P	P	P	P	P	P	P	P	[P]
73	P	P	P	P	P	P	P	P	P	[P]
74	P	P	P	P	P	P	P	P	P	[P]
75	P	P	P	P	P	P	P	P	P	[P]
76	P	P	P	P	P	P	P	P	P	[P]
77	P	P	P	P	P	P	P	P	P	[P]
78	P	P	P	P	P	P	P	P	P	[P]
79	P	P	P	P	P	P	P	P	P	[P]
80	P	P	P	P	P	P	P	P	P	[P]
81	P	P	P	P	P	P	P	P	P	[P]

187	178	192	194	192	193	194	194	: Consistency metric (*100)
122	108	97	298	92	83	129	92	: Mahalanobis metric (*10)
6	8	4	2	2	2	4	2	: Number of "impossible" PostGuo's
5	6	6	3	3	3	3	2	: Disagree with Emerson (excluding M's)

Table 1 – 3-voice results

Of the low-majority chapters, [8 22 30 59] were assigned to their voice because a 5/8 majority agreed with Emerson (or at least one of his split voices). Chapters [39 49] were assigned to their voice because of a 5/8 majority and they matched the pattern (across all training sets) observed in chapters [8 22]. Chapter 44 was assigned to Early because of a 5/8 majority and it matched the pattern (across all training sets) observed in chapters [30 59].

There are seven "problem" chapters : five which are so variable across the training sets that they cannot be classified [2 24 45 46 62], chapter 40 which is consistently classified as PostGuo but exists in the Guodian manuscript, and chapter 50 which is consistently classified as PostGuo but which Emerson says is Early.

Ignoring those chapters for the moment, the best ("majority") 3-voice assignments are :

Early : [1 4 5 6 7 9 10 13 14 15 16 20 21 23 25 28 30 31 32 34 35 37 39 41 44 51 52 55 56]
Late : [3 8 12 17 18 19 22 27 29 38 47 48 49 53 54 57 58 59 60 61 63 64 65 66]
PostGuo : [11 26 33 36 42 43 67-81]

Of these 74 chapters, 42% of these chapters were not always in a training set, yet were classified the same across all or all-but-one of the training sets, indicating a great deal of internal consistency. Excluding chapters

Emerson labels as Middle, this assignment agrees with him completely (or at least with one of his split-voice assignments), except for 54 which he says is Added.

In order to determine the classifications of the seven problem chapters, two other analysis methods were used : looking at the distributions of symbols between the 74 "known" chapters, and averaging all nine individual features across all eight training sets.

Using the majority assignments above, the distribution of all symbols and about a dozen symbol pairs was analyzed to determine how often each showed up in Early, Late, and PostGuo chapters. Symbols that show up much more often in one voice over the others are summarized in Table 2 below. "Weak" symbols are 3 to 6 times more likely to show up in the indicated voice than the other two, while "Strong" symbols are more than 6 times more likely to be in that voice. "Unique" symbols show up only in that voice, and not even once in the others. Phrases (two or more symbols) are shown in italics. Weak and Strong symbols must show up in at least five different chapters to be included, while Unique symbols must be in at least three different chapters.

	Mostly Early	**Mostly Late**	**Mostly Post-Guodian**
Weak	valley call, say, speak return self deep mystery, deep and mysterious mother child, children fill, full maintain, protect	*do not act* truth, honest, trust* affairs, duties, trouble honest and just *non-action* kind, kindness* wise, wisdom* heart/mind	strong, inflexible, try originally, undoubtedly, firm, strong soft, softness, yielding dare, daring *Dao of*
Strong	! (#1) name, fame, reputation	choose, take, take hold of 100	weak, weakness victory, conquer die, death
Unique	? (#1) army blended, mingled pure and clear clear, pure, bright *newborn infant* hard work unfortunate, bad nobles (specifically Marquis) within, middle exist, survive, keep disgrace *without-name* same, sameness stop, rest, stay ? (#2) *no danger* gateway image *heaven & earth*	*rare goods* abandon *non-interference* family morality* clever, skillful bandit, evil thief, robbery confuse, confusion	treasure hard teach, teaching

Table 2 – 3-voice symbols predominately in one voice

Symbols that show up much more often in two voices but rarely in the other voice are summarized in Table 3 below, using the same format and criteria as for Table 2.

	Rarely in Early	Rarely in Late	Rarely in Post-Guodian
Weak	govern behavior, perform, travel virtue, good(ness), skilled ! (#2) *sage*	weapons who, which ready, would, about to, general create, life, produce one	De
Strong	benefit, profit, sharp nation citizens *thus the sage*		
Never	*virtuous person* misfortune serious, double easy *do not strive* difficult, hard	Qi good fortune scholar since, once maintain, protect master depend on, concerned with rare, few *do not know*	*100 families* female ocean, sea subtle mystery spirit bright, brightness empty arise, make move, movement, action begin, beginning arise, produce, go out !,? uncarved block not (非)

Table 3 – 3-voice symbols mostly absent in one voice

Analyzing the distribution of the "Unique" and "Never" symbols is problematic, because while there can be examples where the distributions are consistent, there can also be many inconsistencies. As an example of the former, chapter 14 has a unique Early symbol and symbols never used by Late or PostGuo, thus strengthening the confidence in labeling 14 as Early. But while 40 and 50 have symbols in them that never appear in other PostGuo chapters, every training set classified both of them as PostGuo because of the frequency distribution of many other symbols. In addition, chapters 24 and 45 contain symbols in them unique to both Early and Late, and 62 has symbols in it unique to all three voices. At the same time, 45 and 62 also contain symbols that never appear in other Late chapters. Finally, chapter 2 has symbols in it that never appear in Early chapters, another symbol which never appears in Late chapters, and still another symbol that never appears in PostGuo chapters. This demonstrates another reason why unique symbols were not used in the classification metrics, although *it is interesting to note that every chapter that had contradictory "Unique" and "Never" symbols is a problem chapter*.

