Covariance Estimation for Small Sample Data with Applications to Forensic Glass

Karen Deanna Huang Pan

M.S., University of Virginia, United States, 2016
B.S., University of Virginia, United States, 2015

A Dissertation Presented to the Graduate Faculty
of University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May 2020

Abstract

This study is motivated by a trace element analysis procedure often used in forensic science for the analysis of glass and other trace evidence. Existing glass data sets used to analyze error rates are of small sample size (n) relative to dimension (p); and potential benefits may arise from the consolidation of multiple small data sets into one. However, while forensic data is generally not readily available, information concerning their covariance (or correlation) matrices may be. This research proposes two methods for detecting similarity between covariance matrices and their true effective dimension as well as a method for combining covariance matrices based on the subspaces they span. These metrics and methods may apply to similar situations where data is inaccessible or otherwise unavailable due to privacy or other reasons. They may also apply if the data is believed or known to contain many outliers, as analysis of the covariance matrix may be more robust in reducing some of the effect of outliers. Although motivated by small n, we believe these methods may scale to the case where both n and p are large or $p > n$.

Acknowledgements

I would like to express my sincere and deepest gratitude to my adviser, Professor Karen Kafadar, for her constant support, guidance, and encouragement throughout the course of my graduate research and study. From her I have learned statistics, research, life lessons, and much more.

I would like to thank the rest of my committee members, Professor Jianhui Zhou, Professor Jordan Rodu, and Dr. Hari Iyer, for their continued support, time, and help on my dissertation.

My gratitude goes to all the faculty in the department. For courses, discussions, time, expertise, and their eagerness and willingness to guide and support. In particular, to Professor Tingting Zhang, Professor Jianhui Zhou, Professor Tianxi Li, Professor Daniel Keenan, and Professor Jeff Holt. To Karen Dalton, who greets me with a smile no matter how many receipts or questions I have for her.

To my friends, who brighten my life and are always there for me.

To my family, who have supported me through all my endeavours.

Contents

Chapter 1

Introduction

1.1 Introduction

Every contact leaves a trace.

– Dr. Edmond Locard, 20th century French forensic science pioneer

Wherever he steps, whatever he touches, whatever he leaves, even unconsciously, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibers from his clothes, the glass he breaks, the tool mark he leaves, the paint he scratches, the blood or semen he deposits or collects. All of these and more, bear mute witness against him. This is evidence that does not forget. It is not confused by the excitement of the moment. It is not absent because human witnesses are. It is factual evidence. Physical evidence cannot be wrong, it cannot perjure itself, it cannot be wholly absent. Only human failure to find it, study and understand it, can diminish its value.

– Paul Kirk, forensic scientist [58]

Locard's exchange principle and Kirk's statement formulated the basis of and paved the way for forensic science today. Although physical evidence "does not forget," it may be easily contaminated or degraded in its natural environment. Most importantly, as Kirk himself recognized, physical evidence may be misinterpreted by well-meaning forensic scientists. The research presented in this paper is motivated by a procedure known as trace element analysis, which has been used to evaluate the source of trace evidence such as bullets, glass, paint, and copper wire. While our data specifically concerns glass fragments, the procedure for all trace element analysis is similar. The "working hypothesis" is that concentrations of certain elements – those presumably highly specific to different pieces of evidence – provide a distinctive "signature" that allows one to conclude whether "analytically indistinguishable" evidence found at a crime scene and that in possession of a suspect share a common source.

The statistical issues surrounding this approach are evident. Consider a glass manufacturer who produces float glass for windowpanes. If the batch of material from which many windows are manufactured is extremely homogeneous, measurements on many windows from the same batch may be deemed "not distinguishable" (depending on the level of error in the measurements themselves), leading one to erroneously conclude different windows "came from the same source" (are actually one window) and hence to potential false positives. Conversely, if the windowpane in question is rather inhomogeneous, measurements from two different fragments of the same window may be different, leading to false negatives. Trace element analysis of forensic evidence may be unsatisfactory for both inclusion and exclusion purposes.

Recently, three American Society for Testing and Materials (ASTM) International standards have been proposed detailing the full procedure for trace evidence

analysis of forensic glass evidence. The standards, which describe the process of determining if two glass fragments are "analytically indistinguishable," do not explicitly use the terms "inclusion" and "exclusion"; however, evidence has shown that jurors, at least in the U.S., may understand "analytically indistinguishable" to mean "identification" or "from the same source" [34]. Additional phrasing from the standards, "[i]f the samples are distinguishable... in any of these observed and measured properties [for example, color, refractive index, density, elemental composition], it may be concluded that they did not originate from the same source of broken glass," provides conditions where "distinguishable" evidence samples may be used for exclusion (ASTM E2330-12 Section 1.1 [51], ASTM E2927-16 Introduction [53], ASTM E2926-13 Introduction [52]). ASTM E2927-16 [53] further asserts that the described technique "yields high discrimination among sources of glass" and "provides high discriminating value in the forensic comparison of glass fragments." In addition to jurors potentially misunderstanding the meaning of "not distinguishable," the standards themselves describe the probative value of this determination in a manner that may well be highly misleading.[1]

Up until recently, there has been minimal use of statistics in forensic science disciplines. The 2009 National Academy of Sciences (NAS) report, *Strengthening Forensic Science in the United States: A Path Forward* (National Research Council (NRC) 2009) [22], recognized this and proposed a number of recommendations for the future direction of the field. An earlier report provided detailed findings on a trace evidence analysis procedure previously in use at the U.S. Federal Bureau of Investigation (FBI) known as Compositional Analysis of Bullet Lead (CABL) [21], which is described due to its similarity to the procedure in use for analysis of glass fragments. Some background on forensic glass analysis and the specific analysis

[1]The author thanks Professor William Thompson for this remark.

procedure are provided before introducing the data currently available to us.

1.2 Compositional Analysis of Bullet Lead (CABL)

In 2004, the NAS published its findings on the Compositional Analysis of Bullet Lead (CABL), a procedure for comparing the "signatures" of two bullets (or bullet fragments): one from the crime scene (CS) and one from the potential suspect (PS), often referred to in the forensics discpline as the "known" (K) and "questioned" (Q) or "recovered" (R) [21]. Using the working hypothesis described above, a vector of seven measured trace elemental concentrations (silver Ag, antimony Sb, arsenic As, bismuth Bi, cadmium Cd, copper Cu, lead Pb) was assumed to provide a unique "signature" [with the ability] to distinguish samples from differing and common source. Elemental concentrations from the K and Q bullets were measured in triplicate, and the sample mean and standard deviation were calculated for each bullet (fragment). The "2-SD-overlap" procedure then involved forming "mean $\pm$ 2·SD" intervals. If the K and corresponding Q intervals overlapped for all seven elements, they were deemed "analytically indistinguishable." The FBI often went further in the courtroom by testifying that K and Q "likely originated from the same manufacturer's source of lead" or "must have come from the same box" [21, pp. 91-92]. A "false positive error rate" (probability of claiming "same source" when the samples came from different sources) by counting the number of pairs between any two of 1,837 samples in their "data base" that resulted in a "false match" by their procedure; i.e., they found 693 among their 1,686,366 pairs that resulted in a "false match." In this paper, we refer to the rule used to determine "analytically indistinguishable" as the "match rule" and the proportion of times that the rule is satisfied as the "match rate."

The NAS Committee concluded that:

1. The "data base" of 1,837 samples including "one specimen from each combination of bullet caliber, style, and nominal alloy class" [59, 62] cannot be viewed as a representative sample of bullets – they were "selected" in hopes of spanning the space of all possible bullet types, thereby resulting in pairs of bullets that could be expected to be more different than might be seen in a real case.

2. A procedure for estimating CABL's error rate is based more properly on statistical modeling of the covariance (or correlation) matrix among the seven elements in the proposed "signature," from which more valid estimates of sensitivity (given that the true concentrations differ by less than a prescribed "difference threshold," the probability that the FBI procedure properly concludes "same source") and specificity (given that the true concentrations differ by more than a prescribed "difference threshold," the probability that the FBI procedure properly concludes "different source") can be calculated.

The Appendices in the NRC (2004) report concluded that the CABL procedure error rates would be much higher than the claimed 0.04%. The Committee found no fault with the elemental concentration measurement technique (inductively coupled plasma optical emission spectrometry, or ICP-OES) and had few recommendations on the laboratory procedures; rather, the concerns centered around the claimed [stated] error rates and the documented claims of "same source" identifications using the FBI's "match" procedure.

Another major concern that was raised was the existence of thousands or even millions of bullets with similar chemical "signatures" simply due to the consistency in the lead manufacturing process: large homogenized batches of lead are likely to

yield very similar concentrations for the thousands or millions of bullets created from a homogeneous batch. Further, once made into bullets and packaged into boxes of 25 or 50 bullets per box, no box could be guaranteed to have bullets originating from only one batch of lead. Hence, a definitive statement such as "this bullet must have come from this box of 50 bullets" could not be supported, knowing that hundreds of other boxes likely contained bullets with the same signature. Moreover, bullets that did *not* satisfy the "match rule" did not guarantee that the two bullets came from different boxes, as bullets from different batches might have ended up in the same box. Thus, the procedure was useful for neither "inclusion" nor "exclusion" (for more about CABL, see [94, 37]).

1.3 Forensic Glass Analysis

1.3.1 Background and History

Glass is defined as "an inorganic product of fusion which has cooled to a rigid condition without crystallizing" [50], of which a typical melting tank may produce several hundred tons per day [13]. Different types of glass are manufactured for various purposes, including but not limited to flat/sheet glass, float glass, toughened/tempered glass (safety glass which shatters without sharp edges or points [50] and may be four to five times stronger than non-tempered glasses [106]), and laminated glass (required for vehicle windshield glass in the U.S. for safety reasons [13]).

When glass shatters, fragments may eject in all directions and land on any person within a few feet [13, 74]. Glass fragments may be retained on clothing but tend to fall off over time. These fragments, which may be extremely small, are collected by forensic examiners for use in analysis.

> *It is clear that the only conclusive proof of the origin of a fragment of glass is the finding among the "comparison" glass of a piece or pieces with which the "exhibit" fragment shows a perfect fit.*
> – F. G. Tryhorn [103]

The early analyses of forensic glass evidence examined how glass physically broke. Glass can be regarded "as a slightly elastic medium of uniform composition" [103] that bulges, causing one side becomes convex and the other concave, upon the application of force. When the force exceeds the tensile forces of the glass' surface, the glass breaks. If possible, broken fragments were pieced together – the only "conclusive" proof of fragments having a common origin [103, 36, 73]. This method becomes impossible if fragments are too small or missing.

Physical, optical, and chemical properties of glass

Before chemical analysis of glass was performed on a large scale, examiners compared the physical and optical properties of glass. These include properties of glass such as thickness, color, edge marks and hatch marks, fluorescence, specific gravity, and refractive index (RI). In the 1980s, refractive index was considered among the most discriminating methods for comparing glass fragments. However, following improvements in the manufacturing process, the difference in refractive index between glasses from manufacturers and machines decreased, rendering refractive index less discriminating.

At the same time, chemistry equipment was becoming cheaper and required less labor and time to operate. The widespread availability of such machinery pushed forensic glass analysis in the direction of chemical concentration analysis for forensic glass discrimination, which was shown to be effective when RI and other physical

or optical properties were inadequate [12, 19, 20, 63, 109, 64, 28, 60, 61]. Although challenges of using chemical analysis remain today, Koons and Buscaglia (1999) [60] state that:

> The forensic scientist should use the most discriminating technique available in the examination of glass or other form of trace evidence because it is the most effective means of both avoiding false associations and excluding two similar, but separate, sources. *It is in the best interest of the court for the scientist to use the most discriminating analytical technique even if this means that exact probability figures for a conclusion cannot be calculated.* In cases where the analytical discrimination is very good, as in compositional measurements of glass, factors such as manufacturer distribution of products and age and breakage of glass objects in the crime scene and suspect environments are more significant than the probability of two randomly selected sources from a large glass population having coincidentally indistinguishable characteristics. These factors can either be determined by standard investigative techniques or they involve everyday experiences of the nonscientist. *As a result, their significance can be readily weighed by the trier of fact without resorting to statistical calculations.* [emphasis added]

The job of a statistician is to understand the validity of the claims made in this paragraph; specifically: probabilities for a conclusion are needed in order to assure that (1) the most discriminative analytical technique has been used; (2) experiments can and should be designed to quantify the magnitude of effects of factors such as "manufacturer distribution of products and age and breakage of glass objects in the crime scene and suspect environments" to determine whether these factors are

indeed "more significant than the probability of two randomly selected sources from a large glass population having coincidentally indistinguishable characteristics"; and (3) the trier of fact cannot judge their significance without resorting to statistical calculations.

The measurements of chemical concentrations in glass can be conducted using any of several methods, including spectrographic analysis [36, 12, 93], neutron activated analysis (NAA) [42, 49, 19, 43], spark source mass spectrometry [42, 12, 19], dilution spark source mass spectrometry [42], flame atomic emission spectrometry [43], energy dispersive X-ray (EDX), inductively coupled plasma atomic emission spectrometry (ICP-AES) [20, 43, 63, 64, 16, 60, 61, 29], inductively coupled plasma mass spectrometry (ICP-MS) [116, 28, 29], and laser ablation ICP-MS (LA-ICP-MS) [102, 7, 65, 101, 9].

1.3.2 ASTM Standards

> *When small fragments of glass are recovered in the investigation of breakings or of motor accidents the problem presented is usually that of determining whether the fragments came from a given source...* ***conclusive*** *proof of the origin of the fragments is difficult to obtain. As is so often the case with circumstantial evidence the result of the examination usually indicates the* ***probable*** *rather than the* ***actual*** *source of the glass.*
>
> – F. G. Tryhorn [103]

An approach similar to CABL's "2-SD-overlap" procedure has been recommended for comparing glass samples found at a crime scene (the "known" K fragment) with those found on or in connection with a potential suspect (the "questioned" Q or "recovered" R fragment). Using measured trace elemental concentra-

tions from these glass fragments, the standards provide a method for determining if two fragments are "analytically distinguishable" and "it may be concluded that they did not originate from the same source of broken glass" [51, 52, 53]. These three ASTM standards outline the same general process but differ in the instruments and techniques used to measure and process glass samples and the number of elements that should be analyzed (8-17).[2]

- ASTM E2330-12, *Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*
- ASTM E2927-16, *Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons*
- ASTM E2926-13, *Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*

For simplicity, these methods are denoted by ICP-MS, LA-ICP-MS, and μ-XRF, respectively. Where applicable, samples from completely different panes of glass will be referred to as "samples," and pieces from the same pane of glass as "fragments," because more consistency is expected in fragments from a single pane than between samples from different panes. Indeed, this difference in consistency forms the basis of the ASTM standards on forensic glass evidence.

[2]The elements in the "signature" differ for each standard. Standard ASTM E2330-12 for ICP-MS recommends 14 elements: magnesium (Mg), aluminum (Al), iron (Fe), titanium (Ti), manganese (Mn), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba), lanthanum (La), cerium (Ce), neodymium (Nd), samarium (Sm), and lead (Pb). ASTM E2927-16 for LA-ICP-MS recommends all of the same except Sm, plus lithium (Li), potassium (K), calcium (Ca), and cerium (Ce) for a total of 17 elements. The standard for XRF is less specific; see ASTM E2926-13 Section 10.6.2.1 [52].

Each standard includes a section entitled "Calculation and Interpretation of Results." The steps in this section are similar in each standard; below are those for ASTM E2330-12 (ICP-MS):

10.1.1 *For the Known source fragments, using a minimum of 3 measurements, calculate the mean for each element.*

10.1.2 *Calculate the standard deviation for each element. This is the Measured SD.*

10.1.3 *Calculate a value equal to 3% of the mean for each element. This is the Minimum SD.*

10.1.4 *Calculate a match interval for each element with a lower limit equal to the mean minus 4 times the SD (Measured or Minimum, whichever is greater) and an upper limit equal to the mean plus 4 times the SD (Measured or Minimum, whichever is greater).*

10.1.5 *For each Recovered fragment, using a minimum of 3 measurements, calculate the mean concentration for each element.*

10.1.6 *For each element, compare the mean concentration in the Recovered fragment to the match interval for the corresponding element from the Known fragments.*

10.1.7 *If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not "match" and the glass samples are considered distinguishable.*

For ASTM E2927-16 (LA-ICP-MS), "Calculation and Interpretation of Results" appears as Section 11, also with a "4-SD match interval"; for E2926-13, Section 10.7.3.2 uses a "3-SD match interval":

> *"For each elemental ratio, compare the average ratio for the questioned specimen to the average ratio for the known specimens* $\pm 3s$*. This range corresponds to 99.7% of a normally distributed population. If, for one or more elements, the average ratio in the questioned specimen does not fall within the average ratio for the known specimens* $\pm 3s$*, it may be concluded that the samples are not from the same source."*

It should be noted that the "99.7%" coverage applies only if the standard deviations were known, not estimated – as it is here, from possibly as few as three measurements – and only if the measurements come from a Gaussian (normal) distribution.

Justification for these procedures [26, 108, 102, 61] appears to be based on empirically observed "error rates" calculated from all possible pairs of glass samples from different sources. For example, Weis et al. [108] measured 62 different glass samples, mostly from different manufacturers, but some from the same manufacturer produced from different batches at different time periods. The "error rate" was then calculated as the proportion of all pairs that satisfied the "match" criterion, even though the two came from different sources. Comparing each one of the 62 samples as the K with any one of the other 61 samples as the Q, they found two of the 1,891 pairs satisfied their "modified n-sigma criterion with fixed relative standard deviations (FRSDs)," giving a Type II error rate of 0.11%, where the FRSDs varied between 3.0% and 8.9%. Dorn et al. [26] used a similar "4-SD match criterion," but with an RSD_{min} set to 3% for the concentrations of the 10 elements in their study.

They found similarly small error rates: 0.27% (6/2256, 48 same-source samples)[3] for their Type I error rate (two samples from same source failed to satisfy the "match" criterion), and 0.11% (7/6642, 82 different-source samples) for their Type II error rate (two different samples satisfied the "match" criterion).

Reported error rates from four commonly referenced papers are very low, typically less than 1% [26, 108, 102, 61]. These error rates are calculated, as above, by comparing two different-source samples from among all possible pairs in the data set, and marking a comparison as a "false positive" if sample means for all 17 elements from the R fragment fall within the "mean ± 4·SD" interval created from the K fragment. In this process, the same sample is used for multiple different-source comparisons, and the match or no match conclusion may differ if samples i and j are the R or K fragments, respectively. False positive rates (FPR) calculated in this manner also depend heavily on the specific data set used, some of which are purposefully created to be diverse. If samples in the data set are all highly similar samples (e.g., Toyota windshields) manufactured at similar times, the expected mean difference between these samples will be lower than if samples were very different (e.g., car windshields and baby food jars). The estimated FPR for the first data set may well be higher than that of the second data set.

While the FBI's CABL procedure involved calculating "mean ± 2·SD" for both the K and R specimens, the glass standards calculate instead "mean ± 4·SD" for only the K fragment and check to see if the means for the R fragment fall in that interval. In other words, only one set of standard deviations (for the K fragment) is calculated, and variability in the R fragment is not taken into consideration. The glass standards also recommend the use of 8–17 elements, not just seven. Correlations between the

[3]Note that Dorn et al. (2015) [26] actually measured 24 fragments 9 times each, and a 25^{th} fragment 24 times, which is quite different from 48 fragments. See page 87, "*Group 1*", for details.

elements, which are not insignificant, are also not accounted for; and exploratory data analysis plots have indeed shown that certain corresponding pairs of elements are highly correlated. Thus, individual "match intervals" cannot be treated as independent.

Ideally, a multivariate version of Student's t such as Hotelling's T^2 is conducted to account for the issues mentioned above. Unfortunately, sample sizes are usually not large enough to allow for such testing. ASTM E2330 above requires a minimum of three measurements per sample, while ASTM E2927 requires "a minimum of 9 measurements (from at least 3 fragments, if possible)" [53]. None of the data sets available to us for analysis has more than nine replicates, yet the number of elements varies from 8–17 depending on the standard. Thus, Weis et al. (2011) [108] dismiss the use of Hotelling's T^2 statistic for assessing the significance of the measurement difference in elemental concentrations in two samples:

> *Hotelling's T^2-test, a multivariate equivalent of Student's t-test, has the disadvantage that for mathematical reasons the number of factors (replicate measurements on the control sample plus replicate measurements on the recovered sample) must be at least larger by two than the number of dimensions (i.e. the number of element concentrations, in our case 18). Therefore, at least 10 replicate measurements of both samples to be compared must be conducted for the Hotelling's T^2-test to be applicable. If only six replicate measurements are carried out for each of the two samples to be compared, the number of elements used for the comparisons has to be reduced to 10, which leads to a loss of evidential value. Hence, Hotelling's T^2-test calculations will not be addressed in this paper.*

In fact, having fewer replicates than elements does not relieve us of the need for more replicates, because we still need to estimate the correlations in the measurements among the different elements. Forensic glass experts are well aware of the correlations among certain elements, based on their chemical properties.[4] The correlation (or covariance) matrix is used explicitly in Hotelling's T^2 statistic, but even if not used explicitly, knowing the correlations between pairs of elements removes any temptation to treat individual "match intervals" as independent.

1.4 Data Introduction

Our motivation stems from roughly normal (or lognormal) glass data which tends to contain outliers, is of small sample size n, and has dimension p of similar magnitude to n. However, forensic science is a discipline where data may likely be inaccessible; and the proposed methods may extend not only to data sets of higher dimensionality (large n and large p), but also to cases where only covariance (or correlation) information is available.

For a few reasons, we are interested in logarithms of the glass data. First, chemists tend to refer to the "relative standard deviation" (RSD) rather than raw SD, as the SD of elemental concentrations tends to be related to the mean. For example, six measurements of ^{7}Li might be very different from six measurements of ^{90}Zr, which has means and SDs around 11 times larger, but whose RSDs are very similar (see Table 1.1). Second, elemental concentration measurements may have slightly skewed distributions, while the distributions of the logarithms tend to be more symmetric. The data we receive from laboratories has typically undergone

[4]For example, the very high correlation between hafnium and zirconium is well known: not surprising to chemists, as Hf and Zr are near each other in the periodic table, as are other pairs of elements.

some basic manipulation (e.g., background correction), on which we further take logarithms.

raw data	1	2	3	4	5	6	mean/SD	RSD
^{7}Li	4.56	4.68	4.79	4.25	4.33	4.49	0.205/4.517	**4.54%**
^{90}Zr	54.16	55.25	51.93	50.13	49.97	49.44	2.416/51.813	**4.66%**
log data	1	2	3	4	5	6	mean	SD
log(^{7}Li)	1.517	1.543	1.567	1.447	1.466	1.502	1.507	**4.54%**
log(^{90}Zr)	3.992	4.012	3.950	3.915	3.911	3.901	3.947	**4.62%**

Table 1.1: Means, SDs, and RSDs for six measurements of ^{7}Li and ^{90}Zr. The estimated SDs are approximately the RSDs.

A toy data set containing 10 elements is shown in Table 1.2. The first three rows indicate three measurements taken on a known glass sample. The next sets of rows are calculations of, as the standards instruct, the means, standard deviations, and 3% of the mean to give a lower bound for the SD. The lower and upper bounds of the created interval can then be calculated using "mean ± 4·max{0.03*mean, SD}" to create the match interval. The last row shows elemental concentrations means from three measurements of a recovered (or questioned) glass fragment. As concentrations for two of the elements fall outside the interval (in this case, larger than the upper interval bound), according to step **10.1.7**, these glass samples "are considered distinguishable" [51].

1.5 Data Sets

In this section we describe the four glass data sets that formed the basis for our research motivation and some preliminary exploratory analyses. The first three data sets are measured using LA-ICP-MS while the fourth uses ICP-MS. Since the elements measured using these methods differ, we use these two groups of data

	Mg	Al	Fe	Ti	Mn	La	Ce	Nd	Sm	Pb
K fragment 1	30500	2217	4169	206	112	2.994	5.728	2.54	0.542	1.086
K fragment 2	30110	2150	4213	194	111	3.034	5.648	2.81	0.68	1.056
K fragment 3	30580	2226	4155	208	115	2.954	5.69	2.58	0.89	1.13
$\text{mean}_K(x) = \bar{x}_K$	30396.67	2197.67	4179	202.67	112.67	2.99	5.69	2.64	0.7	1.09
$\text{sd}_K = \sigma_K$	251.46	41.53	30.27	7.57	2.08	0.04	0.04	0.15	0.18	0.04
$0.03\bar{x}_K$	911.9	65.93	125.37	6.08	3.38	0.09	0.17	0.08	0.02	0.03
$\max\{0.03\bar{x}_K, \sigma_K\}$	911.9	65.93	125.37	7.57	3.38	0.09	0.17	0.15	0.18	0.04
lower bound	26749.07	1933.95	3677.52	172.39	99.15	2.63	5.01	2.04	-0.02	0.93
upper bound	34044.27	2461.39	4680.48	232.95	126.19	3.35	6.37	3.24	1.42	1.25
$\text{mean}_R(x)$	30400	2321	4590	**240**	112	2.4	5.8	3	0.7	**1.3**

Table 1.2: An illustration of ASTM E2330 with a toy data set. The measured concentrations in two elements from the recovered glass fragment, Ti and Pb, fall outside the upper bound of the $4 \pm \max\{0.03\bar{x}_K, \sigma_K\}$ interval; thus these glass samples "are considered distinguishable."

separately in analyses.

The data provided by each laboratory may have previously undergone some basic manipulation (e.g., background correction, usually involving subtracting a lower limit). Logarithms are further taken to help normalize the data, especially given the small sample sizes.

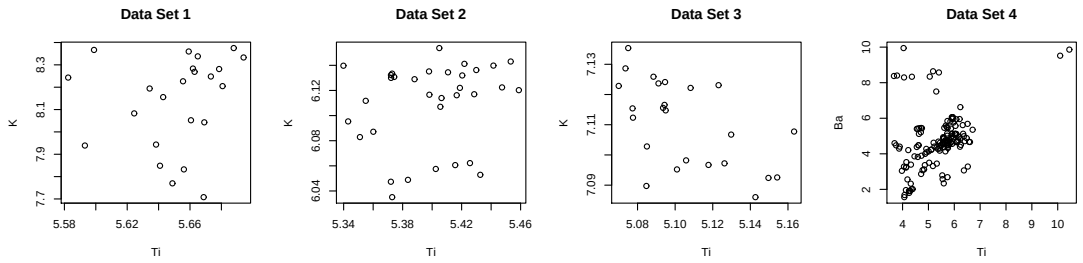


Figure 1.1: Outliers. Points indicate mean log elemental concentrations averaged across replicate measurements per fragment (ranging from 3 to 9 with the exception of certain fragments in Data Set 3). The first three plots show Ti against K with $n = 24, 33, 24$ (one pane chosen at random from Data Set 3). The plot for Data Set 4, which did not include K, is of Ti against Ba for Container glass ($n = 160$). Outliers can be seen in almost all elements.

1.5.1 Data Set 1 – Canadian data set

Dr. David Ruddell (Centre of Forensic Sciences, Toronto, Canada) kindly shared the data from Dorn et al. (2015) [26]. The data from the "pane study" consisted of 48 "samples" taken from a single 4' × 6' pane of glass: 24 fragments were cut from the glass pane and measured 9 times each; a 25^{th} fragment was measured 24 times (see page 87, "*Group 1*," for details). These fragments spanned the entire pane of glass. The 23 elements measured include all 17 elements cited in E2927-16, plus silicon (^{29}Si), cobalt (^{59}Co), tin (^{118}Sn), thorium (^{232}Th), and uranium (^{238}U). These data can be used to estimate within-fragment variability (from the 25^{th} fragment measured 24 times).

1.5.2 Data Set 2 – German data set

Dr. Peter Weis (Bundeskriminalamt/Federal Criminal Police Office, Forensic Science Institute, KT 42 – Inorganic Materials and Microtraces, Coatings, Wiesbaden, Germany) kindly shared the data that were published in Weis et al. (2011) [108]. Each fragment was measured 6 times, which allows for reliable estimates of within-fragment variability for all 20 elements that were measured. The elements include all 17 of the elements cited in E2927-16, plus sodium (^{23}Na), tin (^{118}Sn), and silicon (^{29}Si, as the constant standard). In this collection, data set (A) "Same" consisted of 33 fragments from the upper triangular half of the same pane of glass, plus a 34^{th} fragment that was measured 6 times on each of 11 consecutive days, permitting rough estimates of between-fragment variability (among the 33 fragments) and between-day variability (among the 11 days). We also can verify that the within-fragment variability (measurement variation among the 6 replicates) is consistent with the within-day variability (also 6 replicates each on 11 days).

1.5.3 Data Set 3 – ISU data set

Dr. Soyoung Park and Dr. Alicia Carriquiry of Iowa State University (ISU) kindly shared data on float glass samples from two manufacturers [77, 78]. The 18 elements measured included all 17 cited in ASTM E2927-16 as well as sodium (^{23}Na). Data from 48 glass panes were collected, 31 from Company A produced over a three week period, and 17 from Company B produced over a two week period. From each pane, 24 fragments were randomly sampled, with 21 fragments having five replicate measurements and three fragments having 20 measurements. These data provide valuable information on within- and between-manufacturer variability. Exploratory analyses suggest that for within-manufacturer panes, at least half the elements vary by at least $\delta = 0.1$, or roughly 10% on the log scale. Between-manufacturer panes differ in certain elemental concentrations for just over half of the 18 elements (see Figure 1.2).

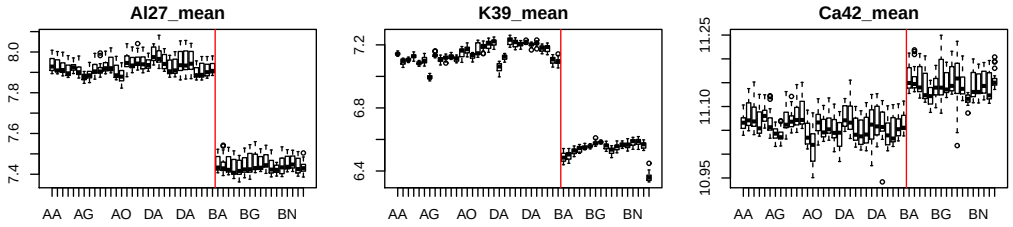


Figure 1.2: Log concentrations for three elements from Company A (left of vertical red line) and Company B (right of vertical red line). Each vertical boxplot represents concentrations from a single pane (48 panes total). Relatively clear splits between manufacturers similar to those shown here can be observed in just over half of the 18 elements; concentrations of ^{27}Al, ^{39}K, ^{42}Ca (and other elements) from the 31/17 different panes within-manufacturer tend to be (but are not always) very similar.

