STATISTICAL INTEGRATION OF INFORMATION

Qing Feng

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2016

Approved by:

Shankar Bhamidi

Jan Hannig

Katherine Hoadley

Yufeng Liu

J. S. Marron

Andrew Nobel

ABSTRACT

QING FENG: Statistical Integration of Information
(Under the direction of J. S. Marron and Jan Hannig)

Modern data analysis frequently involves multiple large and diverse data sets generated from current high-throughput technologies. An integrative analysis of these sources of information is very promising for improving knowledge discovery in various fields. This dissertation focuses on three distinct challenges in the integration of information.

The variables obtained from diverse and novel platforms often have highly non-Gaussian marginal distributions and therefore are challenging to analyze by commonly used methods. The first part introduces an automatic transformation for improving data quality before integrating multiple data sources. For each variable, a new family of parametrizations of the shifted logarithm transformation is proposed, which allows transformation for both left and right skewness within the single family and an automatic selection of the parameter value.

The second part discusses an integrative analysis of disparate data blocks measured on a common set of experimental subjects. This data integration naturally motivates the simultaneous exploration of the joint and individual variation within each data block resulting in new insights. We introduce Non-iterative Joint and Individual Variation Explained (Non-iterative JIVE), capturing both joint and individual variation within each data block. This is a major improvement over earlier approaches to this challenge in terms of both a new conceptual understanding and a fast linear algebra computation. An important mathematical contribution is the use of score subspaces as the principal descriptors of variation structure and the use of perturbation theory as the guide for variation segmentation. Furthermore, this makes our method robust against the heterogeneity among data blocks, without a need for normalization.

The last part proposes a Generalized Fiducial Inference inspired method for finding a robust consensus among several independently derived confidence distributions (CDs) for a quantity of interest. The resulting fused CD is robust to the existence of potentially

discrepant CDs in the collection. The method uses computationally efficient fiducial model averaging to obtain a robust consensus distribution without the need to eliminate discrepant CDs from the analysis.

To my parents

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisors Professor J. S. Marron and Professor Jan Hannig. Without their guidance and persistent help, this dissertation would not have been possible.

Dr. Marron, thank you for all the guidance, patience and help! I had a tremendous fun to work with you in the past five years. Your great insights in handling various data challenges help me to develop many different perspectives for data analysis and also to improve my ability as a statistician. Besides, I have learned many good habits from you such as in writing, organization, communication and attention to details.

Dr. Hannig, thank you very much for your generosity with your time and knowledge! I cannot express how much I enjoy doing research with you. I am always amazed by your novel ideas and your ways to tackle problems. Your straightforward explanations and interpretations make the abstract theorems even approachable. I will continue carrying the enthusiasm in research and seek all possibilities to apply them in my future job!

In addition, I would like to thank Dr. Hari K. Iyer for the collaboration in the fiducial projects! I appreciate all your time and effort! Besides, I want to thank Professor Shankar Bhamidi, Professor Katherine A. Hoadley, Professor Yufeng Liu and Professor Andrew Nobel, for participation on my dissertation committee. Thank you all for providing the insightful comments and suggestions. I would also thank all other faculties and staff members in our department. Thank you for making my Ph.D. study so wonderful!

Moreover, I would like to deliver my thanks to my friends who make my Ph.D. life so colorful. Siliang, Siying, Minghui, Dongqing, Jie, Yu, Dong and Dan, thank you for all kinds of support! I would say having lunch with you guys are the happiest time of each day. I will remember all the wonderful, fun time we have spent together. No matter where we will be, we are always friends! Guan and Eunjee, thank you for the friendship and especially all the help during the graduate study! I cannot express how much I like

to discuss various questions with you. Yanni, Lijiang, Xingyi and Ning, thank you all for the support and encouragement even when we are apart from each other!

Finally, I want to thank my dear parents. Thank you for unconditional support and encouragement to every decision I've made! I appreciate all the sacrifices you have made for me. I love you more than I could express!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Current high-throughput technologies are powering the generation of large and diverse data sets. A collection of different types of data can be obtained from multiple platforms and a major challenge to integrate them for meaningful analysis. Such integration of multiple sources of information is very promising for improving knowledge discovery in various fields. For example, data integration methodologies become more and more important in the life sciences research. One well-known data intense biological context is The Cancer Genome Atlas Project (TCGA). It aims at generating insights into the heterogeneity of different cancer subtypes by analyzing various data types from high-throughput technologies. For instance, Network et al. (2012) characterized the breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, microRNA sequencing and reverse phase protein arrays. Through integrating information across platforms, Network et al. (2012) provided key insights into previously-defined gene expression subtypes and demonstrated the existence of four main breast cancer classes when combining data from five platforms, each of which shows significant molecular heterogeneity.

Another popular application of data integration analysis is *precision medicine* (PM). This is a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products being tailored to the individual patient. Instead of classifying individuals into a particular disease, PM targets at classifying individuals into subpopulations based on integrating personal multi-OMIC data (Binder et al., 2014), imaging data and clinical data. These methods enable characterization of the genotype and/or molecular phenotype on a personalized basis with the aim of increasing our understanding of disease genesis and progression and, in final consequence, improvement of diagnosis and treatment options (Binder et al., 2014).

Data integration analysis presents many challenges to traditional analytical tools considering the large volume and complexity of the collections of data sets. This dis-

sertation addresses some *statistical* challenges raised by integration of information, in particular three aspects described in the following sections.

1.1 Automatic Data Transformation

The multiple data sets generated from different and novel platforms frequently have highly non-Gaussian marginal distributions. However commonly used analysis methods are most effective with roughly Gaussian data. Therefore, an appropriate data transformation can be very useful for improving the data quality before any further analysis or development of new methodologies. One great challenge of transforming an integrated data set comes from the massive amount of variables and their heterogeneity. Considering the high dimensionality and the big difference in magnitudes, it is crucial to automate the transformation and make it robust for each individual variable.

This dissertation introduces an automatic data transformation technique in Chapter 2. This method proposes a new family of parametrizations of the shifted logarithm transformation in which the parameter selection is invariant to the magnitudes of variables. This new family thus allows an automatic selection of parameters by minimizing the Anderson–Darling test statistic of the transformed data.

1.2 Non-iterative Joint and Individual Variation Explained

One common and important data integration task is the combination of diverse information from disparate data sets measured on a common set of experimental subjects. This type of integrated data set is also known as *multi-block* data. Each data block from distinct platforms provides important information of commonly measured subjects, such as the TCGA example mentioned above. A unified and insightful understanding of the set of data blocks is expected from simultaneously exploring the joint variation representing the inter-block associations and the individual variation specific to each block.

The first challenge is interpretability. More insightful data analysis comes from careful understanding of the common pattern across the blocks and the unique pattern of each individual data set. To achieve this, it requires a model framework having

meaningful and rigorous definitions of each type of variation together with constraints to obtain identifiability.

The second challenge is in heterogeneity. The block means could be quite different. This can be addressed by subtracting each block mean. After mean normalization, scaling should be considered. This needs more careful consideration because the number of features can be vastly different. Taking the breast cancer study in the TCGA project as an example, while one block contains about 100 features extracted from reverse phase protein arrays, the other has approximately 14,000 gene expression features. Thus, any important information from the reverse phase protein arrays may be swamped out by the large amount of gene expression information.

This dissertation proposes a computationally efficient method, *Non-iterative Joint and Individual Variation Explained* (Non-iterative JIVE), to obtain an identifiable and insightful decomposition to a set of heterogeneous data blocks. Non-iterative JIVE is based on the model framework developed by Lock et al. (2013) and improved the interpretation of variation components and identifiability using the row space (assuming columns are the data objects) of the data matrices. The new concept of *latent score vector* provides many new insights. Furthermore, this method applies a new *principal angle analysis* in the row space to resolve the heterogeneous challenge when segmenting the variation into joint and individual components. Chapter 3 provides the details of the population model, estimation method and Chapter 4 shows the applications to a mortality data set and a TCGA breast cancer data set.

1.3 Fusion Learning for Interlaboratory Comparison

The other data integration challenge comes from *interlaboratory trials* which are often conducted by leading metrology laboratories in the world to compare measurements of various fundamental properties of substances. Such a trial typically involves two or more participants each of whom measures the (nominally) same unknown value (called measurand) and provides the result along with an assessment of the uncertainty in the result. When facing discrepant measurements, statistical modeling and analysis is needed for determining the consensus (reference) value and its associated uncertainty.

One major difficulty in integration stems from the sometimes big discrepancy of the interlaboratory trials being combined. It is generally the case that the results from

one or a few laboratories differ noticeably from the rest in that there is no overlap among the derived confidence intervals. Simply eliminating them from the analysis is often not an acceptable approach, particularly so in view of the fact that the true value being measured is not known and a discrepant result from a lab may be closer to the *true value* than the rest of the results. Additionally, eliminating one or more labs from the analysis can lead to political complications since all labs are regarded as equally competent. These considerations make the proposed method well suited for the task since no laboratory is explicitly eliminated from consideration.

To appropriately incorporate these considerations, a *Generalized Fiducial Inference* inspired method is proposed in this dissertation to derive a robust *consensus value* from a collection of interlaboratory trials. This method does not eliminate any discrepant laboratory measurements and uses a fiducial model based weight to achieve a robust average. More backgrounds and details are discussed in Chapter 5.

CHAPTER 2: AUTOMATIC DATA TRANSFORMATION

2.1 Introduction

Technological developments have led to methods for generating complex data objects such as DNA chip data and digital images of tumors. These new types of data objects frequently strongly violate the approximate normality assumption which is commonly made in statistical techniques. Therefore, an appropriate data transformation can be very useful for improving the closeness of the data distribution to normality.

Many transformation techniques have been proposed. Sakia (1992) provided a comprehensive review of the Box-Cox (Box and Cox, 1964) and related transformations. Various methods have been developed for selecting the transformation parameters, including the maximum likelihood method (Box and Cox, 1964), robust adaptive method (Carroll, 1980), Kullback-Leibler information based method (Hernandez and Johnson, 1980), and Kendall's rank correlation-based method (Han, 1987).

A commonly used member of the Box-Cox family is the logarithm transformation, which is useful for tackling data sets generated by a multiplicative process. Furthermore, the logarithm transformation can stabilize the asymptotic variance of data. One important application is to transform some types of microarray data. A shift parameter was further introduced to make the logarithm transformations more flexible and useful. See Section 3 of Yang (1995) for a good overview of the shifted logarithm transformation. The parameterizations of the shift parameter strongly depend on knowledge of the data e.g. data range, data distribution, so user intervention is usually required. However, modern high-output data sets usually have a very large number of variables, i.e. features, so there is a strong need to automate the selection of shift parameter, which is an important contribution of this paper.

We propose a new automatic data transformation scheme for making various types of marginal distributions close to being normally distributed. In particular, we aim at addressing certain types of departures from normality, for example, strong skewness.

Our proposed method focuses on the family of shifted logarithm transformations and introduces a new parametrization which treats the data as lying on the entire real line. Besides, our parametrization makes the selection of shift tuning parameter independent of data magnitude which is an advantage for automation. This algorithm is designed to automatically select a parameter value such that the transformed data has the smallest *Anderson–Darling test statistic*. Furthermore, this transformation scheme includes a winsorization of influential observations based on the extreme value theorem. The transformation is univariate in nature and thus cannot guarantee multivariate normality. However, we have seen many real data sets where bivariate normality is a clear consequence.

2.1.1 Data Example

A motivating data example is digital image analysis in a study of mutant types of melanocytic lesions (Miedema et al., 2012). Image features are constructed as mathematical descriptions of cell and nuclei shape, color and relationship, capturing image aspects such as the regularity of nuclear shape, nuclear area and stain intensity. A set of 33 features are extracted for each cell (approximately 1,425,000), describing both nuclei and surrounding cytoplasm. A table of a summarized description of these features can be found in Miedema et al. (2012).

Many of the raw features extracted from digital images contain excessive skewness. For example, the marginal distributions of two of the image features, Hu4 and Eccentricity, are visualized by the *kernel density estimated plots* (KDE plots) in the top row of Figure 2.1. The blue curves are the Gaussian kernel density estimate i.e. smoothed histograms, using Sheather-Jones plug-in bandwidths (See Chapter 3 of Wand and Jones (1994) for the comparison of bandwidth selection methods). The green dots are jitter plots of the data. Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is based on data ordering for visual separation. As can be seen, these distributions are highly skewed. For such data sets with substantial skewness, an analysis based on a Gaussian assumption would tend to generate poor results. The bottom plots of Figure 2.1 display the KDE plots of each feature vector after our automatic transformation. The kernel density estimates (blue curves) are approximately symmetric.

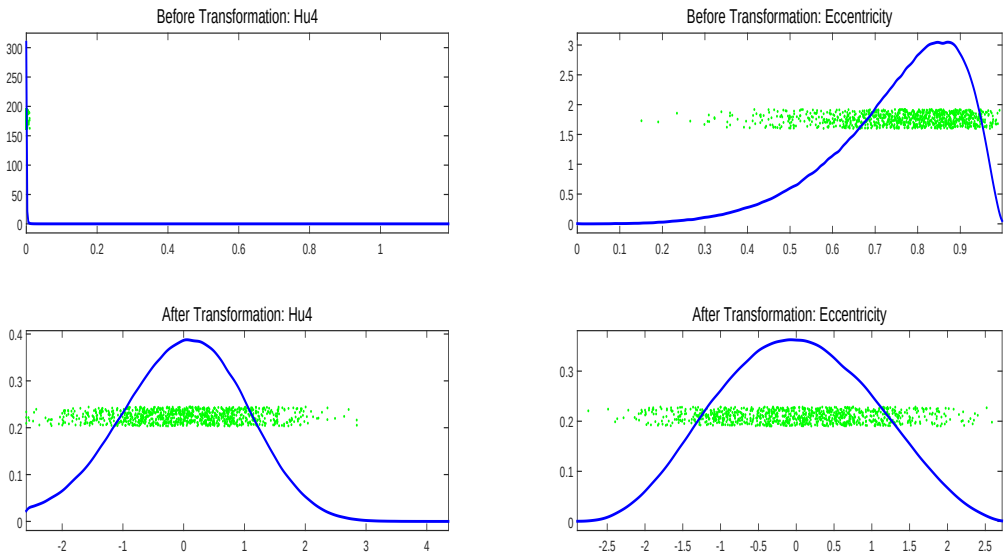


Figure 2.1: Comparison of the KDE-plots of two image feature vectors before (top row) and after (bottom row) transformation. This shows that the transformed distributions are much closer to Gaussian for data with both positive (Hu4, left column) and negative (Eccentricity, right column) skewness.

The Quantile–Quantile (Q–Q) plots in Figure 2.2 give a more precise measure of closeness to the standard normal distribution. The left panel shows the Q–Q plots for Hu4 applied with standardization only (blue plus signs) and for Hu4 after automatic transformation (green stars). The symbols are the quantiles of 1000 randomly selected data points against the theoretical quantiles of the standard normal distribution. For comparison, we also show the 45° red dashed line. The blue plus signs clearly depart from this line, while the green stars approximately lie on the line. This contrast suggests a dramatic improvement in normality by our automatic transformation of Hu4. A similar improvement in normality of Eccentricity is also shown in the right panel. Although there are slight departures at each tail of the transformed data, an overall improvement can be seen as the majority of the quantiles approach the theoretical quantiles of the standard normal distribution.

Even though our transformation acts only on the marginal distributions, it often results in major improvement of the joint distribution of the features. In Figure 2.3, the scatter plot on the left shows a strong non-linear relationship between the Hu4 and Eccentricity that were studied in Figures 2.1 and 2.2. After transformation, the scatter plot on the right shows a bivariate Gaussian relationship which is much more amenable to analysis using standard statistical tools.