Chapters that are present in the Guodian document but often classified as PostGuo here (such as 2 and 40) may not be as "impossible" as they seem. For example, if perhaps the post-Guodian "style" was already being developed at the time the Guodian document was created, then the presence of these chapters in it could mean that they had already been written in this style, perhaps shortly before the Guodian document was created, whereas chapters 68-81 had yet to be written (or the Guodian compiler was not aware of their existence). There are probably many other scenarios that could explain such a result.

The other analysis technique that was tried was to *first* average all 9 features from the InCommon, Rare, and AllButOne metrics across all eight training sets, and then sum them as usual down to 3 features, thus taking the average of all training sets at the lowest feature level. Using this approach, all chapters were assigned their majority voice from above, excluding the problem chapters. At this point, the LOO Mahalanobis distance from each chapter to the center of all three voices was calculated, and the chapter with the smallest distance was always the same as the majority assignment. Recalling that the features basically represent the frequencies of various symbols in each voice, another classifier was created which simply returned the voice with the largest feature value, and it also always agreed with every majority assignment, except for two of the problem chapters. Both of these results strengthen the confidence in the assignments of the low-majority chapters [8 22 30 39 44 49 59] from above.

In addition to the Mahalanobis distance classifier, the K-nearest-neighbor classifier was also used for the problem chapters (using the Mahalanobis distance, since the voices are relatively Gaussian and definitely not circular in shape, as will be seen below). Both these classifiers produced results that agreed with each other and the majority vote for all the problem chapters except for 2, which still appears as a mixture of all three voices.

Table 4 summarizes the information from all these sources for each of the problem chapters. The first three columns indicate the number of "votes" for that voice that each chapter received from Table 1.

Ch#	Early	Late	PostGuo	Maha,KNN	Emerson
2	1	2	5	E or P	E/M/L
24	4	1	3	E	M
40	0	0	8	P	M
45	4	4	0	←L	M
46	0	5	3	L	M
50	0	0	8	P	E or P
62	4	3	1	E	←L

Table 4 – 3-voice problem chapters

In Emerson's post-Guodian discussion[3], he speculated that chapter 50 might be PostGuo, in agreement with the results found here. Although 40 is unanimously PostGuo, it is present in the Guodian manuscript, but that is not considered a problem for the reasons discussed above. Chapter 45 is tied between Early and Late, so the Mahalnobis distance is used as the tie-breaker, because it is the most accurate measurement of the distance to a voice in the feature space. Even though 62 is most often classified as Early, it is often considered Late, and because of Emerson's Late classification and the fact that 62 is right in the middle of a run of Late chapters, it seems reasonable to re-label it as Late (remember that 62 has symbols in it unique to <u>all three</u> voices, *and* symbols that <u>never</u> appear in the other Late chapters – which is why it is a problem chapter). Finally, while chapter 2 is predominantly classified as PostGuo, it is also sometimes classified (by all the different classifiers) as Early and Late, and the symbol distribution simultaneously indicates that it cannot belong to any of them! If any chapter is truly a mixture of voices, it is chapter 2 (which basically agrees with Emerson, although he never considered it as being PostGuo).

Based on these decisions, the final best 3-voice classification of all chapters is :

Early : [1 4 5 6 7 9 10 13 14 15 16 20 21 23 24 25 28 30 31 32 34 35 37 39 41 44 51 52 55 56]
Late : [3 8 12 17 18 19 22 27 29 38 45 46 47 48 49 53 54 57-62-66]
PostGuo : [11 26 33 36 40 42 43 50 67-81]
Unknown : [2]

Except for the chapters Emerson identified as Middle and Added, <u>this result agrees with Emerson completely</u>. The supposed Middle chapters are fairly evenly spread throughout these voices : five are classified as Early, seven as Late, and six as PostGuo. Except for 40, none of the Post-Guo chapters are present in the Guodian manuscript.

One interesting result that emerged independently from this work is that Emerson thought that chapters [22 24 33 38 39 42] might also be post-Guodian. Compare that to the results found here where [11 26 33 36 40 42 43] were classified as PostGuo, agreeing with Emerson for chapters [33 42 50].

<u>4-Voice Results</u>

The results for 4 voices are shown in Table 5. Due to the number of variably classified chapters between the original seven reduced training sets, another initial training set ("constant") was created from only those chapters that were labeled the same across the seven training sets (whether because they were in the training set

or they were classified as such) and agreed with Emerson, and pruned as usual. However, in this case chapters were only removed if when classified they remained in the chapter they had originally been assigned to.