1.5.4 Data Set 4 – FIU data set

Data were obtained from Florida International University (FIU) in which concentrations of 16 elements were measured on multiple glass samples via ICP-MS [5].

The elements include 13 of the 14 cited in E2330-12 (with two isotopes of strontium, ^{86}Sr and ^{88}Sr) minus neodymium (Nd), plus antimony (^{121}Sb and ^{123}Sb), gallium (^{71}Ga), and hafnium (^{178}Hf). Each sample had 3 measurements, and the collection of 590 samples included seven types of glass: 160 Container glass samples, 189 Float Architecture, 46 Float Autowindow (CFS), 97 Float Autowindow (non-CFS), 45 Headlamp, 10 Laboratory, and 43 "Rare." Not all types of glass had elemental concentration measurements for all 16 elements.

Because the types of glass are so different, from container to decorative architectural to automotive, and because the measurement technique (ICP-MS) differs from the previous three data sets, these data cannot be combined with the previous three data sets, and any analyses will be performed separately. However, this data set confirms the high (generally positive) correlation between elements. Table 1.3 shows several pairs of elements with consistently high correlations across all glass types.

	Ce-La	Ce-Sm	La-Sm	Mn-Sm	Ba-Mn	Ba-Sm	Mn-Ti	La-Mn	Sm-Ti
Container	**0.98**	**0.92**	**0.94**	**0.83**	0.47	0.65	0.63	**0.83**	*0.74*
Float Arch*	**0.96**	**0.92**	**0.95**	*0.76*	*0.70*	**0.82**	*0.77*	*0.70*	0.43
Float Auto (CFS)*	—	0.37	—	**0.87**	**-0.83**	*-0.75*	*-0.77*	—	*-0.73*
Float Auto (non CFS)*	**0.89**	**0.92**	**0.95**	**0.83**	**0.83**	**0.92**	*0.79*	**0.81**	**0.87**
Headlamp	**0.98**	**0.96**	**0.92**	-0.32	0.17	0.37	-0.23	-0.29	0.48
Lab†	**0.98**	**1.00**	**0.98**	*0.71*	**0.97**	**0.86**	**0.88**	**0.82**	**0.96**
Rare	**0.99**	**0.92**	**0.95**	0.42	**0.90**	*0.72*	0.54	0.41	*0.80*

Table 1.3: Robust correlations for FIU data (*Italic:* $0.7 \leq |x| < 0.8$, **Bold:** $0.8 \leq |x| \leq 1$).
*Float Auto (CFS) does not have measurements for La; all three Float glass types do not have measurements for Sb. †Lab has only 10 samples (and some missing values), not enough to calculate robust correlations. Classical correlation values are shown.

1.6 Goals and Outline of Dissertation

All four of these data sets to which we have access are all of small sample size yet relatively large dimension (e.g., 24×17). A combined overall data set with increased sample size would likely produce much better and more stable estimates. So, we would like to consider situations in which data sets are sufficiently similar enough to combine – then combine them. Two questions thus arise: (1) what does it mean for data sets to be "similar"; and (2) assuming adequate "similarity" has been determined, what is the best way to "combine" (pool) data sets?

One issue with combining data sets is lack of access to data. While we were kindly granted permission to use the data sets mentioned above, there are many situations in which data would likely be inaccessible. Forensic scientists and laboratories may be extremely hesitant to share case data on information concerning or related to criminals or criminal activity that may be highly classified; or out of reluctance for statisticians to calculate error rates for their methodology and practices. Researchers in psychology and related areas with human experimental data likewise may be restricted from distributing data due to privacy concerns. Companies may be unwilling to provide access to large, unique data sets that they have arduously collected over long periods of time. In other cases, one may wish to conduct a meta-analysis of multiple papers, in which full data sets may not be published but covariance or correlation matrices may be. (This was the case with Data Set 2 [108]; the paper contained correlation matrix information, but we had to request and were almost denied the data.) Thus, we propose combining covariance matrices directly instead of data sets. As the data motivating our work are of small sample sizes, we believe that combining covariance matrices may also help to mitigate some of the variability and effect of outliers.

The question that naturally follows is what it means for covariance matrices to be "similar" or "equal." We are interested in something analogous to the F-test for equal variances, but for covariance matrices. Furthermore, assuming two (or more) covariance matrices are similar enough to be combined, (1) how should they be combined; and (2) how can we tell if the resulting combination is a good estimator?

This thesis explores approaches to address these questions, which led to additional considerations regarding effective dimension of subspaces spanned by our covariance matrices.

Chapter 2 reviews existing literature on testing for equality and similarity of covariance matrices, combining covariance matrices, robust covariance estimators. We will consider the limitations created by restrictive assumptions and computational complexities, and discuss why robust covariance estimators are unsuitable for our data. We will then discuss methods for analyzing the subspaces matrices span instead of the matrices themselves, and argue this is the preferred way for comparing and combining covariance matrices.

In Chapter 3 we will introduce our proposed methods. Two simple metrics for comparing similarity of covariance matrices and determining true effective dimension based on the singular value decomposition (SVD) are introduced, as well as a method for combining covariance matrices based on the subspaces they span. These metrics and methods are illustrated with simulation data.

In Chapter 4, the proposed methods are applied to our motivating forensics glass data and simulation data based on these motivating data. Our results indicate that the proposed methods can be used for small sample data as well as larger data sets.

Chapter 5 concludes and discusses potential future research directions concerning our proposed metrics and methods.

Chapter 2

Literature Review

The first section focuses on methods based on covariance matrices, and the second section introduces methods based on the subspaces these covariance matrices span.

2.1 Covariance Based Methods

First, we review several tests that have been proposed for testing the equality of covariance matrices. Next, we look at literature that attempts to combine covariance matrices; many originate from DNA microarray data, where data collection is expensive. Lastly we cover robust covariance estimators which downweight the effect of outliers, and a brief discussion on the limitations of these existing methods.

Robust covariance estimates would be useful for combining covariance matrices if they are known to come from the same Σ, especially in elemental composition data where outliers have been observed and are known to exist. However, we must first decide if matrices are similar enough to combine; and if so, whether to calculate robust estimates using the full matrices or a dimension-reduced version of them. For example, if the true dimension of a 17×17 covariance is only six (the other

11 dimensions are effectively "noise"), then it would be more practical to consider a robust estimator of the reduced-dimension covariance matrix.

The end of this section examines robust covariance estimates that could be applied to similar matrices, and leads into the next section discussing methods for determining similarity and effective dimension based on subspaces that provide more meaningful information on covariance structure.

2.1.1 Equality of Covariance Matrices

The assumption of equal covariance matrices is common in certain multivariate analysis tests such as Hotelling's T^2 test, multivariate analysis of variance, and discriminant analysis. In other cases, testing for equality may be used as an initial test for determining if further analysis is needed. While many tests and test statistics have been proposed for testing the equality of covariance matrices, they are generally very complicated and have many assumptions.

The likelihood ratio statistic for testing $H : \Sigma_1 = ... = \Sigma_k = \Sigma$ rejects the null hypothesis if test statistic

$$\Lambda = \frac{\prod_{i=1}^{k} det(A_i)^{N_i/2}}{det(A)^{N/2}} \cdot \frac{N^{nM/2}}{\prod_{i=1}^{k} N_i^{mN_i/2}} \quad \leq \quad c_\alpha$$

where

$$\bar{X}_i = \frac{1}{N_i} \sum_{i=1}^{N} X_{ij}, \quad A_i = \sum_{i=1}^{N} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)', \quad A = \sum_{i=1}^{k} A_i, \quad N = \sum_{i=1}^{k} N_i$$

and c_α is chosen so the test is of size α [72]. Although this test may be considered asymptotically unbiased as $n \to \infty$, it is known to be biased when sample sizes are unequal [75, 6, 17, 80, 99]. Bartlett's modified likelihood ratio test statistic

uses the degrees of freedom instead of number of observations as weights [8, 81, 99, 72]. The test statistic itself may be reasonable for elliptical or other moderately nonnormal distributions, but the asymptotic χ^2 distribution may not hold [114]. While unbiased, these likelihood ratio based tests are very sensitive to violations of normality and "may not be appropriate to use the likelihood ratio test if $[p]$ is much larger than $n_i/2$" [91]. Several more robust procedures were proposed to combat this sensitivity [100, 76, 114].

Schott 2001 [90] developed several Wald tests for $H_0 : \Sigma_1 = ... = \Sigma_k = \Sigma$ where Σ is an unknown common covariance matrix. Schott considered four assumptions for the k populations with m parameters (dimension) and their corresponding test statistics. Letting S_i denote the unbiased sample covariance matrix given by $S_i = \frac{1}{n_i}\sum_{j=1}^{N_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$ and $S = \sum_{i=1}^{k}\frac{n_i}{n}S_i$, these test statistics take on the general form

$$T_2 = n\left(\sum_{i=1}^{k}\left\{\frac{1}{2}\hat{\delta}_1\gamma_i tr(S_iS^{-1}S_iS^{-1}) - \hat{\delta}_2\gamma_i tr(S_iS^{-1})^2\right\}\right.$$
$$\left. - \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{1}{2}\hat{\delta}_1\gamma_i\gamma_j tr(S_iS^{-1}S_jS^{-1}) - \hat{\delta}_2\gamma_i\gamma_j tr(S_iS^{-1})tr(S_jS^{-1})\right)$$

and follow an asymptotic χ^2_ν distribution with degrees of freedom $\nu = \frac{(k-1)m(m+1)}{2}$ (see [90] for specifics on test statistics T_1, T_3, T_4). Each of the test statistics, respectively, has differing assumptions:

1. T_1. Populations follow multivariate normal distributions. This test statistic is an alternative yet asymptotically equivalent (i.e., $M = T_1 + o_p(1)$) test to the likelihood ratio test, and thus also suffers from sensitivity to violations of normality.

2. T_2. Populations have elliptical distributions with common kurtosis parameter κ. This statistic should perform better than M and T_1 when the underlying distributions are not normal, and holds where $\kappa > \frac{-2}{m+2}$.

3. T_3. Populations have elliptical distributions but different kurtosis parameters.

4. T_4. Distributions have finite fourth moments.

Simulation evaluations of these statistics suggest that, as expected, T_1 performs best when the assumption of normality holds, although T_2 and T_3 perform reasonably well when $n \geq 20$. T_2 is fairly robust to violation of common kurtosis and is more and less powerful than Zhang and Boos [114] for the normal and multivariate t distributions, respectively. T_4 converges slowly to the asymptotic distribution and has inflated significance levels. T_2 appears to be the most useful of the tests, as it is more robust under normal theory, computationally simpler, and performs better than T_3. T_4 may not be appropriate unless m is very small or n very large. However, these test statistics are complicated and in addition to the restrictive assumptions mentioned above, require kurtosis parameters (T_2 and T_3) or estimates of M_{4i} (T_4); and the choice of estimates may significantly impact the significance level for small samples. The statistic mentioned as most useful, T_2, assumes elliptically distributed data with common kurtosis. Schott (2007) recognized the need for a similar statistic that would function in "[a] more realistic situation" where "[p] is not particularly large, but the n_i's are very small" [91]. This is a common occurrence in DNA microarray data, where data may involve thousands of gene expressions on around 100 people. However the proposed statistic [91], which performs when the likelihood ratio test cannot and is asymptotically normally distributed, assumes populations are multivariate normal.

Srivastava (2005, 2007) [95, 96] and Srivastava and Hirokazu (2010) [97] proposed multiple tests concerning covariance matrices. The two tests proposed in [97] are most relevant to the situation arising from glass data, and are roughly based on differences between the traces of two squared covariance matrices ($tr\{\Sigma_1^2\}$ and $tr\{\Sigma_2^2\}$), which may "throw a light on the differences between the two covariances." These tests can be used in all situations (does not require $p >> n$), but assume populations are normally distributed and are rather complicated; requiring use of the Moore-Penrose inverse and consistent estimators. Proposed test statistics (denoted by T_k^2 and Q_k^2) are compared to Schott (2007) [91] (denoted by J_k). All three perform well if n, p are large, and J_k and Q_k^2 perform well (power-wise) for small n, p. Srivastava and Hirokazu (2010) define the Attained Significance Level (ASL) metric to measure how close the empirical and asymptotic distributions of test statistics are. The ASL of J_k fluctuates substantially, thus Q_k^2 may be preferred.

More recent proposed tests include Li and Chen (2012) [67] based on the squared Frobenius norm $tr\{(\Sigma_1 - \Sigma_2)^2\}$, Cai et al. (2013) [18] based on the maximum of standardized differences between entries in two covariance matrices, and Bilgrau et al. (2018) [10] for high inter-study homogeneity.

2.1.2 Combining Covariance Matrices

The simplest method for combining covariance matrices is a pooled or weighted average estimate of the sample covariance matrices, such as that used in Hotelling's T^2. This method implicitly assumes the existence of a global covariance matrix and may not be applicable if subgroups are clearly distinct or batch effects exist [10]; cf. Data Set 3 in Figure 1.1. While the classical sample covariance estimate S is consistent and efficient under the normal model, it is not at all robust to outliers

[82]. Our data sets of interest are known to have outliers and small sample sizes; thus pooling unstable estimates is not ideal.

However, there are many circumstances where it would be ideal to "pool" or combine covariance matrices (or data, if available). Pooling data across studies or labs can increase sample sizes, which may in turn increase power [54]. Even if data is unavailable, the information contained in "similar" covariance matrices can be combined to obtain as much information as possible. A common issue with DNA microarray data is the existence of batch effects, meaning data cannot be directly combined unless first processed in some manner [54, 66]. However, this involves some background knowledge or assumptions that batch or lab effects exist as well as a general idea of how they should be treated. Unfortunately, we lack this knowledge for glass data, and hope to propose a method not requiring such existing knowledge.

Bilgrau et al. (2018) [10] propose a hierarchical random covariance model (RCM) for the meta-analysis of gene correlation networks from different studies. Individual p-dimensional covariance matrices are assumed to come from a common inverse Wishart distribution, and study data is then generated from a multivariate normal distribution given each individual covariance matrix.

$$\boldsymbol{\Sigma_i} \sim W_p^{-1}(\boldsymbol{\Psi}, \nu), \qquad \boldsymbol{x}|\boldsymbol{\Sigma_i} \sim N_p(\mathbf{0}_p, \boldsymbol{\Sigma_i}), \quad i = 1, ..., k$$

The maximum likelihood estimator for the underlying common covariance matrix is then estimated by an EM algorithm. The proposed method is complex, not easily interpretable, and "computationally demanding and only feasible when p is sufficiently small" (and not generalizable to $p \gg n$). Furthermore, the estimator "perform[s] better or not worse than a simple pooled estimator" [10].

2.1.3 Robust Covariance Estimators

The classical sample covariance estimate S is not robust to outliers and has an asymptotic and explosion breakdown value of zero [82]. Because "REAL DATA OFTEN FAIL to be Gaussian IN MANY WAYS" [14] and "we know that the parametric model is not quite true" [40], alternative, robust estimators are necessary for "[safeguarding] against unsuspectedly large amounts of gross errors [and] putting a bound on the influence of hidden contamination and questionable outliers" [41].

Robust covariance estimators fall into two general families: (1) high-breakdown affine equivariant estimators and (2) non-affine equivariant estimators [82, 47]. Generally, it is recommended to choose a high breakdown affine equivariant estimator if $p < 10$ and a non-affine estimator (OGK, DetMCD) for higher dimensions.

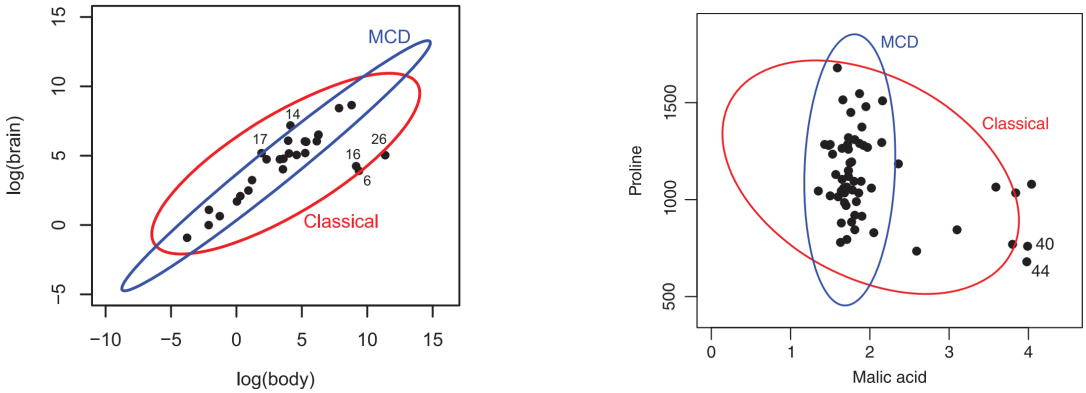
Affine Equivariant Estimators

Affine equivariant scatter estimators by definition behave well under linear transformations of the data. They have a sharp upper bound finite-sample breakdown value (FSBV) of $\lfloor (n-p+1)/2 \rfloor / n$ [23, 85], and begin calculations with a large numbers of random initial subsets. This causes computation time to increase exponentially as dimension increases – meaning these procedures may not be ideal if $p \geq 10$ [47, 82]. The Stahel-Donoho estimator [98, 25, 69], calculated as a weighted mean and covariance matrix based on Stahel-Donoho outlyingness (univariate and multivariate),

$$SDO_i = SDO^{(1)}(x_i, X_n) = \frac{|x_i - med(X_n)|}{MAD(X_n)}$$

$$SDO_i = SDO(\boldsymbol{x}_i, X_n) = \sup_{\boldsymbol{a} \in \mathbb{R}^p} SDO^{(1)}(\boldsymbol{a}\boldsymbol{x}_i, X_n\boldsymbol{a})$$

was the first affine equivariant estimator with a breakdown point of 50%. This estimator performs best if non-outlying values are roughly elliptical or symmetrically distributed, as the mean absolute deviation (MAD) will not capture asymmetry [89]. Highly contaminated data may result in the estimator being singular; and modifications have been proposed to address these and improve performance for skewed or outlier-containing data [15, 46, 24, 105, 82].



(a) (b)

Figure 2.1: Tolerance ellipses for the classical (red) vs. robust MCD (blue) estimators. Figure (a) comes from [87] and is of the Animal data set of brain vs. body weights. Figure (b) from [45] is of two variables from the Wine data set. In both figures, outliers largely affect the classical estimate whereas the robust MCD estimate ignores such outliers.

More often used in this family estimators are the Minimum Covariance Determinant (MCD, the "mean of the h points of X for which the determinant of the covariance matrix is minimal") and Minimum Volume Ellipsoid (MVE, the "center of the minimal volume ellipsoid covering (at least) h points of X") estimators [83, 84]. The parameter h, constrained to $[(n+p+1)/2] \leq h \leq n$, is the number of observations and controls the breakdown value [82]. As the MCD estimator involves calculation of the determinant of a covariance matrix, h must be larger than dimension p to avoid a determinant of zero; which is satisfied by $n \geq 2p$, but $n \geq 5p$

is recommended to avoid excessive noise [82]. Despite fast implementations of these algorithms (FastMVE [71] and FastMCD [86]), they remain computationally intensive and not recommended for p much larger than 10 or 12. Between the two, the MCD estimator is preferred, as the MVE estimator is not asymptotically normal and harder to calculate than the MCD. The MCD has a finite-sample breakdown value (FSBV) of $\lfloor (n-p+1)/2 \rfloor / n$ [85].

Non-Affine Equivariant Estimators

Non-affine equivariant estimators sacrifice affine equivariance for computation time; thus they scale much better as dimension increases [47].

Maronna and Zamar [70] applied their methodology for obtaining approximately affine equivariant, robust positive definite scatter matrices from pairwise robust covariance matrices to Gnanadesikan and Kettenring's robust covariance estimate [38] to obtain the orthogonalized Gnanadesikan-Kettenring (OGK) estimator. The Deterministic MCD (DetMCD) algorithm, as the name suggests, is a deterministic (permutation invariant) alternative to Fast-MCD with lower runtime [48, 86]. The deviance of both of these estimators from affine equivariance is very small [82].

2.1.4 Shortcomings of Existing Approaches

Many of the existing tests for equality of covariances assume normality or at least elliptically distributed data, which, "however, is still quite strong and hard to verify in small samples" [114]. Although the elliptically distributed assumption may not be an issue for our motivating glass data (because logarithms symmetrized the distributions), other data sets may violate this assumption. Other assumptions concern moderate dimensionality, relating to either n, p, or p/n (while our glass data is

of small to moderate dimensions, we hope methods can scale to larger dimensionality as well). Tests geared towards more high dimensional covariance matrices may have more specific assumptions (e.g., sparsity) or more particular goals (e.g., testing equality of off-diagonal sub-matrices or joint distribution of rows between covariance matrices). In addition to restrictive assumptions, many tests are computationally intensive or intuitively difficult to understand for non-statisticians.

Methods for combining covariance matrices originate mainly from DNA microarray data, and are not ideal for our purposes because (1) while batch or laboratory effects likely exist in our data, the existence and amount of such effects is unknown (and likely of a different manner than that of DNA microarray data) and therefore microarray data preprocessing methods cannot be applied directly to our (and other) data sets; (2) ideally methods do not require such assumptions or existing knowledge; (3) methods should be simpler and more intuitive to facilitate understanding between disciplines.

Although robust covariance estimators have high breakdown points and our data is known to have outliers, they generally assume multivariate normal or elliptically distributed data (the MCD estimator assumes "elliptically symmetric unimodal distribution" data [45]). Large sample sizes ($n > 2p$ or $n > 5p$ [82]) are required to be effective, and the algorithms are computationally intensive especially as dimension increases. While our motivating data is of small dimension, we hope to propose a method that can scale to large dimensionality. Such data might include social media, marketing, or retail companies with data constantly streaming in. These data can be expected to be of very large n with possibly even larger p.

2.2 Subspace Based Approaches

Subspace methods arise largely from image recognition and pattern matching techniques, where the goal is to discover similarity or patterns between subspaces. These techniques are widely used in facial recognition applications, where a subject may be represented by a set of vectors – a subspace – and compared to subspaces generated from other subjects [3, 107]. The Mutual Subspace Method (MSM) [112] and Constrained Mutual Subspace Methods (CMSM) [33] are extensions of subspace methods that use the cosine of the smallest principal (or canonical) angle to compare subspaces for similarity.

Principal angles, also called canonical angles, were introduced by Jordan in 1875 and formalized by Hotelling in 1936 [57, 44, 35, 113]. Intuitively, these angles are the "minimal angles between all possible bases of two spaces" [115]. The principal angles, $0 \leq \theta_1 \leq ... \leq \theta_k$, between two subspaces U_1 and U_2 can be recursively defined for $k = 1, ..., p$ by [44]

$$\cos\theta_k = \max_{u_k \in U_1} \max_{v_k \in U_2} u_k^T v_k, \quad ||u_k||_2 = 1, ||v_k||_2 = 1$$

subject to

$$u_j^T u_k = 0, v_j^T v_k = 0 \text{ for } j = 1, ..., k-1$$

where vectors u_i and v_i are the principal vectors of the subspaces. Principal angles are always uniquely defined (although the principal vectors may not be), and can be computed using the singular value decomposition where the singular values of $U_1^T U_2$ are the cosines of the principal angles [11, 4].

Distance metrics based on the principal angles [110, 113] have been proposed to quantify the distance between subspaces, mainly in mathematics and computer

Metric	Principal angles
Asimov	$d^{\alpha}(U,V) = \theta_k$
Binet-Cauchy	$d^{\beta}(U,V) = \left(1 - \prod_{i=1}^{k} \cos^2\theta\right)^{1/2}$
Chordal	$d^{\kappa}(U,V) = \left(\sum_{i=1}^{k} \sin^2\theta\right)^{1/2}$
Fubini-Study	$d^{\phi}(U,V) = \cos^{-1}\left(\prod_{i=1}^{k} \cos\theta\right)$
Grassmannian	$d^{\gamma}(U,V) = \left(\sum_{i=1}^{k} \theta_i^2\right)^{1/2}$
Martin	$d^{\mu}(U,V) = \left(\log \prod_{i=2}^{k} 1/\cos^2\theta_i\right)^{1/2}$
Procrustes	$d^{\rho}(U,V) = 2\left(\sum_{i=1}^{k} \sin^2\theta_i/2\right)^{1/2}$
Projection	$d^{\pi}(U,V) = \sin\theta_k$
Spectral	$d^{\sigma}(U,V) = 2\sin\theta_k/2$

Table 2.1: Table 2 from [113] including Grassmannian distance.

science literature. Many of these are defined in terms of the Grassmannian. In mathematics, the Grassmannian, sometimes denoted as $Gr(n,k)$, is the set of k-dimensional subspaces in n-dimensional vector space [2]. The Grassmannian manifold $g_{n,k}$ is the space of k-dimensional subspaces in $\mathbb{R}^n$ [1, 104]. Subspaces of k dimension can be considered "points on the Grassmannian $Gr(k,n)$, a Riemannian manifold, and the geodesic distance between them gives [an] intrinsic distance" [113]. Some commonly cited distances are listed in Table 2.1. Other distance metrics include the product angle cosine, Friedrich's angle, Dixmier angles, and the maximum and minimum correlations [35, 39].

Distances and metrics must be invariant under different representations of subspaces. While a "distance" quantifies the length of the path between two points, a "metric" is a distance that satisfies a few additional axioms for all $u_1, u_2, u_3 \in \mathcal{U}$ [39]:

1. $d(u_1, u_2) \geq 0$

2. $d(u_1, u_2) = 0$ iff $u_1 = u_2$

3. $d(u_1, u_2) = d(u_2, u_1)$

4. $d(u_1, u_2) + d(u_2, u_3) \leq d(u_1, u_3)$

In the following, we refer to "distance" and "metric" interchangeably.

Some of these distances are based only on the largest canonical angle while others are functions of all the angles. Although there may be cases where a metric based on all of the principal angles is ideal [39], we focus on the Asimov, or largest principal angle, as it is the simplest to interpret and sufficient for our applications (see additional discussion in Section 3.1.2).

Absil et al. 2006 [4] derive the probability distribution and density functions for the largest principal angle (d^{α} in Table 2.1) between two subspaces of the same dimension from the uniform distribution on the Grassmannian manifold, Grass(p, n). The uniform distribution on the Grassmannian is defined as "the distribution with probability measure invariant under the transformation of Grass(p, n) induced by orthogonal transformations of $\mathbb{R}^n$" [4]. To begin, Absil et al. derive the joint PDF of the eigenvalues of the Cholesky decomposition of the orthonormalized U_2, which has multivariate Beta distribution $\text{Beta}_p\left(\frac{1}{2}n_1, \frac{1}{2}n_2\right)$. The PDF is rewritten in terms of the largest principal angle (which corresponds to the smallest singular value or eigenvalue) x, where $x = \lambda_p$. Since the cosines of the principal angles are the singular values of $U_1^T U_2$, a change of variable using $x = \cos^2\theta_p$ can be made. Then, the probability density function for θ_p is given by

$$\text{dens}\,(\theta_p) = p(n-p)\frac{\Gamma(\frac{p+1}{2})\Gamma(\frac{n-p+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})}(\sin\theta_p)^{p(n-p)-1}\,{}_2F_1\left(\frac{n-p-1}{2}, \frac{1}{2}; \frac{n+1}{2}; \sin^2\theta_p I_{p-1}\right)$$

where ${}_2F_1$ is the Gaussian hypergeometric function of matrix argument. The correspond-

ing probability distribution function is

$$\Pr\left(\theta_p < \hat{\theta}_p\right) = \frac{\Gamma(\frac{p+1}{2})\Gamma(\frac{n-p+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})}(\sin\hat{\theta}_p)^{p(n-p)}{}_2F_1\left(\frac{n-p}{2}, \frac{1}{2}; \frac{n+1}{2}; \sin^2\hat{\theta}_p I_p\right)$$

Distribution functions for distances based on the largest principal angle can be calculated in a similar manner.

Chapter 3

Approach

We propose straightforward, easily-computable methods to compare covariance matrices and identify sources or cases where it makes sense to pool data (specifically, the covariance matrices themselves). These methods are based on the singular value decomposition (SVD), which for symmetric positive definite covariance matrices is equivalent to the eigenvalue decomposition. The SVD is defined as $S = UDV^T$, and in this case $U = V^T$. The U and V matrices are orthonormal ($U^TU = I, V^TV = I$) and capture direction (the left and right eigenvectors) while the diagonal D matrix captures scale (the non-singular, square roots of non-zero eigenvalues). We believe that analysis of covariance matrices should involve comparison and combination of the subspaces they span, which the SVD can provide.

Ideally in testing these methods we have a set of covariance matrices spanning a known range from equality or extremely similar, i.e. $\hat{\Sigma} = \Sigma$, to completely different, i.e. $\hat{\Sigma} \neq \Sigma$; with similarity and differences confirmed by multiple metrics. We did not find such a set of matrices in the literature; defining such a set of matrices is a potential future direction. Here, simulations were performed under varying conditions in attempt to cover a wide range of "similar" and "different" covariance matrices.

3.1 Similarity and Effective Dimension

We propose two metrics for determining the similarity and effective dimension of covariance matrices. While previous approaches have considered batch and lab effects or modeling covariances from an inverse Wishart distribution, our approach is more intuitive. The more similar two covariance matrices are, they closer the subspaces they span and thus overlap should be. Subspace information is obtained using the singular value decomposition, which decomposes a covariance into direction (U, V) and scale (D) matrices.

3.1.1 Xia Metric

Xia et al. (2010) propose a dimension reduction method in the context of survival analysis data for estimating the entire central subspace consistently and exhaustively [111]. The goal is to find a $p \times q_0$ matrix B, where $q_0 < p$, such that all information regarding the relationship between T (the true and sometimes unobservable failure time) and X ($(x_1, ..., x_p)^T$ covariates) is included in $B^T X$. In addition to a dimension reduction technique, they propose a metric for quantifying asymptotic results for their dimension-reduced estimate.