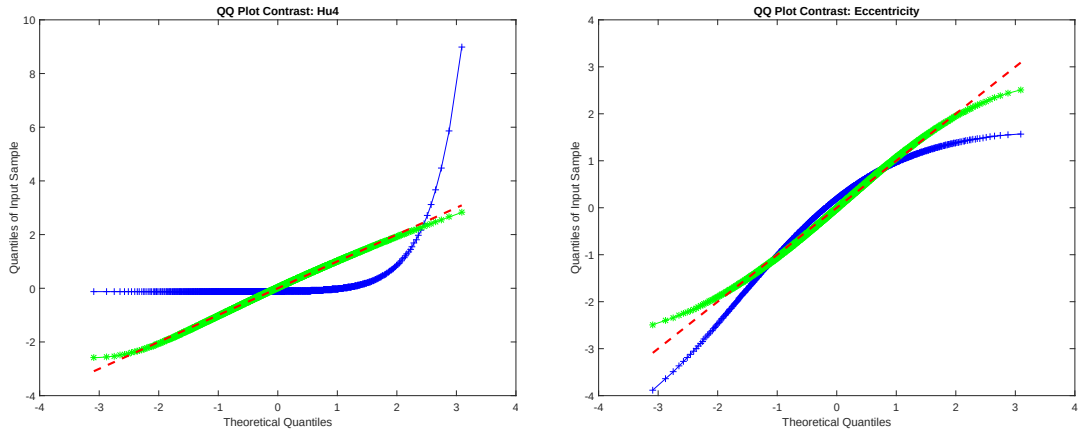


Figure 2.2: The Q–Q plots of Hu4 (left) and Eccentricity (right). The comparison between before (blue plus signs) and after (green stars) indicates a major overall improvement in closeness to normality made by transformation.

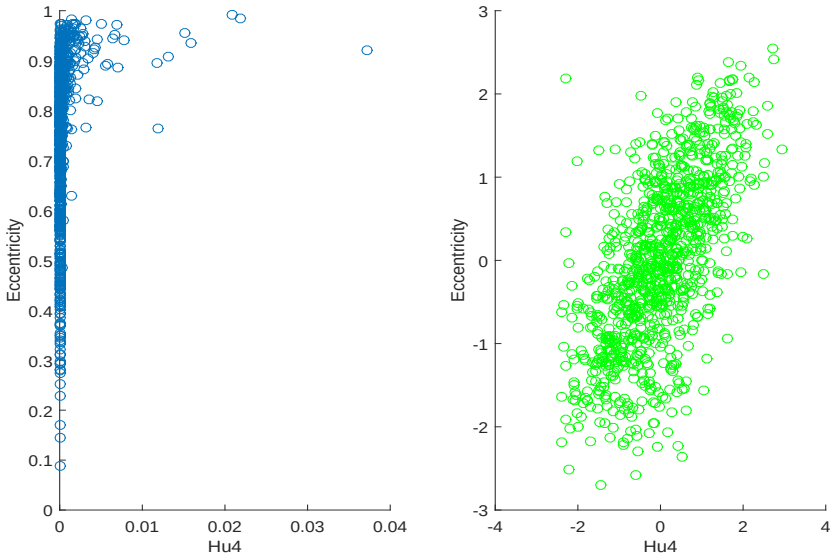


Figure 2.3: Comparison of the scatter plots, showing the joint distributions from Figure 2.1, before (left) and after (right) transformation. Relationship after transformation is much closer to linear.

2.2 Methodology

In this section, a novel automatic data transformation scheme is proposed for general data sets to achieve approximate normality. For any given data set, the transformation works feature by feature. In other words, for a data matrix with columns considered as data objects and rows considered as features, the transformation is applied to each row.

The transformation scheme consists of three components: a family of shifted logarithm transformation functions indexed by a parameter β, standardization with an option for winsorization of extreme observations and an evaluation of the transformation with a given parameter value. The key steps will be introduced in the following subsections.

The transformation scheme is a grid search based on three components to determine the optimal value of β for each feature, which is outlined as

- *Initialization:* Construct a grid of parameter values $\beta = \{\beta_i, i = 1, \ldots, m\}$
- *Step 1:* Apply the transformation function to the feature vector for each parameter value β_k.
- *Step 2:* Standardize the transformed feature vector and winsorize any existing extreme observations. Re-standardize the feature vector if winsorization has been done.
- *Step 3:* Calculate the Anderson–Darling test statistic.

Lastly, select β to minimize the Anderson–Darling test statistic for normality.

2.2.1 Transformation Function

A new parametrization of the family of shifted logarithm functions, $\{\phi_\beta, \beta \in \mathbb{R}\}$, is proposed for addressing both left and right skewness in each individual feature. For a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the sample skewness of $\mathbf{x}$ is

$$g(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2)^{\frac{3}{2}}}$$

where $\bar{\mathbf{x}}$ is the sample mean of the vector $\mathbf{x}$.

As convex transformation functions tend to increase the skewness of data while concave transformations reduce it (van Zwet, 1964), the transformation functions are chosen to be concave for $g(\mathbf{x}) > 0$ and convex for $g(\mathbf{x}) < 0$. As logarithm functions are concave, the transformation function can be a logarithm for the case $g(\mathbf{x}) > 0$ i.e. $\log(x_i)$. While for the other case $g(\mathbf{x}) < 0$, the transformation function should be made convex by inserting negative signs within and before a logarithm function i.e. $-\log(-x_i)$.

A limitation of sample skewness is its lack of robustness, i.e. sensitivity to outliers. Our algorithm only uses the sign of the skewness, and not its magnitude. Hence it works well in the typical case where outliers go in the same direction as the skewness. There can be exceptions to this, which will be detected during the recommended visualization of the results of our transformation.

Because logarithm functions require positive inputs, it is important to modify the functions for both cases to be valid for any element x_i. For example, in the case $g(\mathbf{x}) > 0$, this concern can be resolved by subtracting the minimal value of the feature vectors from x_i and adding a positive shift parameter η. That is,

$$\log(x_i - \min(x_1, x_2, \ldots, x_n) + \eta), \tag{2.1}$$

Similarly for the negative skewness $g(\mathbf{x}) < 0$, the function is

$$-\log(\max(x_1, x_2, \ldots, x_n) - x_i + \eta), \tag{2.2}$$

The shift parameter η is further parameterized in terms of the multiples of the range of the feature vectors i.e. $R = \max(x_1, x_2, \ldots, x_n)$ - $\min(x_1, x_2, \ldots, x_n)$. This makes the selection of parameter values independent of the data magnitude. In particular, set

$$\eta = |\frac{1}{\beta}|R. \tag{2.3}$$

By tuning the value of β, the effect of the transformation varies. In particular, the transformation together with standardization is equivalent to standardization only, when the parameter β approaches 0. In order to make the resulting transformation function $\phi_\beta(x_i)$ continuous over $\beta \in \mathbb{R}$, we define our transformation to be standardization only for $\beta = 0$.

Incorporating all these elements, the formal representation of the family of transformation functions is

$$\phi_\beta(x_i) = \begin{cases} \log(x_i - \min(x_1, x_2, \ldots, x_n) + |\frac{1}{\beta}|R), & \beta > 0 \\ -\log(\max(x_1, x_2, \ldots, x_n) - x_i + |\frac{1}{\beta}|R), & \beta < 0 \end{cases} \tag{2.4}$$

in which $\beta \in \mathbb{R}$, R and $g(\mathbf{x})$ are as defined above.

2.2.2 Standardization and Winsorization

After logarithm transformation, standardization is applied to the transformed feature vector i.e.$[\phi_\beta(x_1), \ldots, \phi_\beta(x_n)]$, by subtracting its median and dividing by $\sqrt{\frac{\pi}{2}}$ of the mean absolute deviation from the median.[1] The median is used considering the lack of robustness of sample mean. The mean absolute deviation is choose over the median absolute deviation as it is less likely to be zero and therefore is computationally preferable. Denote the vector after standardization as $\mathbf{x}^\dagger$.

While the shifted logarithm is frequently very successful at eliminating skewness, in some situation there can still be influential outliers. Here a winsorization of $\mathbf{x}^\dagger$ at an appropriate threshold is further applied to reduce the impact of extreme observations. Because standardization and winsorization are both based on roughly Gaussian data, it is important to apply these operations after log transformation.

2.2.2.1 Winsorization

Extreme value theory provides reasonable choices of thresholds for winsorization. A fundamental result of that area is the *Fisher–Tippett–Gnedenko Theorem*, also known as the *Three Types Theorem*. This theorem was first developed by Fisher and Tippett (1928); Gnedenko (1943). See De Haan and Ferreira (2007) and Leadbetter et al. (2011) for detailed discussion.

Theorem 2.1 (Fisher–Tippett–Gnedenko Theorem)**.** *Suppose* $X = (X_1, X_2, \ldots, X_n)$ *are independent random variables with the underlying distribution* F*. Define* $M_n = \max(X_1, X_2, \ldots, X_n)$*. Assume there exist constants* $a_n > 0$*,* b_n *such that*

$$P(\frac{M_n - b_n}{a_n} \leq x) = F(a_n x + b_n)^n \to G(x) \tag{2.5}$$

If a non-degenerate G *exists, it belongs to the family of generalized extreme value distributions* $G_\zeta(ax + b)$ *with* $a > 0$ *and* $b \in \mathbb{R}$*, where*

$$G_\zeta(x) = e^{-(1+\zeta x)^{-1/\zeta}}, 1 + \zeta x > 0 \tag{2.6}$$

[1] If the mean absolute deviation is zero, return a vector of zeros.

The parameter ζ is a real number named as extreme value index and governs the tail behaviors of each type of distribution.

The $G_\zeta(x)$ has three types of distribution defined by $\zeta > 0$, $\zeta = 0$ and $\zeta < 0$. When $\zeta = 0$, $G_\zeta(x)$ is represented as $e^{-e^{-x}}$, which is known as the standard *Gumbel* distribution or *type I* extreme value distribution. The other two types, $\zeta > 0$ and $\zeta < 0$, respectively correspond to *Fréchet* and *Weibull* distributions. For our purpose a common case of Theorem 2.1 is where the underlying distribution F is standard normal distribution and the generalized extreme value distribution is the standard Gumbel distribution as discussed in De Haan and Ferreira (2007).

Theorem 2.2. *Suppose $X = (X_1, X_2, \ldots, X_n)$ are independent, identically distributed standard normal random variables, there exist real constants $a_n > 0$ and b_n such that*

$$P(\frac{M_n - b_n}{a_n} \leq x) \rightarrow G(x) \tag{2.7}$$

where $G(x)$ is the cumulative distribution function of the standard Gumbel distribution i.e. $G(x) = e^{-e^{-x}}$ and

$$b_n = \sqrt{2\log n - \log\log n - \log 4\pi} \tag{2.8}$$

$$a_n = (2\log n)^{-\frac{1}{2}} \tag{2.9}$$

From this extreme value theory, the threshold of the standardized vector $\mathbf{x}^\dagger$ is computed based on the 95th percentile of the standard Gumbel distribution (p_{95}), that is

$$L = p_{95}a_n + b_n. \tag{2.10}$$

When the absolute value of the element in $\mathbf{x}^\dagger$ is greater than L i.e. $|x_i^\dagger| > L$, the element value is winsorized (i.e pulled back) to the value $\text{sign}(x_i^\dagger)L$. After the winsorization, the feature vector will be standardized again, using the sample mean and standard deviation, since the impact of the outliers has been mitigated.

2.2.3 Evaluation

The evaluation of the stated transformation procedure is based on measuring the distance between the empirical distribution function (EDF) of the transformed data and

the cumulative distribution function (CDF) of the standard normal. Commonly used EDF statistics are the Kolmogorov–Smirnov test statistic, the Cramér-von Mises test statistic, the Watson statistic and the Anderson–Darling test statistic. Stephens (1974) conducted power studies of these statistics under different specifications of hypothesized distributions. Based on this study, the Anderson–Darling test statistic is considered as powerful for detecting most common departures from normality. Therefore, that is used here as the criterion for evaluation. The Anderson–Darling test statistic is constructed based on measuring a distance between the empirical distribution function of observations $\{x_i, i = 1, \ldots, n\}$ i.e. F_n and the CDF of the standard normal distribution Φ. The empirical distribution function F_n is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i < x) \tag{2.11}$$

The Anderson–Darling test statistic aims to give appropriate weight to the tails using a weighted L^2 metric. In particular, the Anderson–Darling test statistic is based on

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - \Phi(x))^2}{\Phi(x)(1 - \Phi(x))} \, d\Phi(x). \tag{2.12}$$

A simply computable form of the Anderson–Darling test statistic is defined in terms of the order statistics (Anderson and Darling, 1954) i.e.

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i - 1}{n} [\log \Phi(x_{(i)}) + \log(1 - \Phi(x_{(n+1-i)}))] \tag{2.13}$$

Larger values of this indicate stronger departures from Gaussianity. Thus, by searching for a parameter value minimizing this statistic, an optimal transformation for improving the closeness of the distributions of features to normality is obtained.

2.3 Discussion

2.3.1 Limitations of the Shifted Log Transformation

The newly proposed shifted log transformation has useful power for improving the closeness to normality of a data distribution with strong skewness. However, as with all transformations, it has some limitations. For example, the shifted logarithm is not

useful in modifying a symmetric, but highly kurtotic distribution. It also provides no benefit for binary variables.

2.3.2 Computation

In our proposed transformation, a grid search is used to search for the parameter value β to minimize the Anderson–Darling test statistic. This algorithm performs an exhaustive search through a manually specified subset of the parameter space, guided by the statistic as performance metric i.e. objective function. In general, a grid search is most appropriate when objective functions are cheap to compute or the number of parameters is small. Considering the complication of optimizing the Anderson–Darling test statistic, a grid search is preferable in this context. Besides, there is only one parameter for searching and the algorithm can be easily computed in parallel as the metric evaluations are independent of each other.

The parameter β is real-valued for arbitrary data sets because of our reparametrization. The search region is among either positive or negative values based on the sign of the skewness. Take the positive values as an example for illustration. The search candidates for this side are the exponential values of equally-spaced points in the interval $[0,\ 9]$ with step 0.01. This search space contains more candidates with small values. This is because the shift parameter η ($\eta = \frac{R}{|\beta|}$) in the function 2.4 is more sensitive to changes in small values of β. Thus, more candidates near zero lower the chance of missing optima. As discussed in Section 2.2.1, when the parameter β approaches 0, the transformation together with standardization is equivalent to standardization only. When the parameter β increases to a very large number e.g. the upper bound e^9 of the search region, the shift parameter η tends to be zero and the transformation works similarly as the conventional shifted logarithm transformation.

CHAPTER 3: NON-ITERATIVE JOINT AND INDIVIDUAL VARIATION EXPLAINED

3.1 Introduction

A major challenge in modern data analysis is data integration, combining diverse information from disparate data sets measured on a common set of experimental subjects. A unified and insightful understanding of the set of data blocks is expected from simultaneously exploring the joint variation representing the inter-block associations and the individual variation specific to each block.

Lock et al. (2013) formulated this challenge into a matrix decomposition problem. Each data block is decomposed into three matrices modeling different types of variation, including a low-rank approximation of the joint variation across the blocks, low-rank approximations of the individual variation for each data block, and residual noise. Definitions and constraints were proposed for the joint and individual variation together with a method named *JIVE* for obtaining a target decomposition.