	Pure		Consist		Maximal			Const		
Ch	#1	#2	#1	#2	1a	1b	2		Emer	Final
1	E	E	E	E	E	E	E	E	E	[E]
2	P*	P*	P*	P*	P*	P*	P*	P*	E/M/L	???
3	L	L	L	L	L	L	L	L	L	[L]
4	E	E	E	E	E	E	E	E	E	E
5	E	E	E	E	E	E	E	E	E	[E]
6	E	E	E	E	E	E	E	E	E	[E]
7	E	E	P	P	L	L	E	P	E/L	???
8	L	L	L	L	L	L	P	L	M/L	L
9	E	E	E	E	E	E	E	E	M	E
10	E	E	E	E	E	E	E	E	E	E
11	M	M	M	M	M	M	M	M	M	[M]
12	L	L	L	L	L	L	L	L	L	[L]
13	M	M	M	M	E	E	M	M	E	???
14	E	E	E	E	E	E	E	E	E	E
15	E	E	E	E	E	M	E	E	E	E
16	E	E	E	E	E	E	E	E	E	E
17	L	L	L	L	L	L	L	L	L	[L]
18	L	L	L	L	L	L	L	L	L	[L]
19	L	L	L	L	L	L	L	L	L	[L]
20	E	E	P*	E	E	E	E	L	E	(E)
21	E	E	E	E	E	E	E	E	E	E
22	L	L	P	P	L	L	E	L	M/L	???
23	E	E	E	E	E	E	E	E	E/M	E
24	M	M	M	M	M	M	M	M	M	[M]
25	M	M	M	M	E	E	E	M	E	???
26	M	M	M	M	M	M	M	M	M	[M]
27	L	L	P	P	P	M	L	L	M/L	???
28	E	E	E	E	E	E	E	E	E	E
29	L	L	P	P	E	E	M	L	M/L	???
30	L	M	P*	P*	E	E	E	P*	E	???
31	E	E	E	E	E	E	E	E	E	[E]
32	E	E	E	E	E	E	E	E	E	E
33	P	P	P	P	M	M	P	P	M	???
34	L	E	E	E	E	E	E	E	E/L	E
35	**E**	**E**	**E**	**M**	**E**	**E**	**E**	**M**	**E**	**(E)**
36	P	P	P	P	M	M	M	P	M	???
37	E	E	E	E	E	E	M	E	E	E
38	L	M	P	L	L	L	M	L	M	???
39	M	M	M	M	M	M	M	M	M	[M]
40	M	M	M	M	M	M	M	M	M	[M]
41	M	M	M	M	M	M	M	M	M	[M]
42	M	M	M	M	M	M	M	M	M	M
43	M	M	M	M	M	M	M	M	M	[M]
44	M	M	M	M	M	M	M	M	M	[M]
45	M	M	M	M	M	M	M	M	M	M
46	M	M	P*	P*	P*	L	M	P*	M	???
47	M	M	M	M	M	M	M	M	M	[M]
48	M	M	M	M	M	M	M	M	M	[M]
49	L	L	L	L	M	M	L	L	M	???
50	M	M	P	P	P	P	E	M	E	???
51	E	E	E	E	E	E	E	E	E	E
52	E	E	E	E	E	E	E	E	E	E
53	L	L	L	L	P	P	L	L	L	(L)
54	L	L	L	L	E	E	L	L	Add	(L)
55	P*	P*	E	E	E	E	E	E	E	(E)
56	E	E	E	E	E	E	E	E	E	[E]
57	L	L	L	L	L	L	L	L	L	L
58	P	P	L	P	L	L	L	P	L	???
59	P*	P*	P*	P*	L	L	L	E	L	???
60	L	L	L	L	L	L	L	L	L	L
61	L	L	L	L	L	L	L	L	L	L
62	L	L	P	P	L	L	L	L	L	(L)
63	L	L	P*	P*	L	L	L	L	L	(L)
64	L	L	P*	P*	L	L	L	L	L	(L)

65	L	L	L	L	L	L	L	L	L	L	
66	L	L	L	P*	L	L	L	L	L	L	
67	P	P	P	P	P	P	P	P	P	P	[P]
68	P	P	P	P	P	P	P	P	P	P	[P]
69	P	P	P	P	P	P	P	P	P	P	[P]
70	P	P	P	P	P	P	P	P	P	P	[P]
71	P	P	P	P	P	P	P	P	P	P	[P]
72	P	P	P	P	P	P	P	P	P	P	[P]
73	P	P	P	P	P	P	P	P	P	P	[P]
74	P	P	P	P	P	P	P	P	P	P	[P]
75	P	P	P	P	P	P	P	P	P	P	[P]
76	P	P	P	P	P	P	P	P	P	P	[P]
77	P	P	P	P	P	P	P	P	P	P	[P]
78	P	P	P	P	P	P	P	P	P	P	[P]
79	P	P	P	P	P	P	P	P	P	P	[P]
80	P	P	P	P	P	P	P	P	P	P	[P]
81	P	P	P	P	P	P	P	P	P	P	[P]

184	175	191	190	189	183	187	189	: Consistency metric (*100)
128	120	125	121	117	133	133	129	: Maha metric (*10)
3	3	7	7	2	1	1	3	: "Impossible" PostGuo's
12	11	15	17	5	6	6	14	: Disagree with Emerson

Table 5 – 4-voice results

There are seventeen "problem" chapters : 2 which is consistently classified as PostGuo even though it exists in the Guodian manuscript (and which Emerson says is a mixture of everything *but* PostGuo), 9 which is consistently classified as Early despite Emerson's Middle classification, [13 25 33 36 49] which have a large majority that disagrees with Emerson, and [7 22 27 29 30 38 46 50 58 59] which are too variable to determine.