> *If the working dimension q is correctly specified, that is, $q = q_0$... the largest singular value of $B_0 B_0^T - \hat{B}_q \hat{B}_q^T$ tends to 0... If the working dimension q is greater than the true dimension, then the true central subspace is consistently contained in the estimated space; if q is smaller than the true dimension, then the estimated space is consistently a subspace of the central subspace.*

The metric of interest is the largest singular value of $B_0 B_0^T - \hat{B}_q \hat{B}_q^T$, which should tend to zero as n increases or as B_0 and $\hat{B}_q$ become more similar. Intuitively, this metric is easy to understand. The more similar two matrices are, the smaller the differences between them will be, and thus the largest singular value of those differences will tend towards

zero. Rewritten in terms of the U matrix from the SVD, this metric becomes

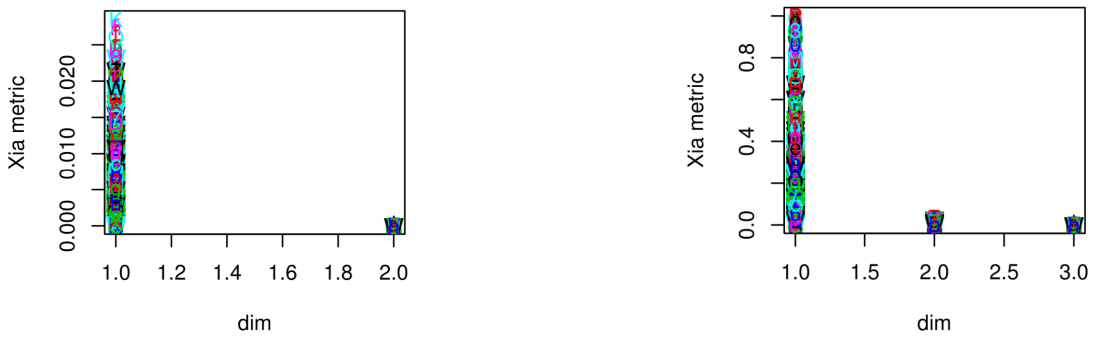
$$U_0 U_0^T - \hat{U}_q \hat{U}_q^T \to 0 \tag{3.1}$$

where U_0 is the matrix of directions of a "true" covariance matrix and $\hat{U}_q$ denotes directions of an estimated covariance matrix. This metric should tend towards zero at q, assuming q is correctly specified. The reasoning behind using the U matrices for comparison in the metric is to see how well the space spanned by the U directions of two covariance matrices match.

This metric can be applied in one of two ways. First, assume a covariance matrix Σ is known to be the true covariance with effective dimension (or rank) q lower than dimension p. Let U_0 be of dimension $p \times p$ and $\hat{U}_q$ of dimension $p \times q$, then if true effective dimension q is correctly specified, the metric should indicate this by tending towards zero at q. This allows us to determine if a $p \times q$ subset of the covariance matrix is adequate for capturing all information the full matrix contains. This metric may tend towards zero at true effective dimension q even when the scree plot does not show a clear drop. Second, if the goal is to test similarity between two covariance matrices (to find something in common between the two; or perhaps it is known there are similarities but not exactly where or what), one can let both U and $\hat{U}$ be of dimension $p \times q, q = 1, ..., p$.

To illustrate the use of this metric for analyzing effective dimension, two low dimensional toy data sets of dimension two and three were generated from a standard Gaussian distribution with a signal to noise ratio (SNR) of 10. Figure 3.1 indicates the Xia metric drops towards zero at the true effective dimension as expected.

Two toy data set of dimension 17 and true effective dimension five were generated with varying signal to noise ratios to analyze the effect of column permutation within the signal and noise spaces. The Xia metric was calculated using these toy data sets as the true Σ compared against $\hat{\Sigma}$ created by combining all possible 120 permutations of the five signal columns and a random subset of 120 permutations of the noise columns. Under these



(a) Xia metric for 2D toy data set.

(b) Xia metric for 3D toy data set.

Figure 3.1: Xia metric for two low dimensional toy data sets with signal to noise ratios of 10. (a) has one signal and noise column; (b) has one signal and two noise columns.

situations, the Xia metric still indicated (effective) similar matrices (apart from "noise"), with a drop at true effective dimension of five, cf. Figure 3.2. The almost-zero ordinate values further indicate the metric is able to detect that the matrices are essentially the same despite permutation within the signal and noise subspaces.

We then analyze the performance of the metric under situations closer to our motivating data sets. We generate a true covariance matrix Σ with $p = 17$ and a true effective dimension of $q = 5$ (the remaining 12 dimensions are essentially noise). First, matrices U and D were generated with five effective signal dimensions and 12 noise dimensions were generated from various normal distributions. These were then combined using $\Sigma = UDU^T$ to create a true covariance matrix Σ. We denote these matrices used to generate Σ by `gen_u` and `gen_d`, respectively, and the U and D resulting from the singular value decomposition of Σ by `true_u` and `true_d`. In Figure 3.3, 10,000 estimates $\hat{\Sigma}$ were generated from a Wishart distribution with true Σ and varying degrees of freedom. Xia metric scores are calculated from each of the simulated vs. true covariance matrices, and the average scores are plotted in Figure 3.3. As expected, fewer degrees of freedom result in more variable $\hat{\Sigma}$ being sampled and thus higher Xia metric scores. A drop to zero can be seen at dimension five as expected.

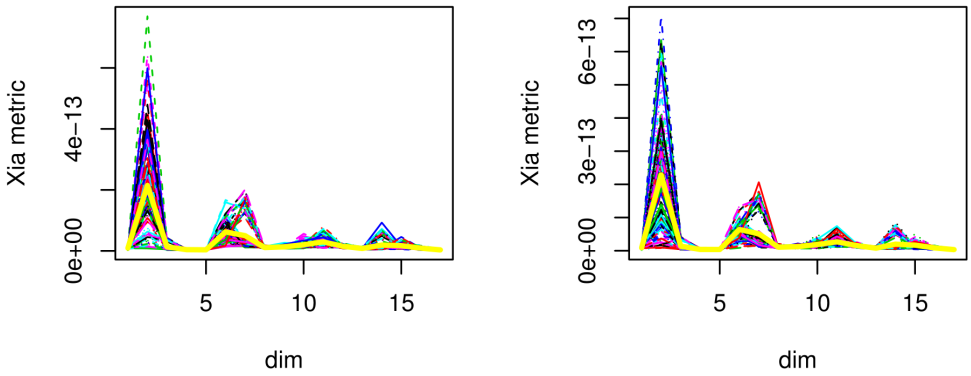


Figure 3.2: Xia metric for two toy data sets of dimension 17. The left figure has a signal to noise ratio of 2, and the right has a signal to noise ratio of 1.33. All possible permutations of signal columns and a random subset of permutations for noise columns were used to calculate the Xia metric. The Xia metric is able to detect that the matrices are essentially the same despite permuting within the signal and noise subspaces, as can be seen from the almost-zero ordinate values. Signal to noise ratios of 10 and 4 were also analyzed and show similar results. Yellow line indicates the mean over 10,000 simulation runs.

Figure 3.4 shows the effects of changing D on the Xia metric. While the Σ and $\hat{\Sigma}$ used in the two subfigures is different, the simulation process is the same. The `true_u` matrix obtained from the SVD of Σ is used as is. A new D matrix is generated assuming effective dimensions of either $q = 4, 5, 6$. This D and `true_u` are combined to create the estimate $\hat{\Sigma}$, from which Xia metrics are calculated (averaged over 10,000 simulations). Subfigures labeled (1) use the same distribution to generate D as `true_d`, while the following subfigures (2) through (5) increase (ˆ) or decrease (v) signal or noise variation. The main difference between (a) and (b) is that (a) uses a `true_u` generated from two normal distributions: $N(0, 1)$ (first five columns) and $N(0, 0.1)$ (last 12 columns), while `true_u` for (b) (all 17 columns) is generated from a $N(0, 2)$ distribution. (However, the D matrices vary slightly as well.)

Figure 3.5 uses `gen_u` in generation of $\hat{\Sigma}$ instead of `true_u` as the previous images did. This has a relatively large influence on the resulting Xia metric scores. It is clear intuitively that an estimate $\hat{\Sigma}$ generated from `true_u` will be much closer to the true Σ for reasonable values of D as opposed to one generated from `gen_u`, which is a few steps

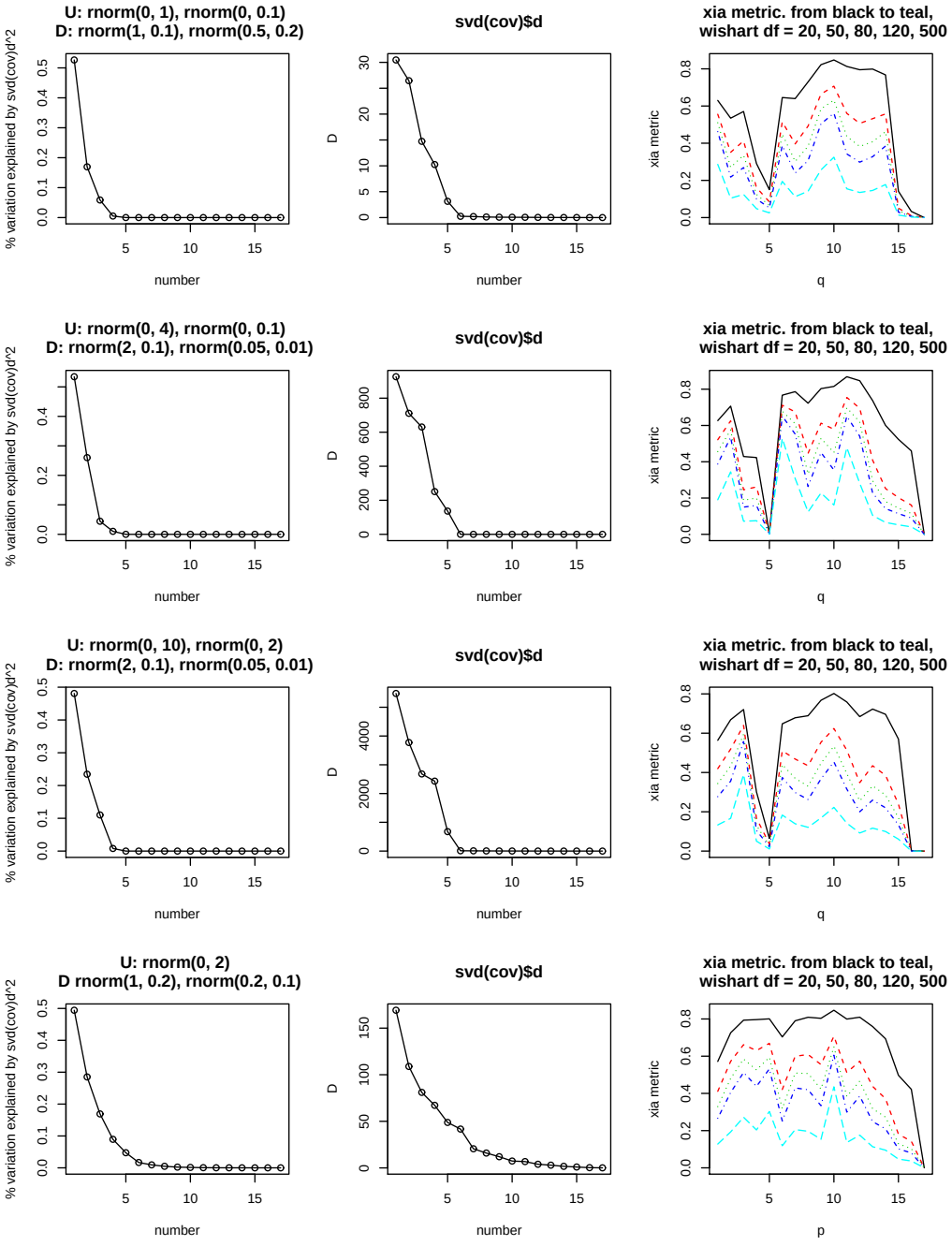
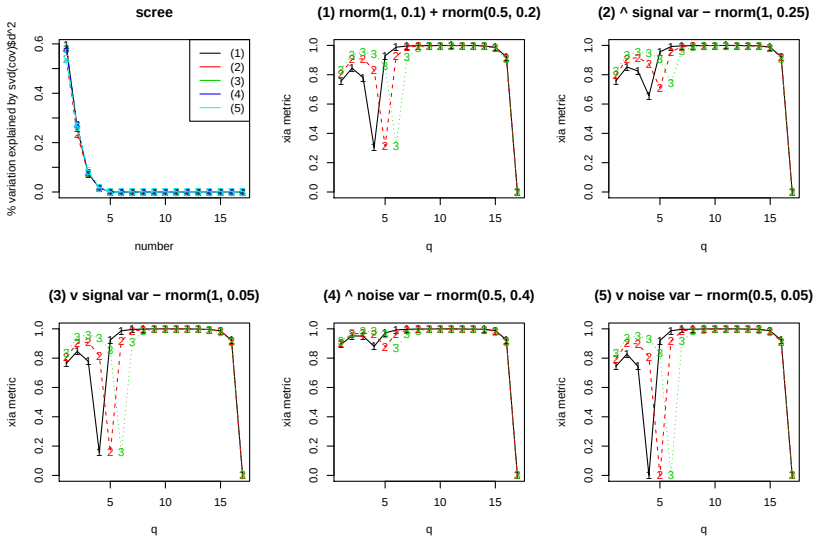
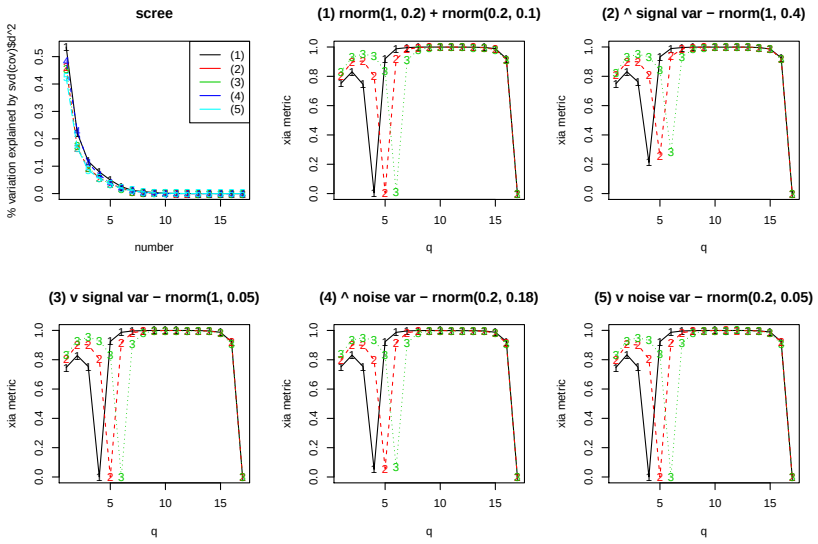


Figure 3.3: Eigenvalues and Xia metric scores (averaged over 10,000 runs). Using a true Σ created from the Gaussian distributions specified in the first column, estimates $\hat{\Sigma}$ are generated from a Wishart distribution with varying degrees of freedom. As expected, Xia metric scores are lower for higher degrees of freedom (more stable estimates).



(a) `true_u`: first five columns generated from $N(0, 1)$ and next 12 from $N(0, 0.1)$



(b) `true_u`: all 17 columns generated from $N(0, 2)$

Figure 3.4: Eigenvalues and Xia metric scores. Given a true Σ from a Gaussian distribution, estimates $\hat{\Sigma}$ are generated using `true_u` and a modified D matrix. Scenarios (1) through (5) indicate increases (denoted by ˆ) or decreases (v) in the signal or noise variation for generating new D. Solid black lines (1) indicate $q = 4$, dashed red lines (2) $q = 5$, and dotted green lines (3) $q = 6$.

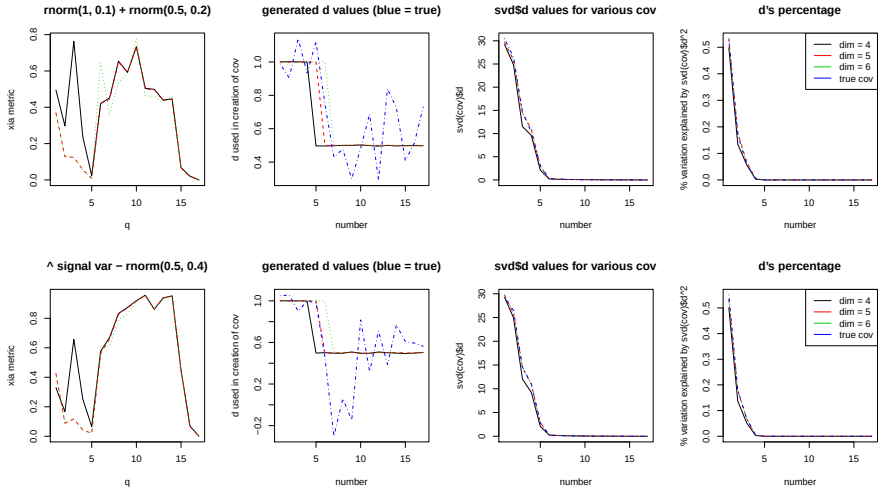


Figure 3.5: Xia metric scores and eigenvalues give a true Σ and an estimated $\hat{\Sigma}$ generated with `gen_u` instead of `true_u` in addition to modifying D. The three lines in the first column indicate effective dimension $q = 4, 5, 6$ (solid black line, dashed red line, dotted green line) used to generate D. Xia metric scores drop to zero at five regardless of how many effective dimensions D was generated with.

further removed from `true_u`. In contrast to the previous simulations using `true_u`, the Xia metric in this situation is much more variable, and the drop towards zero is not quite as steep. Furthermore, Figure 3.4 showed clear troughs at the effective dimension used in generating $\hat{\Sigma}$ ($q = 4, 5, 6$), while Figure 3.5 shows troughs only at around $q = 5$, regardless of the effective dimension actually used to generate $\hat{\Sigma}$.

The simulations in Figure 3.6 modified the effective dimensions ($q = 4, 5, 6$) of the U in addition to the D matrix. It is evident this has a much larger effect on the metric, as modifying the U matrix greatly increases the variability. The red dotted lines in the first column, which correspond to $\hat{\Sigma}$ with effective dimension of five (same as Σ), still show drop to zero at $q = 5$ as expected. However, the black and green lines with dimensions $q = 4, 6$ no longer tend towards zero until around $p = 17$.

The Xia metric can be considered as the largest singular value of a random matrix. If we consider the square or the eigenvalue instead, this can be considered to follow the Tracy-Widom distribution for the largest eigenvalue of a random matrix, cf. Section 4.1.3.

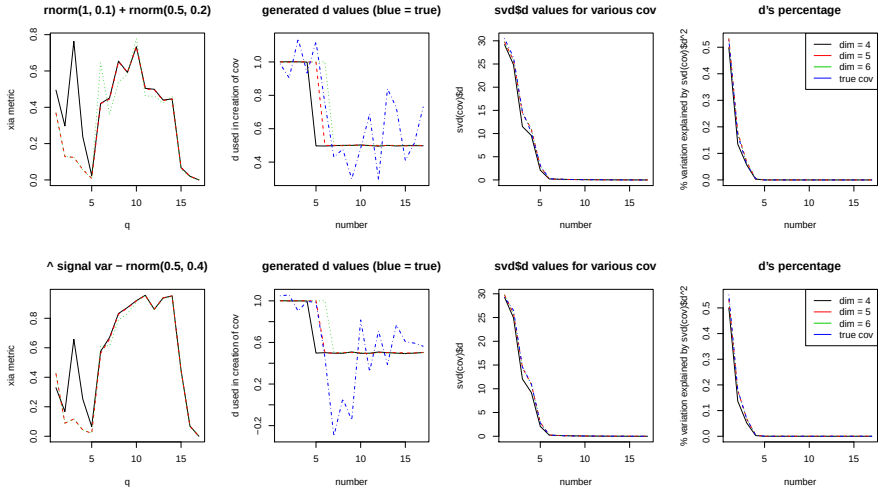


Figure 3.6: Xia metric scores and eigenvalues. These simulations change the the effective dimensions ($q = 4, 5, 6$) of U as well as D, and has a much larger effect on Xia metric scores. Solid black lines indicate $q = 4$, dashed red lines $q = 5$, and dotted green lines $q = 6$. The Xia metric drops to zero at $q = 5$ only when the effective dimensions was also equal to five. The caret in the second row indidates an increase in noise variation.

3.1.2 Largest Principal Angle

We discuss a simple visual comparison that can be performed on the matrix of eigenvectors U, especially for low dimensional examples, before moving onto the largest principal or canonical angle method.

U-Based Visual Method

In estimating the similarity between two covariance matrices, we can consider a visual method based only on the U direction matrices from an SVD. The dot product between two U matrices,

$$U_1^T U_2 \tag{3.2}$$

where U_1 is from the "true" covariance matrix and U_2 is from an estimated covariance matrix, essentially calculates the cosine of angles between two columns at a time, and can

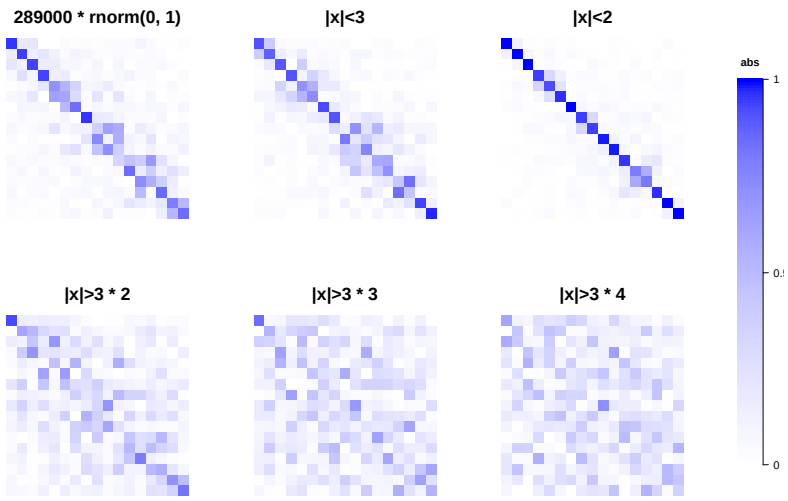


Figure 3.7: Data generated from a $N(0,1)$ distribution. A diagonal structure can be seen clearly in the top row as outlying observations are removed, which gets harder to detect (bottom row) as the effect of outliers is increased.

clearly visually us how similar two U matrices are. A heatmap of the dot product should show a diagonal structure if the U's are similar, and roughly block diagonal structure if the U's also have similar effective dimensions less than p. The effectiveness of this metric can be illustrated by comparing classical and robust estimates of covariance matrices where data sets have known outliers. Letting each of the U matrices be from SVDs of the estimates, absolute value of the heatmaps are plotted.

Data from a $N(0,1)$ distribution were simulated and the classical and robust (MCD) estimators calculated for this data. Figure 3.7 shows heatmaps of $U_1^T U_2$ from the classical and robust estimates for (first row) all data and data with slightly "outlying" values ($abs(x)$ less than 3 or 2) removed. A diagonal structure can be seen, which gets clearer as more "outlying" values (outliers) are removed. The second row exaggerates these "outlying" values by multiplying them by a factor of 2, 3, or 4. The diagonal structure, once clear, gets increasingly hard to detect (adding outliers causes the difference between the classical and robust estimates to increase).

This can further be illustrated by some well known data sets containing outliers. We

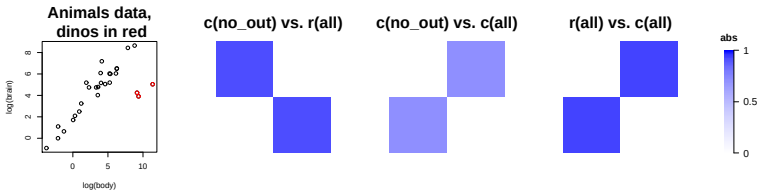


Figure 3.8: The Animals data set contains average brain (g) and body weight (kg) data for 28 species of land animals, including three species of dinosaurs, indicated by the three red dots in the scatterplot (a) [88, p. 57]. Despite only having two variables, the heatmaps behave as we expect. The diagonal structure in heatmap (b) indicates that a classical estimate of the data without outliers and a robust estimate of all the data are very similar. Heatmaps (c) and (d) indicate that the two estimators differ; the classical estimate is affected by as few as three outliers.

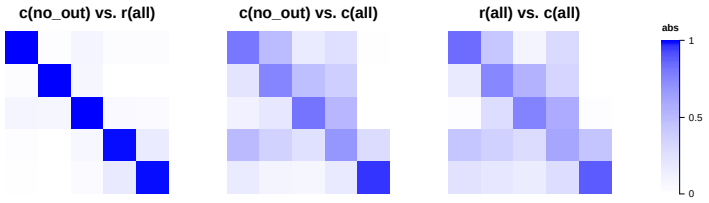


Figure 3.9: The original Wood Specific Gravity data set contained 20 observations with 6 parameters [27, p. 227]. Four rows were purposefully contaminated by Rousseeuw and Leroy [88, p. 243] in all dimensions with bad leverage points not outlying in any individual variable (masking effect, so these outliers are hard to detect using traditional methods). The first heatmap on the left shows a clear diagonal structure indicating the two estimates are similar; and that the MCD estimator is correctly identifying outliers. The blurry diagonals and off-diagonals in heatmaps (b) and (c) indicate the estimators are not as similar as in (a).

let `all` denote the entire data set including outliers and `no_out` denote the data set with outliers removed, `c()` denote the classical covariance estimate, and `r()` denote the robust MCD estimator. These are shown in Figures 3.8 and 3.9.

Largest Principal Angle

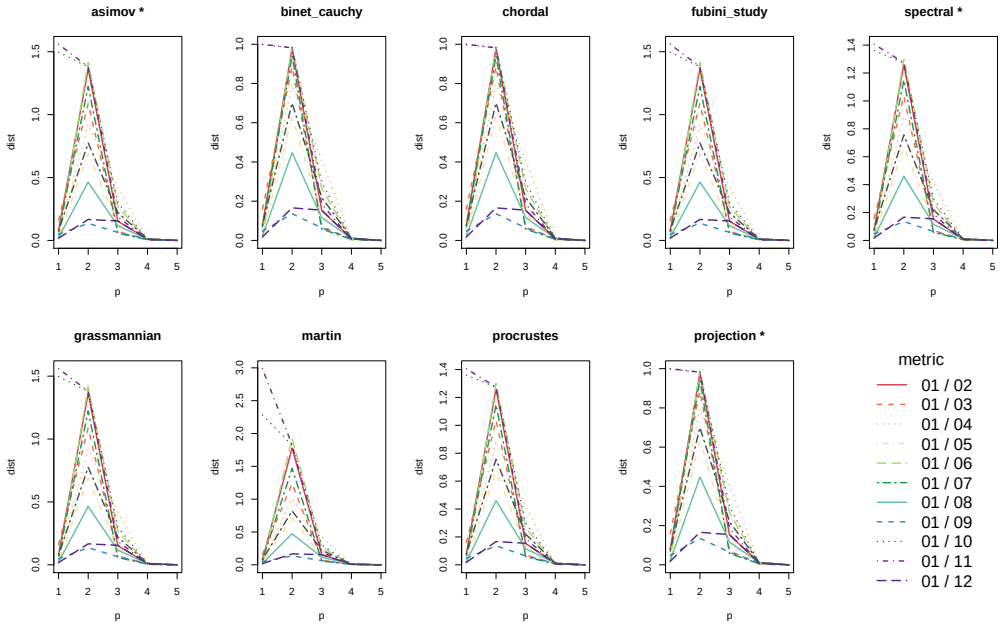
Table 2.1 listed many possible distance metrics for comparing subspaces, some of which use only the largest principal angle while others are functions of all angles. We can relate this

to the use of Hotelling's T^2 multivariate test. There may be scenarios where Hotelling's test is clearly a better choice (e.g., correlations between variables exist, and ideally all variables and their interactions are taken into account together); this might be sensitive to small deviations in all variables, but at the same time has the disadvantage of potentially not capturing large deviation in just one variable. In contrast, univariate t tests could capture deviation in variables independently, but not correlation across multiple variables.

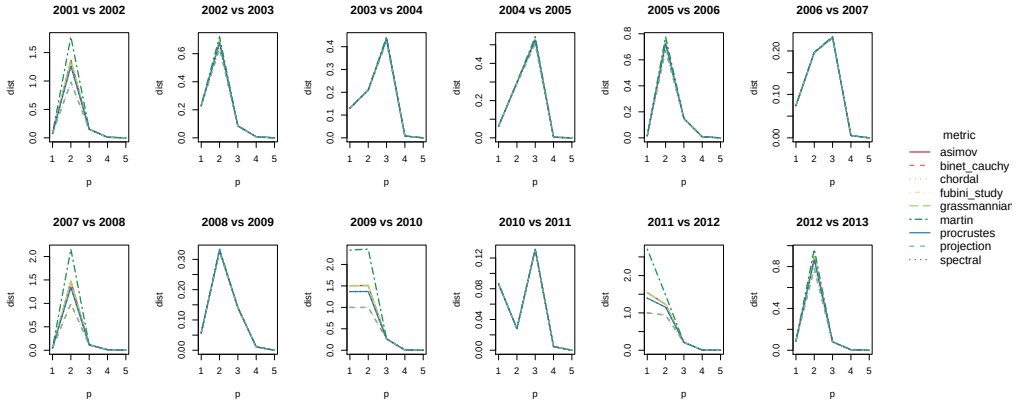
The nine metrics in Table 2.1 were calculated and compared for an oceanography data set and shown in Figure 3.10. The Martin metric had the largest range and variability in comparison to all others, unsurprisingly – it is the only metric taking the logarithm of the inverse of the cosine of the principal angles. The Chordal and Procrustes metrics followed similar trends, confirmed by the similarity in their formulas (both involve the sine of the angles). Although modern computers are capable of extremely quick calculations, their formulas indicate the Martin and Procrustes distances may be most computationally expensive, based simply on the number of operations required. We conclude the Asimov distance (or the largest principal or canonical angle), which is simplest to calculate and interpret, is sufficient for the purposes of our analyses. Situations may exist where a metric taking into account all the angles may be preferred, and can be analyzed in future research.

These metrics were also calculated for our motivating glass data sets to ensure they behave as expected (see Figure 3.11).

We propose the use of the largest principal angle as a metric for similarity between covariance matrices. First, the singular value decomposition of two covariance matrices is taken to obtain the eigenvectors. Then, principal angles between these two matrices of eigenvectors are calculated using the singular value decomposition and the cosines of the singular values. This is done in cumulative fashion adding columns one by one. We expect larger angles (up to $\hat{\theta} \leq \frac{\pi}{2}$) when subspaces are very different or even orthogonal ($\hat{\theta} = \frac{\pi}{2}$), and smaller angles for more similar subspaces.