The method JIVE developed a very promising framework for studying multiple data matrices. However, the concepts of joint and individual variation were neither fully understood nor well defined. That lack of understanding of variation led to problems in computation. The Lock et al. (2013) algorithm was iterative (thus slow) and had no guarantee of achieving a solution that satisfied the definitions of JIVE. The example in Figure 3.4 below shows this is a serious issue. Another drawback of that approach includes a need for arbitrary normalization of the data sets which can be hard to choose in some complicated contexts. A related algorithm was developed by Zhou et al. (2015), which consider a JIVE type decomposition as a quadratic optimization problem with restrictions to ensure identifiability. But it still has some drawbacks in terms of interpretation. Besides, the Zhou et al. (2015) algorithm also requires iterations and an additional tuning parameter for distinguishing joint and individual variation.

A novel solution is proposed here for addressing this matrix decomposition problem. This provides a relatively very efficient non-iterative algorithm ensuring an identifiable

decomposition and also an insightful new interpretation of extracted variation structure. The key insight is the use of row spaces, i.e. a focus on scores, as the principal descriptor of the joint and individual variation, assuming columns are the n data objects, e.g. vectors of measurements on patients. This focuses the methodology on variation patterns across data objects, e.g. patient signatures, which gives straightforward definitions of the components and thus provides identifiability. These variation patterns are captured by the *row patterns* living in the row space, defined as *score subspaces* of $\mathbb{R}^n$. Segmentation of joint and individual variation is based on studying the relationship between these score subspaces and using perturbation theory to quantify noise effects (Stewart and Sun, 1990).

Using score subspaces to describe variation contained in a matrix not only empowers the interpretation of analysis but also improves the correctness and efficiency of the algorithm. An identifiable decomposition can now be obtained with all definitions and constraints satisfied. Moreover, the selection of a tuning parameter to distinguish joint and individual variation is eliminated based on theoretical justification using perturbation theory (Stewart and Sun, 1990). A consequence is a fast linear algebra based algorithm which no longer requires any iteration. The algorithm achieves an overall speedup factor around 16 compared with JIVE, when analyzing the data described in section 3.1.1. A further benefit of this new approach is that a very problematic data normalization to handle data scaling and widely differing numbers of features is no longer needed as variation patterns are now quantified by score subspaces.

Other methods that aim to study joint variation patterns and/or individual variation patterns have also been developed. Westerhuis et al. (1998) discusses two types of methods. One main type extends the traditional Principal Component Analysis (PCA), such as Consensus PCA and Hierarchical PCA first introduced by Wold et al. (1987, 1996). An overview of extended PCA methods is discussed in Smilde et al. (2003). This type of method computes the block scores, block loadings, global loadings and global scores based on an iterative procedure. The other main type of method are extensions of Partial Least Squares (PLS) (Wold, 1985) or Canonical Correlation Analysis (CCA) (Hotelling, 1936) that seek associated patterns between the two data blocks by maximizing covariance/correlation. For example, Wold et al. (1996) introduced multi-

block PLS and hierarchical PLS (HPLS) and Trygg and Wold (2003) proposed *O2-PLS* to better reconstruct joint signals by removing structured individual variation.

A connection between extended PCA and extended PLS methods is discussed in Hanafi et al. (2011). Both types of methods provide an integrative analysis by taking the inter-block associations into account. These papers make the recommendations to use normalization to address potential scale heterogeneity, including normalizing by the Frobenius norm, or the largest singular value of each data block etc. However, there are no consistent criteria for normalization and some of these methods have convergence problems. An important point is that none of these approaches provide simultaneous decomposition highlighting joint and individual modes of variation with the goal of contrasting these to reveal new insights.

3.1.1 Practical Motivation

Simultaneous variation decomposition has been useful in many practical applications, e.g., cancer genomic research. For example, Lock and Dunson (2013), Kühnle (2011), Mo et al. (2013) performed integrative clustering on multiple sources to reveal novel and consistent subtypes based on understanding of joint and individual variation. Other types of application include analysis of multi-source metabolomic data (Kuligowski et al., 2015), extraction of commuting patterns in railway networks (Jere et al., 2014), recognition of braincomputer interface (Zhang et al., 2015) etc.

The Cancer Genome Atlas (TCGA) (Network et al., 2012) provides a prototypical example for the application of JIVE. TCGA contains disparate genomic data types generated from high-throughput technologies. Integration of these is fundamental for studying cancer on a molecular level. As a concrete example, we analyze gene expression, copy number variations, reverse phase protein arrays (RPPA) and gene mutation for a set of 616 breast cancer tumor samples. For each tumor sample, there are measurements of 16615 gene expression features, 24174 copy number variations features, 187 RPPA features and 18256 mutation features. Thus, these data sources have very different dimensions. Additionally, the various data sources have different scalings, e.g., the gene expression data are continuous with range between -20 and 20 while the mutation data are binary valued.

The tumor samples are classified into four molecular subtypes: Basal-like, HER2, Luminal A and Luminal B. An integrative analysis targets the association among the features of these four disparate data sources that jointly quantify the differences between tumor subtypes. In addition, identification of driving features for each source and subtype is obtained from studying loadings.

3.1.2 Toy Example

A toy example provides a clear view of multiple challenges brought by potentially very disparate data blocks. This toy example has two data blocks, X (100×100) and Y (10000×100), with patterns corresponding to joint and individual structures. Figure 3.1 shows colormap views of matrices, with the value of each matrix entry colored according to the color bar at the bottom of each subplot. The signals have row mean 0. Therefore mean centering is not necessary in this case. A careful look at the color bar scalings shows the values are almost 4 orders of magnitude larger for the top matrices. Each column of these matrices is regarded as a common data object and each row is considered as one feature. The number of features is also very different as labeled in the y-axis. Each of the two raw data matrices (X and Y in the left panel) is the sum of joint, individual and noise components shown in the other panels.

The joint variation for both blocks presents a contrast between the left and right halves of the data matrix, thus having the same rank one score subspace. If for example the left half columns were male and right half were female, this joint variation component can be interpreted as a contrast of gender groups which exist in both data blocks for those features where color appears.

The X individual variation partitions the columns into two groups of size 50 that are arranged so the row space signature is orthogonal to that of the joint score subspace. The individual signal for Y contains two variation components, each driven by the half of the features. The first component, displayed in the first 5000 rows, partitions the columns into three groups. The other component is driven by the bottom half of the features and partitions the columns into two groups, both with row spaces orthogonal to the joint. Note that these two individual score subspaces for X and Y are different but not orthogonal. The largest principal angle between the individual subspaces is 48°.

This example presents several challenging aspects, which also appear in real data sets, such as TCGA. First, the values of the features are orders of magnitude different between X and Y. There are two standard approaches to handle this, both having drawbacks. Feature by feature normalization loses information in X because Y has so many more features. Total power normalization tends to underweight the signal in Y because each feature then receives too little weight.

The noise matrices are standard Gaussian random matrices (scaled by 5000 for X) which generates a very noisy context for both data blocks and thus a challenge for analysis.

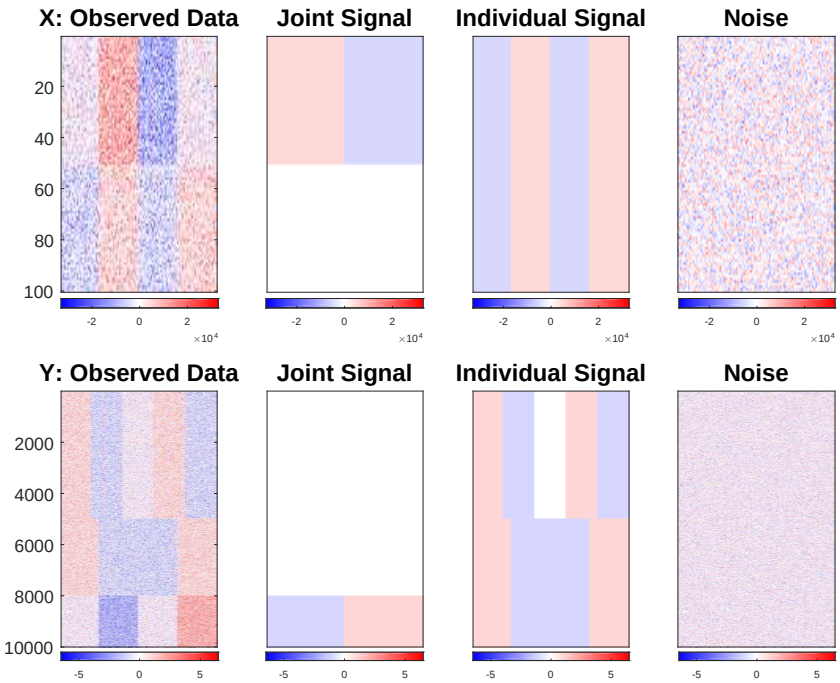


Figure 3.1: Data blocks X (top) and Y (bottom) in the toy example. The left panel of figures present the observed data matrices with each type of signal and noise matrices depicted in the remaining panels. Color bar at the bottom of each sub-plot. These structures are challenging to capture using conventional methods due to very different orders of magnitude and numbers of features.

A first attempt at integrative analysis can be done by concatenating X and Y on columns and performing a singular value decomposition on this concatenated matrix. Figure 3.2 shows the results for 3 choices of rank. The rank 2 approximation essentially captures the joint variation component and the individual variation component of X. This can be clearly seen in the rank 2 approximation of Y. The bottom 2000 rows show a contrast of two groups as the joint variation and the top half reveals differences

of four groups as the individual component of X. Considering the magnitude of the X matrix, the rank 2 approximation gives a reasonable result. One might hope that the Y individual components would show up in the rank 3 and rank 4 approximations. However, because the noise in the X matrix is so large, a noise component from X dominates the Y signal, so the important latter component disappears from this low rank representation. In this example, this naive approach completely fails to give a meaningful joint analysis.

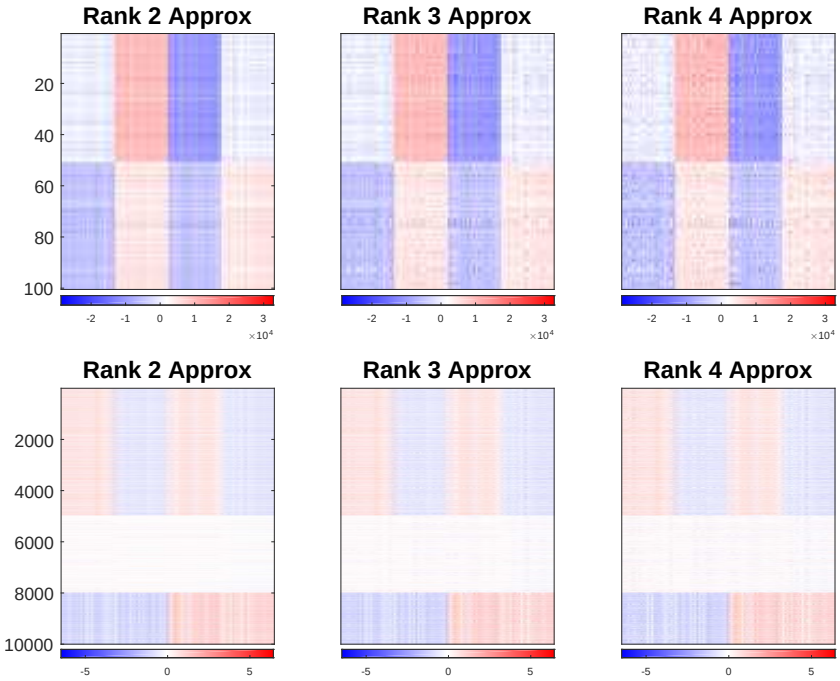


Figure 3.2: Shows the concatenation SVD approximation of each block for rank 2 (left), 3 (center) and 4 (right). Although block X has a relatively accurate approximation when the rank is chosen as 2, the individual pattern in block Y has never been captured due to the heterogeneity between X and Y.

PLS and CCA might be used to address the magnitude difference in this examples and capture the signal components. However, they target at finding common relationships between two data matrices and therefore are not able to simultaneously extract and distinguish the two types of variation. Figure 3.3 presents the PLS approximations with different number of components selected. The first PLS component shown in the left panel mainly captures the individual component in X. Although the joint variation is expected to be in the one component PLS approximations, it is later captured by the two components PLS approximations displayed in the middle. Therefore PLS completely fails to distinguish the joint and individual variation structure. The right

panel shows the three component PLS approximations which only include more noise. The two individual components in Y are not captured by any of these selected number of components. More detailed studies of SVD, PLS and CCA are given in Section 3.2.

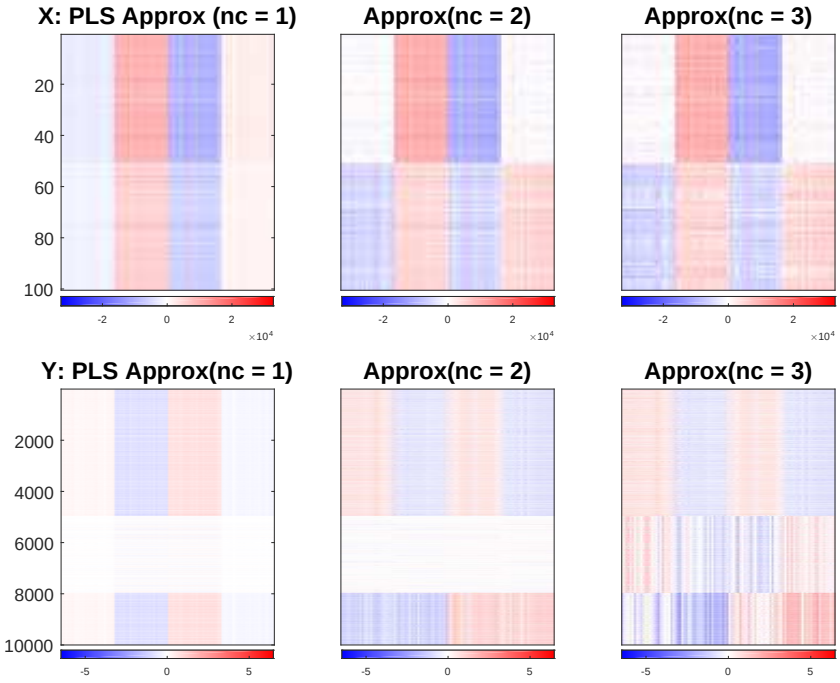


Figure 3.3: Shows the PLS approximations of each block for numbers of components as 1 (left), numbers of components as 2 (center) and numbers of components as 3 (right). PLS fails to distinguish the joint and individual variation structure as the one component PLS approximation is driven by the individual component in X.

The Lock et al. (2013) method, called old JIVE here, is applied to this toy data set. The left panel of Figure 3.4 shows a reasonable JIVE approximation of the total signal variation within each data block. However, the Lock et al. (2013) method gives rank 2 approximations to the joint matrices shown in the middle panel. The approximation consists of the real joint component together with the individual component of X. Following this, the approximation of the X individual matrix is a zero matrix and a wrong approximation of the Y individual matrix is obtained shown in the top half of the right panel. We speculate that failure to correctly apportion the joint and individual variation is caused by either the iterative algorithm that cannot guarantee the satisfaction of JIVE definitions, and/or the Frobenius norm normalization of the individual components.

The left panel of Figure 3.5 shows our JIVE approximation of each data block which well captures the signal variations within both X and Y. What's more, our

method correctly distinguishes the types of variation showing its robustness against heterogeneity across data blocks. The approximations of both joint and individual signal are depicted in the remaining panels.