Ignoring the problem chapters, the best ("maximum") 4-voice assignments are :

Early = [1 4 5 6 10 14 15 16 20 21 23 28 31 32 34 35 37 51 52 55 56]
Middle = [11 24 26 39 40 41 42 43 44 45 47 48]
Late = [3 8 12 17 18 19 53 54 57 60 61 62 63 64 65 66]
PostGuo = [67-81]

where 37% of non-problem chapters were classified the same across all training sets. This agrees with Emerson completely (or at least with one of his split voice assignments).

Comparing the 3-voice and 4-voice final results (ignoring the problem chapters from both for the moment), they have the following chapter assignments in common :

Early = [1 4 5 6 10 14 15 16 20 21 23 28 31 32 34 35 37 51 52 55 56]
Late = [3 8 12 17 18 19 53 54 57 60 61 63 64 65 66]
PostGuo = [67-81]

where underlined chapters were consistently classified or always ended up in a training set across all or all-but-one training sets for both voices. This list accounts for 51 out of the 81 chapters (63%), or considering that the twelve Middle chapters by definition cannot match any 3-voice assignments, 51 out of 69 (74%).

Symbols that show up much more often in one voice over the others are summarized in Table 6, using the same format as before. Symbols that show up much more often in three voices but rarely in the other voice are summarized in Table 7. The phrases "do not strive", "do not dare", and "virtuous person" are common in both Late and PostGuo, but never appear in Early or Middle. The phrase "100 families" appears in both Early and Late, but never in Middle or PostGuo.

Ratio	Mostly Early	Mostly Middle	Mostly Late	Mostly Post-Guodian
Weak	ever-constant happy, pleased, music deep mystery, deep and mysterious return ? (#1) can, able maintain, protect	day hear	affairs,duties,trouble govern nation 100	remainder, excess victory, surpass there is no, no one soft, yield weapons weak strong, try, inflexible dare *now : only*
Strong	return !(#3)	one		die, death
Unique	army blend, confused child, children obstruct, block up unfortunate, bad early follow, submit infant lose, lost, left behind loosen, clarify open oppress, dampen, subdue sharp subtle, mysterious *heaven and earth*	boast boast die, lose straight, straightforward teaching	abandon family easy to, fond of front, before righteousness thief, traitor, evil wealth	misfortune *the way of (Dao of)*

Table 6 – 4-voice symbols predominately in one voice

Ratio	Not in Early	Not in Middle	Not in Late	Not in Post-Guodian
Weak	*thus*	man, now: we, I	use, useful create, life	obtain, gain, get thing, creature
Strong				
Never	behavior, act, perform, travel few, lonely, widowed honest, just certainly, must choose, take(hold of) *thus the sage*	dwell/reside ancient depart, leave, remove, distance truth, honest, trust back, behind, after, later can, able, capable, competent all (to cause)harm peaceful, calm, content, how two, both beauty fear, afraid just like, to scheme before, first each other, mutually hold, grasp, maintain small, insignificant *this is called*	who, which long time, long master love substantial rare/infrequent *do not understand*	not(非) valley whole, entire, in the end stillness self move(ment)/action as good as, as if, like produce/arise, go out embrace also *non-action* *do not act*

Table 7 – 4-voice results mostly absent in one voice

As before, summing all 12 features across all eight training sets and combining them down to 4 features also produced the same Mahalanobis-distance classifications as the majority-vote list above for the non-problem chapters, except for chapters [53 55 67]. In addition, the "largest feature value" classifier did not agree with the majority-vote for 55. Together these results indicate less internal consistency in the 4-voice model. In addition,

the Mahalanobis-distance and KNN classifications for the problem chapters did not always agree with each other. Table 8 shows the combined information from all sources for each of the problem chapters.

Ch#	Early	Middle	Late	PostGuo	Maha,KNN	Emerson
2	0	0	1	9*	L,P	E/M/L
7	3	0	4	3	L	E/L
9	10	0	0	0	E	M
13	2	8	0	0	M	E
22	1	0	7	2	L	M/L (or P)
25	5	5	0	0	E	E
27	0	1	6	3	L	M/L
29	2	1	5	2	L	M/L
30	3	1	2	4*	L	E
33	0	3	0	7	M	M (or P)
36	0	4	0	6	M	M
38	0	2	7	1	L	M (or P)
46	0	3	1	6*	M	M
49	0	2	8	0	L	M
50	1	4	0	5	M	E (or P)
58	0	0	6	4	L	L
59	1	0	5	4*	L	L

Table 8 – 4-voice problem chapters

It can be seen that [9 22 38 49] are all in strong agreement with their assignments, although all but 22 *disagree* with Emerson. [27 29] have a moderate agreement with their assignments, and agree with Emerson (or at least one of his split assignments). [7 25 58 59] have only a slight majority for their assignment, but agree with Emerson. And the Mahalanobis and KNN classifications agree with the magority votes for all these chapters except 25, which was evenly split Early/Middle and where the Maha/KNN vote broke the tie. Finally, chapters [9 25 27 49 58 59] are all <u>well</u> within their class as evidenced by their Mahalanobis distances (see Figure 4). As will be seen below, even though chapter 13 is strongly assigned to the Middle voice, it is actually an outlier (like point B in Figure 3).