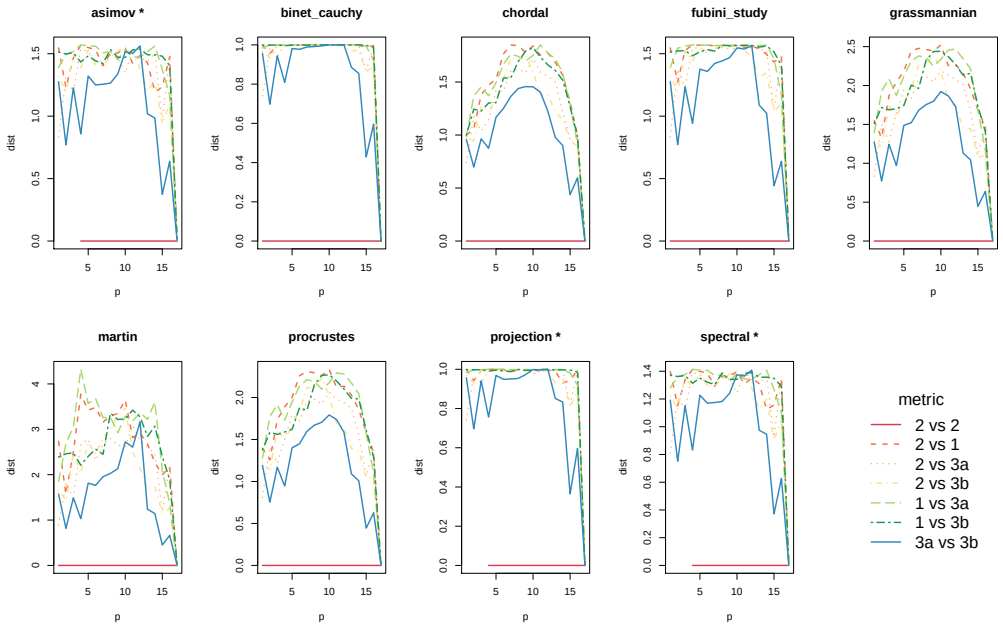


(a) Subspace distance metrics comparing oceanography data from 2001 to years 2002 through 2012. The asterisk indicates distances involving only the largest principal angle.

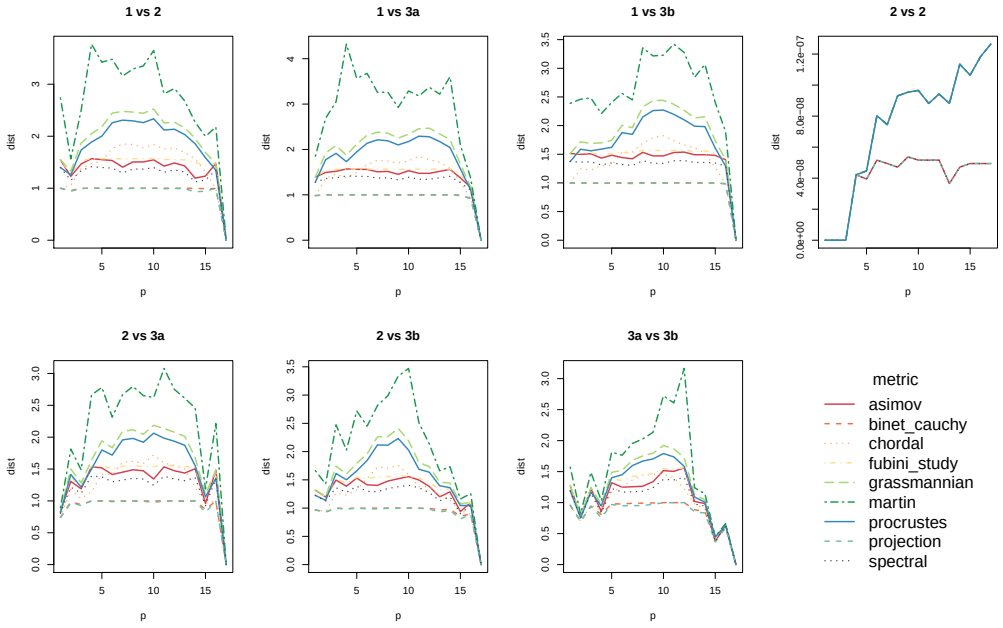


(b) Subspace distance metrics comparing oceanography data for consecutive years from 2001 to 2013.

Figure 3.10: Nine metrics from Table 2.1 for oceanography data with five variables (temperature, salinity, density, chlorophyll, and nitrate).



(a) Subspace distance metrics for four motivating data sets. The metrics are lower for comparisons between data from the two companies in Data Set 3, indicating these are more similar to each other than to Data Sets 1 or 2. The asterisk indicates distances involving only the largest principal angle.



(b) Subspace distance metrics by data set. The top right plot compares Data Set 2 to itself; the corresponding metrics are essentially zero. The Martin distance is consistently larger than the others.

Figure 3.11: Nine metrics from Table 2.1 for three motivating glass data sets.

If we have an idea of what the true effective dimension q is, we can observe the largest principal angle in a small range around q (e.g., $q \pm 2, 3$) to find at which specific q the subspaces are most "similar." For dimensions $\hat{q} < q$, the principal angle may be larger or more variable due to the dimensions not capturing enough common signal. Another possibility is that the eigengaps are too small, resulting in neighboring eigenvectors exchanging positions. For dimensions $\hat{q} > q$, the largest principal angle may start to increase as more noise is included in addition to the signal.

We can analyze the effect of variability on this metric through simulation (see Section 4.2). First, we can vary the eigengaps and magnitude of eigenvalues (or singular values) for the true covariance matrix Σ. Second, $\hat{\Sigma}$ can be simulated from various distributions (Wishart if simulating covariance matrices directly; Gaussian or t if data is first simulated) with varying degrees of freedom. From these data, for each dimension, we can estimate a robust regression line between the degrees of freedom and quantiles of interest – mainly, the median and 95th percentile. This linear relationship may break down for smaller degrees of freedom (or n) where p and n are of similar magnitude (our simulations used $p = 17$ and n ranging from 25 to 100,000); values of $n \leq 50$ were excluded in fitting the line; cf. Figure 4.22. Theoretical quantiles can also be calculated from the density functions [4].

For each quantile, the intercept and slope are plotted against dimension (analogous to using the slope or intercept to estimate parameters for discrete distributions, such as Poisson or Binomial). While either intercept or slope may be used, the slope values tend to be more stable, as our simulations (see Section 4.2.2) confirm. Confidence intervals using the estimated standard error can also be calculated for the coefficients at each dimension. Sudden jumps in expected intercept or slope values may indicate the value corresponding to the true effective dimension.

3.1.3 Other Considered Metrics

Li's 1991 paper on sliced inverse regression is a method for evaluating the effectiveness of an estimated dimension reduction (e.d.r.) direction [68].

$$R^2(b) = \max_{\beta \in B} \frac{(b\Sigma_{xx}\beta'))^2}{b\Sigma_{xx}b \cdot \beta\Sigma_{xx}\beta'}$$

This R^2 metric is a squared multiple correlation coefficient between bx, the projected variable, and $\beta_1 x, ..., \beta_k x$, the ideally reduced variables. While this was considered a potential metric for finding true effective dimension, we focused on the previous metrics which seemed more suitable for the problem at hand.

3.2 Combining Covariance Matrices

Three approaches were considered for combining multiple covariance matrices.

1. Pooled or weighted average
2. Use the U matrix calculated from the pooled data, and average D matrices from individual covariance matrices
3. Average the U and D matrices

The arithmetic mean in the first method suffers from non-robustness from placing too much weight on outliers, but it may be appropriate if data are believed or known to be "similar" (e.g., produced from the same machine, data from same person, etc.). The second method results in estimates very similar to the first, and as it requires actual data in addition to the covariance matrix, was not further considered. The third method, averaging the U (and $V = U^T$) and D matrices, is the proposed method. As information concerning the subspace spanned by a covariance matrix can be captured by the U and D matrices from an SVD, combining this information from multiple covariances essentially

combines information on the subspaces of multiple covariances. Instead of pooling individual elements from a covariance matrix, we propose a method to pool the spaces they span.

The proposed estimator has properties of robustness (more so than a simple pooled covariance) and is positive definite (PD) without any modifications (U is orthogonal with orthonormal columns). Other desirable properties are further discussed in Section 5.

Chapter 4

Simulations and Applications

In this section we will apply our proposed methods to the motivating glass data sets and simulations based on realistic scenarios based on these data sets.

4.1 Xia Metric

Simulations in Section 3.1.1 illustrated the use of the Xia metric for checking effective dimension of a covariance matrices, showing plots with a sudden drop at the true q. We see a different pattern when attempting to use this metric to determine similarity between covariances for Data Sets 1, 2, and 3. Data Sets 1 and 2 measure only one pane of glass; for consistency, one pane was selected at random from Data Set 3 (which contains information on 48 different panes). Recall that calculation of the Xia metric involves comparing the first q directions (abcissa in Figure 4.1) from the SVD of two covariances for $q = 1, ..., p$, and will tend towards zero as the two matrices become more similar (smaller scores indicating higher similarity). Figure 4.1 indicates that comparing Data Sets 2 and 3 gives the lowest score at $q = 1$ of 0.66, as well as lower scores from $q = 14$ to 16. This shape, which resembles a downward-facing parabola, suggests that the first few eigenvalues are very close (i.e., the eigengap between eigenvectors is very small), especially

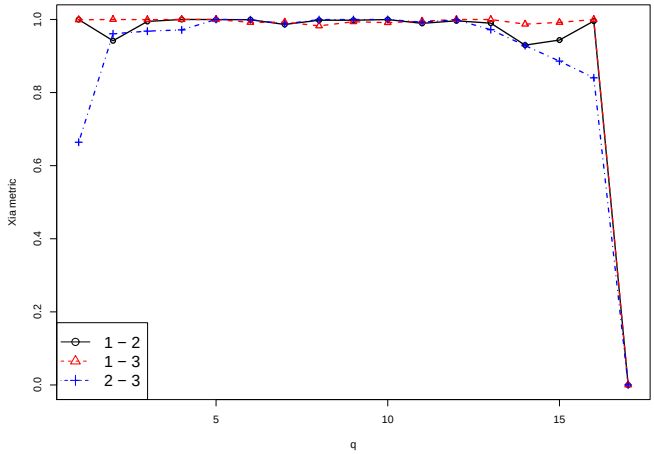
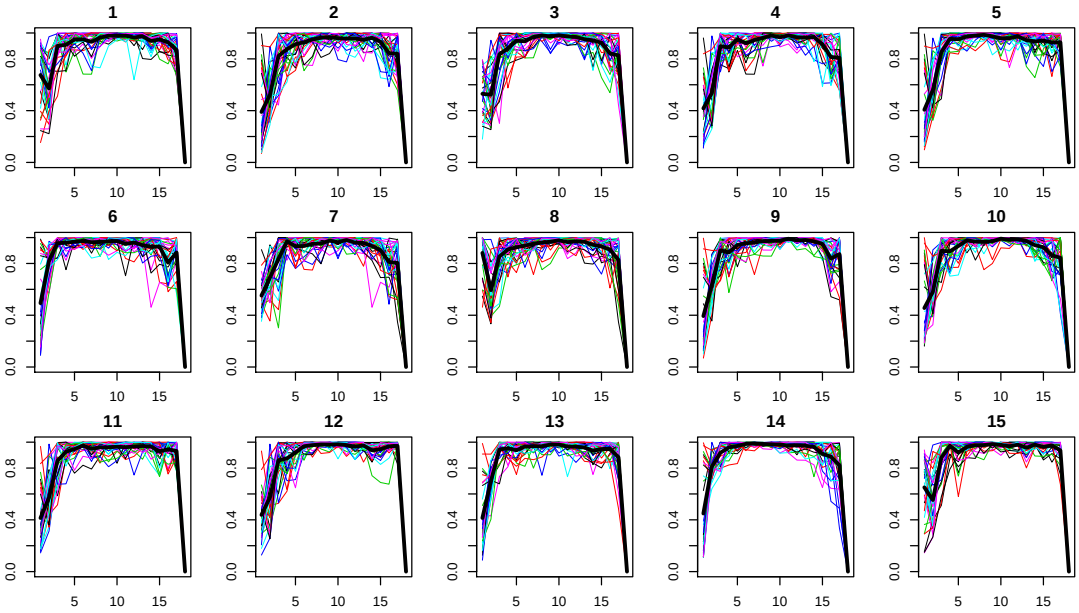


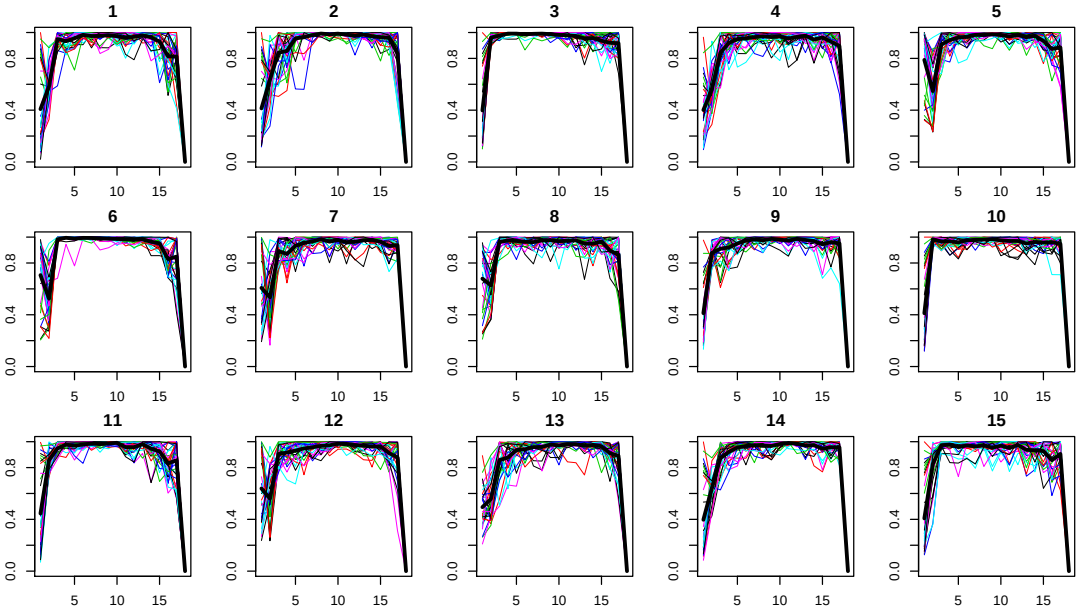
Figure 4.1: Xia metric for comparing similarities between Data Sets 1, 2, and 3 (a single pane). The legend indicates pairs of comparisons by "x – y" (comparing data set x to data set y).

taking into account small sample sizes, resulting in the eigenvectors swapping positions (or ordering). Indeed, if this is the case, heatmaps of the correlation (covariance) matrices (see top row of Figure 4.24) of Data Sets 2 and 3 are more similar than either of those with Data Set 1. This is examined further with simulations in Section 4.2.2.

Focusing on the 48 glass panes from Data Set 3, individual glass panes (31 from Company A and 17 from Company B) were compared against each other to analyze within- and between-manufacturer variability. Subfigure (a) in Figure 4.2 shows within-manufacturer comparisons – each individual pane compared to the remaining 30 or 16 from the same company. Subfigure (b) shows between-manufacturer comparisons (i.e., each pane in Company B compared to the 31 from Company A and vice versa). The thin, colored lines indicate Xia metrics resulting from individual pane comparisons, and the thick black line gives the average score across all comparisons. Curves for the missing panes closely resemble those plotted. Subfigure (b), which show Xia metric scores for within-manufacturer panes of glass, have not only overall lower scores than subfigure (a), but also smoother, "rounder" downward facing parabolic shapes (as opposed to the more



(a) Within-manufacturer comparisons. Each pane compared to against the other 16.



(b) Between-manufacturer comparisons. Each pane compared against the 31 from Company A.

Figure 4.2: Company B pane comparisons. Thin colored lines indicate Xia metrics for individual pane comparisons, and the black line indicates the average across all comparisons.

square ones in subfigure (a)). However, these differences are not very clear; the Xia metric is better suited for determining true effective dimension rather than similarity of covariance matrices.

4.1.1 Performance Under Varying Signal to Noise Ratios and Covariance Structures

In Section 3, the effect of permuting columns within the signal and noise spaces was analyzed, as well as the effects of specific U and D matrices. Since this is applied to covariance matrices, we are also interested in the effect of covariance matrix noise and structure on the Xia metric. First, a true Σ was generated from a toy data set of dimension 17 with five signal and 12 noise columns. Signal columns were generated from a standard Gaussian $N(0, 1)$ distribution and noise columns generated from Gaussian distributions with standard deviations of 0.1, 0.25, 0.33, 0.5, and 0.75, resulting in signal to noise ratios (SNR) of 10, 4, 3, 2, and 1.33. $\hat{\Sigma}$ was generated by adding various types of noise to Σ, either by simulating from a Wishart with varying degrees of freedom or by adding additional noise observations to data simulated from Σ then calculating the covariance matrix. Overall, the metric is relatively robust for high signal to noise ratios of 10 and 4, but begins to break down at lower ratios of 2 (depending on other factors) and 1.33.

We also consider different covariance structures, namely, the case where all elements are equal to some σ, a block diagonal covariance structure, and a Toeplitz matrix structure. These are chosen as they appear relatively often in applications – time series models (AR, ARMA, MA) may have Toeplitz structured covariance matrices, and Gaussian processes or block Gaussian processes may have block structured covariances.

Varying Signal to Noise Ratios

Variations in data measurement for different batches, labs, or countries may result in different signal to noise ratios being present in covariance matrices. Simulations were

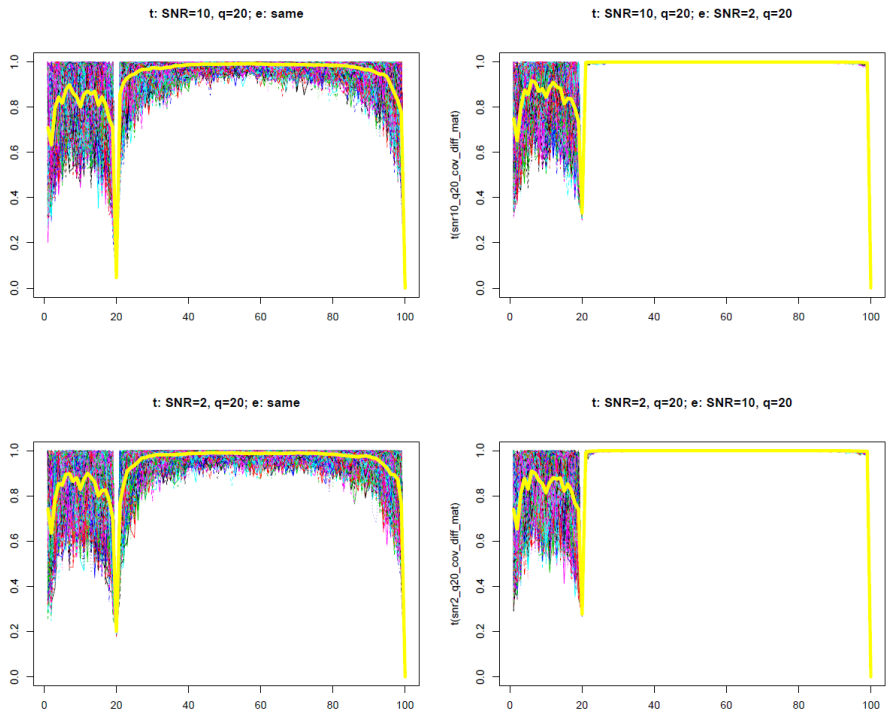


Figure 4.3: Xia metric for covariances with SNR of 10 and 2. Left column: the true (t) and estimated (e) covariances have equal SNR; right column: different SNR. Effective dimension $q = 20$ is detected in all cases and clearest when the SNRs are the same. Yellow line indicates means.

performed to analyze if the Xia metric can detect true effective dimension between two covariances with differing signal to noise ratios. Covariance matrices were generated with dimension 100 and effective dimension $q = 20$, and signal to noise ratios of 10 and 2. The Xia metric was calculated pairwise for these four covariance matrices and shown in Figure 4.3. The true dimension is detected for all cases, but the minimum value the Xia metric reaches varies based on SNR. It is lowest when the two matrices have the same SNR.

Multiplying Σ by a factor

This set of simulations (Figures 4.4 and 4.5) adds noise to Σ by combining the original data (denote by X) used to calculate Σ with randomly generated data (denote by ε) from a Gaussian distribution with varying covariance matrices. This random data is added into the true data to generate $\hat{\Sigma} = cov\{X + \varepsilon\}$.

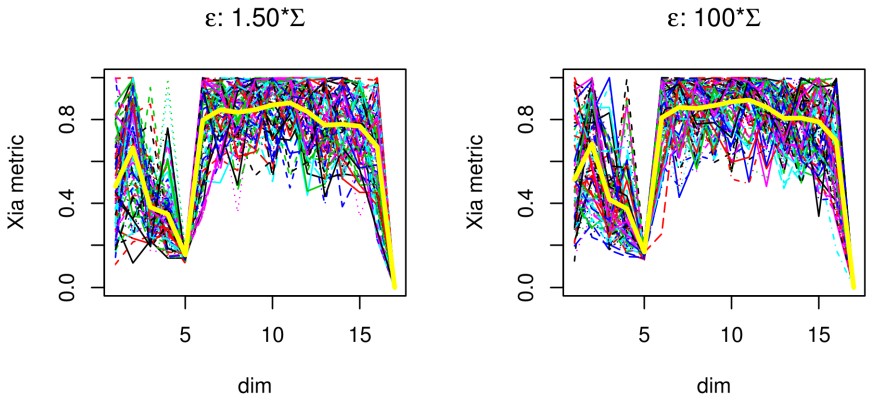


Figure 4.4: Xia metric for $\hat{\Sigma} = c \cdot \Sigma$, where $c = 1.5, 100$. Xia metric is unaffected. Yellow line indicates means of 10,000 simulation runs; other colored lines display certain individual simulation results.

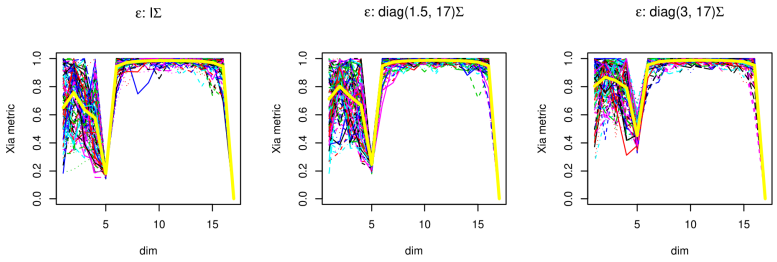
Figure 4.4 indicates that adding noise of the form $\varepsilon \sim N(0, c \cdot \Sigma)$ does not affect the Xia metric, which still detects an effective dimension of five. These results are shown for the lowest SNR tested, 1.33, and are very similar for the other ratios.

Figure 4.5 show results from simulations generating ε from varying $N(0, \Sigma)$ with signal to noise ratios of 10 and 1.33. Overall, the dip in Xia metric at $p = 5$ decreases as $\hat{\Sigma}$ varies more from Σ and as signal to noise decreases. As Figure 4.5(a) and the first row of Figure 4.5(b) indicate, when the SNR is high, the Xia metric can detect effective dimension when ε varies further from the true covariance matrix.

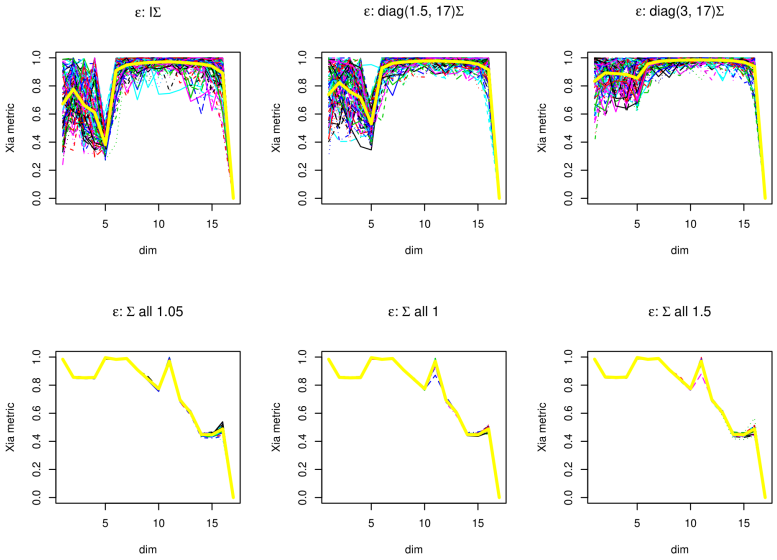
If Σ is merely multiplied by a factor or a similar diagonal structured covariance matrix, the Xia metric is still able to detect effective dimension. However if ε comes from a distribution of the opposite extreme where all elements in the covariance matrix are the same, this metric breaks down very quickly (Figure 4.5b).

Block Structure Covariance Matrices

A block diagonal covariance structure was created by adding a 5×5 matrix of constant c to a 17×17 identity matrix. Constant values c ranging from 1 to 14 were tested for the four signal to noise ratios, but the magnitude of c did not have a noticeable effect.



(a) Signal to noise ratio of 10



(b) Signal to noise ratio of 1.33

Figure 4.5: Xia metric for noisy covariance matrices; $\varepsilon \sim N(0, \Sigma)$ is indicated by individual plot titles. Simulation results for SNRs of 4 and 2 fall between the values shown for SNR of 10 and 1.33. Simulation results with ε as in the second row of (b) look the same for all SNRs. The yellow line indicates means.

Error data ε were generated from this block covariance structure to calculate $\hat{\Sigma}$. Signal to noise ratio seems to have a larger effect on the Xia metric than block structure. The Xia metric proves to be relatively robust to block diagonal covariance structure, which is reasonable – effective dimension remains relatively clear in a block covariance structure. An interesting pattern for this covariance structure is that in addition to dipping at $p = 5$, the Xia metric also has a slight dip at $p = 2$.

Figure 4.6 shows simulation results for a subset of c values ranging from 2.25 to 3.25. Figure 4.7 shows a similar pattern for $p = 100$ and $q = 10$. Here, the Xia metric clearly breaks down with a signal to noise ratio of 1.33.

Toeplitz Structure Covariance Matrices

Error data ε were then simulated from Gaussian distributions with Toeplitz structured covariance matrices with varying number of diagonals. All diagonals contained the same constant "multiplier" – 1.05, 1.10, 1.15, 1.20, or 1.25, essentially adding in 5% to 25% variation to the original data X and true covariance Σ.

The Xia metric breaks down very quickly using a Toeplitz covariance structure as the variability and number of diagonals increase. Each combination of simulation parameters (number of diagonals, amount of variability, and signal to noise ratio) resulted in a plot of the Xia metric with some identified effective dimension. Since the drop towards zero is not always clearly a minimum, this p was identified manually for each case and summarized in Figure 4.8. Green indicates the correct dimension ($p = 5$) was identified and red indicates no reduced effective dimension detected, i.e. the entire space $p = 17$ was identified. We note there may be minor variations as to where the border where green transitions to red may lie if varying simulations parameters are used not captured by this specific simulation data set; however, this is adequate in giving a general range of breakdown.

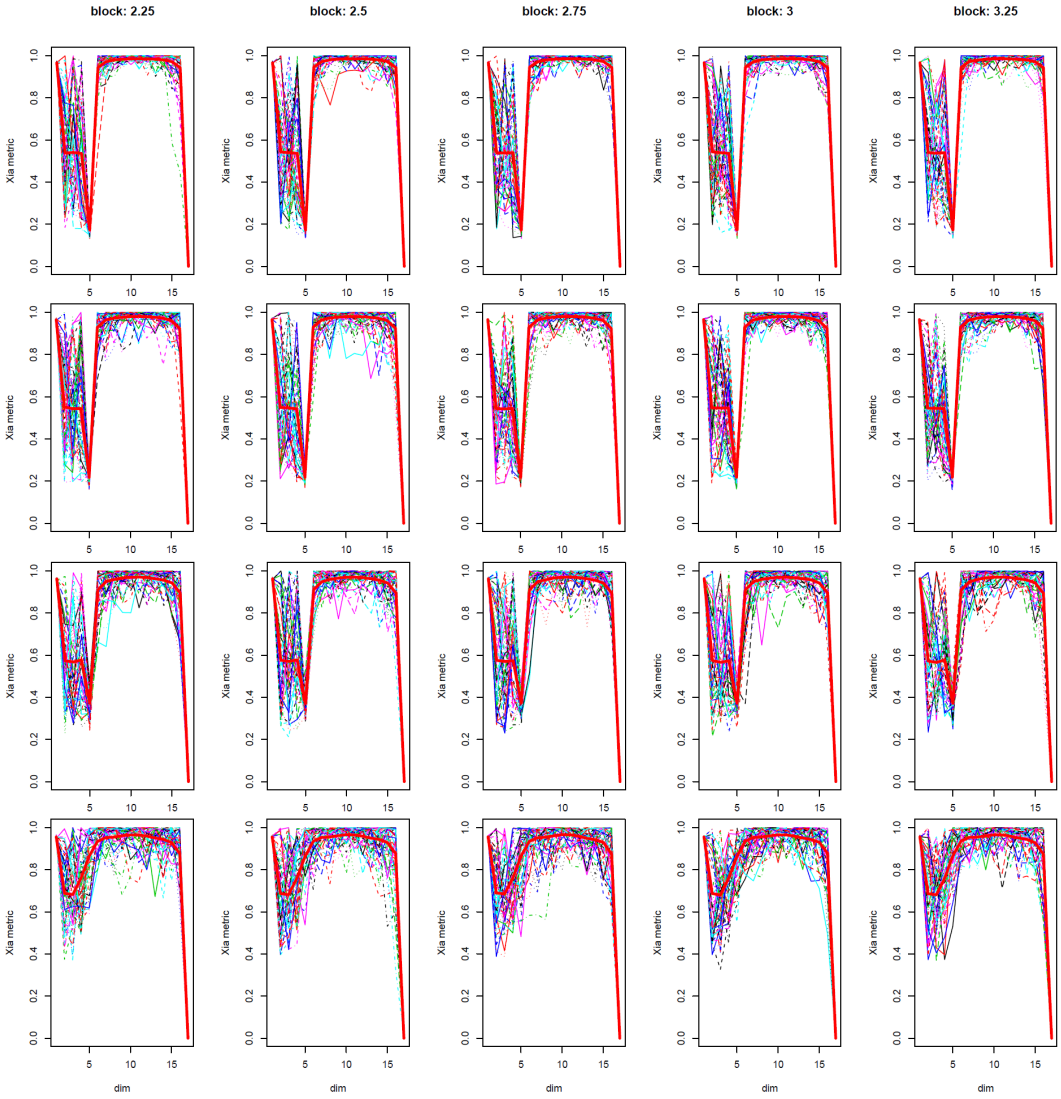


Figure 4.6: Xia metric for simulations with block covariance structure with $p = 17$. Results for all constant values from $c = 1, ..., 14$ were similar thus only a subset are shown. The Xia metric breaks down as signal to noise ratio decreases to 1.33. Red line indicates the mean over 10,000 simulations.