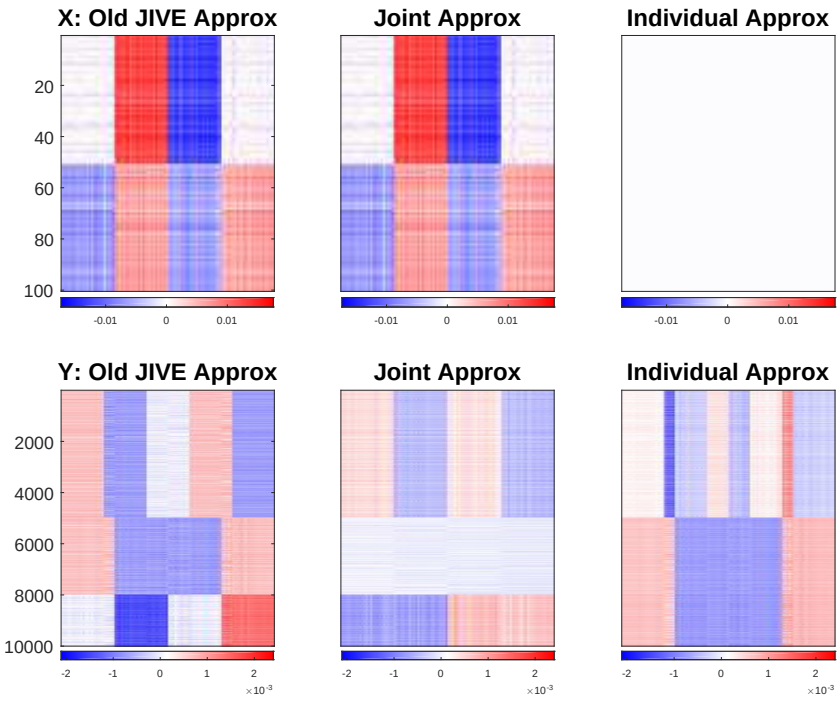


Figure 3.4: The Lock et al. (2013) JIVE method approximation of the data blocks X and Y in the toy example are shown in the first panel of figures. The joint matrix approximations (middle panel) incorrectly contain the individual component of X caused by the problematic algorithm and inappropriate normalization.

The rest of this chapter is organized as follows. Section 3.2 introduces related methods. Section 3.3 describes the population model and the estimation approach. Results of application to a mortality data set and a TCGA breast cancer data set are presented in Chapter 4.

3.2 Related Methods

3.2.1 Singular Value Decomposition (SVD)

SVD is a fundamental tool since it simultaneously provides the *principal components* (PC) for both the row space and the column space of a data matrix, after appropriately subtracting the feature means. However, unlike PCA, it does not necessarily have to be centered and allows more choices of mean centering. A brief introduction to SVD is

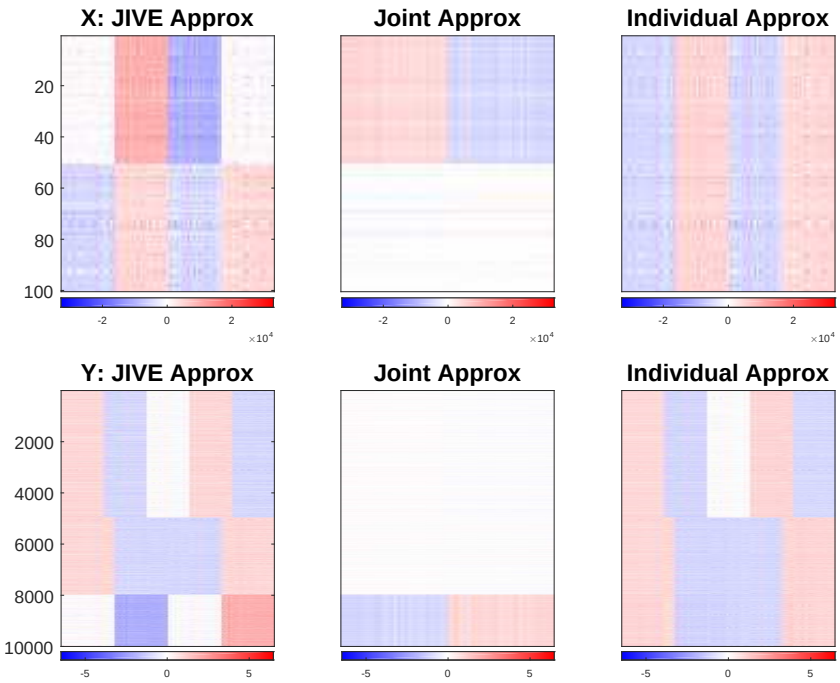


Figure 3.5: Our JIVE method approximation of the data blocks X and Y in the toy example are shown in the first column of figures, with the joint and individual signal matrices depicted in the remaining columns. Both quite diverse types of variations are well captured for each data block by new proposed JIVE.

given in this section to highlight its role in the newly developed method. The discussion of five types of mean centering of SVD in Zhang et al. (2007) is also summarized here.

Let A be a $d \times n$ matrix of rank r. The columns of A are often viewed as data objects in an experiment, and the rows of A are thought of as the features. Then, a full SVD of A can be represented as

$$A = U\Sigma V^T = \sum_{k=1}^{\min(d,n)} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

where the columns of the unitary matrices $U = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ are respectively the *left and right singular vectors* of A. The diagonal entries of the $d \times n$ matrix Σ i.e. $\text{diag}(\sigma_1, \ldots, \sigma_{\min(d,n)}, 0, \ldots, 0)$ are corresponding *singular values* with ordered non-negative numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(d,n)} \geq 0$. When the rank $r < \min(d, n)$, singular values from σ_{r+1} to $\sigma_{\min(d,n)}$ are equal to zero. By eliminating the zero components, an economy version of the SVD is represented as

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

Figure 3.6 visualizes the economic SVD representation of a rank-r matrix A. As can be seen in the figure, the row space of the matrix A is spanned by the r right singular vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ in the matrix V which are also known as the *score vectors*. These score vectors can be understood as the hidden variation patterns whose importance are indicated by the corresponding singular values in Σ. Similarly $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ span the column space of the matrix A, known as the *loading vectors*. The loading vectors give linear combinations of observed features, that is, each row of the matrix A, to generate the latent features that is the score vectors.

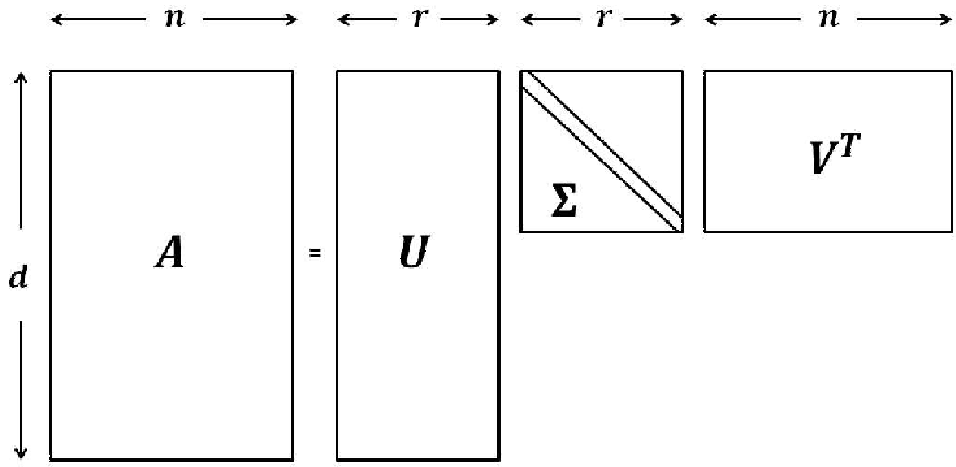


Figure 3.6: SVD decomposes the rank r data matrix A into three parts: the unitary matrix U on the left and V^T on the right respectively contain left and right singular vectors of A. The first r of them correspond to the r positive singular values in decreasing order. The right singular vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ span the row subspace of the matrix A. The right singular vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ span the column space of the matrix A, which are also called loading vectors and gives linear combinations of rows in A to generate the score vectors

.

The SVD factorization has an important approximation property. In the equation (3.1), A is expressed as sum of orthogonal layers $\sigma_k \mathbf{u}_k \mathbf{v}_k^T$ of decreasing importance indicated by the singular value σ_k. It is common to keep the layers with larger σ_k and treat the rest as noise, especially for high dimensional low sample size data sets. This property enables SVD to work as a signal extraction device by providing lower rank approximation of a noisy data matrix. For any selected integer $l \leq r$, the matrix

$$A^{(l)} = \sum_{k=1}^{l} \sigma_k \mathbf{u}_k \mathbf{v}_k^T \tag{3.1}$$

is the closest *rank l approximation* to A in terms of Frobenius norm, that is

$$A^{(l)} = \underset{\tilde{A}:rank(\tilde{A})=l}{\operatorname{argmin}} \|A - \tilde{A}\|_F^2$$

The sum of the remaining layers can be defined as the *residual matrix* $R(A) = A - A^{(l)}$. And the sum of squares of the elements in $R(A)$ is denoted as $RSS(A)$.

Mean Centering Process

Zhang et al. (2007) provided a comprehensive discussion of five types of mean centering of SVD: *no centering*, *overall centering*, *column centering*, *row centering* and *double centering* which is centering in both row and column directions.

Let $\bar{a}$ be the sample overall mean of all the elements in A; $\bar{a}_c$ be the sample column mean vector of the columns as data objects and $\bar{a}_r$ be the sample row mean vector of the rows as data objects. These mean vectors can be respectively written as

$$\bar{a} = \frac{1}{nd} 1_{1\times d} A 1_{n\times 1}$$
$$\bar{a}_c = (\frac{1}{d} 1_{1\times d} A)^{'}$$
$$\bar{a}_r = \frac{1}{n} A 1_{n\times 1}$$

The sample mean matrices of each type can be correspondingly defined as

- No centering: $0_{d\times n}$
- Overall centering: $\bar{a} I_{d\times n}$
- Column centering: $1_{d\times 1}\bar{a}_c^{'}$
- Row centering: $\bar{a}_r 1_{1\times n}$
- Double centering: $DM = 1_{d\times 1}\bar{a}_c^{'} + \bar{a}_r 1_{1\times n}$ - $\bar{a} 1_{d\times n}$

These sample mean matrices can be considered as the projections of matrix A onto a set of special subspaces which either or both have a same value of column or row vector. Denote P_d and P_n as two projection matrices $P_d = \frac{1}{d} 1_{d\times d}$, $P_n = \frac{1}{n} 1_{n\times n}$. Then, the projection representations of each sample mean matrix are respectively written as

$P_d A P_n$ for overall centering, $P_d A$ for column centering, $A P_n$ for row centering and $P_d A + A P_n - P_d A P_n$ for double centering.

Zhang et al. (2007) stated that the overall mean centering might cause data sets to lose some good features, for example, the orthogonality of the curves in functional data set. Therefore, the overall mean centering was not recommended. The comparison was made for the other four types of mean centering in terms of approximation performance. Denote $A^{(N)}$, $A^{(C)}$, $A^{(R)}$ and $A^{(D)}$ respectively as the best approximation matrices of SVD after each type of mean centering. Depending on the ranks of these matrices, the comparisons of approximation performances are different. The main results are

1. With a same selected rank of SVD after column or row centering, double centering gives a better approximation than either of column centering or row centering alone.

2. No centering is better than either column or row centering, if $rank(A^{(N)}) - 1 = rank(A^{(C)}) = rank(A^{(R)})$.

3. Column and row centering is better than no centering if the SVDs have the same number of components after centering, i.e, $rank(A^{(N)}) = rank(A^{(C)}) = rank(A^{(R)})$.

4. If $rank(A^{(D)}) + 1 = rank(A^{(C)}) = rank(A^{(R)})$, column and row centering is better than double centering.

5. In terms of the Frobenius norm of residual matrix, there is no clear relationship between column centering and row centering (either could be better), nor between double centering and no centering.

As stated in Zhang et al. (2007), the choice of the centering depends on the specific context of a data set and trying all the options was recommended. The decision could be made following criteria such as small Frobenius norm of the residual matrix, few components and straightforward interpretation. Therefore, the mean centering should be determined according to properties of each data set. In the following sections, Non-iterative JIVE will be developed on a set of data blocks, assuming all of them have been appropriately mean-centered.

3.2.2 Partial Least Squares (PLS)

Whereas SVD maximizes variation explained within a dataset, Partial Least Squares (PLS) seeks to maximize covariation explained between two datasets, originally introduced by Wold (1985). See Mateos-Aparicio (2011) for a historical overview of PLS.

Consider X and Y as two datasets measured on the same set of samples. The covariance matrix of their vertical concatenation can be expressed as

$$\text{Cov} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \Sigma_{XX}, & \Sigma_{XY} \\ \Sigma_{YX}, & \Sigma_{YY} \end{bmatrix},$$

in which Σ_{XX} and Σ_{YY} are the covariance matrices of X and Y. The cross-product matrices Σ_{XY} and Σ_{YX} contain the information of relationship between X and Y.

The goal of PLS is to find two unit vectors $\mathbf{b}_x$ and $\mathbf{b}_y$ such that

$$\max_{\|\mathbf{b}_x\|=\|\mathbf{b}_y\|=1} \text{Cov}(X^T\mathbf{b}_x, Y^T\mathbf{b}_y),$$

This goal can be achieved by performing SVD on the $d_X \times d_Y$ matrix Σ_{XY} i.e.

$$\Sigma_{XY} = U\Sigma V^T.$$

The left singular vectors in U and right singular vectors in V provide weights of original features, respectively for X and Y, such that the covariances between their linear combinations are sorted in a decreasing order. The covariance of each pair of linear combinations is indicated by the squared singular values σ_i^2.

PLS was extended to a predictive scheme and is known as PLS regression, that is, one data matrix, e.g., Y is taken as a set of response variables and the other matrix, e.g., X is a set of predictor variables. PLS regression is particularly suitable when the matrix X is ill-conditioned i.e. X has more predictors than the observations or contains multi-collinearity (Wold et al., 1984). Following the mechanism of PLS, PLS regression finds principal components that explain X and are also the best for explaining Y.

3.2.3 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) introduced by Hotelling (1936) maximizes correlation between two sets of variables for globally examining their relationship. Take X and Y as two datasets on the same set of samples. Similarly as PLS, CCA tries to find a pair of basis vectors $(\mathbf{b}_x, \mathbf{b}_y)$ such that the respective projections of X and Y onto them have a maximal correlation, that is,

$$\max_{\|\mathbf{b}_x\|=\|\mathbf{b}_y\|=1} \text{Corr}(X^T\mathbf{b}_x, Y^T\mathbf{b}_y),$$

The projections with maximal correlation i.e. $X^T\mathbf{b}_x$, $Y^T\mathbf{b}_y$ are the first pair of *canonical variates*. Similarly, this can be obtained by performing SVD on the matrix $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$. When the covariance matrices Σ_{XX} or Σ_{YY} are not invertible, the pseudo-inverse can be used.

3.3 Proposed Method

In this section the details of the new proposed JIVE are discussed. The population model is proposed in Section 3.3.1. The theoretical foundations based on matrix perturbation theory from linear algebra (Stewart and Sun, 1990) are given in Section 3.3.4. These theoretical results motivate our estimation approach which is proposed in Section 3.3.5.

3.3.1 Population Model - Signal

Matrices $\{X_k,\ k = 1, \ldots, K\}$ $(d_k \times n)$ are a set of data blocks for study. The columns are regarded as data objects, one for each experimental subject, while rows are considered as features. All X_k therefore have the same number of columns and perhaps a different number of rows .