The remaining chapters [30 33 36 46 50] are *still* problematic, either with no clear majority, a majority that does not agree with the Maha/KNN assignments or Emerson, or combinations of all three. For [33 36 46], the Mahalanobis classification agrees with Emerson, so those are the ones that were used, even though they disagree with the majority votes. Note that 33 is predominately PostGuo in the separate classifiers, but definitely Middle using Maha and KNN, and Emerson says it has elements of both. Chapter 30 is pretty evenly tied between Early, Late, and PostGuo, but is closest to Late in terms of Mahalanobis distance. Chapter 50 is pretty evenly tied between Middle and PostGuo, but is clearly much closer to Middle in terms of the Mahalanobis distance.

Chapter 2 is consistently identified as PostGuo using the individual training sets even though it exists in the Guodian manuscript, is classified as both Late and PostGuo using the Mahalanobis and KNN classifiers, and Emerson says it has elements from every voice *except* PostGuo. So again, chapter 2 seems to be unidentifiable.

Based on these decisions, the final best 4-voice result is :

Early = [1 4 5 6 9 10 14 15 16 20 21 23 25 28 31 32 34 35 37 51 52 55 56]
Middle = [11 (13) 24 26 33 36 39 40 41 42 43 44 45 46 47 48 50]
Late = [3 7 8 12 17 18 19 22 27 29 30 38 49 53 54 57-58-59-66]
PostGuo = [67-81]
Unknown = [2]

Of these, chapters [9 13 30 38 49 50] disagree with Emerson, and [7 30] changed voices from the 3-voice set (other than those chapters that were assigned to Middle). Interestingly, 7 went from Early to Late, and Emerson says it has elements of both. In addition, *every PostGuo chapter below 67 in the 3-voice set moved to Middle* (although this does not always agree with Emerson). Of the problem chapters, [9 25] in Early and [22 27 29 38 49 58 59] in Late match their 3-voice assignments.

The chapters that the final 3-voice and 4-voice assignments have in common are :

Early = [1 4 5 6 9 10 14 15 16 20 21 23 25 28 31 32 34 35 37 51 52 55 56]
Late = [3 8 12 17 18 19 22 27 29 38 49 53 54 57-66]
PostGuo = [67-81]

accounting for 75% of all chapters (and excluding Middle chapters, 95%).

Graphical Results

A graphical display of the results was also generated for each case. A Linear Discriminant Analysis (LDA) plot takes multi-dimensional data and finds the two-dimensional ("2-D") projection that provides maximum visual separability between all the classes[1]. The features used for these graphs came from the summed features of the eight training sets. Chapters on or just outside an oval are still definitely part of that voice.