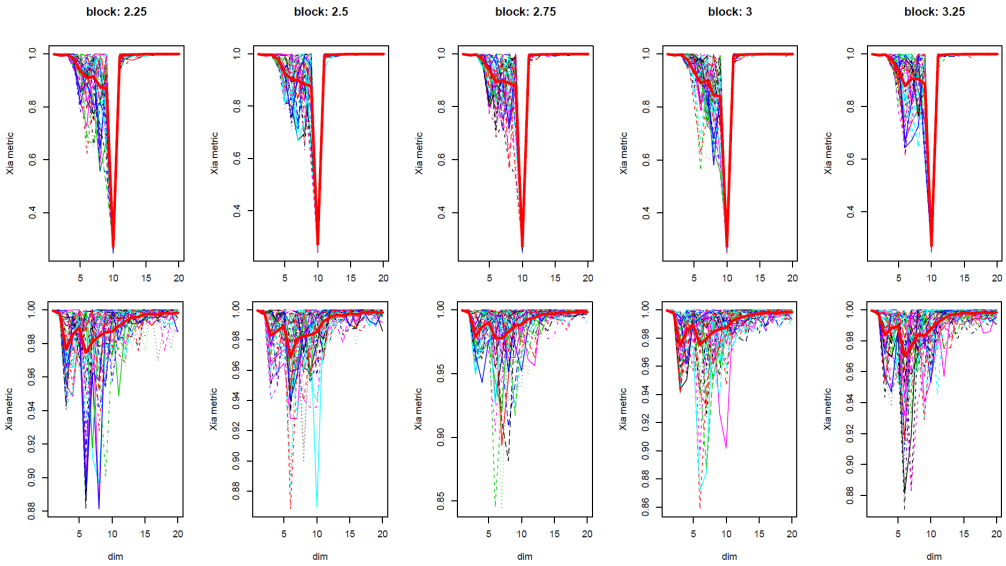


Figure 4.7: Xia metric for simulations with block covariance structure with $p = 100$. Results for all constant values from $c = 1, ..., 14$ were similar thus only a subset are shown. The Xia metric is stable for all signal to noise ratios but 1.33; thus results for SNR of 4 and 2 are omitted. Red line indicates the mean over 10,000 simulations.

4.1.2 Bootstrap Approach

While we have the luxury of running simulations tens if not hundreds of thousands of times, this is not realistic for real world applications. Usually, we will be given one true covariance and one estimated covariance – or perhaps just a true covariance matrix. Bootstrap approaches can be used to estimate the Xia metric under these conditions.

True and Estimated Covariance

We consider the case where two covariance matrices are given: the true and estimated. There are two methods of bootstrapping for this approach.

First, for the estimated covariance $\hat{\Sigma}$, we bootstrap by simulating additional covariance matrices $\hat{\Sigma}'$ from the Wishart distribution with degrees of freedom ranging from 20, 50, 80, 120, 150, and 500. Figure 4.9 shows that for high signal to noise ratios of 10 and 4, the Xia metric with bootstrap is able to detect the correct effective dimension. However, it breaks down for signal to noise ratios of 2 or less (figures omitted).

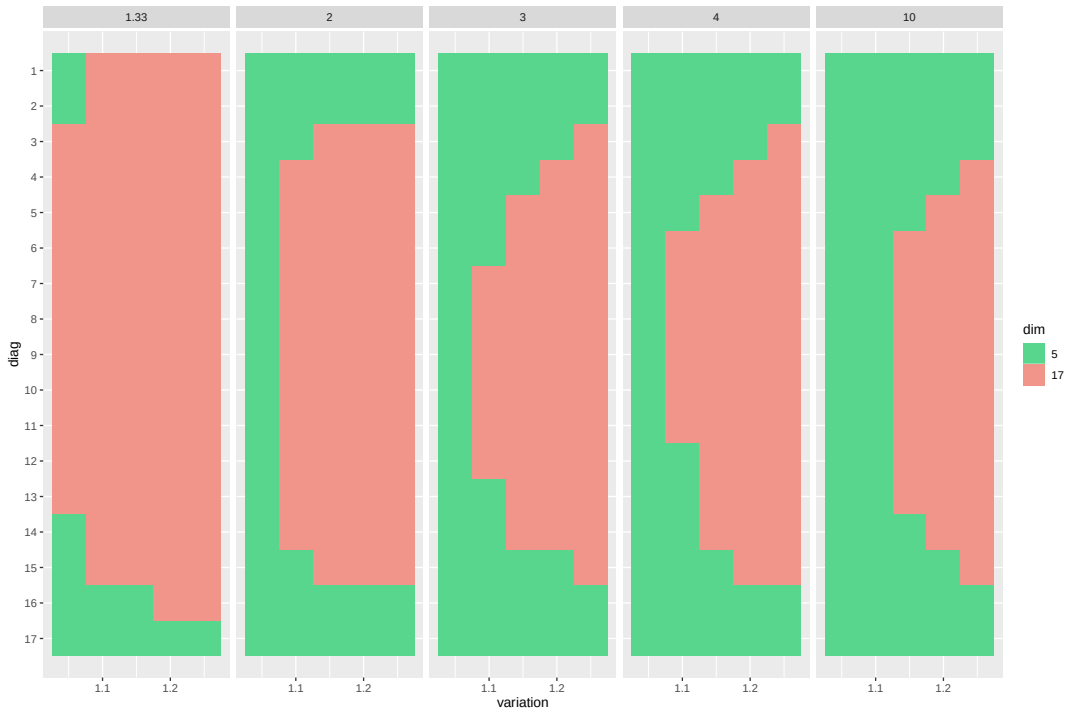
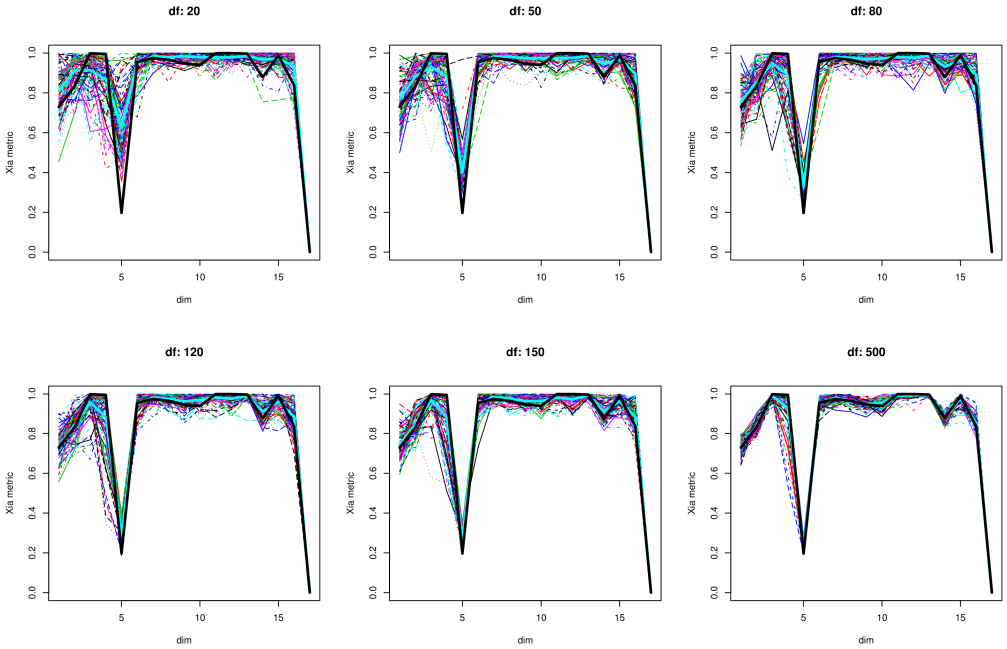
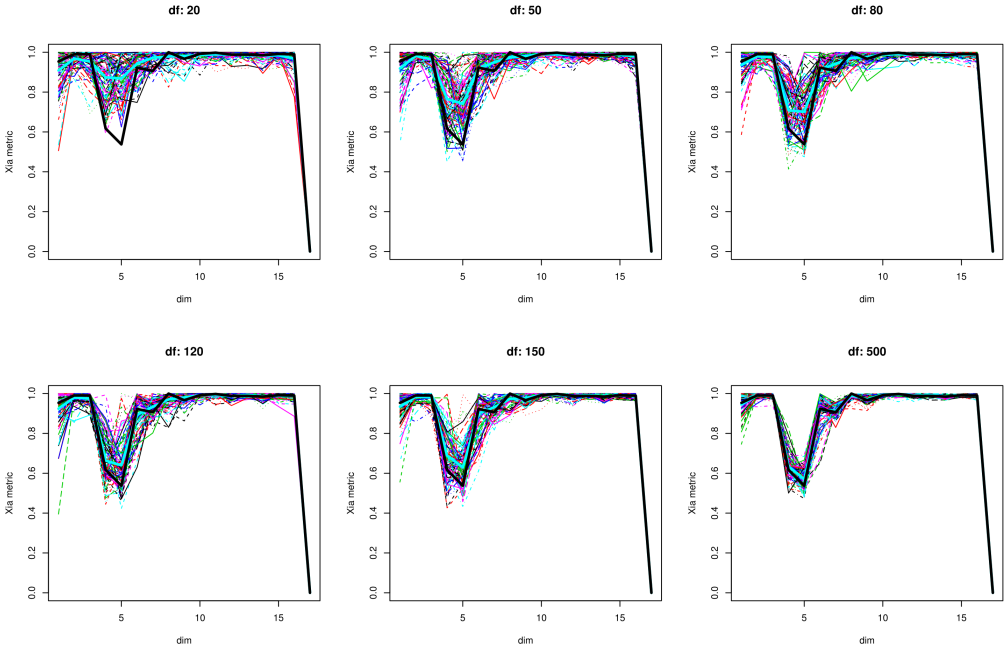


Figure 4.8: Heatmap visualizing the breakdown of the Xia metric under various simulation parameters (signal to noise ratio, number of diagonals, and amount of variation). Green indicates the Xia metric correctly dropped at $p = 5$, and red indicates no reduction in effective dimension detected ($p = 17$).



(a) Signal to noise ratio of 10



(b) Signal to noise ratio of 4

Figure 4.9: Xia metric simulation results for the bootstrap approach generating $\hat{\Sigma}'$ from the Wishart distribution with varying degrees of freedom. This approach can correctly detect effective dimension with just 20 degrees of freedom for higher signal to noise ratios. The black line indicates the Xia metric calculated from Σ and $\hat{\Sigma}$; the blue line plots the mean over all simulation runs.

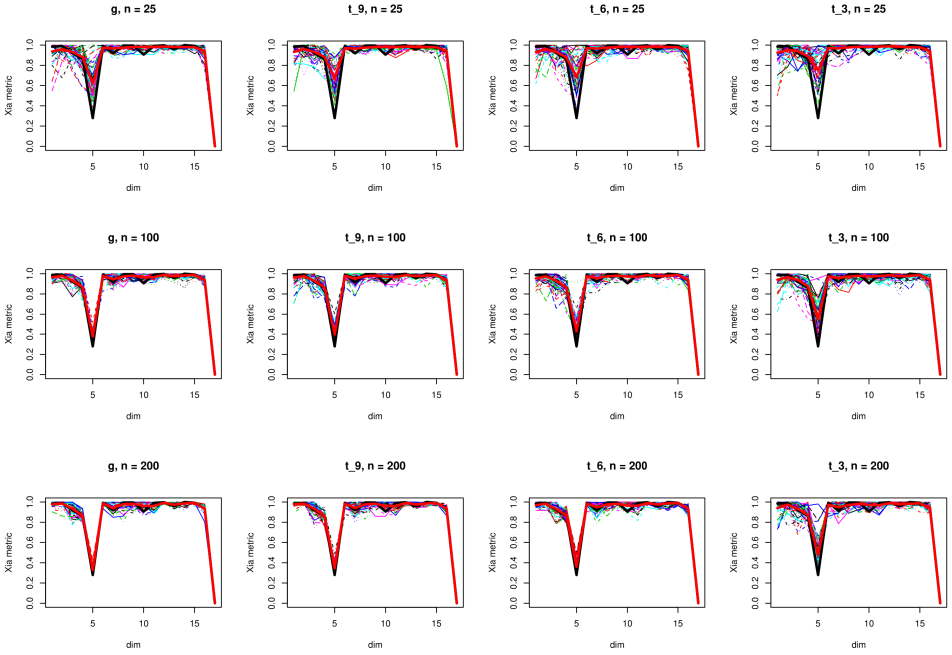
Second, from $\hat{\Sigma}$, we generate data from four potential distributions (Gaussian and t with 9, 6, and 3 degrees of freedom) and three sample sizes (25, 100, 200), from which a $\hat{\Sigma}'$ is calculated and compared to the true Σ. Similar to the previous method, this approach is able to capture the true effective dimension well for higher signal to noise ratios (10, 4), larger degrees of freedom, and lighter tailed distributions. Figure 4.10 shows results for these simulations. For a 17 dimensional data set, bootstrapping only 25 samples breaks down very quickly as the amount of noise increases.

True Covariance

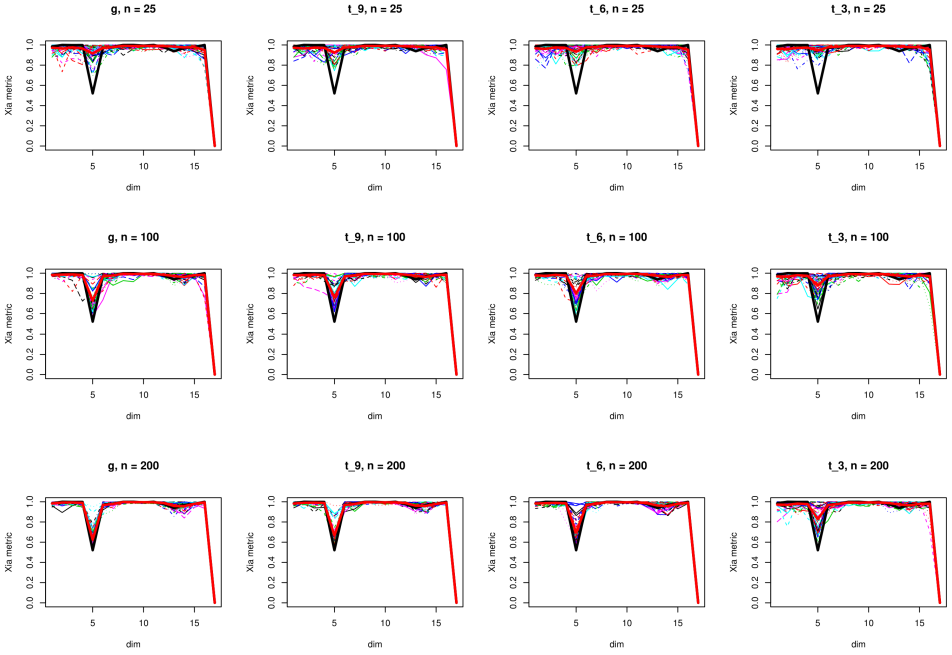
Second, we consider the case where only one true covariance matrix Σ is given. The simulation procedure is similar to the second scenario discussed previously, but data is generated (from four potential distribution using three sample sizes) using Σ as $\hat{\Sigma}$ is not given. Figure 4.11 shows a portion of these simulations results, which are further summarized in Figure 4.12.

Figure 4.12 shows the breakdown of the Xia metric using this bootstrap approach under varying simulation parameters. Green indicates the correct effective dimension was identified and red indicates no reduction in effective dimension detected.

Many real world applications will involve larger dimensions and differing effective dimensions. This bootstrap approach is tested for covariance matrices of dimension 100 and 200. Although detailed simulation results are not shown, the heatmaps in Figure 4.13 give an idea of the breakdown of the Xia metric.

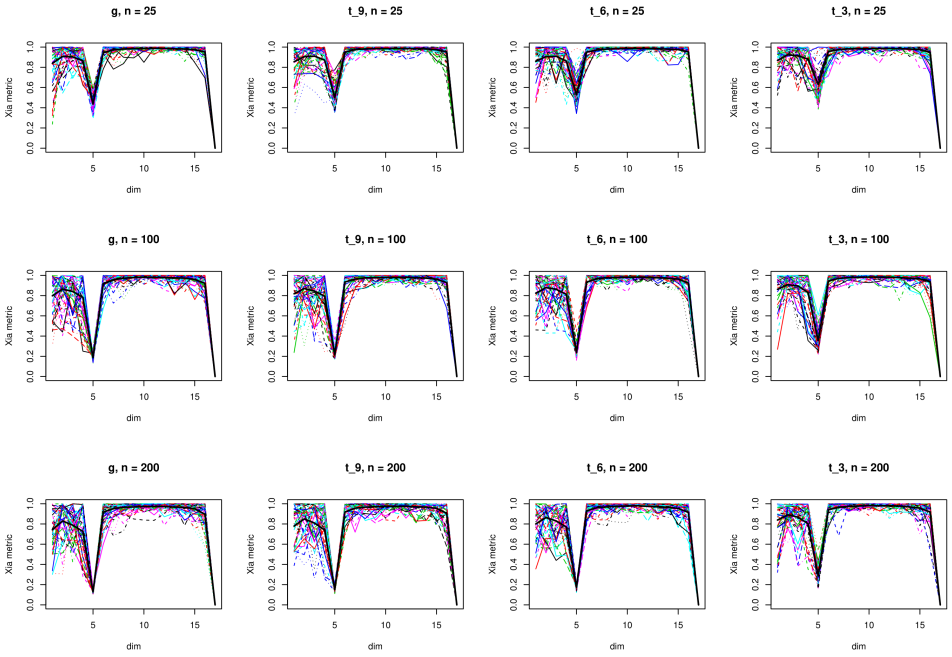


(a) Signal to noise ratio of 10

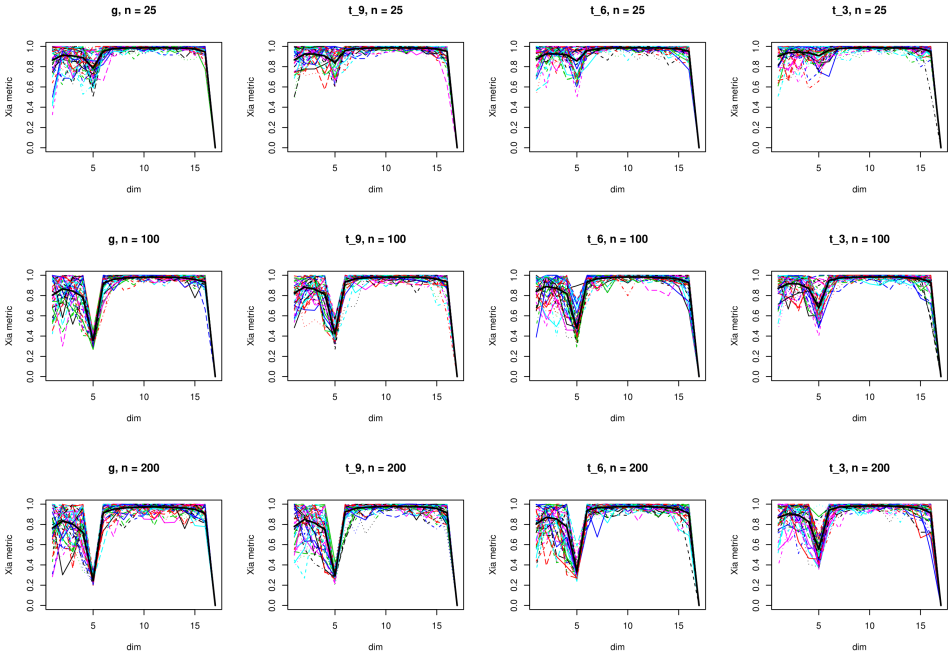


(b) Signal to noise ratio of 4

Figure 4.10: Xia metric simulation results for the bootstrap approach where data generated from varying distribution with covariance $\hat{\Sigma}$ is used to estimate $\hat{\Sigma}'$. This approach works well for high signal to noise ratios in all conditions (see (a)), but breaks down very quickly if sample sizes are too small or distributions too heavy tailed (b). The black line indicates the Xia metric calculated from Σ and $\hat{\Sigma}$; the red line plots the mean over all simulation runs using $\hat{\Sigma}'$.



(a) Signal to noise ratio of 10



(b) Signal to noise ratio of 4

Figure 4.11: Xia metric simulation results for the bootstrap approach where data is generated from the only given Σ. This approach works well for high signal to noise ratios in all conditions (see (a)), but breaks down very quickly if sample sizes are too small or distributions too heavy tailed (b). The black line plots the mean over all simulation runs using.

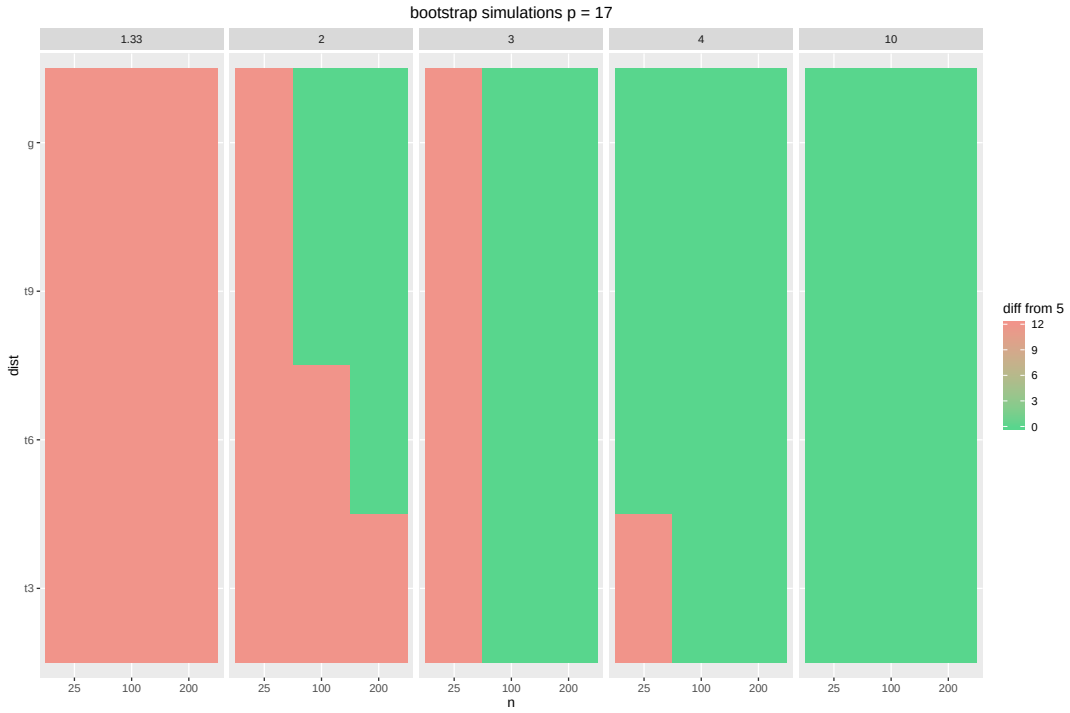
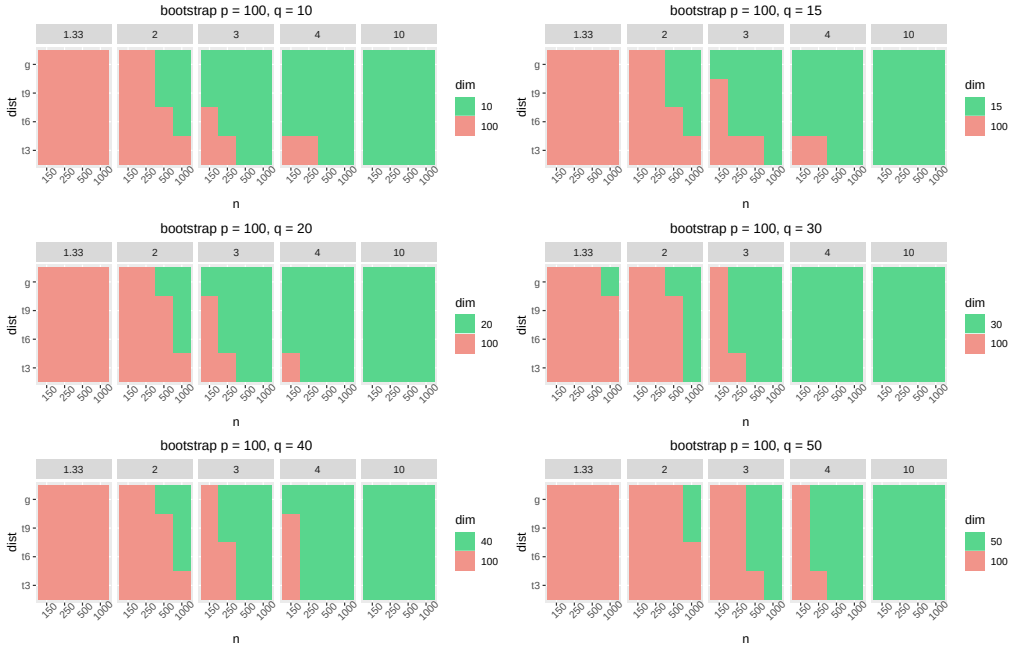
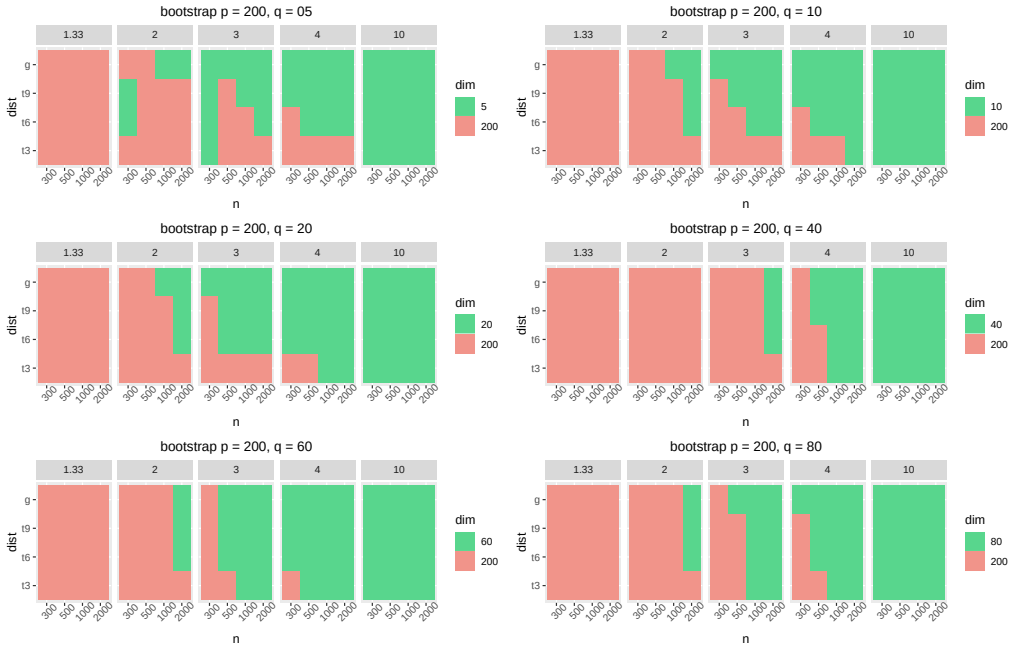


Figure 4.12: Heatmap visualizing the breakdown of the Xia metric under various simulation parameters (signal to noise ratio, number of diagonals, and amount of variation). Green indicates the Xia metric correctly dropped at $p = 5$, and red indicates no reduction in effective dimension detected ($p = 17$).



(a) bootstrap simulations with dimension $p = 100$.



(b) Bootstrap simulations with dimension $p = 200$.

Figure 4.13: Heatmaps visualizing the breakdown of the Xia metric for larger dimensions with varying true effective dimension q.

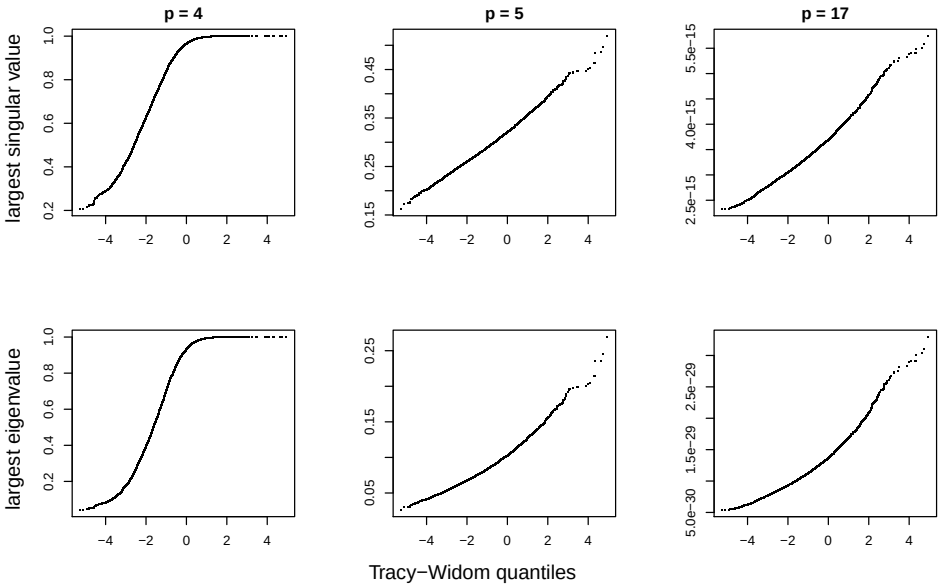


Figure 4.14: QQ plots for the largest singular value and eigenvalue against Tracy-Widom quantiles ($\beta = 1$) at $p = 4, 5, 17$.