Each X_k is modeled as low rank signals A_k perturbed by additive noise matrices E_k. Each low rank signal is the summation of two matrices containing joint and individual

variation, denoted as J_k and I_k respectively for each block

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_K \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_K \end{bmatrix} = \begin{bmatrix} J_1 \\ J_2 \\ \vdots \\ J_K \end{bmatrix} + \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_K \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_K \end{bmatrix} \tag{3.2}$$

Our approach focuses on *score vectors*, e.g., patient *signatures*, which are determined by the *row patterns* living in the row space, $\mathbb{R}^n$. These row patterns are essentially represented by the right basis vectors of appropriate SVDs. These score vectors generate the *score subspace*($\subset \mathbb{R}^n$). Therefore, the matrices capturing joint variation i.e. joint matrices are defined as sharing a common score subspace denoted as $\text{row}(J)$

$$\text{row}(J_k) = \text{row}(J), \quad k = 1, \ldots, K.$$

The individual matrices are individual in the sense that the intersection of their score subspaces is the zero vector space, i.e.

$$\bigcap_{k=1}^{K} \text{row}(I_k) = \{\mathbf{0}\}, \quad k = 1, \ldots, K.$$

This can be understood as there is no non-trivial common row pattern living in the individual score subspaces across blocks. To ensure an identifiable variation decomposition, orthogonality between the score subspaces of matrices containing joint and individual variation is assumed. In particular, $\text{row}(J) \perp \text{row}(I_k)$, $k = 1, \ldots, K$. Note that orthogonality between individual matrices $\{I_k,\ k = 1, \ldots, K\}$ is *not* assumed as it is not required for the model to be uniquely determined. The relationship between individual matrices, to some extent, has an impact on the estimation accuracy which will be discussed in Section 3.3.5.

Under these assumptions, the model is identifiable in the sense:

Theorem 3.1. *Given a set of matrices* $\{A_k,\ k = 1, \ldots, K\}$*, there are unique sets of matrices* $\{J_k,\ k = 1, \ldots, K\}$*, and* $\{I_k,\ k = 1, \ldots, K\}$ *so that:*

1. $A_k = J_k + I_k$, $k = 1, \ldots, K$
2. $\text{row}(J_k) = \text{row}(J)$, $k = 1, \ldots, K$

3. $\text{row}(J) \perp \text{row}(I_k),\ k = 1, \ldots, K$

4. $\bigcap_{k=1}^{K} \text{row}(I_k) = \{\mathbf{0}\}$.

The proof is provided in the Appendix. This model has enhanced the matrix decomposition idea proposed in Lock et al. (2013) by providing a clearer mathematical framework and precise understanding of the different types of variation. In particular, Lock et al. (2013) imposed rank constraints on the joint matrices i.e. $rank(J_k)$ are the same for all data blocks but did not clearly formulate the definition of a common row pattern. Furthermore, the orthogonality constraint was formulated on matrices instead of score subspaces i.e. $J_k I_k^T = \mathbf{0}$, which tended to obscure the role of row spaces in defining variation structure. An unnecessary orthogonality among individual matrices was further suggested, although not explicitly enforced in the estimation, for ensuring a well defined decomposition.

3.3.2 Population Model - Noise

The additive noise matrices are assumed to follow an isotropic error model where the energy of projection is invariant to direction in both row and column spaces. Important examples include the multivariate standard normal distribution and the multivariate student t–distribution (Kotz and Nadarajah, 2004). The singular values of each noise matrix are assumed to be smaller than the smallest singular values of each signal to give identifiability.

The assumption on the noise distribution here is less strong than the classical i.i.d. Gaussian random matrix, and only comes into play when determining the number of joint components. Other than that, the estimation approach given in Section 3.3.4 reconstructs each signal matrix based on SVD and thus is quite robust against the error distribution.

3.3.3 Principal Angel Analysis

Principal angles (PAs) introduced by Jordan (1875) provide useful language for the subsequent discussion and are defined as

Definition 3.2. *Suppose $\mathcal{P} \subset \mathbb{C}^n$ and $\mathcal{Q} \subset \mathbb{C}^n$ are two subspaces with dimensions p and q. Let $l = min(p, q)$. The principal angles between $\mathcal{P}$ and $\mathcal{Q}$ are*

$$\Theta(\mathcal{P}, \mathcal{Q}) = (\theta_1, \ldots, \theta_l)$$

in which $\theta_i \in [0, \frac{\pi}{2}]$, $i = 1, \ldots, l$, are recursively defined by

$$\cos(\theta_i) \triangleq \langle \mathbf{p}_i, \mathbf{q}_i \rangle = \max_{\mathbf{p} \in \mathcal{P}, \mathbf{q} \in \mathcal{Q}} \langle \mathbf{p}, \mathbf{q} \rangle$$

subject to $\mathbf{p} \perp \mathbf{p}_k$, $\|\mathbf{p}\| = 1$ and $\mathbf{q} \perp \mathbf{q}_k$, $\|\mathbf{q}\| = 1$, for $k = 1, \ldots, i-1$. The l pairs of unitary vectors

$$(p_i, q_i) \in P \times Q, \ i = 1, \ldots, l$$

are the principal vectors *corresponding to each principal angle.*

There are two methods based on SVD to compute principal angles between two subspaces. Let the columns of the matrix $M_P \in \mathbb{R}^{n \times p}$ and the matrix $M_Q \in \mathbb{R}^{n \times q}$ be orthonormal bases for the subspaces $\mathcal{P}$ and $\mathcal{Q}$ respectively. For historical records, the first method was introduced by Björck and Golub (1973), shown in Proposition 1. SVD is performed on $M_P^{'} M_Q$ and the singular values are the cosine value of the principal angles between the subspaces $\mathcal{P}$ and $\mathcal{Q}$.

Proposition 1 (Björck and Golub (1973))**.** *Represent the SVD of $M_P^{'} M_Q$ as $U_P S U_Q^{'}$. The first $l = min(p, q)$ singular values are $s_1 \geq s_2 \geq \cdots \geq s_l \geq 0$, then the principal angles of subspaces $\mathcal{P}$ and $\mathcal{Q}$ as defined, $\Theta(\mathcal{P}, \mathcal{Q}) = [\theta_1, \ldots, \theta_l]$, are*

$$\cos \theta_i = s_i, \ i = 1, \ldots, l$$

and $s_1 = \cdots = s_k = 1 > s_{k+1}$ if only if $\dim(\mathcal{P} \cap \mathcal{Q}) = k$. Moreover, the principal vectors are the first l columns of matrices $M_P U_P$ and $M_Q U_Q$.

The second method was later proposed by Miao and Ben-Israel (1992), as seen in Proposition 2. This method performs SVD on the vertical concatenation of the matrices $M_P^{'}$ and $M_Q^{'}$, which fits more easily into to our framework. Thus, the second method will be utilized for computing principal angles in the later sections.

Proposition 2 (Miao and Ben-Israel (1992)). *Denote a* $(p+q) \times n$ *matrix* M *as*

$$M \triangleq \begin{bmatrix} M_P' \\ M_Q' \end{bmatrix}.$$

Let the SVD of the matrix M *be* USV'*. Then the singular values on the diagonal of* S *are equal to*

$$\sqrt{1+\cos\theta_1}, \ldots, \sqrt{1+\cos\theta_l}, 1, \ldots, 1, \sqrt{1-\cos\theta_l}, \ldots, \sqrt{1-\cos\theta_1} \tag{3.3}$$

in which $\theta_i, i = 1, \cdots, l$ *are the principal angles between subspaces* $\mathcal{P}$ *and* $\mathcal{Q}$ *as in the definition. There are* $\max(p,q) - l$ *number of singular values taking on the value* 1 *in the middle.*

Let the matrix $U^{(l)}$ *be the first* l *columns of* U *and write it as the vertical concatenation of the* $p \times l$ *matrix* $U_P^{(l)}$ *and the* $q \times l$ *matrix* $U_Q^{(l)}$ *i.e.* $U^{(l)} = [U_P^{(l)}; U_Q^{(l)}]$*. Then, the principal vectors are the* l *columns of the matrices* $M_P U_P^{(l)}$ *and* $M_Q U_Q^{(l)}$*.*

Both methods can compute both principal angles and principal vectors for the two subspaces. However, the second method also provides l right singular vectors in V pointing in the same direction as the sum of corresponding principal vector pairs. In particular, consider the principal vectors p_i of subspace $\mathcal{P}$ and q_i of subspaces $\mathcal{Q}$, which correspond to the principal angle θ_i shown in Figure 3.7. The sum of the vectors p_i and q_i, denoted as v_i, is depicted as a red dashed line. The vector v_i also points to the same direction as the right singular vector in V with the singular value being $\sqrt{1+\cos\theta_i}$.

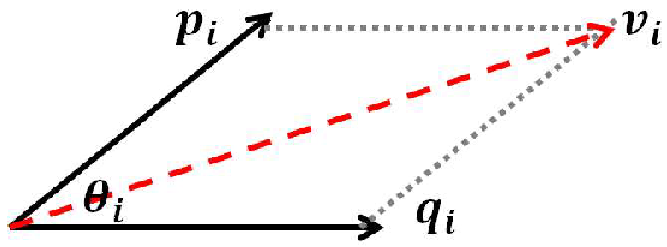


Figure 3.7: A diagram displaying the relationship between each pair of principal vectors and the right singular vector corresponding to the same principal angle.

3.3.4 Theoretical Foundations

The main challenge is segmentation of the joint and individual variation in the presence of noise which individually perturbs each signal. Let $\{\tilde{A}_k, k = 1, \ldots, K\}$ be noisy approximations of $\{A_k, k = 1, \ldots, K\}$ respectively. The subspaces of joint variation within the approximations $\tilde{A}_k$, while expected to be similar, are no longer exactly the same due to noise. If some subspaces of $\{\tilde{A}_k, k = 1, \ldots, K\}$ are very close, they can be considered as estimates of the common score subspace under different perturbations. Application of the results of the *Generalized* $\sin\theta$ *Theorem* (Wedin, 1972) is proposed to decide when a set of subspaces are close enough to be regarded as estimates of the joint score space. Based on this theorem, the number of joint components can be determined resulting in an appropriate segmentation.

Take the approximation $\tilde{A}_k$ of A_k as an example of perturbation of each matrix's score space. For consistency with the Generalized $\sin\theta$ Theorem, a notion of distance between theoretical and perturbed subspaces is defined as a measure of perturbation. Let $\mathcal{Q}$, $\tilde{\mathcal{Q}}$ be the l dimensional score subspaces of $\mathbb{R}^n$ respectively for the matrix A_k and its approximation $\tilde{A}_k$. The corresponding symmetric projection matrices are $P_{\mathcal{Q}}$ and $P_{\tilde{\mathcal{Q}}}$. The distance between the two subspaces is defined as the difference of the projection matrices under the L^2 operator norm, i.e $\rho(\mathcal{Q}, \tilde{\mathcal{Q}}) = \|P_{\mathcal{Q}} - P_{\tilde{\mathcal{Q}}}\|$ (Stewart and Sun, 1990).

An insightful understanding of this defined distance $\rho(\mathcal{Q}, \tilde{\mathcal{Q}})$ comes from a principal angle analysis (Jordan, 1875; Hotelling, 1936) of the subspaces $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$. Denote the principal angles between $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ as $\Phi(\mathcal{Q}, \tilde{\mathcal{Q}}) = \{\phi_1, \ldots, \phi_l\}$ with $\phi_1 \geq \phi_2 \ldots \geq \phi_l$. The distance ρ is equal to the sine of the maximal principal angle, i.e. $\sin\phi_1$. This suggests that the largest principal angle between two subspaces can indicate their closeness, i.e. distance. Under a slight perturbation, the largest principal angle between $\mathcal{Q}$ (a theoretical subspace) and $\tilde{\mathcal{Q}}$ (its perturbed subspace) is expected to be small.

The distance $\rho(\mathcal{Q}, \tilde{\mathcal{Q}})$ can be also written as

$$\rho(\mathcal{Q}, \tilde{\mathcal{Q}}) = \|(I - P_{\mathcal{Q}})P_{\tilde{\mathcal{Q}}}\| = \|(I - P_{\tilde{\mathcal{Q}}})P_{\mathcal{Q}}\|$$

which brings another useful understanding of this definition. It measures the relative deviation of the signal variation from the theoretical subspace. Accordingly, the similarity/closeness between the subspaces and its perturbation can be written as $\|P_{\mathcal{Q}}P_{\tilde{\mathcal{Q}}}\|$ and

is equal to the cosine of the maximal principal angle defined above, i.e. $\cos\phi_1$. Hence, $\sin^2\phi_1$ indicates the percentage of signal deviation and $\cos^2\phi_1$ tells the percentage of remaining signal in the theoretical subspace.

The generalized $\sin\theta$ theorem provides a bound for the distance between a subspace and its perturbation, e.g., the subspaces $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$. This bound quantifies how the theoretical subspace $\mathcal{Q}$ is affected by noise. In particular,

Theorem 3.3 (The Generalized $\sin\theta$ Theorem (Wedin, 1972))**.** *Signal matrix A_k is perturbed by additive noise E_k. Let ϕ_k be the largest principal angle for the subspace of signal A_k and its approximation $\tilde{A}_k$. Denote the SVD of $\tilde{A}_k$ as $\tilde{U}_k\tilde{\Sigma}_k\tilde{V}_k^T$. The distance between the subspaces of A_k and $\tilde{A}_k$, $\rho(\mathcal{Q},\tilde{\mathcal{Q}})$ i.e. sines of ϕ_k, is bounded*

$$\rho(\mathcal{Q},\tilde{\mathcal{Q}}) = \sin\phi_k \leq \frac{\max(\|E_k\tilde{V}_k\|, \|E_k^T\tilde{U}_k\|)}{\sigma_{\min}(\tilde{\Sigma}_k)}, \tag{3.4}$$

where $\sigma_{\min}(\tilde{\Sigma}_k)$ is the smallest singular value of $\tilde{A}_k$.

This bound measures how far the perturbed space can be away from the theoretical one. The deviation is bounded by the maximal value of noise energy on column and row spaces and also the smallest signal singular values. This is consistent with the intuition that a deviation distance, i.e. a largest principal angle, is small when the signal is strong and perturbations are weak.

Notice that the bound in Theorem 3.3 is applicable but cannot be directly used for data analysis since the error matrices E_k are not observable. As the error matrices are assumed to be isotropic, we propose to re-sample noisy directions from the residuals of the low rank approximations. The L^2 norm of these error related terms can thus be estimated by projecting the observed data onto the subspace spanned by re-sampled directions. This re-sampling based method can also provide confidence intervals for these perturbation bounds. More details of estimating the perturbation bound will be discussed in Section 3.3.5.

3.3.5 Estimation Approach

The algorithm uses SVD as a building block to find an estimate of the targeted decomposition . A three-step algorithm is outlined below.

1. Obtain an initial estimate of the signal score space of each data block by thresholding the singular values.

2. Extract the joint score space from the signal score spaces using a threshold derived from Theorem 3.3.

3. Decompose each data matrix into joint and individual variation matrices using projections onto the score space in Step 2.

As a basic illustration for each step we use the toy example described in Section 3.1. Details for each step appear in the following subsections.

Signal Space Initial Extraction

Even though the signal components $\{A_k,\ k = 1, \ldots, K\}$ are low rank, the data matrices $\{X_k,\ k = 1, \ldots, K\}$ are usually of full rank due to corruption by noise. SVD works as a signal extraction device in this step, keeping components with singular values greater than selected thresholds individually for each data block. These thresholds are selected using a multi-scale perspective. For example, by finding relatively big jumps in a scree plot. Figure 3.8 shows the scree plots of each data block for the toy example in Section 3.1. The left scree plot for X suggests a selection of rank as 2 and the right one for Y suggests the rank being 3, since in both cases those components stand out while the rest of the singular values decay slowly showing no clear jump.