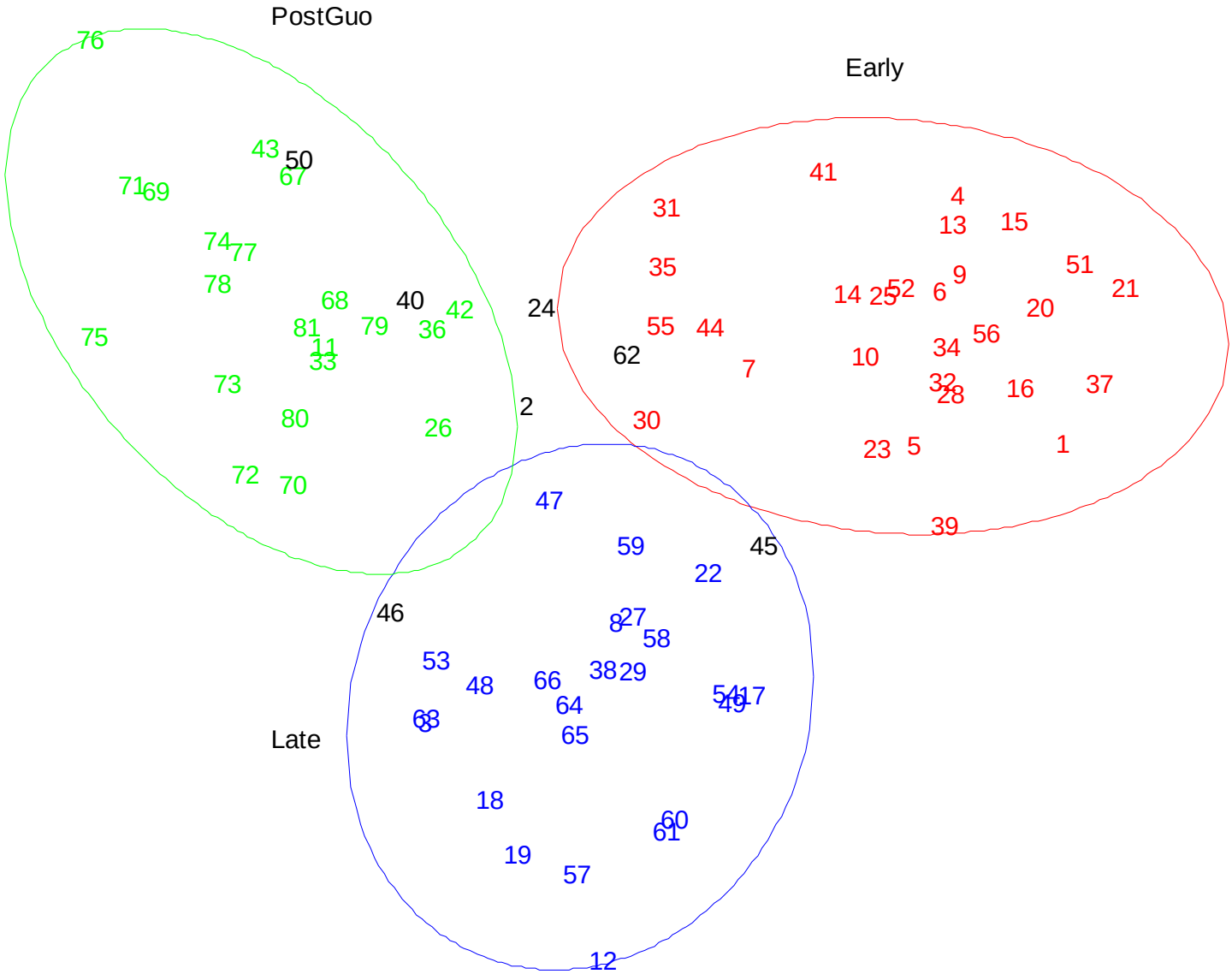

Figure 5 – 3-voice distribution of chapters

To some degree, the relative positions of the chapters in these pictures can demonstrate why the "problem" chapters are problems. Looking at the 3-voice plot in Figure 5, note that even though chapters 45 and 46 are within the Late oval, 45 could reasonably belong to Early, and 46 could belong to PostGuo. Likewise, while 24 is closest to Early it is also close to Late, and although chapter 2 is closest to PostGuo, it could reasonably belong to any of the three. Also note the internal clusterings, indicating chapters that are in some way very similar to each other – for example : [28 32] and [25 52] in Early; [3 63] and [17 49 54] in Late; [11 33], [69 71], and [74 77] in PostGuo.

Looking at the 4-voice plots in Figure 6, note that the overlapping ovals are an artifact – in the higher-dimensional space of the original data, any three of the ovals (actually 3-D football-shaped ellipsoids) are in one plane, and the fourth oval is perpendicularly above that plane (like the four vertices of a skewed triangular pyramid), so that they are actually completely separated (in this view, PostGuo is elevated above the plane of the paper that the Early, Middle, and Late ovals are in). Chapter numbers for the training sets that created the ovals are left out for clarity. Due to the loss of the third dimension in this image, the only thing that can be claimed with certainty is that chapter 13 is significantly different than <u>any</u> of the other chapters (but closest to Middle). Looking at 3-D graphs of the data developed during this research, it was confirmed that 9 and 25 are *within* the Early oval, and [49 58 59] are *within* the Late oval. In addition, 36 and 50 are both right on the edge of the Middle oval.