4.1.3 Distribution of Xia Metric

In Section 3, it was mentioned that the largest singular value could be thought of in terms of its square – the largest eigenvalue, which is known to follow a Tracy-Widom distribution. To confirm this, quantile-quantile (QQ) plots of quantiles from the Tracy-Widom distribution were plotted against quantiles of the largest singular value and eigenvalues. The Tracy-Widom distribution parameter can take on three values, $\beta = 1$, 2, or 4, corresponding to Gaussian Orthogonal Ensembles, Gaussian Unitary Ensembles, and Gaussian Symplectic Ensembles, respectively [56]. QQ plots using $\beta = 1$ and 2 both show linear trends, and results for $\beta = 1$ are shown below.

For both the eigenvalues and singular values, QQ plots show linear trends at larger signal to noise ratios of 10 and 4 at both $p = 5$ and 17. Those for smaller signal to noise ratios are only linear at $p = 17$, with QQ plots at all other dimensions showing heavy right tails. This trend is also seen for data up to dimension 100 with effective dimension ten. Figure 4.14 shows the QQ plots for SNR = 4 at certain dimensions.

4.2 Largest Principal Angle

4.2.1 U-Based Visual Method

Section 3.1.2 illustrated the diagonal structure of the $U_1^T U_2$ metric when comparing classical and robust estimates of covariance matrices of data sets with known outliers for a series of simulated and small yet well known data sets. This is further illustrated with Data Set 3.

Data Set 3 contained data on float glass panes from two manufacturers, Company A and Company B. Figure 1.2 showed that log concentrations of certain elements (about half of the 18 elements) varied rather largely between the two manufacturers. Rather than manually contaminating data sets by creating outliers ourselves, which may be unrealistic as the distribution of trace elements in float glass is unknown (at least to us), we can use data from one of the two companies as the "true" data and slowly introduce data from the other company as "outliers." Here we will use panes from Company A as the "true" data and panes from Company B as "outliers"; where the "outliers" are outlying from the "true" data by at least 10% in half of the 18 elements.

Each float glass pane had around 24 fragments (some had fewer due to missing data). Glass panes were randomly selected from Company A and contaminated by an increasing number of "outlying fragments" from a Company B pane. The algorithm is as follows:

1. Randomly select Company A pane, X, to represent "true" data
2. Obtain U_X from $SVD(cov(X))$
3. `for num_outliers = 1, ..., 11:`
    (a) `for num_panes = 1, ..., 100:`
        i. Randomly select Company B pane, Y, from which "outliers" will be draw
        ii. `for num_samples = 1, ..., 100:`

A. From Y, randomly select `num_outliers` fragments as "outliers"

B. Create contaminated data set Z ($n = 24 +$ `num_outliers` fragments)

C. Obtain U_Z from $SVD(cov\_rob(Z))$

D. Calculate $U_X^T U_Z$

(b) Average 10,000 resulting $U_X^T U_Z$ matrices and plot heatmap of absolute values

where `num_outliers` indicates the number of fragments to select from Company B pane Y and `num_samples` and `num_panes` indicate 100 samples of `num_outliers` sampled for 100 panes. `SVD` denotes the singular value decomposition, `cov` calculation of the classical covariance matrix, and `cov_rob` calculation of Rousseeuw's robust MCD estimator.

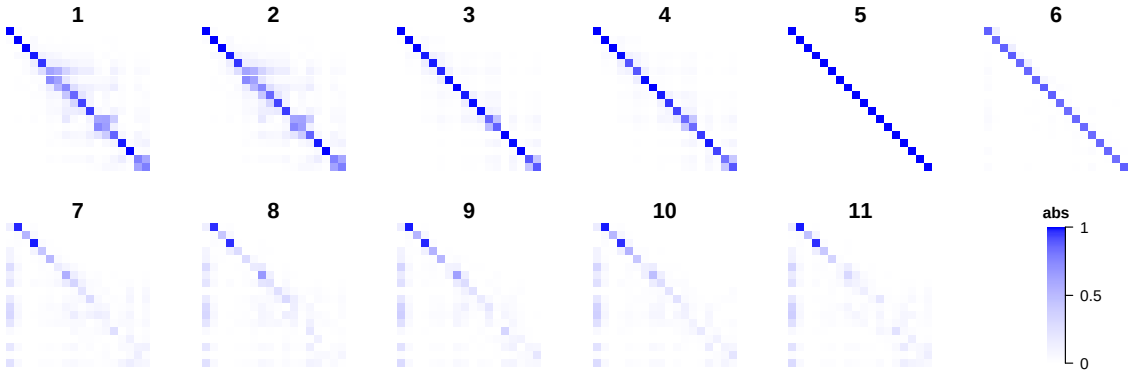
This algorithm was run for all 31 panes from Company A, and the majority display a similar trend (see Figure 4.15 (a) and (b)). The strong diagonal structure can be seen for up to `num_outliers = 6`, after which point the MCD estimator breaks down. The finite-sample breakdown value (FSBV) of the MCD estimator, given by $\lfloor (n-p+1)/2 \rfloor / n$ [85], would be around

$$\begin{aligned} \frac{\lfloor (24-18+1)/2 \rfloor}{24} &= \frac{3}{24} = 12.5\% \\ \frac{\lfloor ((24+6)-18+1)/2 \rfloor}{24+6} &= \frac{6}{30} = 20\% \\ \frac{\lfloor ((24+7)-18+1)/2 \rfloor}{24+7} &= \frac{7}{31} = 22.58\% \end{aligned}$$

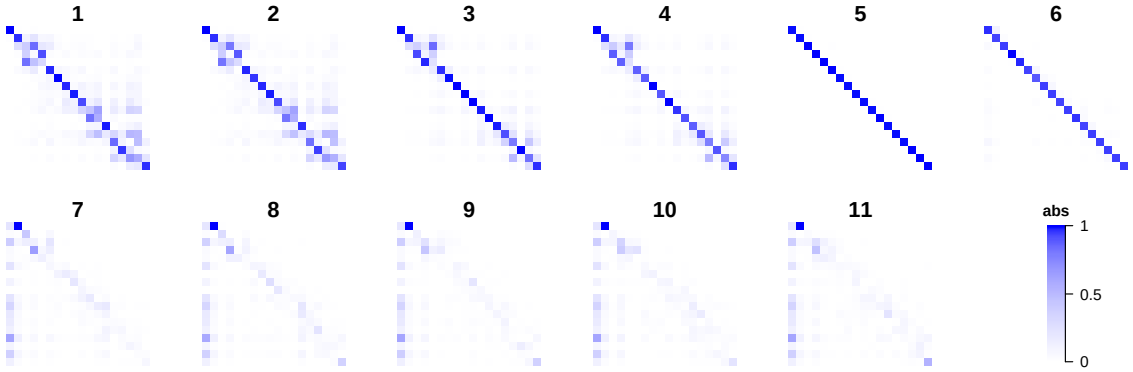
for six and seven outlying fragments. Figure 4.15 (c), Pane 20, was the only pane to show a breakdown at only five outliers, likely due to variability in that single pane.

4.2.2 Largest Principal Angle

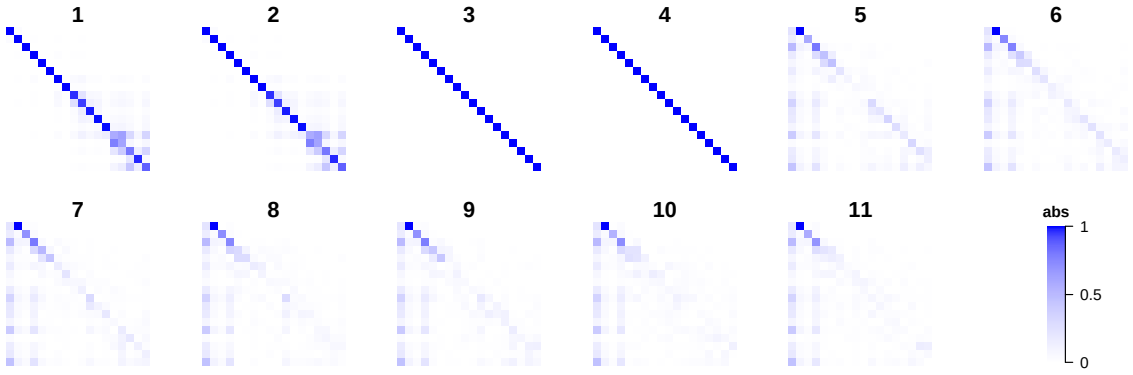
In addition to simply calculating the Asimov distance, we are interested in the 95th percentile to get an idea of what a cutoff value and what the tail end of the distribution looks like. Simulations using the covariance matrices estimated from three of our motivating



(a) Pane 24



(b) Pane 30



(c) Pane 20

Figure 4.15: $U_1^T U_2$ metric applied to Data Set 3, where U_1 is obtained from the classical covariance of data with no outliers (only Company A data) and U_2 is obtained from the robust MCD estimator of data containing outlying fragments (from Company B). Of the 31 Company A panes, all but one (Pane 20, subfigure (c)) showed strong diagonal structure with varying levels of off-diagonal noise until seven outlying fragments were introduced, at which point the MCD estimator broke down. Pane 20, (c), was the only pane to break down after four outlying fragments.

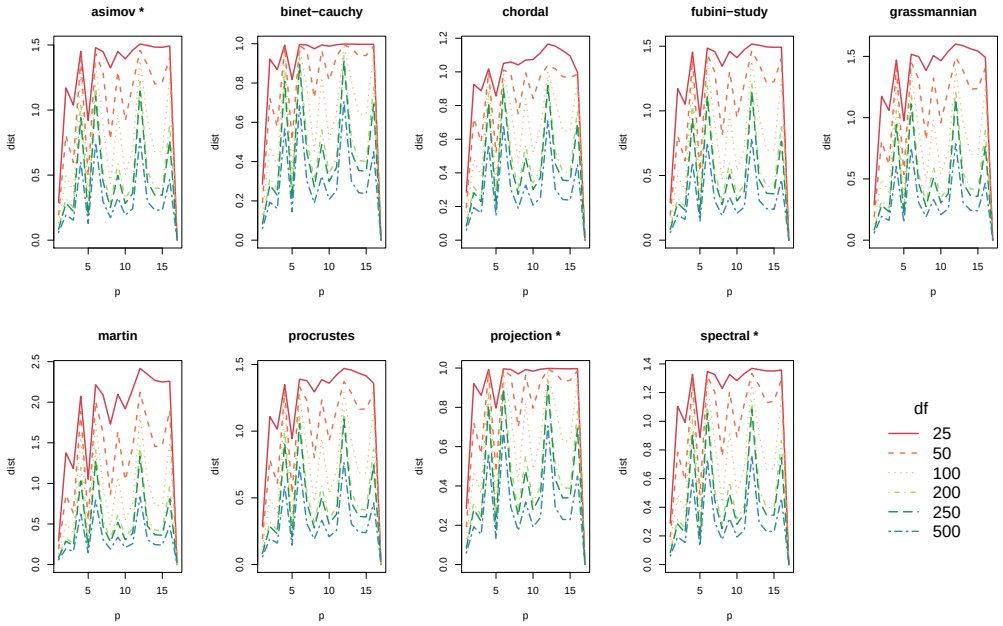
data sets were performed with varying degrees of freedom ($p = 17$ for all). Results from Data Set 1 are shown in Figure 4.16. Simulations from Data Sets 2 and 3 display similar oscillating patterns (see Appendix A). As the degrees of freedom increases, the range in which the metrics oscillate increases. This stabilizes when the degrees of freedom is 100 or more, indicating that less than 100 degrees of freedom may be far too few and unable to capture enough information for data of dimension 17.

This oscillating pattern may be due to small eigengaps. As the eigengaps decrease towards zero, it becomes increasingly difficult to distinguish between consecutive subspaces and may result in eigenvectors swapping positions – causing the overall subspaces at times to look more (or less) similar with the presence of just one additional eigenvector. There are two ways to test this hypothesis: (1) set pre-specified eigengaps of known size; and (2) largely increase the degrees of freedom to mitigate the effect of small eigengaps. Simulations were performed with modified versions of the Data Set 1 covariance matrix with specified eigengaps and larger degrees of freedom.

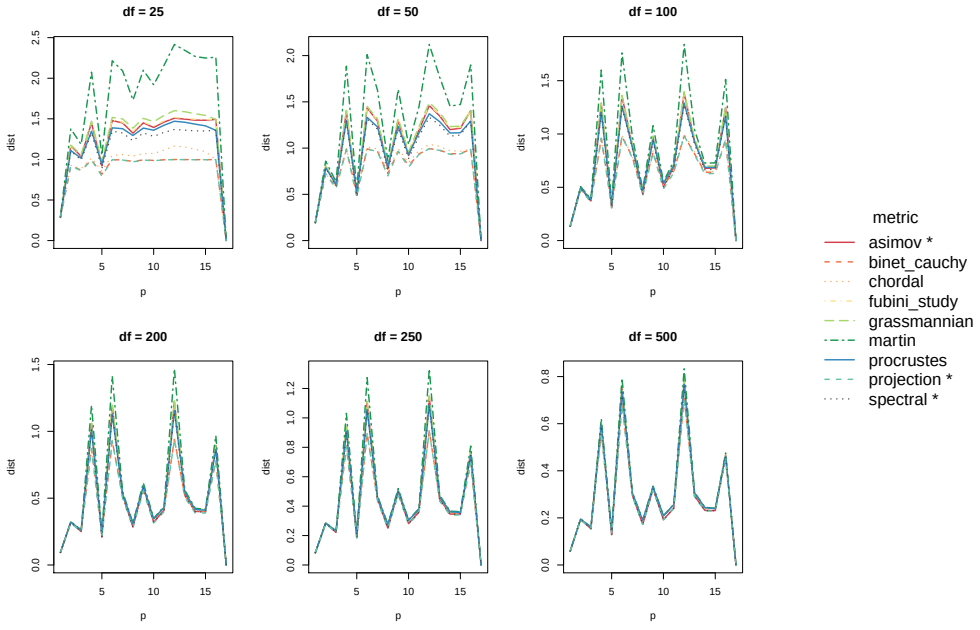
Table 4.1 shows some of the singular values and corresponding eigengaps used in simulations, labeled 5 through 10. True covariance matrices Σ were created by decomposing the Data Set 1 covariance matrix into the corresponding eigenvectors and singular values and reassembling the eigenvectors with using the singular values in Table 4.1. The eigengaps for simulation 5 and 6 are relatively similar, and eigengaps 7 and 8 are generated by multiplying eigengap 5 by a factor of 100 and 10, respectively. These are referred to in the following as "gap5" through "gap10."

Figure 4.17 shows the largest principal angle for five of the simulation scenarios above, as well as the logarithms of the distances. The distribution of the largest principal angle at various dimensions seem to vary widely, with some having considerably longer and heavier tails. These look much more normal after taking logarithms.

Figure 4.18 shows simulations results from gap5 in more detail. The vertical bars below each dimension correspond to a scaled version of the eigengap. As eigengap decreases,



(a) Subspace distance metrics for Data Set 1. The metrics all show an oscillating pattern as p increases, though it is not as clear when the degrees of freedom is close to p.



(b) Subspace distance metrics by degrees of freedom. Variability in the metrics decreases as the degrees of freedom increases. An oscillating pattern can still be seen.

Figure 4.16: Nine metrics from Table 2.1 for Data Set 1. An oscillating pattern can be seen for each of the metrics over varying dimensions p. The asterisk indicates distances involving only the largest principal angle.

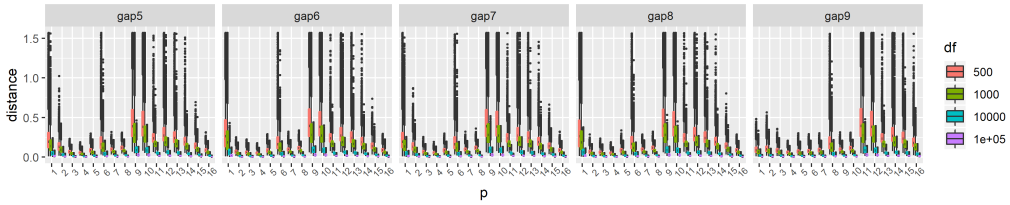
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
5	12	10	7	3	1	0.5	0.4	0.2	0.1	0.09	0.08	0.06	0.05	0.04	0.03	0.02	0.01
6	18	16	7	3	1	0.5	0.4	0.2	0.1	0.09	0.08	0.06	0.05	0.04	0.03	0.02	0.01
7	1200	1000	700	300	100	50	40	20	10	9	8	6	5	4	3	2	1
8	180	1600	70	30	10	5	4	2	1	0.9	0.8	0.6	0.5	0.4	0.3	0.2	0.1
9	45	28	18	10	4	2	1	0.5	0.4	0.2	0.1	0.09	0.08	0.06	0.05	0.04	0.03
10	100	60	35	20	10	4	2	0.5	0.4	0.2	0.1	0.09	0.08	0.06	0.05	0.04	0.03
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
5	-	2	3	4	2	0.5	0.1	0.2	0.1	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
6	-	2	9	4	2	0.5	0.1	0.2	0.1	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
7	-	200	300	400	200	50	10	20	10	1	1	2	1	1	1	1	1
8	-	20	90	40	20	5	1	2	1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1
9	-	17	10	8	6	2	1	0.5	0.1	0.2	0.1	0.01	0.01	0.02	0.01	0.01	0.01
10	-	40	25	15	10	6	2	1.5	0.1	0.2	0.1	0.01	0.01	0.02	0.01	0.01	0.01

Table 4.1: Singular values (top) and gaps (bottom) used to create covariance matrices for six different simulations.

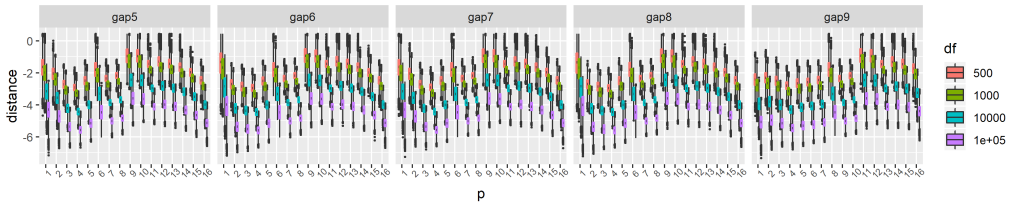
notably when $p \geq 6$, an oscillating pattern can be seen to emerge with long as heavier tails in the distribution of the Asimov distance. The colored lines indicate the 95th percentile of the Asimov distance. Together, Figures 4.17 and 4.18 indicate that the size of the eigengaps seems to be the determining factor causing the oscillating pattern rather than the magnitude of the singular values.

Following figures and approaches focus on data from simulations using gap5 and gap10, which have eigengaps of relatively different magnitude and pattern. Figure 4.19 shows the largest angle and logarithm of the largest angle for these simulations, as well as the corresponding plot of eigengaps for smaller degrees of freedom (from 25 to 250).

While data sets exist where having 100,000 or more degrees of freedom for 17 dimensions is realistic, it is improbable this will ever be the case for forensic glass evidence. Nonetheless, the approach may apply to "big data" scenarios. Figure 4.20 shows the Asimov distance for ten different degrees of freedom ranging from 25 to 100,000. There is a relatively clear decreasing trend between the largest principal angle and increasing degrees of freedom. These are confirmed by the QQ plots of the logarithms of the Asimov distances in Figure 4.21, especially if smaller degrees of freedom, namely 25 and 50, are overlooked. QQ plots for simulations using gap5 have similar trends so are not shown; the only difference being the specific dimensions at which heavier tails exist (which is dependent on the true effective dimension of the covariance matrix). The QQ plots which



(a) The largest principal angle (Asimov distance) for five different simulated eigengaps.



(b) Logarithms of the values in Figure (a).

Figure 4.17: The largest principal angle for five different simulated eigengaps (a) as is and (b) after taking logarithms.

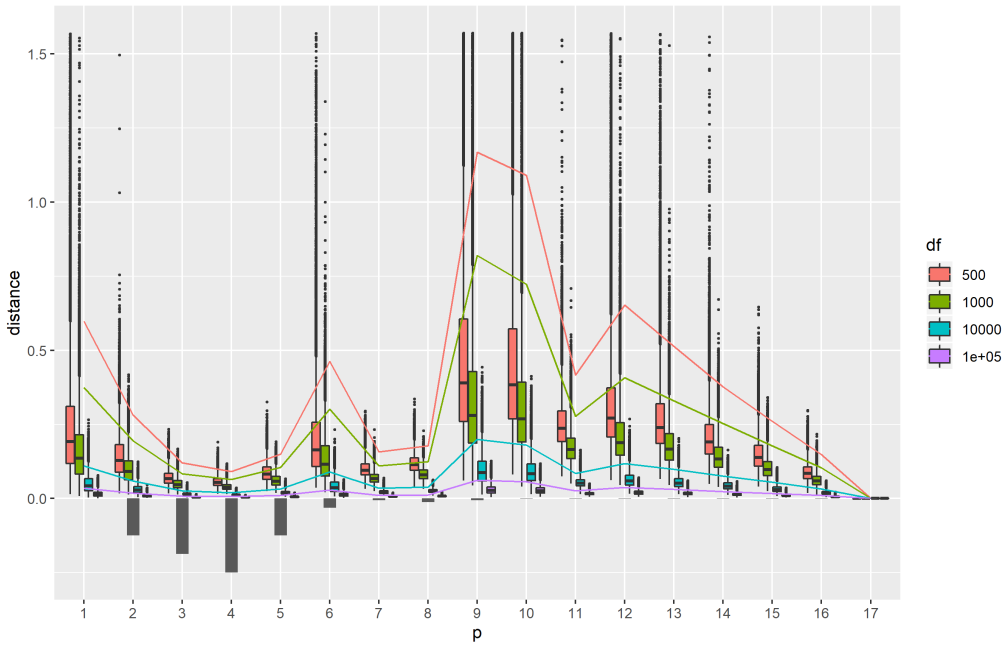
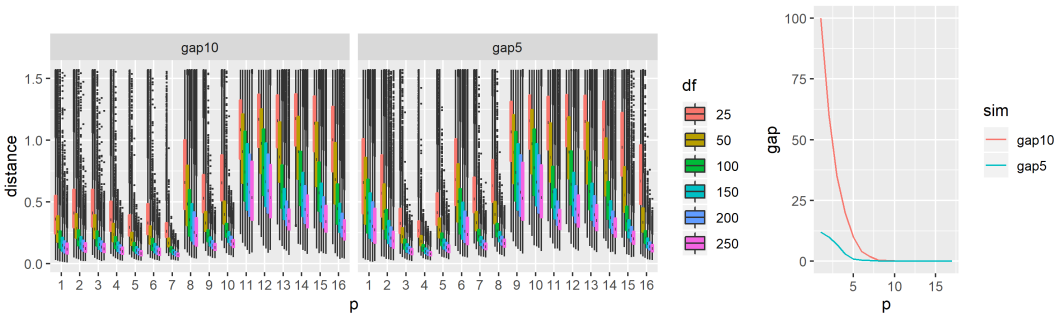
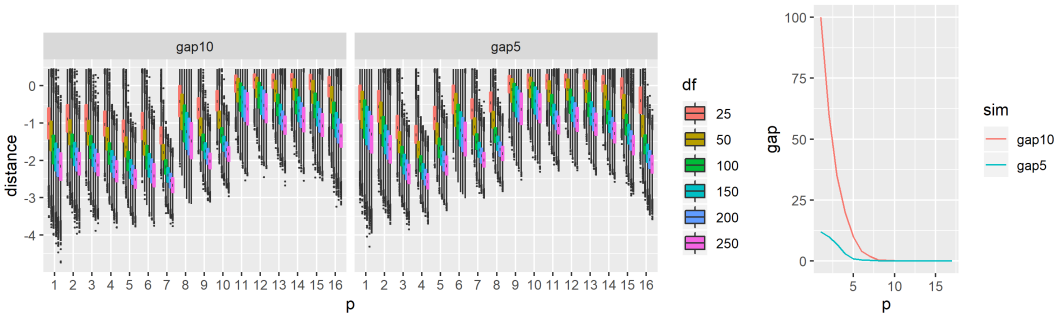


Figure 4.18: Simulations with increased degrees of freedom using gap5. The bar plot on the bottom indicates a scaled version of the eigengaps between dimensions, and the colored lines mark the 95th percentile of the largest principal angle.



(a) Largest principal angle (Asimov distance).



(b) Logarithms of the largest principal angle.

Figure 4.19: The largest principal angle for simulations using gap5 and gap10 with smaller degrees of freedom and their corresponding eigengaps.

show deviation from normality in the tails correspond to the heavy tails seen in previous figures (e.g., Figure 4.19).

From the distributions of the largest principal angle, we are interested mainly in two quartiles: the median and 95th percentile. These quartiles will give an indication of how stable and what the tail of the Asimov distance could look like, and help in determining when two subspaces can be considered "similar." In real world applications, it is likely a single data set exists or only a small number of simulations would be able to performed. Using our simulated data, we will model the log of the Asimov distance as a function of the degrees of freedom (less values 25 and 50, which are too small considering the dimension is of 17). Ideally, knowing the value of the median or 95th percentile at one specific degree of freedom will allow us to infer the rest. Figure 4.22, using values from the 95th percentile, indicates a clear linear relationship between the logarithm of the Asimov distance and the logarithm of the degrees of freedom exists for most p, corresponding to the p mentioned in Figure 4.21. Figures for the median are not shown as they are almost perfectly linear for all p.

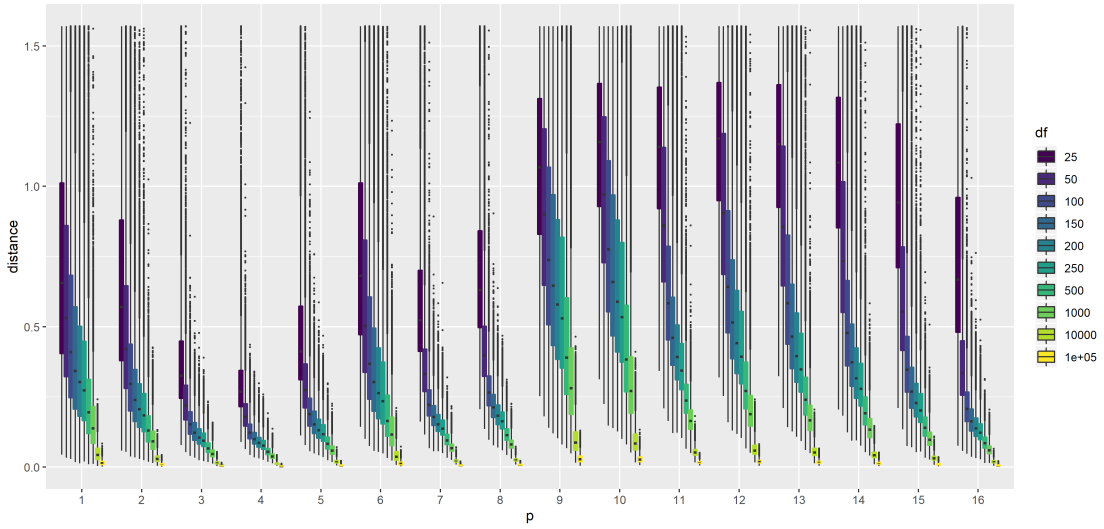
A robust regression line

$$\log(df) \sim \log(quantile)$$

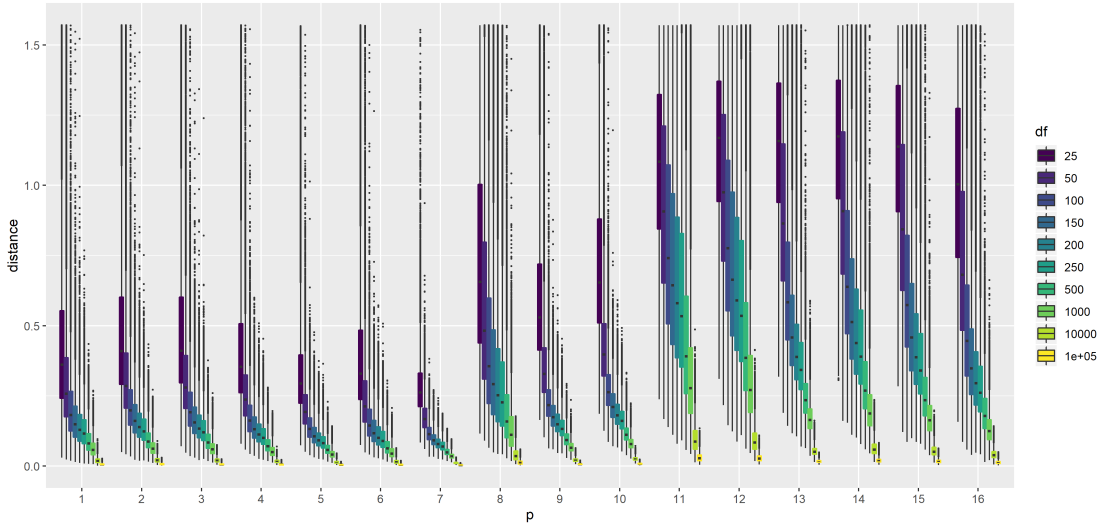
is fit for all dimensions $p < 17$ with eight degrees of freedom ($n = 100$, 150, 200, 250, 500, 1,000, 10,000, 100,000). Smaller degrees of freedom (25 and 50) were considered too variable for 17 dimensions, but theoretical quantiles may be calculated from the distribution and density functions derived in [4].

For each dimension, we estimate an intercept and slope, as well as biweight regression weights w for each point using the residuals r.

$$w_i = \left(1 - u^2\right)^2, \text{ where } u = \begin{cases} \frac{r}{4 \cdot \mathrm{sd}(r)} & |u| \leq 1 \\ 1 & \text{otw} \end{cases}$$



(a) Simulations using gap5



(b) Simulations using gap10

Figure 4.20: The largest principal angle for simulations using gap5 and gap10 at different values of degrees of freedom ranging from 25 to 100,000.