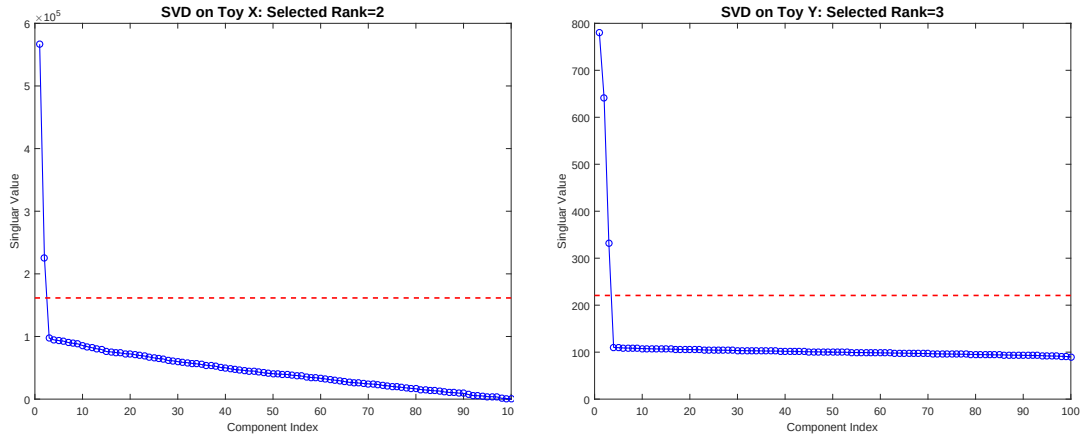


Figure 3.8: Scree plots for the toy data sets X (left) and Y (right). Both plots display the singular values associated with a component in descending order versus the index of the component. The components with singular values above the dashed red threshold line are regarded as the initial signal components in the first step of JIVE.

Let $\{\tilde{r}_k,\ k = 1, \ldots, K\}$ be the initial estimates of the signal ranks $\{r_k,\ k = 1, \ldots, K\}$. In the toy example $\tilde{r}_1 = 2$ (for X) and $\tilde{r}_2 = 3$ (for Y). Each data block has a low rank approximation, $\tilde{A}_k$, which is the initial estimate of the signal matrix A_k, $k = 1, \ldots, K$. The estimate is decomposed as

$$\tilde{A}_k = \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^T \tag{3.5}$$

where $\tilde{U}_k$ contains the left singular vectors corresponding to the largest $\tilde{r}_k$ singular values respectively for each data block. The initial estimate of the signal score space, denoted as $\text{row}(\tilde{A}_k)$, is spanned by the right singular vectors in $\tilde{V}_k$.

Score Space Segmentation: Two-Block

For a clear introduction of the basic idea of score space segmentation, the two-block special case ($K = 2$) is first studied. The goal is to use the low rank approximations $\tilde{A}_k$ from equation (3.5) to obtain estimates of the common joint and individual score subspaces. Due to the presence of noise, the components of $\text{row}(\tilde{A}_k)$ corresponding to the underlying joint space, no longer are the same, but should have a relatively small angle. Similarly, the components corresponding to the underlying individual spaces are expected to have a relatively large angle. This motivates the use of principal angle analysis as discussed in Section 3.3.3 to separate the joint from the individual components. The following Lemma 1 provides a bound on the largest allowable principal angle of the joint part of the initial estimated spaces.

Lemma 1. *Let θ be the largest principal angle between two subspaces that are each a perturbation of the common row space within* $\text{row}(\tilde{A}_1)$ *and* $\text{row}(\tilde{A}_2)$*. That angle is bounded by*

$$\sin\theta \leq \sin(\phi_1 + \phi_2) \tag{3.6}$$

in which ϕ_1 and ϕ_2 are the angles given in Theorem 3.3.

The proof is provided in the Appendix. As mentioned in Section 3.3.4, the perturbation bounds of each θ_k require the estimation of terms $\|E_k \tilde{V}_k\|$, $\|E_k^T \tilde{U}_k\|$ for $k = 1, 2$. These terms are the measurements of energies of noise matrices projected onto the signal column and row spaces. Since an isotropic error model is assumed, the energy

of the noise matrices in arbitrary directions are supposed to be equal. Denote $\tilde{V}_k^{\perp}$ $(n \times (\min(d_k, n) - \tilde{r}_k))$ and $\tilde{U}_k^{\perp}$ $(d_k \times (\min(d_k, n) - \tilde{r}_k))$ as the respective orthonormal bases of the row and column subspaces of the residual matrices from the low rank approximations in equation (3.5). Thus, we propose to resample noisy directions, i.e. column vectors, from the matrices $\tilde{V}_k^{\perp}$ and $\tilde{U}_k^{\perp}$.

Take the term $\|E_k\tilde{V}_k\|$ as an example for illustration. Given the $\tilde{r}_k$ number of column vectors resampled from $\tilde{V}_k^{\perp}$, denoted as $V^{\star}$, the observed data block X_k is projected onto the subspace spanned by $V^{\star}$, written as $X_kV^{\star}$. The L^2 norm, $\|X_kV^{\star}\|$, is taken as the estimate of the term $\|E_k\tilde{V}_k\|$. This can be similarly applied to $\|E_k^T\tilde{U}_k\|$ for $k = 1, 2$. A typical number of resamples is 1000. The quantiles of this distribution provide both a point estimate and a simulated confidence interval for terms $\|E_k\tilde{V}_k\|$, $\|E_k^T\tilde{U}_k\|$, resulting in a confidence interval for the perturbation bound. Typically the median is chosen as the estimate of the angle bound for exploratory analysis. This will result in at least 50% confidence that all joint components are included. For certain cases that no loss of joint components is desired, the 95^{th} percentile of these estimated terms can be used to derive a conservative angle threshold, resulting in at least 95% confidence of finding all joint components.

The principal angles between $\text{row}(\tilde{A}_1)$ and $\text{row}(\tilde{A}_2)$ are computed by performing SVD on a concatenation of their right singular vector matrices (Miao and Ben-Israel, 1992), i.e.

$$M \triangleq \begin{bmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{bmatrix} = U_M \Sigma_M V_M^T.$$

where the singular values Σ_M determine the principal angles, $\Theta(\text{row}(\tilde{A}_1), \text{row}(\tilde{A}_2)) = \{\theta_1, \ldots, \theta_l\}$ as

$$\theta_i = \arccos((\sigma_{M,i})^2 - 1), \ i = 1, \ldots, \min(\tilde{r}_1, \tilde{r}_2). \tag{3.7}$$

Given a left singular vector $U_{M,i}$ denoted as $\mathbf{u}$, a pair of principal vectors $\{\mathbf{p}_i, \mathbf{q}_i\}$ in each subspace can be constructed by projecting $\tilde{V}_1$ and $\tilde{V}_2$ onto the vector $\mathbf{u}$. Denote $\mathbf{u}$ as the concatenation of $[\mathbf{u}_1; \mathbf{u}_2]$. Note that the length of $\mathbf{u}_1$ is equal to the number of columns of $\tilde{V}_1$ and similarly for the other part. The principal vectors in each subspace can be written as $\mathbf{p}_i = \tilde{V}_1\mathbf{u}_1$ and $\mathbf{q}_i = \tilde{V}_2\mathbf{u}_2$ respectively. The angle between the pair

of principal vectors θ_i is equal to the principal angle computed from the singular value corresponding to $\mathbf{u}$, as illustrated in Figure 3.7.

As seen in Miao and Ben-Israel (1992), the vector $\mathbf{v}_i$, the corresponding right singular vector of V_M, points in the same direction as the sum of principal vector pairs of each subspace. When the principal angle θ_i is smaller than the perturbation bound θ, this right singular vector can be taken as an estimate of the theoretical joint direction to assure the definition of joint variation.

This SVD decomposition can be understood as a tool sorting pairs of directions within the two subspaces in increasing order of the angle between each pair. When the corresponding principal angle is smaller than the perturbation bound θ, the pair of principal vectors can be considered as noisy versions of the same joint direction. Assume there are $\hat{r}_J$ principal angles smaller than the bound θ. The first $\hat{r}_J$ singular vectors $\mathbf{v}_i$ are used as the natural orthonormal basis of the estimated joint score subspace i.e. $\text{row}(\hat{J})$.

The left panel of Figure 3.9 depicts the principal angles of the concatenated right singular vector matrices for the toy example in Section 3.1.2. Since the initial estimates of r_x and r_y are 2 and 3, there are only two potential components for joint variation. The associated principal angles between the initially estimated signal row spaces are labeled next to the first two component as 10.99° and 47.11°. The estimated bound on the principal angle in Lemma 1 is 31.29° for this toy example. The 5% and 95% one-sided confidence intervals of the angle bound are $[-\infty,\ 30.00]$ and $[-\infty,\ 32.92]$ degree. Each provides a respective 5% and 95% chance for including all the joint components. This provides a clear indication that the number of joint components should be $\hat{r}_J = 1$. The corresponding first right singular vector of M will be taken as the joint score vector.

Score Space Segmentation: Multi-Block

To generalizing the above idea to more than two blocks, they key is to focus more on singular values than on angles in equation (3.7). In other words, instead of finding an upper bound on an angle, we will focus on a lower bound on the remaining energy as expressed by the sum of the squared singular values. Hence, an analogous SVD will be used for studying the closeness of multiple initial signal score subspace estimates. Similarly, for the vertical concatenation of right singular vector matrices $\{\tilde{V}_k^T,\ k =$

$1, \ldots, K\}$, we have

$$M \triangleq \begin{bmatrix} \tilde{V}_1^T \\ \vdots \\ \tilde{V}_K^T \end{bmatrix} = U_M \Sigma_M V_M^T.$$

Once again, SVD sorts the directions within these K subspaces in increasing order of amount of deviation from the theoretical joint direction. The squared singular value $\sigma_{M,i}^2$ indicates the total amount of variation explained in the common direction $V_{M,i}^T$ in the score subspace $\subset \mathbb{R}^n$. A large value of $\sigma_{M,i}^2$ (close to K) suggests that there is a set of basis vectors within each subspace close with each other and thus are potential noisy versions of a common joint score vector. A threshold on singular values is needed to segment the joint components. This is done in Lemma 2.

Lemma 2. *Let ϕ_k be the bound on the principal angles between the theoretical subspace row(A_k) and its perturbation row$(\tilde{A}_k)$ for K data blocks from Theorem (Wedin, 1972). The squared singular values ($\sigma_{M,i}^2$) corresponding to the estimates of joint components should satisfy*

$$\sigma_{M,i}^2 \geq K - \sum_{k=1}^{K} \sin^2 \phi_k \geq K - \sum_{k=1}^{K} \Big(\frac{\max(\|E_k \tilde{V}_k\|, \|E_k^T \tilde{U}_k\|)}{\sigma_{\min}(\tilde{\Sigma}_k)}\Big)^2. \tag{3.8}$$

The proof is provided in the Appendix. This lower bound is independent of the variation magnitudes. This property gives some robustness against heterogeneity across each block when extracting joint variation information.

As above, the terms $\|E_k \tilde{V}_k\|$, $\|E_k^T \tilde{U}_k\|$ are resampled to derive a point estimate and confidence interval for the threshold. As for the two-block case, if there were $\hat{r}_J$ singular values selected, the first $\hat{r}_J$ right singular vectors are used as the basis of the estimate of row(J).

The right panel of Figure 3.9 depicts the first 2 singular values of the vertical concatenated matrix M for the toy example. This is an analysis of the same data, but performed on the scale of squared singular values instead of principal angles. The associated squared singular values are labeled next to the these two component as 1.98 and 1.68. The estimated threshold (using median) is 1.85 for the toy example. This threshold together with its 5% and 95% one sided confidence intervals, $[1.86, +\infty]$ and $[1.84, +\infty]$ respectively, suggest that the number of joint components $\hat{r}_J$ should be 1.

The corresponding right singular vector is taken as the estimate of the orthonormal basis of the joint score subspace.

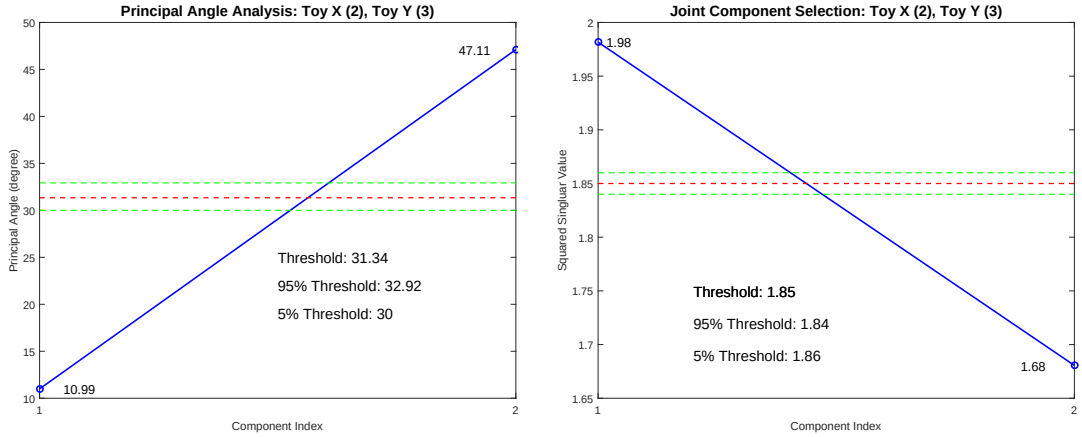


Figure 3.9: Left panel: Principal angles between the initial estimates of signal row spaces. The bound for the largest angle is 31.29 degree, suggesting the existence of one joint component. To indicate the uncertainty, the 5% and 95% one-sided confidence intervals of the angle threshold are also shown. Right panel: Squared singular values plot of the vertical concatenated matrix M for the toy example. Both thresholds correctly capture the underlying structure of this toy example with the selection of one joint component.

Final Decomposition

Based on the estimate of the joint row space, matrices containing joint variation in each data block can be reconstructed by projecting X_k onto this estimated space. Define the matrix $\hat{V}_J$ as $[\mathbf{v}_{M,1}, \ldots, \mathbf{v}_{M,\hat{r}_J}]$, where $\mathbf{v}_{M,i}$ is the i^{th} column in the matrix V_M. The projection matrix is

$$P_J = \hat{V}_J(\hat{V}_J^T\hat{V}_J)^{-1}\hat{V}_J^T$$

and the estimates of joint variation matrices in each block are

$$\hat{J}_k = X_k P_J, \quad k = 1, \ldots, K.$$

The row space of joint structure is orthogonal to the row spaces of each individual structure. Therefore, the original data blocks are projected to the orthogonal space of row$(\hat{J})$. The projection matrix onto the orthogonal space of row$(\hat{J})$ is $P_J^\perp = I - P_J$ and the projections of each data block are denoted as $X_k^\perp$ respectively for each block i.e.

$$X_k^\perp = X_k P_J^\perp$$

Finally we threshold this projection by performing SVD on $\{X_k^{\perp},\ k = 1, \ldots, K\}$. The components with singular values larger than the first thresholds from Section 3.3.5 are kept as the individual components, denoted as $\{\hat{I}_k^{\perp},\ k = 1, \ldots, K\}$. The remaining components of each SVD are regarded as an estimate of the noise matrices.