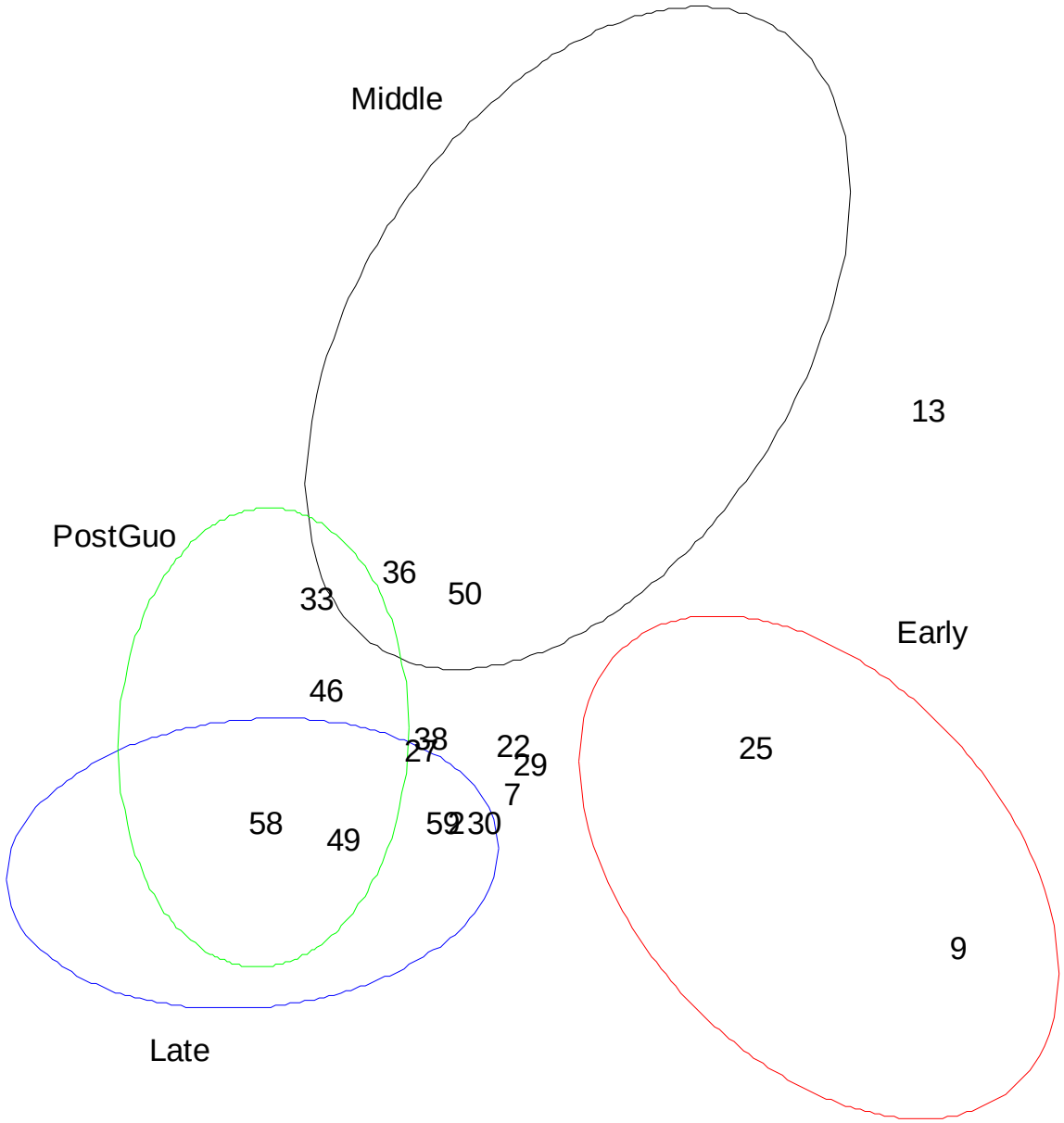

Figure 6 – 4-voice distribution of chapters

On the question of 3 or 4 voices

To determine whether 3 voices or 4 voices is "better", several measures-of-quality were used from *cluster analysis*. Clustering attempts to group similar examples together without any a-priori labeled examples (i.e., without a training set). To determine the optimum number of clusters (which are classes or in this case, voices) in the data, a number of "clustering validity measures" have been developed[5]. One way or another, they all prefer clusterings with the smallest "spread" within each class and the largest distance between classes (see Figure 7).

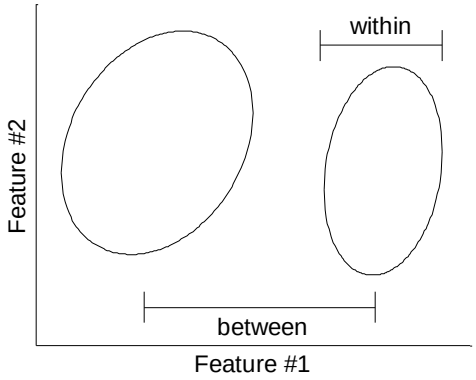

Figure 7

The first metric is the mean of the Mahalanobis distances of all examples in each class to their class center, divided by the mean Mahalanobis distances between all pairs of classes. *Smaller* values of this metric indicate a better fit.

The Variance Ratio Criterion (VRC) calculates the average "scatter" (spread) within all the classes separately (Sw), and the scatter between the means of the classes (Sb), and returns a value

$$\frac{(N-C) * \text{trace}(Sb)}{(C-1) * \text{trace}(Sw)}$$

where N is the total number of examples and C is the number of classes. *Larger* values of this metric indicate a better fit. Both this metric and the Mahalanobis metric assume that the data has a Gaussian (multi-dimensional "bell curve") distribution, which this data does.

The Davies-Bouldin index (DB) finds the average (across all classes) of the maximum of $(d_i + d_j)/d_{ij}$, where d_i is the average Euclidean distance of all examples in class *i* to their center, and d_{ij} is the Euclidean distance between the centers of class *i* and *j*. *Smaller* values of this metric indicate a better fit.

A variation on Dunn's index (Dunn) takes the smallest value (across all pairs of classes) of $d_{ij}/\max(d_k)$, where d_{ij} is the largest Euclidean distance between all examples in classes *i* and *j*, and d_k is twice the average Euclidean distance of all examples in class k to their center. *Larger* values of this metric indicate a better fit. Both this metric and the previous one assume the data has a spherical distribution, which this data does not.

The results are shown in Table 9, with the values from all the training sets sorted in ascending order.

Maha (smaller)	**3-voice**	220	237	258	266	266	296	303	333
	4-voice	275	284	333	334	335	341	354	362
VRC (bigger)	**3-voice**	98	114	117	135	139	141	193	213
	4-voice	60	61	61	63	65	66	73	76
DB (smaller)	**3-voice**	128	128	131	132	132	134	135	141
	4-voice	141	141	145	146	147	148	148	151
Dunn (bigger)	**3-voice**	277	278	289	307	308	332	338	358
	4-voice	201	208	208	208	212	223	223	240

Table 9 – cluster validity metrics

While cluster metrics are designed to compare data with different numbers of classes, comparing cluster metrics in cases with different numbers of features is not straightforward. A simulation was run to determine the relationship between the 3-feature and 4-feature cluster metric values :

- The best 3-voice Maha metrics above are slightly better than the best 4-voice Maha metrics.