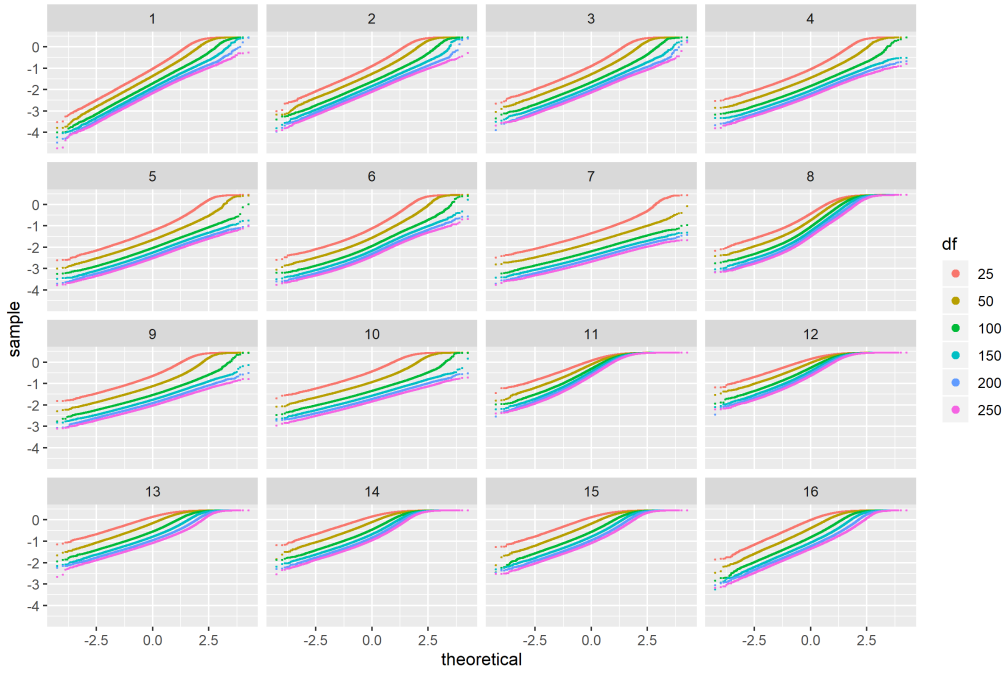


Figure 4.21: QQ plots for gap10 simulation data. Lines for larger degrees of freedom are omitted as the show similar trends to those plotted. Very clear linear patterns can be seen for dimensions up to $p = 10$ (excluding $p = 8$). The remaining dimensions have much heavier tails.

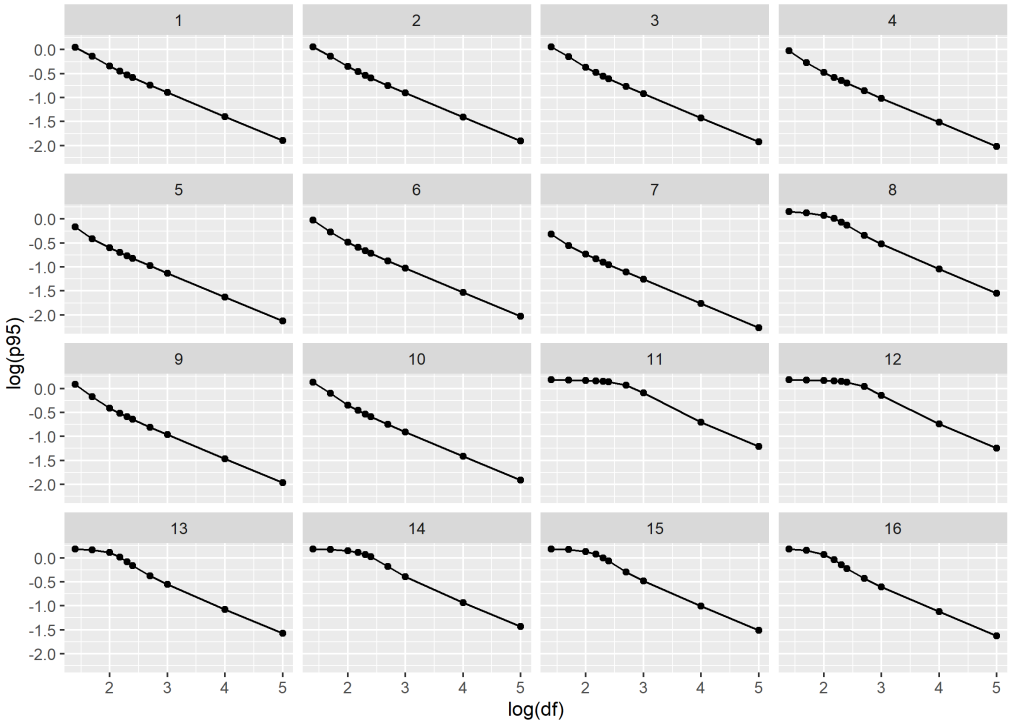


Figure 4.22: log(df) $\sim$ log(quantile) for each dimension p; data from the 95th percentile of simulations using "gap10." A very clear linear relationship exists between the logarithms of the Asimov distance and degrees of freedom.

These weights are used to estimate the standard error, from which an approximate 95% confidence interval is calculated and shown for each dimension. While either the intercept or slope may be used to estimate a threshold, the slope tends to be more stable, and is here as well (Figure 4.23). We can see that around true effective dimension $p = 9$ in (a) and $p = 11$ in (b), the slope spikes above the mean and median. As expected, the 95th percentile is much more variable than the median.

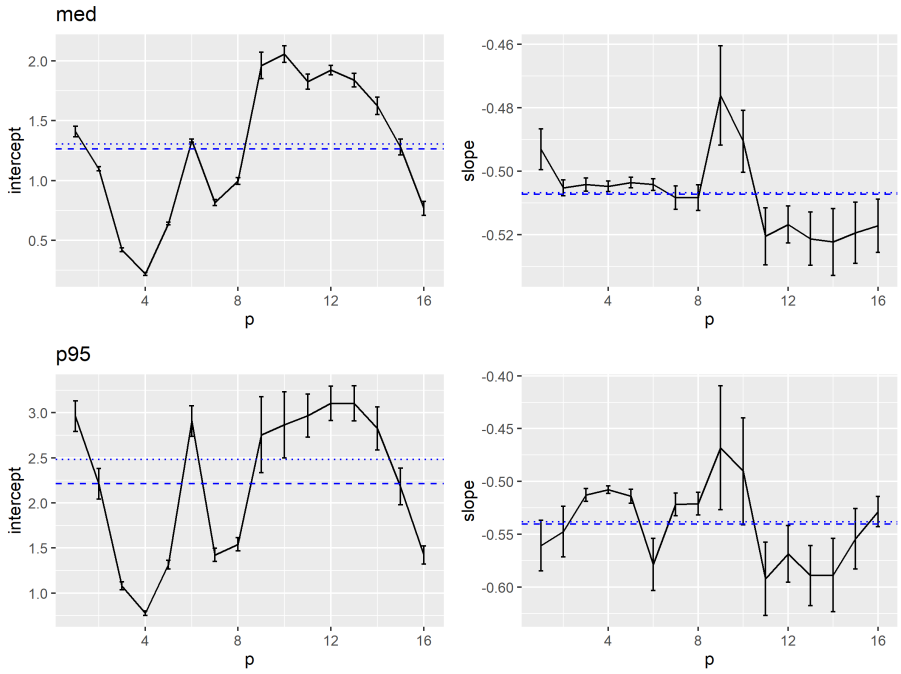
Figure 4.23 indicates the slope values to be around -0.51 for the median (or 50th percentile) and -0.54 for the 95th percentile. However, many of the confidence intervals for the 95th percentile seemed to include -0.51. Future work can be done to determine if the quantile has a significant effect on slope (and intercept) values; and if so, how they are related.

Recalling the eigengaps in Table 4.1, we can see that the eigengaps and singular values for gap10 drop at $p = 8$, and again so (albeit at a smaller magnitude) at $p = 11$ before shrinking to very small values. This seems to be reflected in Figure 4.23(b), as the slope dips down below the mean and median at $p = 8$ before jumping up at $p = 11$. In analyzing a scree plot (see Figure 4.19), one might conclude the effective dimension is $p = 8$. Using this method based on the largest principal angle, one might decide to keep up to $p = 11$ dimensions. In other words, using the largest principal angle may give a conservative estimate of the true effective dimension.

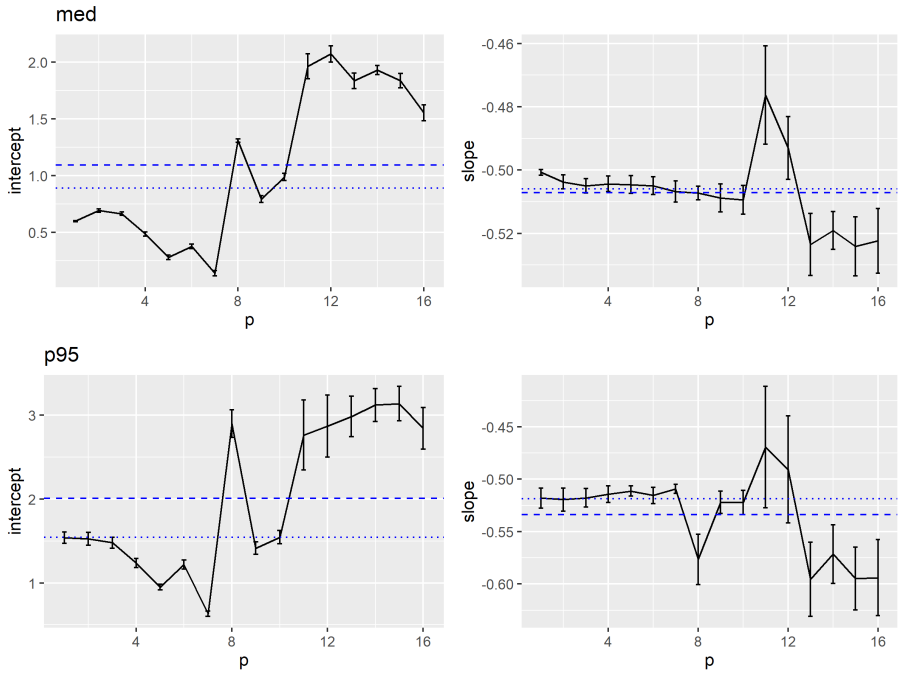
4.3 Combining covariance matrices

Proposed methods for combining covariance matrices can be illustrated via heatmaps. Covariances and the three estimates discussed in Section 3.2 are visualized by plotting absolute values of the heatmaps of the corresponding correlation matrices. In the following figures, proposed estimation methods as numbered as in Section 3.2 and labeled as:

1. Weighted average – “weighted avg”



(a) Simulations using "gap5."



(b) Simulations using "gap10."

Figure 4.23: Intercept and slope values for the median and 95th percentile. Slope is more stable than intercept, and median more stable than the 95th percentile. Blue dashed lines indicate the mean and blue dotted lines indicate the median.

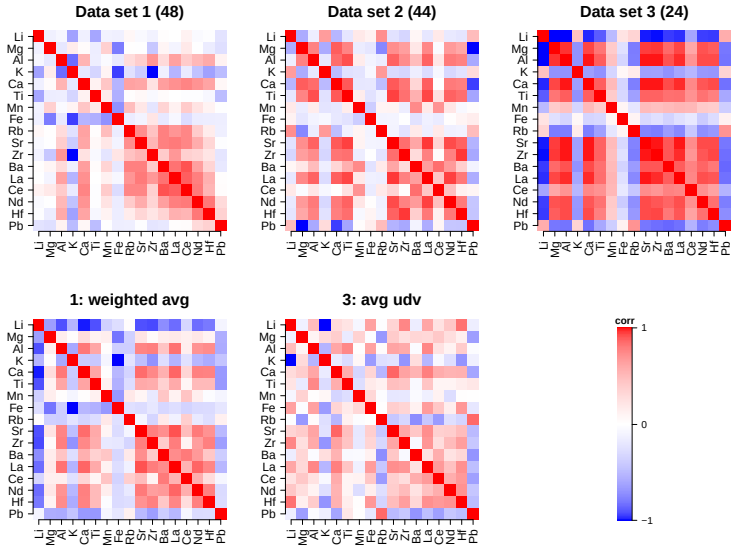


Figure 4.24: Two different estimates, using methods (1) and (3), resulting from combining covariance matrices from Data Sets 1, 2, and 3 (one pane selected at random). Heatmaps plot absolute values of correlations (blue = -1, white = 0, red = 1); and numbers in parentheses indicate sample sizes.

2. U from the pooled data and average individual D matrices – "share uv, avg d"

3. Average U and D matrices – "avg udv"

We first illustrate these estimates using Data Sets 1, 2, and 3, assuming data is unavailable and using only the covariance matrices (thus, only estimates (1) and (3) can be calculated). Only one pane from Data Set 3 was used for consistency.

We can also illustrate these estimates using pairs of the three data sets. Each column in Figure 4.25 shows estimates (1) and (3) for a pair of Data Sets: 1 and 2, 1 and 3, and 2 and 3, respectively.

These estimates are further applied to Data Set 4, in which we use available data to split certain glass categories into subcategories, then combine the covariance matrices of these subcategories. Since data were available, we calculated estimate using method (2) as well for comparison. First we will look at container glass (Figure 4.26). These glass samples under container can be classified into one of three groups: (1) alcoholic beverage bottles; (2) beverage bottles; (3) other (including baby food, olive jars, etc.). The first

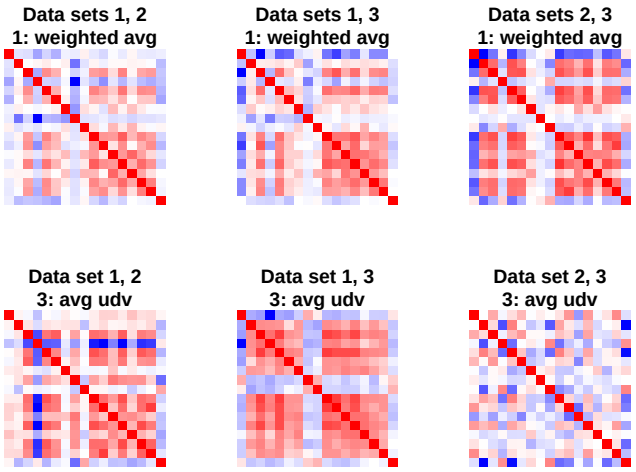


Figure 4.25: Covariance estimates using methods (1) and (3) for all pairs of Data Sets 1, 2, and 3. Each column represents the two estimates, which can be seen to differ. Heatmaps plot absolute values or correlations (blue = -1, white = 0, red = 1).

row of Figure 4.26 indicates that correlation (covariance) matrices between the subgroups are considerably different (the number in parenthesis indicates sample size n). The second row show correlation matrices of the three estimates.

Similar to container glass, float architecture glass comes from four main manufacturers: Guardian, PPG, Temp Glass, and Cardinal, all of which have relatively different correlation matrices (Figure 4.27). Float automotive glass comes from two categories: CFS (Centre of Forensic Sciences in Canada) casework samples and non-CFS (Figure 4.28).

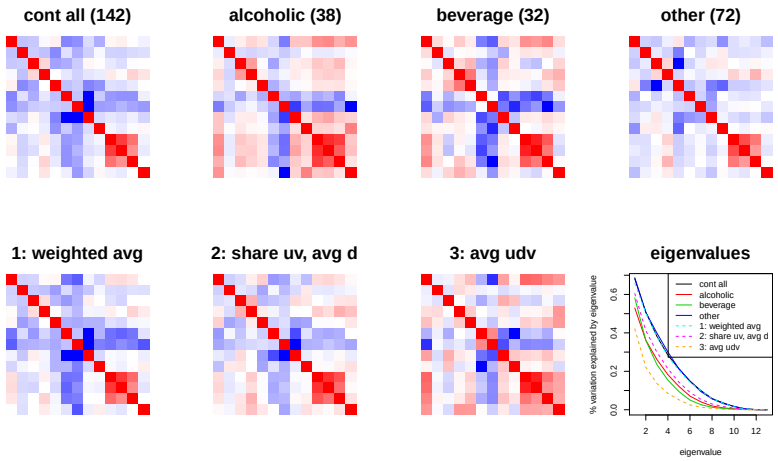


Figure 4.26: Correlation (covariance) matrices of container glass and its three subcategories: alcoholic beverage bottles, beverage bottles, and other (baby food, ketchup jars, etc.). Heatmaps plot absolute values (blue = -1, white = 0, red = 1); and numbers in parentheses indicate sample sizes. The first row shows that correlation matrices of subcategories are considerably different. The second row shows correlation matrices of the three estimates discussed in Section 3.2 and the percentage of variation explained by eigenvalues.

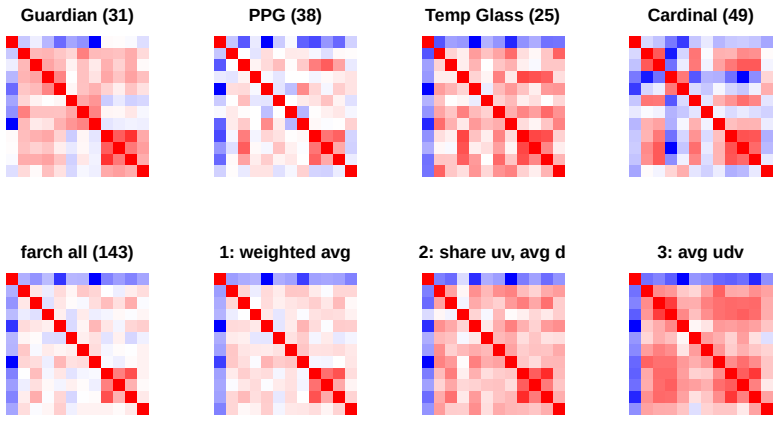


Figure 4.27: Correlation (covariance) matrices of float architecture glass and the four main manufacturers: Guardian, PPG, Temp Glass, and Cardinal. Heatmaps plot absolute values (blue = -1, white = 0, red = 1); and numbers in parentheses indicate sample sizes. Correlation matrices from four manufacturers (first row) are considerably different. The second row shows correlations of all the data and the three estimates. A scree plot is not show, but resembles that of Figure 4.26.

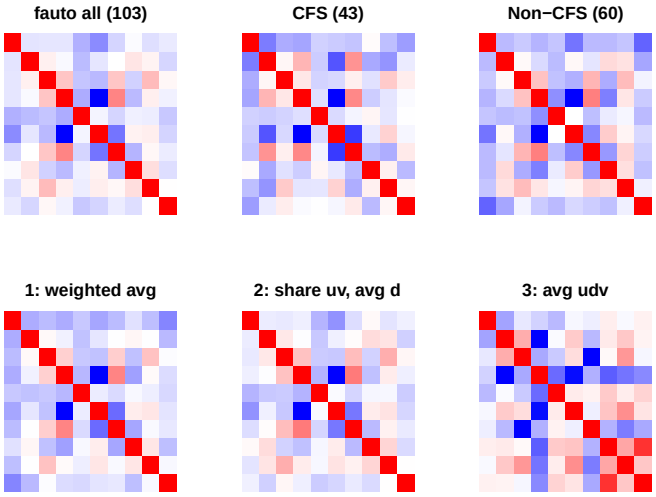


Figure 4.28: Correlation (covariance) matrices of float automotive glass from CFS (Centre of Forensic Sciences in Canaday) casework samples and non-CFS (other) samples. Heatmaps plot absolute values (blue = -1, white = 0, red = 1); and numbers in parentheses indicate sample sizes.

Chapter 5

Conclusions and Future Work

In this chapter, we summarize advantages of the proposed metrics and methods concerning covariance matrices over existing methods. After that, we discuss potential directions for future study of these topics.

5.1 Conclusions

Two methods for comparing the similarity of covariance matrices and estimating true effective dimension were presented, as well as a method for estimating a combined (pooled) covariance matrix. These methods are based on the belief that analysis and comparison of covariance matrices should be based on the subspaces they span, which we obtain through the singular value decomposition. Our motivating forensic science data originate from a field where it is often the case that data is inaccessible or otherwise unavailable; and these methods can be performed given only information on covariances (or correlations). The main advantages over existing methodology are simplicity, lack of distribution assumptions, no sample size requirements, and fast computation time.

The lack of minimum sample size requirement is crucial for forensics glass data, which is of small n, relatively large dimension p, and is known to contain outliers. This is also

the reason many robust covariance estimators, which may require $n > 5p$, are unsuitable. However, the simplicity and fast computation times of the proposed methods allows for applications to much larger data sets (and their covariance matrices) with large sample size and dimensionality. They may be especially applicable to fields (e.g., social media, marketing, etc.) with constantly streaming (yet possibly confidential) data.

The proposed methods for determining effective dimension may be preferable to simple eigenvalue-based methods (e.g., scree plot). When n and p are of similar magnitude, the eigenvalues of a sample covariance matrix are known to be poor estimators of the population covariance (the largest sample covariance eigenvalue tends to overestimate that of the population covariance); and in "large n, large p" asymptotics, sample covariance matrix eigenvectors "are not consistent estimators of the population eigenvectors" [32, 79]. Furthermore, our simulations indicate the largest principal angle method may give a more conservative estimate of effective dimension than a scree plot.

5.2 Practical Implications

Although these methods may suggest two covariance matrices have the same effective dimension, whether or not it is appropriate to combine them depends heavily on the context. Our motivation arises from glass chemical composition data from various labs. The decision to combine or not will benefit greatly from subject matter experts; one can imagine an ideal scenario for combining covariance matrices would be when they have similar effective dimension and similar elements. If the effective dimension is the same yet specific elements vary with little or no overlap, chemists may be able to inform us how meaningful or appropriate (or not) it would be to combine these matrices. Similarly, different detected effective dimensions may indicate the subspaces that data span are very different and should not be combined.

Rotating one subspace into another, regardless of dimension, may result in a loss of

information. Knowing two subspaces have the same (or similar) effective dimension can provide information that one can be rotated into the other, which is in itself useful. For example, continuing with the forensic glass lab motivation, the information may suggest that neither lab needs to measure all 17 elements: five dimensions will suffice, determine by five principal components, but those components may involve different linear combinations of different sets of 7 elements, depending on the lab (one lab may have better facilities for measuring one set of five elements than another). Another aspect for consideration is the scale. If the eigenvalues defining two "similar" subspaces are of very dissimilar magnitude (e.g., diagonal values of 1 versus 20), this seems to indicate lab measurements differ at least with respect to scale, and thus may be inappropriate to combine. The proposed methods may help us conclude the structure of the data is similar if not accounting for scale. Again, in the case of labs, one lab may have far greater precision than another - inspiring the less precise lab to undergo an investigation for its greatly reduced precision relative to the other lab. Either way, the information about effective dimension has been insightful.

Generally, any decision involving combining or rotating subspaces will depend heavily on the application – the actual variables involved, how interpretation may be affected, and the overall aim. The needs, requirements, and guidelines for combining subspaces will differ when considering combining data on glass panes versus those for combining medical data related to detection of disease.

5.3 Future Work

In this section, we discuss potential future directions building on top of the proposed methods for covariance comparison and combination.

5.3.1 Comparison and Effective Dimension Methods

At the beginning of Section 3, we mentioned that ideally, a set of covariance matrices ranging from identical to very different would be ideal for testing our proposed methods on. One direction would be to identify such a set of covariances and analyze the performance of our methods as well as other methods mentioned in Section 2.

We chose to use the Asimov distance, or the largest principal angle, as a metric. However, many potential distances were identified in Table 2.1. Future directions include identifying under which cases metrics based on one or more principal angles is most appropriate.

While visual examination of Xia metric plots may be adequate for determining dimensionality (effective dimensions), a numerical test or threshold is necessary. Recall that the Xia metric is defined as the largest singular value of the difference between two matrices. Although eigenvalues of random matrices are known to follow Tracy-Widom distributions, more theoretical research can be done in this area. Random matrix theory literature, which discusses derivations of asymptotic distributions for singular values of differences between two random matrices [30] and eigenvalues of covariance matrices [55, 79, 31, 32], may provide a basis to confirm the Tracy-Widom distribution is appropriate, as well as potentially detect a threshold (e.g., a 95th percentile) of the Xia Metric score. Potential questions to be considered include:

- Does the distribution depend on number of non-zero singular values?
- What if the matrix is singular?
- What are the assumptions? (matrix singular, effective dimensions, etc.)

Empirically, it may also be possible to calculate a threshold or critical value by simulating and sampling of two $\hat{\Sigma}$ from a known Σ under the same null hypothesis. A procedure similar to that in Section 3.1.2 and 4.2.2 might also be performed.

Largest Principal Angle

Section 4.2.2 and Figure 4.23 analyzed the intercept and slope values of the median and 95th percentile for using the largest principal angle to determine similarity between subspaces. These analyses suggested the slope had a median of around -0.51 for the median or 50th percentile and a median of around -0.54 for the 95th percentile, although many of the confidence intervals for the 95th percentile seemed to include or fall very close to -0.51. Additional simulations with varying parameters can be conducted to see if this effect still holds (our analyses included data from simulations of dimension $p = 17$ and degrees of freedom larger than 100). Further work can be performed to determine if the intercept and slope values are significantly related to the specific quantile of interest.

5.3.2 Proposed Estimator

Detailed properties of the proposed estimator should be determined, including but not limited to,

- Affine equivariance (or close to)
- Breakdown point – perhaps calculated by simulation, or using sensitivity curves (comparing the performance of estimator with and without outliers) or influence functions
- Comparison to MCD or other robust estimators (and their properties)

In line with examining the properties above, it is necessary to show the proposed estimator is robust, which could be done by adding large outliers into covariance matrix (e.g., misplacing decimal places, a rather common occurrence). From this, the breakdown point of our estimator can be calculated to determine what percentage of outliers can exist before the estimator breaks down. Another direction to examine for the proposed estimator is the use of the Fréchet mean instead of a straight average of the U matrices. Subspace distance metrics mentioned in Table 2.1 may be potential distances to use.

Further study can be conducted on how outliers (and different data distributions) influence the estimator, or more specifically the U and D matrices themselves. If outliers are present, does this affect one of the SVD matrices more than the other, or do they equally display effects of outliers? We also must decide what it means to be "sensitive;" e.g., what percentage of the U matrix (number of columns) can be affected, and how this effect establishes itself (perhaps the angles of multiple columns are rotated or directions changed, or one direction is largely affected; possibly measured in degrees of percentage of degrees).

Appendix A

A.1 Asymptotic Distribution of Xia Metric

The Xia Metric in Section 3.1.1 is defined as the largest singular value of

$$U_0 U_0^T - \hat{U}_q \hat{U}_q^T \tag{A.1}$$

where U_0 denotes the matrix of directions of a "true" covariance matrix and $\hat{U}_q$ denotes directions of an estimated covariance matrix. In other words, $\hat{U}_q = U_0 + \varepsilon$ where ε denotes error in the estimation of directions of the "true" covariance matrix. The Xia Metric becomes, in turn, the largest singular value of

$$U_0 U_0^T - (U_0 - \varepsilon)(U_0 - \varepsilon)^T = U_0 U_0^T - \left(U_0 U_0^T + U_0 \varepsilon^T \varepsilon U_0^T + \varepsilon \varepsilon^T\right), \tag{A.2}$$

which we denote by V.

For any given covariance matrix, the eigenvectors U_0 are a constant matrix. [Without loss of generality] we assume that the error term, ε, follows a Gaussian distirbution with zero mean and some covariance Σ, or, $\varepsilon \sim N(0, \Sigma)$.

The expected value of V is then

$$\begin{aligned} E[V] &= E\left[U_0 U_0^{'} - \left(U_0 U_0^{'} + U_0 \varepsilon^{'} \varepsilon U_0^{'} + \varepsilon \varepsilon^{'}\right)\right] \\ &= 0 - E\left[U_0 U_0^{'} + U_0 \varepsilon^{'} \varepsilon U_0^{'} + \varepsilon \varepsilon^{'}\right] \\ &= E[\varepsilon \varepsilon^{'}] \end{aligned}$$

where the expectation of the first and third moments are zero following from the distri-

bution of ε. As the product of two random matrices, this is known to follow a Wishart distribution with expected value given by the degrees of freedom multiplied by the covariance matrix. Letting e_i denote the columns of ε, Schott (2016) [92, p. 488] Theorem 11.28 derives this as:

$$E[\varepsilon\varepsilon^{'}] = \sum_{i=1}^{n} E[e_i e_i^{'}] = \sum_{i=1}^{n} \Sigma = n\Sigma. \tag{A.3}$$

Our expected value will $-n\Sigma$.