By taking a union of the estimated row spaces of each type of variation, the estimated signal row spaces are

$$\text{row}(\hat{A}_k) = \text{row}(\hat{J}) \oplus \text{row}(\hat{I}_k)$$

with rank $\hat{r}_k = \hat{r}_J + \hat{r}_{I_k}$ respectively for $k = 1, \ldots, K$.

Due to the adjustment of directions of joint components, these final estimates of signal row spaces may be different from those obtained in the initial signal extraction step. Note that even the estimates of rank $\hat{r}_k$ might also differ from the initial estimates $\tilde{r}_k$.

3.4 Post JIVE Data Representation

Given the variation decompositions of each data block, several types of post JIVE analyses are available for exploring the joint and individual score variation patterns. The estimates of joint matrices within each data block can be represented by SVD

$$\hat{J}_k = \hat{U}_J^k \hat{\Sigma}_J^k \hat{V}_J^k, \quad k = 1, \ldots, K$$

in which $\hat{V}_J^k$ are the $\hat{r}_J \times n$ score matrices of the estimated joint score space $\text{row}(\hat{J})$. Note that the singular values $\hat{\Sigma}_J^k$ can be completely different, since they are driven by the score variation pattern and can reflect very different amounts of variation between the blocks. The loading matrices $\hat{U}_J^k$ $(d_k \times \hat{r}_J)$ respectively specify distinct $\hat{r}_J$-dimension loading subspaces of $\mathbb{R}^{d_k}$ for each block k.

There are three important matrix representations of the information in the joint score space, with differing uses in post JIVE analyses.

1. *Full Matrix Representation.* For applications where the original features are the main focus (such as finding driving genes) the full matrix representations $\hat{J}_k$ $(d_k \times n)$, $k = 1, \ldots, K$ are most useful. These are shown in Figure 3.5.

2. *Block Specific Score (BSS).* For applications where the relationships between subjects are the main focus (such as discrimination between subtypes) large computational gains are available for using the much lower dimensional representations $\hat{\Sigma}_J^k \hat{V}_J^k$ $(\hat{r}_J \times n)$. This results in no loss of information when rotation invariant methods are used.

3. *Common Normalized Score (CNS).* When it is desirable to study the component of joint behavior that is separate from the within block variation (such as evaluating the relationship between data objects), the analysis should focus on a common basis of row$(\hat{J})$, namely $\hat{V}_J$ $(\hat{r}_J \times n)$ from Section 3.3.5.

The relationship between BSS and CNS is analogous to that of the traditional covariance (i.e PLS) and correlation (i.e CCA) modes of analysis.

Furthermore, different representations have different ways to study the loadings. The full matrix representation and BSS naturally obtain the information from the loading matrix $\hat{U}_J^k$. CNS gives a different representation of the loadings. Given the common basis of row$(\hat{J})$, one can perform regression for $\hat{J}_k$ on each score vector in $\hat{V}_J$, from which the standardized coefficient vector can be taken as the CNS loading. By doing this, there is no guarantee of orthogonality between CNS loading vectors. However, the loadings are linked across blocks by their common scores. Therefore, in this CNS case, the standardized regression coefficients are recommended for use instead of the classical loadings.

The individual variation within blocks can be similarly analyzed resulting in both BSS and CNS analyses for the individual components. When original features are important, the full matrix

$$\hat{I}_k = \hat{U}_I^k \hat{\Sigma}_I^k \hat{V}_I^k, \quad k = 1, \ldots, K$$

with dimension $d_k \times n$ are available. Otherwise large computational savings are available from the BSS version $\hat{\Sigma}_I^k \hat{V}_I^k$ $(\hat{r}_{I_k} \times n)$, $k = 1, \ldots, K$. For studying scale free behaviors, use the *Individual Normalized Score (INS)* $\hat{V}_I^k$ $(\hat{r}_{I_k} \times n)$. For individual components, the matrix $\hat{U}_I^k$ can be taken as loadings for all three representations as INSs cannot be the same.

CHAPTER 4: NON-ITERATIVE JIVE DATA ANALYSIS

In this chapter, we apply Non-iterative JIVE to two real data sets, Spanish mortality as analyzed in Marron and Alonso (2014) and TCGA breast cancer data set (Network et al., 2012). Detailed analyses are given in Section 4.1 and Section 4.2 respectively.

4.1 Spanish Mortality Data

A data set from the Human Mortality Database is studied here, which is consist of both male and female Spanish people. This data set demonstrates the advantage of JIVE in gaining insights. For each gender data block, there is a matrix of *mortality*, defined as the number of people who died divided by the total, for a given age group and year. Because mortality varies by several orders of magnitude, the $\log_{10}$ of the mortality is studied here. Each row represents an age group from 0 to 95, and each column represents a year between 1908 and 2002. In order to associate the historical events with the variations of mortality, columns (i.e. mortality as a function of age) are considered as the common set of data objects of each gender block. Marron and Alonso (2014) performed analysis on the male block and showed interesting interpretations related to Spanish history. Here we are looking for a deeper analysis which integrates both males and females by exploring joint and individual variation patterns.

Non-iterative JIVE is applied to the two gender blocks centered by subtracting the mean of each age group, since the mean structure contains essential variation information. The most interesting JIVE analysis comes from 3 male and 3 female components. The resulting JIVE gives 2 joint components and 1 of each individual component. Since the loading matrices provide important information of the effect of different age groups, BSS analysis together with loading matrices is most informative here.

Figure 4.1 shows a view of the first joint components for the males (left) and females (right) that is very different from the heat map views used in Section 3.1.1 for the toy example. While these components are matrices, additional insights come from plotting

the rows of the matrices as curves over year (top) and the columns as curves over age (bottom). The curves over year (top) are colored using a heat color scheme, indexing age (black = 0 through red = 40 to yellow = 95 as shown in the vertical color bar on the bottom left). The curves over age (bottom) are colored using a rainbow color scheme (magenta = 1908 through green = 1960 to red = 2002, shown in the horizontal color bar in the top) and use the vertical axis as domain with horizontal axis as range to highlight the fact that these are column vectors. Additional visual cues to the matrix structure are the horizontal rainbow color bar in the top panel, showing that year indexes columns of the data matrix and the vertical heat color bar (bottom) showing that age indexes rows of the component matrix. Because this is a single component, i.e. a rank one approximation of the data, each curve is a multiple of a single eigenvector. The corresponding coefficients are shown on the right. In our terminology, the upper BSS coefficients are the *loadings*, and are in fact the entries of the left eigenvectors (colored using the heat color scale on the bottom). Similarly, the lower coefficients are the *scores* and are the entries of the right eigenvectors, colored using the rainbow bar shown in the top.

The scores plots together with the rows as curves plots in Figure 4.1 indicate a dramatic improvement in mortality after the 1950s for both males and females. The scores plots are bimodal indicating rapid overall improvement in mortality around the the 1950s. This is also visible in the rows as curves plot. Thus the first mode of joint variation is driven by overall improvement in mortality. In addition to the overall improvement, the rows as curves and scores plot also show the major mortality events, the global flu pandemic of 1918 and the Spanish Civil war in the late 1930s. The loading plots together with the columns as curves plots present the different impacts of this common variation on different age groups for males and females. The loadings plot for males suggests the improvement in mortality is gradually increasing from older towards younger age groups. In contrast, the female block has a bimodal kernel density estimate of the loadings. This shows that female of child bearing age have received large benefits from improving health care. This effect is similarly visible from comparing the female versus male columns as curves.

The second BSS components of joint variation within each gender are similarly visualized in Figure 4.2. This common variation reflects the contrast between the years

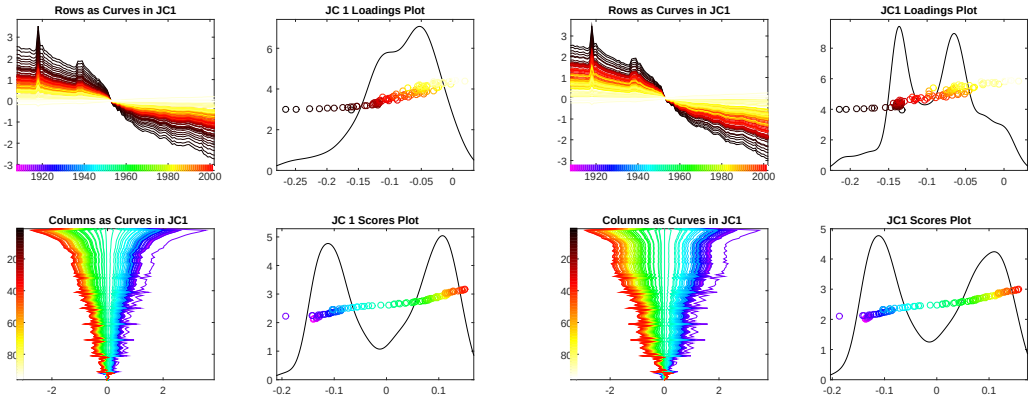


Figure 4.1: The first BSS joint components of male (left panel) and female (right panel) contain the common modes of variation caused by the overall improvement across different age groups, as can be seen from the scores plots in the right bottom of each panel. The dramatic decrease happened around the 1950s shown in the column projection plot. The decrease degrees vary from age groups.

around 1950 and the years around 1980 which can be told from the curves in the left top and the colors in the right bottom subplots in both male and female panel. In the scores plot, the green circles, seen on the left end, represent the years around 1950 when the automobile penetration started. And the orange to red circles on the right end correspond to years around 1980, after seat belt legislation was first introduced in Spain. These modes of variation can be interpreted as the increase in fatalities caused by automobiles and later improvements in safety such as seat belts and safer roads. The upper left loadings plot of males shows that these automobile events had a stronger influence on the 20-45 males in terms of both larger values and a second peak in the kernel density estimate. Although this contrast can also be seen in the loadings plot of females, it is not as strong as for the male block. The JC2 loadings plots show an interesting outlier, the babies of age zero. We speculate this shows an effect in improvement of post-natal care that coincidently happened around the same time.

Another interesting result comes from the studying first individual components (IC1) of males and females, shown in Figure 4.3. In the scores plot of males (left), the blue circles stand out from the rest, corresponding to the years of the Spanish civil war when a significant spike can be seen in the rows as curves plot. Young to middle age groups are affected more than the others which can be found in the loadings plot and columns as curves plot. Such year variation pattern, however, cannot be detected in the individual variation component of females. The columns as curves plot on the lower left suggest

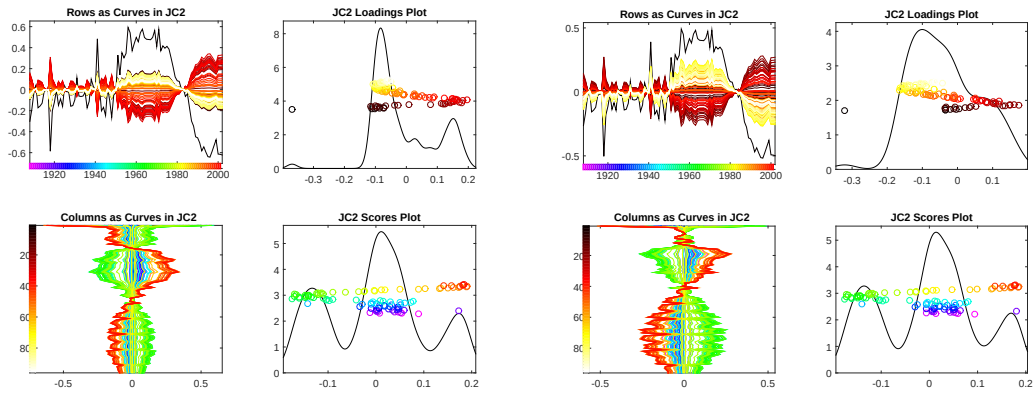


Figure 4.2: The second joint components of male (left) and female (right) contain the common modes of variation driven by the increase in fatalities caused by automobile penetration and later improvement due to safety improvements. This can be seen from the scores plots in the right bottom. The loadings plots show that this automobile event exerted a significantly stronger impact on the 20-45 males.

some type of 5-year age rounding effect, which is seen to occur mostly during the earlier years as indicated both in the rows as curves plot and the colors of the peaks in the columns as curves plot. Note that the plot scales show that the individual female effects are much smaller in magnitude than the male effects.

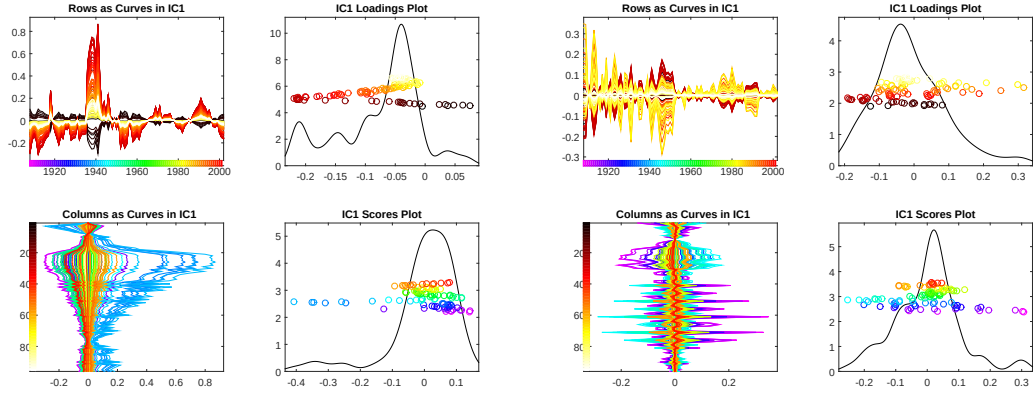


Figure 4.3: The individual component of male (left) contains the variation driven by Spanish civil war which can be seen from the blue circles on the right end of right bottom plot. The Spanish civil war mainly affected the young to middle age male.

4.2 TCGA Data Analysis

A prominent goal of modern cancer research, of which TCGA is a major resource, is the combination of biological insights from multiple types of measurements made on common subjects. JIVE is a powerful new tool for gaining such insights. Here

gene expression, copy number, reverse phase protein arrays and gene mutation features measured on a common set of 616 breast cancer samples are taken as an example. A preliminary analysis and visualization of each data type is given in Section 4.2.1 and the JIVE results are discussed in Section 4.2.2.

4.2.1 Visualization

Gene Expression

Gene expression (GE) is the process by which information from a gene is used in the synthesis of a functional gene product (e.g. protein). The measurement of expression is typically done by detecting *messenger RNA (mRNA)*. Studies of gene expression profile have revealed its power in predicting disease outcome and selecting therapies for individual patients (Van't Veer et al., 2002). For such a high dimensional data set (16615 features), PCA is applied to study important variation components. In Figure 4.4, the diagonal plots display the 1-dimensional distributions of the gene expression data block onto the first 4 principal component (PC) directions i.e. scores using the same format as in Figure 2.1. The off-diagonal plots show the 2-dimensional projections onto the subspaces generated by each pair of these 4 PC directions. Each symbol represents a patient and is colored by subtypes red for Luminal A, magenta for Luminal B, cyan for HER2 and blue for Basal-like.

From the diagonal plots, the first PC shows a clear subtype difference between Basal-like, HER2 and Luminal. The second PC presents a separation between Luminal A and the other subtypes. The third PC contains little subtype information. The fourth shows some but not strong evidence of separation between Basal-like and HER2. This strong connection between gene expression variations and class differences is mainly due to the fact that these subtypes are determined based on gene expression.