- The best 3-voice VRC metrics above are much better than the best 4-voice VRC metrics.
- The best 3-voice DB metrics above are much better than the best 4-voice DB metrics.
- The best 3-voice Dunn metrics above are much better than the best 4-voice Dunn metrics.
- In all cases, the worst of the 3-voice metrics were equally as bad as the worst of the 4-voice metrics.

In addition :

- Six out of eight of the 3-voice consistency metrics were better than all eight of the 4-voice metrics, indicating that it is easier to get self-consistent chapter assignments with 3 voices than with 4 voices
- The more the 4-voice assignments *disagreed* with Emerson and had more "impossible" PostGuo chapters, the better their consistency and clustering metrics, suggesting that classification with 4 voices is not a good fit
- There were only seven problem chapters across all training sets with 3 voices, but seventeen problem chapters with 4 voices, another indication that 3 voices seems to be a better fit
- None of the 3-voice assignments disagreed with Emerson, whereas six of the 4-voice assignments did
- The four kinds of assignments (majority Linear votes, largest-feature-value, Mahalanobis-distance, KNN) all agreed with other much more for 3 voices than for 4 voices
- While pruning the initial training sets, it was noticed that it was often impossible to increase both consistency and compactness for 4 voices, but always easy to do both for 3 voices, another indication of a poor fit for 4 voices

Conclusions

Using pattern recognition techniques for finding voices in the DDJ has the advantage of being free from personal bias, once the desired features have been identified and the training set chapters have been selected. And as has been shown, almost three-quarters of all the chapters are consistently classified as the same voice, independent of the particular training set or even the number of voices.

Regarding the issue of how many voices are present in the DDJ, the 3-voice model seems to be a much better fit than the 4-voice model, according to multiple different measures.

One point that was not pursued, because pruning the training sets takes a very long time (over 1000 hours of CPU time for this article), is how much the classifications would vary if different subsets of features (InCommon, Rare, AllButOne), or different features, were used. Another topic for future research could try to incorporate two-, three-, and four-symbol "phrases" into the classification process. It should also be possible to analytically allow for the identification of chapters with multiple voices, rather than just labeling each chapter as belonging to a single voice.

In summary, the best chapter classifications found here are :

Early = [1 4 5 6 7 9 10 13 14 15 16 20 21 23 24 25 28 30 31 32 34 35 37 39 41 44 51 52 55 56]
Late = [3 8 12 17 18 19 22 27 29 38 45 46 47 48 49 53 54 57-66]
PostGuo = [11 26 33 36 40 42 43 50 67-81]
Unknown = [2]

The following chapters were identified as *strongly* belonging to their voice across all 3-voice and 4-voice training sets :

Early = [1 4 5 6 10 14 15 16 21 23 28 31 34 37 51 52 56]
Late = [3 12 17 18 19 53 57 60 61 65 66]
PostGuo = [67-81]

which accounts for 67% of all chapters not considered to be Middle by Emerson.

Chapters [2 8 22 24 30 39 44 45 46 49 50 59 62] were found to have elements of more than one voice, and should be further studied to see if the lines that contain the symbols that classify them as different voices can be cleanly separated into "sections".

Using the "best" classification above, with all chapters included as examples of their class (except chapter 2, which was left out of all three voices), one last pruning was performed. It was found that by removing [15 41 51 56] from Early and [27 57 66] from Late, a consistency of 1.98 was achieved, better than any of the eight "test" training sets found earlier. In addition, the Maha metric was 6.81, exceptionally smaller than any of the previous training sets, indicating that the above separation of voices is very good, and the placement of the "problem" chapters seems to be correct (all problem chapters were obviously closest to their assigned voice using the Mahalanobis distance, as were 8 and 54 which were originally in question). In addition, leaving out any other chapter or any two chapters from their voices produced worse results. Note that although those seven chapters had to be removed to create a good training set, they themselves were then classified as obviously belonging to the voice they had been removed from. As to *why* they had to be removed to create a good training set could be the basis for future research – perhaps they also contain a mixture of voices.

References

[1] Duda, R.O., Hart, P.E.; "Pattern Classification and Scene Analysis"; John Wiley, 1973.
[2] Emerson, John J.; "A Stratification of Lao Tzu"; The Journal of Chinese Religions,#23, Fall 1995, pp. 1-28.
[3] Emerson, John J.; "Lao Tzu Stratified, II: A Sketch"; http://www.idiocentrism.com/china.strata3.htm.
[4] Linnell, Bruce R; "The Effects of Small Samples on Statistical Pattern Recognition Techniques"; Ph.D. dissertation, ECE Dept., North Carolina State University, 2001.
[5] Lucas Vendramin, Ricardo J. G. B. Campello, Eduardo R. Hruschka; "On the Comparison of Relative Clustering Validity Criteria"; SIAM international Conference on Data Mining, Sparks, NV, USA, 2009, 733-744.