The variance of V can be calculated as in Theorem 11.29 [92, p. 489], which is reproduced below with slightly modified notation. This proof uses vectorized form, where the $vec()$ operator converts a matrix into a vector by stacking the columns, and $vec(\boldsymbol{ab'}) = \boldsymbol{b} \otimes \boldsymbol{a}$ [92, Theorem 8.9, p. 489]. Let e_i denote the columns of ε and μ_i the columns of constant matrix U_0,

$$\begin{aligned} Var\{vec(V)\} &= Var\left\{vec\left(\sum_{i=1}^{n}(e_i+\mu_i)(e_i+\mu_i)^{'}\right)\right\} \\ &= Var\left\{\sum_{i=1}^{n} vec\left((e_i+\mu_i)(e_i+\mu_i)^{'}\right)\right\} \\ &= \sum_{i=1}^{n} Var\left\{(e_i+\mu_i)\otimes(e_i+\mu_i)\right\} \end{aligned}$$

The inside of the summation and variance evaluates to

$$\begin{aligned} (e_i+\mu_i)\otimes(e_i+\mu_i) &= e_i\otimes e_i + e_i\otimes\mu_i + \mu_i\otimes e_i + \mu_i\otimes\mu_i \\ &= e_i\otimes e_i + (I_{mm}+K_{mm})(e_i\otimes\mu_i) + \mu_i\otimes\mu_i \\ &= e_i\otimes e_i + 2N_m(I_{mm}\otimes\mu_i)e_i + \mu_i\otimes\mu_i \end{aligned}$$

Where the matrices I_{mm} denote the Identity matrix of size $m \times m$, K_{mm} denotes the commutation matrix with property $K \cdot vec(A) = vec(A^{'})$, and $N_m = \frac{1}{2}(I_{mm} + K_{mm})$ [92,

p. 380]. The variance of this becomes

$$\begin{aligned} Var\left\{(e_i+\mu_i)\otimes(e_i+\mu_i)\right\} &= var(e_i\otimes e_i) + var\left\{2N_m(I_m\otimes\mu_i)x_i\right\} \\ &= 2N_m(\Sigma\otimes\Sigma) + 4N_m(I_m\otimes +u_i)\Sigma(I_m\otimes\mu_i^{'})N_m \\ &= 2N_m(\Sigma\otimes\Sigma) + 4N_m(\Sigma\otimes\mu_i\mu_i^{'})N_m \\ &= 2N_m(\Sigma\otimes\Sigma + \Sigma\otimes\mu_i\mu_i^{'} + \mu_i\mu_i^{'}\otimes\Sigma) \end{aligned}$$

Substituting this back into the original summation, the variance of the vectorized form of V becomes

$$Var\{vec(V)\} = 2N_m\{n(\Sigma\otimes\Sigma) + \Sigma\otimes U_0^{'}U_0 + U_0^{'}U_0\otimes\Sigma\} \tag{A.4}$$

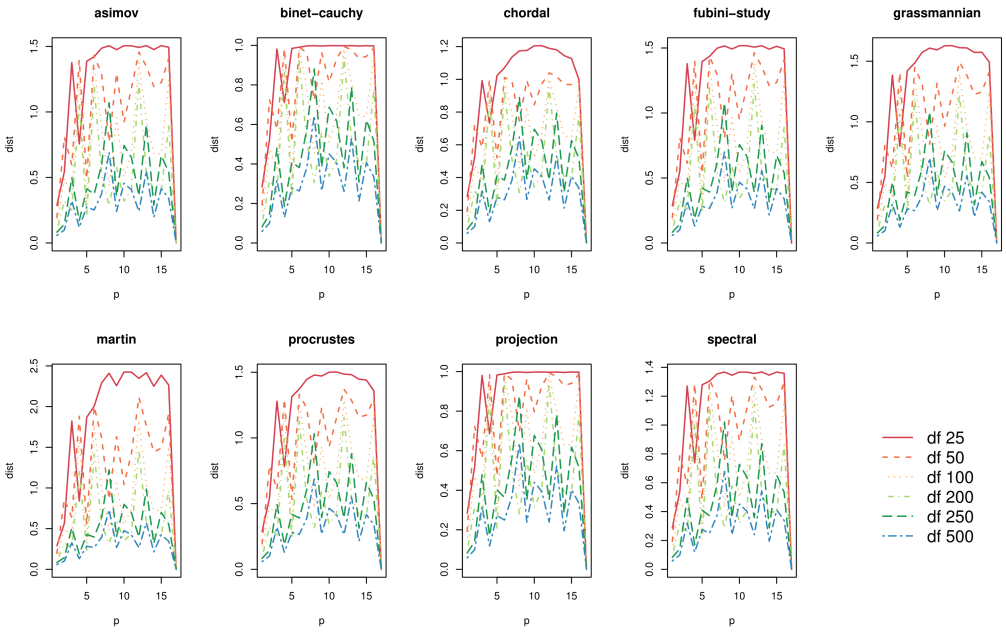
Using Theorem 8.9, $vec(\boldsymbol{ab'}) = \boldsymbol{b}\otimes\boldsymbol{a}$ [92, p. 489], the matrix form is provided below.

$$\begin{aligned} &2N_m\{n(\Sigma\otimes\Sigma) + \Sigma\otimes U_0^{'}U_0 + U_0^{'}U_0\otimes\Sigma\} \\ &\quad= 2N_m\{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} \\ &\quad= (I_{mm}+K_{mm})\{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} \\ &\quad= \{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} + K_{mm}\{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} \\ &\quad= \{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} + \{n(\Sigma\Sigma^{'})^{'} + (U_0^{'}U_0\Sigma)^{'} + (\Sigma U_0^{'}U_0)^{'}\} \\ &\quad= \{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} + \{n\Sigma\Sigma^{'} + \Sigma^{'}U_0^{'}U_0 + U_0^{'}U_0\Sigma^{'}\} \\ &\quad= 2\{n\Sigma\Sigma^{'} + U_0^{'}U_0\Sigma^{'} + \Sigma U_0^{'}U_0\} \end{aligned}$$

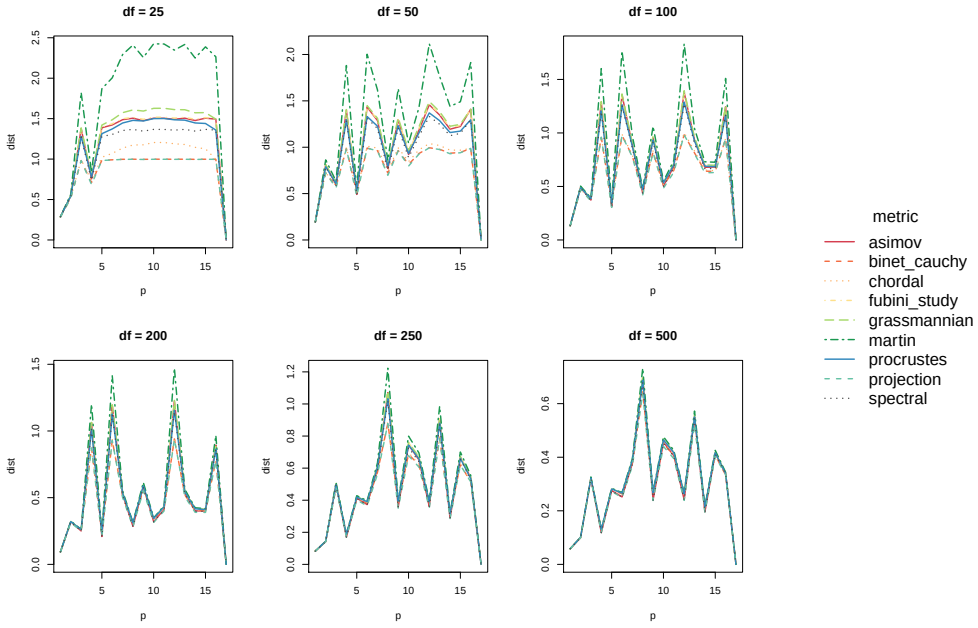
The last inequality holds because covariance matrix Σ is symmetric.

A.2 Additional Simulation Figures

Figure 4.16 which showed subspace metrics for Data Set 1. The following correspond Data Sets 2 and 3.

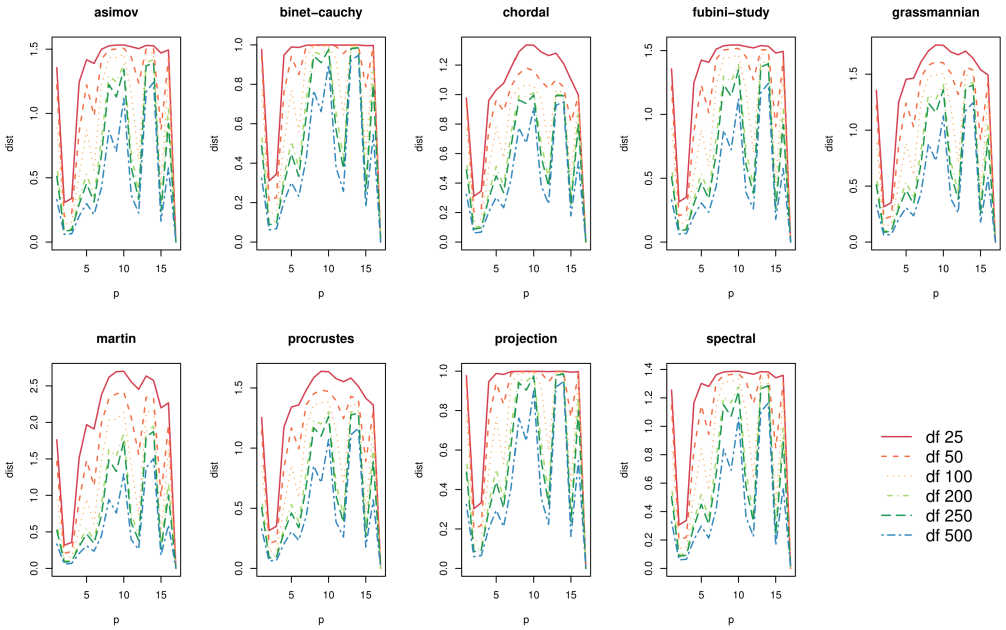


(a) Subspace distance metrics for Data Set 2. The metrics all show an oscillating pattern as p increases, though it is not as clear when the degrees of freedom is close to p.

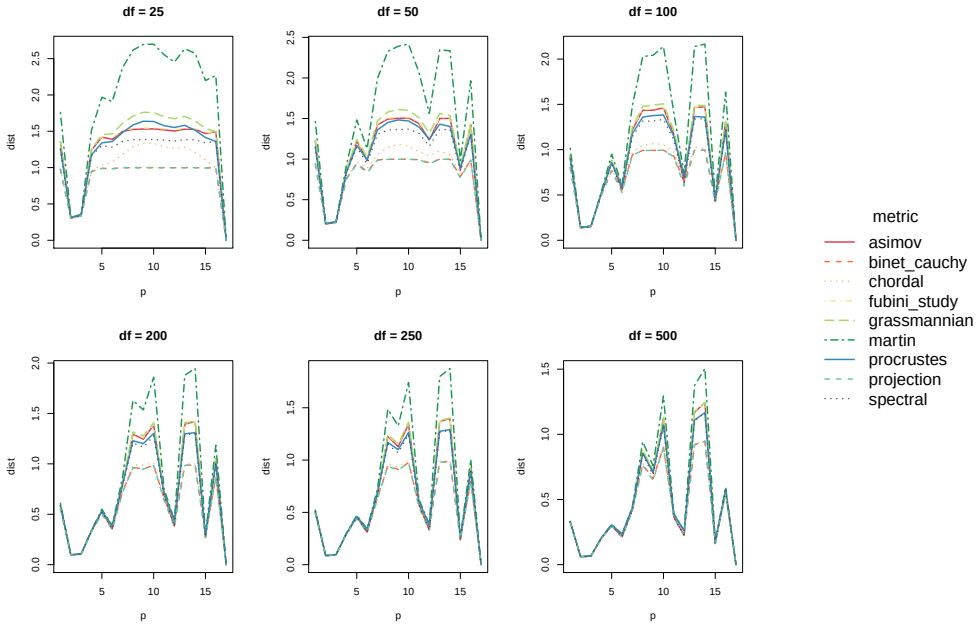


(b) Subspace distance metrics by degrees of freedom. Variability in the metrics decreases as the degrees of freedom increases. An oscillating pattern can still be seen.

Figure A.1: Nine metrics from Table 2.1 for Data Set 2. An oscillating pattern can be seen for each of the metrics over varying dimensions p.

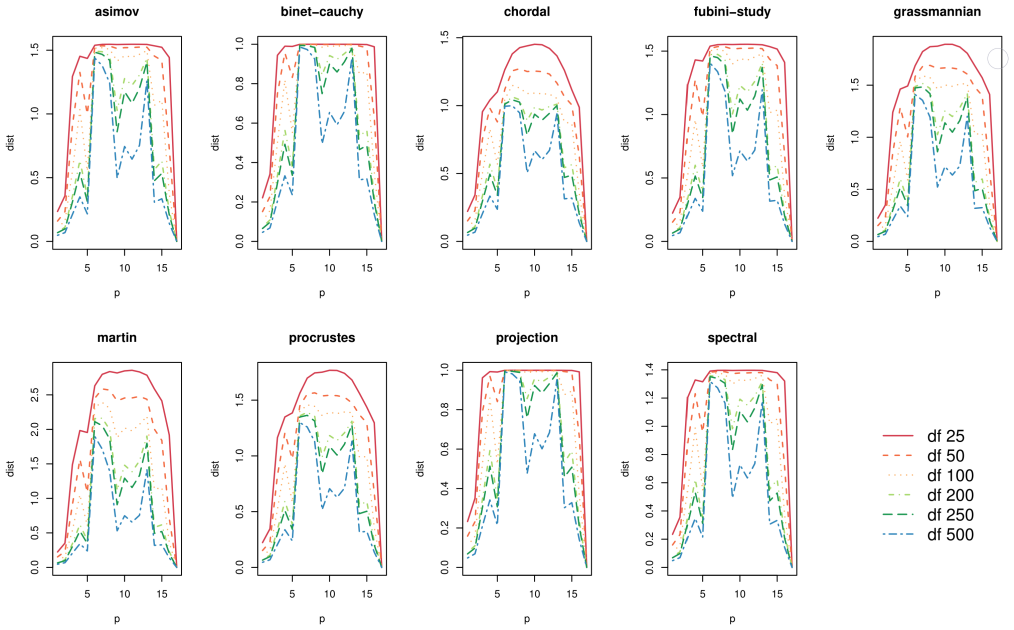


(a) Subspace distance metrics for Data Set 3a. The metrics all show an oscillating pattern as p increases, though it is not as clear when the degrees of freedom is close to p.

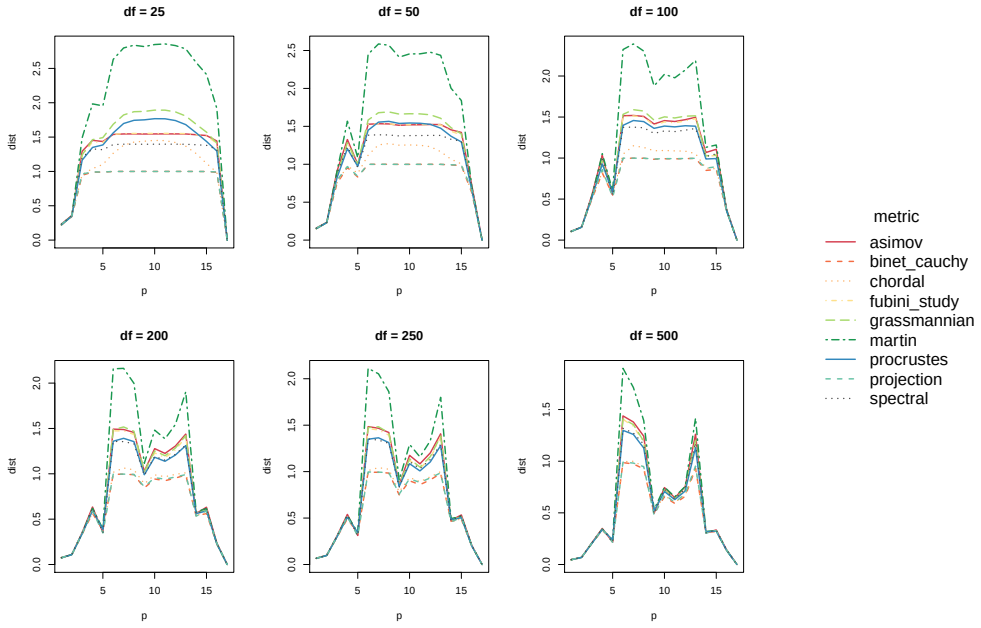


(b) Subspace distance metrics by degrees of freedom. Variability in the metrics decreases as the degrees of freedom increases. An oscillating pattern can still be seen.

Figure A.2: Nine metrics from Table 2.1 for Data Set 3a. An oscillating pattern can be seen for each of the metrics over varying dimensions p.



(a) Subspace distance metrics for Data Set 3b. The metrics all show an oscillating pattern as p increases, though it is not as clear when the degrees of freedom is close to p.



(b) Subspace distance metrics by degrees of freedom. Variability in the metrics decreases as the degrees of freedom increases. An oscillating pattern can still be seen.

Figure A.3: Nine metrics from Table 2.1 for Data Set 3b. An oscillating pattern can be seen for each of the metrics over varying dimensions p.

Bibliography

[1] Grassmann manifold. https://mathworld.wolfram.com/GrassmannManifold.html. Accessed: 2020-03-10.

[2] Grassmannian. https://mathworld.wolfram.com/Grassmannian.html. Accessed: 2020-03-10.

[3] Subspace method. https://www.sciencedirect.com/topics/engineering/subspace-method. Accessed: 2020-03-10.

[4] P. A. Absil, A. Edelman, and P. Koev. On the largest principal angle between random subspaces. *Linear Algebra and its applications*, 414(1):288–294, 2006.

[5] J. Almirall, D. C. Duckworth, C. K. Bayne, and K. Furton. Discrimination of forensic glasses via trace element analysis by inductively coupled plasma mass spectrometry and statistical treatment. Technical report, International Forensic Research Institute, Florida State University, Department of Chemistry, University Park, Miami, FL 33199. Oak Ridge National Laboratory, Oak Ridge, TN 37831-6375., 2016. The Office of Special Technology (OST) Broad Agency Announcement DAAD05-99-T-0734 Technical Support Working Group (TSWG). Mission Area: (R-555) Improved Forensic Glass Analysis and Database Development.

[6] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis (3rd ed.)*. John Wiley & Sons, Inc., 2003.

[7] S. J. Bajic, D. B. Aeschliman, N. J. Saetveit, D. P. Baldwin, and R. S. Houk. Analysis of glass fragments by laser ablation-inductively coupled plasma-mass spectrometry and principal component analysis. *Journal of Forensic Science*, 50(5):JFS2005088–5, 2005.

[8] M. S. Bartlett. Properties of sufficiency and statistical tests. In *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, volume 160 of *901*, pages 268–282, 1937.

[9] S. Berends-Montero, W. Wiarda, P. de Joode, and G. van der Peijl. Forensic analysis of float glass using laser ablation inductively coupled plasma mass spectrometry (la-icp-ms): validation of a method. *Journal of Analytical Atomic Spectrometry*, 21(11):1185–1193, 2006.

[10] A. E. Bilgrau, R. F. Brøndum, P. S. Eriksen, K. Dybkær, and M. Bøgsted. Estimating a common covariance matrix for network meta-analysis of gene expression datasets in diffuse large b-cell lymphoma. *The Annals of Applied Statistics*, 12(3):1894–1913, 2018.

[11] A. Bjorck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[12] E. C. Blacklock, A. Rogers, C. Wall, and B. B. Wheals. The quantitative analysis of glass by emission spectrography: a six element survey. *Forensic science*, 7(2):121–130, 1976.

[13] M. C. Bottrell. Forensic glass comparison: background information used in data interpretation. *Forensic Science Communications [Online]*, 2009.

[14] D. R. Brillinger and J. W. Tukey. *The Collected Works of JW Tukey, 2*, chapter Spectrum analysis in the presence of noise: some issues and examples, pages 1001–1141. 1985.

[15] G. Brys, M. Hubert, and P. J. Rousseeuw. A robustification of independent component analysis. *Journal of Chemometrics*, 19:364–375, 2005.

[16] J. A. Buscaglia. Elemental analysis of small glass fragments in forensic science. *Analytica chimica acta*, 288(1-2):17–24, 1994.

[17] T. Cai, W. Liu, and Y. Xia. On the power of the l_1 test for equality of several variances. *The Annals of Mathematical Statistics*, 10(2):119–128, 1939.

[18] T. Cai, W. Liu, and Y. Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277, 2013.

[19] A. R. Calloway and P. F. Jones. Enhanced discrimination of glass samples by phosphorescence analysis. *Journal of Forensic Science*, 23(2):263–273, 1978.

[20] T. Catterick and D. A. Hickman. The quantitative analysis of glass by inductively coupled plasma-atomic-emission spectrometry: a five-element survey. *Forensic Science International*, 17(3):253–263, 1981.

[21] National Research Council. *Forensic analysis: Weighing bullet lead evidence*. National Academies Press, 2004.

[22] National Research Council. *Strengthening forensic science in the United States: a path forward*. National Academies Press, 2009.

[23] P. L. Davies. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292, 1987.

[24] M. Debruyne and M. Hubert. The influence function of the stahel-donoho covariance estimator of smallest outlyingness. *Statistics & probability letters*, 79(3):275–282, 2009.

[25] D. L. Donoho. Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston, 1982.

[26] H. Dorn, D. E. Ruddell, A. Heydon, and B. D. Burton. Discrimination of float glass by la-icp-ms: assessment of exclusion criteria using casework samples. *Canadian Society of Forensic Science Journal*, 48(2):85–96, 2015.

[27] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, 1966.

[28] D. C. Duckworth, C. K. Bayne, S. J. Morton, and J. Almirall. Analysis of variance in forensic glass analysis by icp-ms: variance within the method. *Journal of Analytical Atomic Spectrometry*, 15(7):821–828, 2000.

[29] D. C. Duckworth, S. J. Morton, C. K. Bayne, R. D. Koons, S. Montero, and J. R. Almirall. Forensic glass analysis by icp-ms: a multi-element assessment of discriminating power via analysis of variance and pairwise comparisons. *Journal of Analytical Atomic Spectrometry*, 17(7):662–668, 2002.

[30] M. Eaton and D. Tyler. The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *Journal of Multivariate Analysis*, 50(2):238–264, 1994.

[31] Noureddine El Karoui. Tracy-widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, 2007.

[32] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(3):2757–2790, 2008.

[33] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Proceedings of the 11th International Symposium of Robotics Research*, pages 192–201, 2003.

[34] J. Gabel-Cino. Expert witnesses and lawyers: Can we all get along? presentation to the second annual conference of the national center for forensic science, oct 2017. Orlando, Florida.

[35] A. Galantai and Cs. J. Hegedus. Jordan's principal angles in complex vector spaces. *Numerical Linear Algebra with Applications*, 13:589–598, 2006.

[36] L. Gamble, D. Q. Burd, and P. L. Kirk. Glass fragments as evidence. a comparative study of physical properties. *Journal of Criminal Law and Criminology (1931-1951)*, 33(5):416–421, 1943.

[37] P. C. Giannelli. Comparative bullet lead analysis: A retrospective. *Criminal Law Bulletin*, 47(306), 2010.

[38] R. Gnanadesikan and J. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124, 1972.

[39] J. Hamm and D. D¿ Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on machine learning*, pages 376–383, 2008.

[40] F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

[41] F. R. Hampel. Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 27(2):87–104, 1973.

[42] M. A. Haney. Comparison of window glasses by isotope dilution spark source mass spectrometry. *Journal of Forensic Science*, 22(3):534–544, 1977.

[43] D. A. Hickman, G. Harbottle, and E. V. Sayre. The selection of the best elemental variables for the classification of glass samples. *Forensic Science International*, 22(2-3):189–212, 1983.

[44] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[45] M. Hubert, M. Debruyne, and P. Rousseeuw. Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3):1–11, 2017.

[46] M. Hubert, P. Rousseeuw, and Vanden Branden K. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[47] M. Hubert, P. Rousseeuw, and K. Vakli. Shape bias of robust covariance estimators: An empirical study. *Statistical Papers*, 55(1):15–28, 2014.

[48] M. Hubert, P. Rousseeuw, and T. Verdonck. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3 (2012)):618–637, 2012.

[49] J. C. Hughes, T. Catterick, and G. Southeard. The quantitative analysis of glass by atomic absorption spectroscopy. *Forensic science*, 8:217–227, 1976.

[50] ASTM International. *ASTM C162-05 Standard Terminology of Glass and Glass Products*. https://doi.org/10.1520/C0162-05, 2005.

[51] ASTM International. *ASTM E2330-12 Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*. https://doi.org/10.1520/E2330-12, 2012.

[52] ASTM International. *ASTM E2926-13 Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*. https://doi.org/10.1520/E2926, 2013.

[53] ASTM International. *ASTM E2927-16e1 Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Abla-*

tion Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons. https://doi.org/10.1520/E2927-16E01, 2016.

[54] W. E. Johnson and C. Li. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[55] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[56] Iain M. Johnstone, Zongming Ma, Patrick O. Perry, and Morteza Shahram. *RMTstat: Distributions, Statistics and Tests derived from Random Matrix Theory*, 2014. R package version 0.3.

[57] C. Jordan. Essai sur la geometrie a n dimensions. *Buletin de la Societe Mathematique de France*, 3:103–174, 1875.

[58] P. L. Kirk. Crime investigation; physical evidence and the police laboratory., 1953.

[59] R. D. Koons. Personal communication to k. kafadar.

[60] R. D. Koons and J. Buscaglia. The forensic significance of glass composition and refractive index measurements. *Journal of Forensic Science*, 44(3):496–503, 1999.

[61] R. D. Koons and J. A. Buscaglia. Interpretation of glass composition measurements: the effects of match criteria on discrimination capability. *Journal of Forensic Science*, 47(3):505–512, 2001.

[62] R. D. Koons and J. A. Buscaglia. Forensic significance of bullet lead compositions. *Journal of Forensic Science*, 50(2):341–351, 2005.

[63] R. D. Koons, C. Fiedler, and R. C. Rawalt. Classification and discrimination of sheet and container glasses by inductively coupled plasma-atomic emission spectrometry and pattern recognition. *Journal of Forensic Science*, 33(1):49–67, 1988.

[64] R. D. Koons, C. A. Peters, and P. S. Rebbert. Comparison of refractive index, energy dispersive x-ray fluorescence and inductively coupled plasma atomic emission spectrometry for forensic characterization of sheet glass fragments. *Journal of Analytical Atomic Spectrometry*, 6(6):451–456, 1991.

[65] C. Latkoczy, S. Becker, M. Dücking, D. Günther, J. A. Hoogewerff, J. R. Almirall, J. Buscaglia, A. Dobney, R. D. Koons, S. Montero, and G.J. van der Peijl. Development and evaluation of a standard method for the quantitative determination of elements in float glass samples by la-icp-ms. *Journal of Forensic Science*, 50(6):JFS2005091–15, 2005.

[66] J. A. Lee, K. K. Dobbin, and J. Ahn. Covariance adjustment for batch effect in gene expression data. *Statistics in Medicine*, 33(15):2681–2695, 2014.

[67] J. Li and S. X. Chen. Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940, 2012.

[68] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):3160327, 1991.

[69] R. Maronna and V. Yohai. The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995.

[70] R. Maronna and R. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.

[71] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[72] R. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., 1982.

[73] D. F. Nelson. Illustrating the fit of glass fragments. *The Journal of Criminal Law, Criminology, and Police Science*, 50(3):312–314, 1959.

[74] D. F. Nelson and B. C. Revell. Backward fragmentation from breaking glass. *Journal of the Forensic Science Society*, 7(2):58–61, 1967.

[75] J. Neyman and E. S. Pearson. On the problem of k samples. *Bull. int. Acad. Cracovie*, A:460–481, 1931.

[76] P. O'Brien. Robust procedures for testing equality of covariance matrices. *Biometrics*, 48(3):819–827, 1992.

[77] S. Park and A. Carriquiry. Glass data description. 2018. Retrieved from https://github.com/CSAFE-ISU/AOAS-2018-glass-manuscript.

[78] S. Park and A. Carriquiry. Learning algorithms to evaluate forensic glass evidence. *The Annals of Applied Statistics*, 2019. In press.

[79] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

[80] M. Perlman. Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *The Annals of Statistics*, 8(2):247–263, 1980.

[81] E. Pitman. Tests of hypotheses concerning location and scale parameters. *Biometrika*, 31(1/2):200–215, 1939.

[82] P. Rousseeuw and M. Hubert. *Robust and Complex Data Structures*, chapter High-breakdown estimators of multivariate location and scatter, pages 49–66. Springer Berlin Heidelberg, 2013.

[83] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[84] P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

[85] P. J. Rousseeuw. Discussion on "breakdown and groups". *The Annals of Statistics*, 33(3):1004–1009, 2005.

[86] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[87] P. J. Rousseeuw and M. Hubert. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):1–14, 2018.

[88] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[89] P. J. Rousseeuw, J. Raymaekers, and M. Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359, 2018.

[90] J. R. Schott. Some tests for the quality of covariance matrices. *Journal of statistical planning and inference*, 94(1):25–36, 2001.

[91] J. R. Schott. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis*, 51(12):6535–6542, 2007.

[92] J. R. Schott. *Matrix Analysis for statistics (3rd ed.)*. John Wiley & Sons, Inc., 2016.

[93] D. Smale. he examination of paint flakes, glass and soils for forensic purposes, with special reference to electron probe microanalysis. *Journal of the Forensic Science Society*, 13(1):5–15, 1973.

[94] C. H. Spiegelman and K. Kafadar. Data integrity and the scientific method: The case of bullet lead data as forensic evidence. *Change*, 19(2):17–25, 2006.

[95] M. S. Srivastava. Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, 35(2):251–272, 2005.

[96] M. S. Srivastava. Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society*, 37(1):53–86, 2007.

[97] M. S. Srivastava and Hirokazu Y. Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate ANalysis*, 101(6):1319–1329, 2010.

[98] W. A. Stahel. Breakdown of covariance estimators. Fachgruppe für Statistik, ETH, Zurich, 1981. Research Report 31.

[99] N. Sugiura and H. Nagao. Unbiasedness of some test criteria for the equality of one or two covariance matrices. *The Annals of Mathematical Statistics*, 39(5):1686–1692, 1968.

[100] M. L. Tiku and N. Balakrishnan. Testing the equality of variance-covariance matrices the robust way. *Communications in Statistics - Theory and Methods*, 14(12):3033–3051, 1985.

[101] T. Trejos and J. R. Almirall. Sampling strategies for the analysis of glass fragments by la-icp-ms: Part i. micro-homogeneity study of glass and its application to the interpretation of forensic evidence. *Talanta*, 67(2):388–395, 2005.

[102] T. Trejos, R. Koons, P. Weis, S. Becker, T. Berman, C. Dalpe, M. Duecking, J. Buscaglia, T. Eckert-Lumsdon, T. Ernst, and C. Hanlon. Forensic analysis of glass by μ-xrf, sn-icp-ms, la-icp-ms and la-icp-oes: evaluation of the performance of different criteria for comparing elemental composition. *Journal of Analytical Atomic Spectrometry*, 28(8):1270–1282, 2013.

[103] F. G. Tryhorn. The examination of glass. *The Police Journal*, 12(3):301–318, 1939.

[104] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011.

[105] S. Van Aelst, E. Vandervieren, and G. Willems. A stahel-donoho estimator based on huberized outlyingness. *Computational Statistics & Data Analysis*, 56(3):531–542, 2012.

[106] A. K. Varshneya and M. Tomozawa. Fundamentals of inorganic glasses. *Journal of Non Crystalline Solids*, 170(1):112, 1994.

[107] T. Wang and P. Shi. *Pattern Recognition Letters*, 30(13):1161–1165, 2009.

[108] P. Weis, M. Dücking, P. Watzke, S. Menges, and S. Becker. Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry*, 26(6):1273–1284, 2011.

[109] K. L. Wolnik, C. M. Gaston, and F. L. Fricke. Analysis of glass in product tampering investigations by inductively coupled plasma atomic emission spectrometry with a hydrofluoric acid resistant torch. *Journal of Analytical Atomic Spectrometry*, 4(1):27–31, 1989.

[110] Y. C. Wong. *Differential geometry of Grassmann manifolds*, 57(3):589–194, 1967.

[111] Y. Xia, D. Zhang, and J. Xu. Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*, 105(489):278–290, 2010.

[112] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image

sequence. In *Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

[113] K. Ye and L. Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applicaions*, 37(3):1176–1197, 2016.

[114] J. Zhang and D. Boos. Bootstrap critical values for testing homogeneity of covariance matrices. *Journal of the American Statistical Association*, 87(418):425–429, 1992.

[115] J. Zhang, G. Zhu, R. W. Heath Jr., and K. Huang. Grassmannian learning: embedding geometry awareness in shallow and deep learning. *arXiv preprint arXiv:1808.02229*.

[116] A. Zurhaar and L. Mullings. Characterisation of forensic glass samples using inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry*, 5(7):611–617, 1990.