Copy Number Variation

The copy number variation (CN) are a form of structural variation of the two copies of a genome. It has been well known that differences in the DNA sequence of genomes have important impacts to personal traits. However, some recent studies have shown that copy number data also play an important role in characterizing individual risk of cancers and drug responses, e.g. Sebat et al. (2007); Xu et al. (2008). In Figure 4.5,

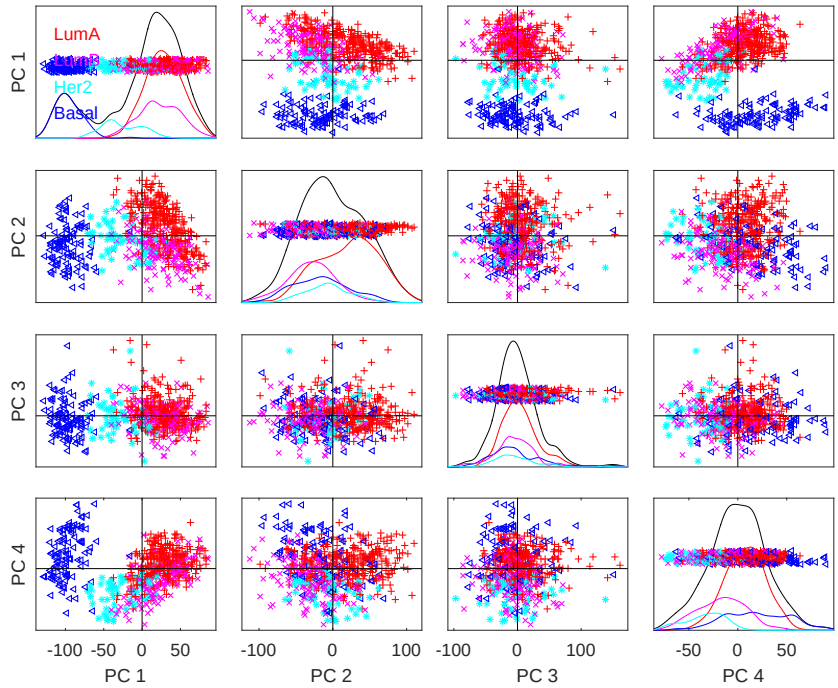


Figure 4.4: The first 4 PC projections of the gene expression data block. The first PC presents strong evidence of subtype differences between Basal-like, and the union of the others. The second PC shows a separation between Luminal A and the other subtypes.

the first 4 PC projections of the copy number data block are displayed similarly as in Figure 4.4. The first two largest variation PCs show little correlation with subtypes differences. The third PC presents strong evidence of differences between Basal-like and the other subtypes and the fourth PC shows some differences between Luminal A and the others.

Reverse Phase Protein Array

Reverse phase protein (micro)arrays (RPPA) is a new, sensitive, high-throughput technology for obtaining protein micro-arrays which provide quantitative profiling of disease associated proteins (Charboneau et al., 2002). A broad assessment of quantitative protein changes in diseased and healthy tissue can be offered by RPPA data. This protein profiling has the potential for detecting meaningful protein and pathway interactions of known proteins (Tibes et al., 2006). Figure 4.6 suggests that the first PC contains useful information for distinguishing the Basal-like from the Luminial B. The second PC presents the differences between Basal-like and Luminal which can be similarly found in the PCs of gene expression and copy number. The fourth displays

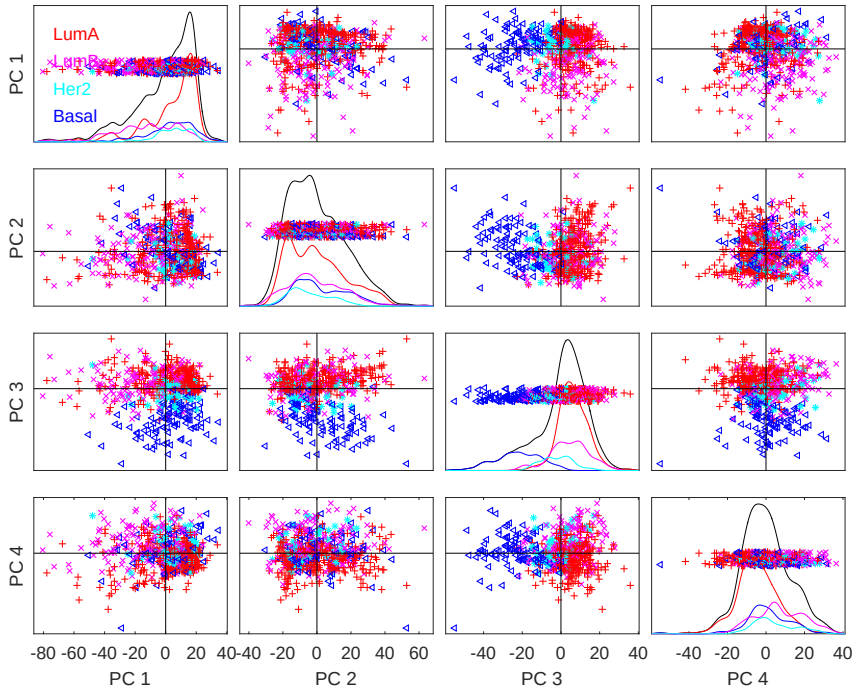


Figure 4.5: The first 4 PC projections of the copy number data block. The third PC presents a strong evidence of differences between Basal-like and the other subtypes and the fourth PC shows some differences between Luminal A and the others.

separation between HER2 and the other subtypes which is not clearly presented in the first 4 PCs of gene expression and copy number.

Gene Mutation

Gene mutation is a permanent alteration of the nucleotide sequence of the genome, which might result in different types of change in sequences and thus alter the product of a gene, or prevent the gene from functioning properly or completely. Mutations in certain genes, described as *high penetrance*, are often associated with high risk of developing some types of cancer e.g. breast cancer (Tung et al., 2015). Figure 4.7 visualizes the first 4 PC projections of the Mutation data block. It can be seen that each PC tends to be driven by several influential data points and is not useful for revealing their associations with subtype differences.

Such observation is due to the low frequency of mutations of most patients, which can be observed in the left panel of Figure 4.8. Each dot corresponds to one patient, with colors based on breast cancer subtypes as in Figure 4.7. Less than 10 patients carry more than 5% mutations among the 18256 genes and these patients are the samples

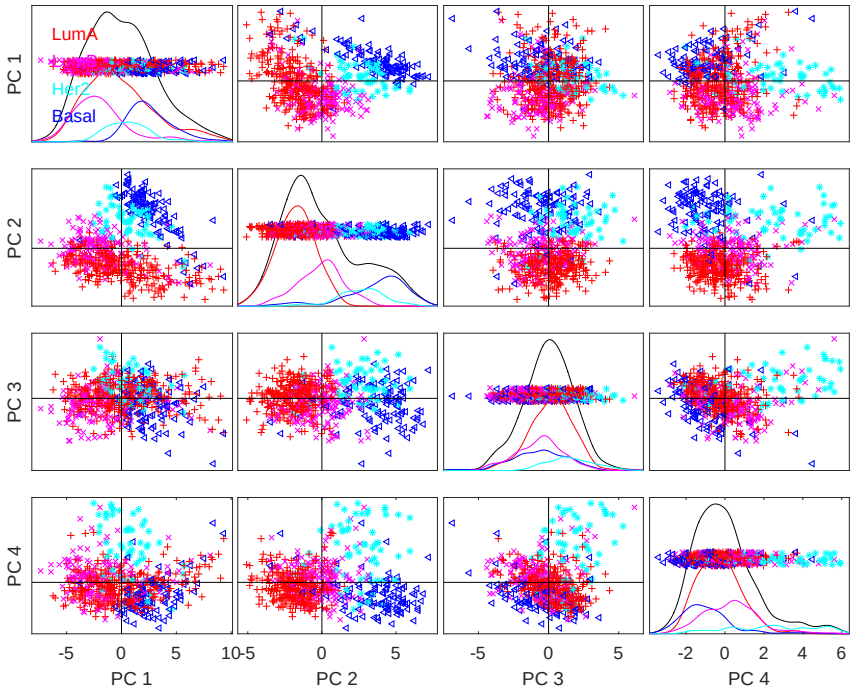


Figure 4.6: The first 4 PC projections of the RPPA data block. The second PC presents differences between Basal-like and Luminal. The fourth PC displays separation between Her2-like and the other subtypes.

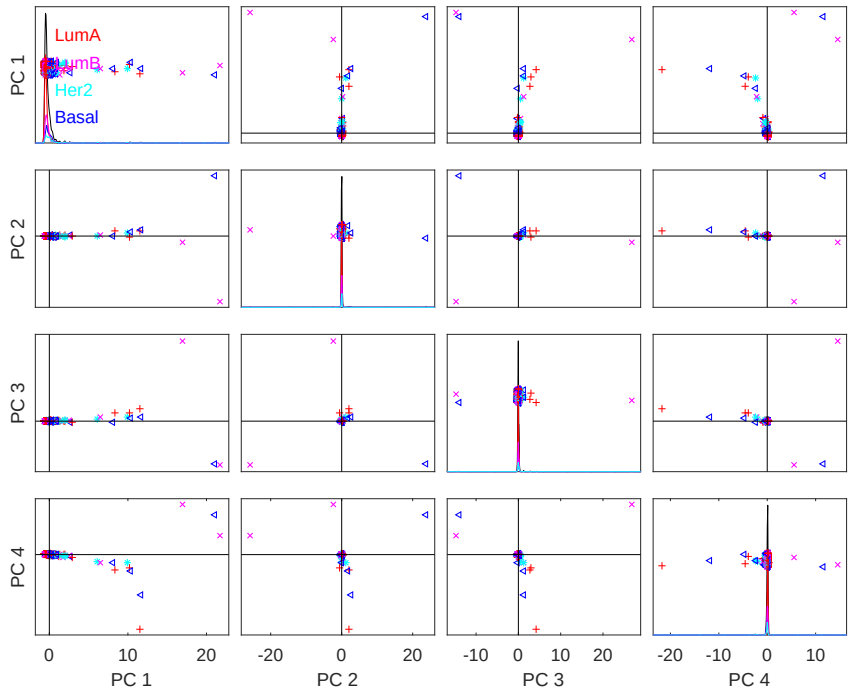


Figure 4.7: The first 4 PC projections of Mutation data block. Each PC is driven by several influential data points and is not useful for revealing their associations with subtype differences.

driving the largest 4 variations presented in Figure 4.7. To reveal additional insights of this gene mutation data, a bar plot in the right panel of Figure 4.8 presents the top 25 genes with highest chance of having mutations. The height of each bar represents the percentage of patients having a mutation in the corresponding gene with the name labeled in the bar. As can be seen, TP53 and PIK3CA are the major players which are known for greatly increasing the risk of developing breast cancer (Network et al., 2012).

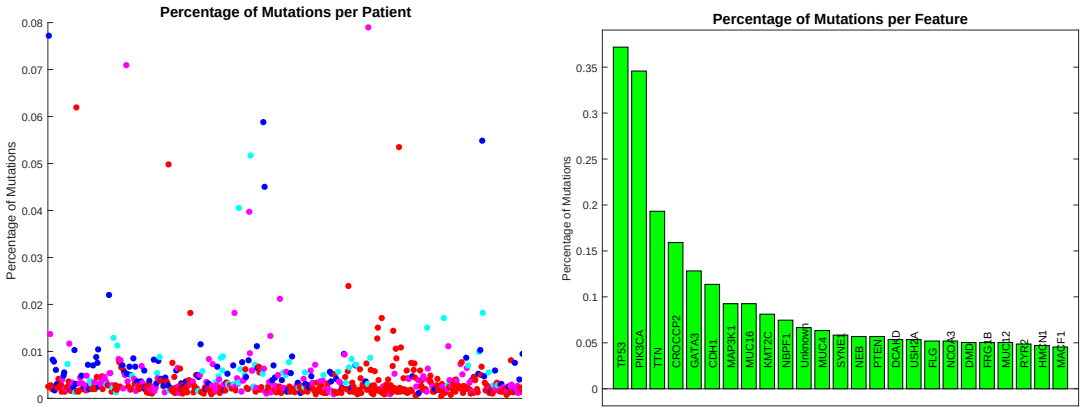
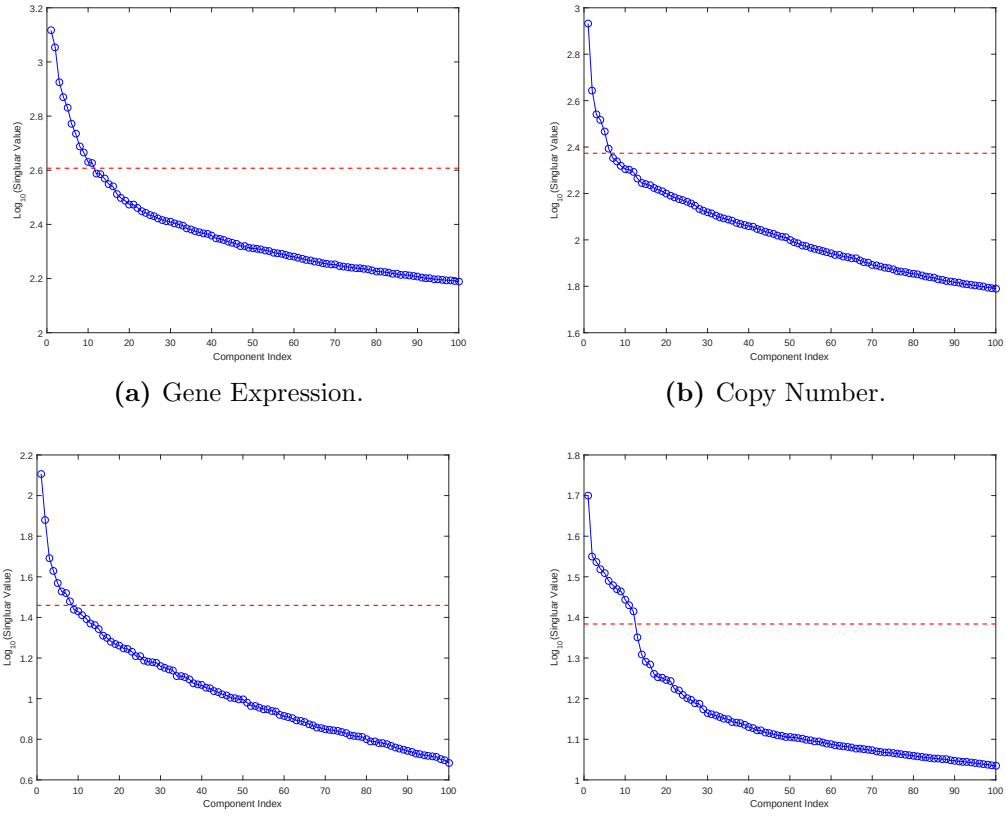


Figure 4.8: The left panel shows percentages of mutations among 18256 genes for each patient. The right panel presents the major set of genes having mutations.

4.2.2 Multi-Block JIVE Analysis

Here we perform our JIVE on gene expression, copy number, reverse phase protein arrays (RPPA) and gene mutation (Mutation) measured on a common set of 616 breast cancer samples. A most interpretable and insightful analysis is generated from low rank approximations of dimensions 11 (gene expression), 6 (copy number), 8 (RPPA) and 12 (Mutation) selected in the first step of JIVE. Figure 4.9 displays the scree plots with singular values in $\log_{10}$ scale for each data block. The red dashed lines indicate the thresholds for selecting initial signal components. These thresholds are mainly determined by the size of jump between adjacent singular values as discussed in Section 3.3.5.

Figure 4.10 presents the second JIVE step. The one sided 95% confidence interval suggests to select two joint components, resulting at least 95% to include all the joint signals. The threshold from the one-sided 5% confidence interval is 2.58 which is very close to the second squared singular value 2.61. For a conservative selection of joint components, this 5% threshold suggests to select only one joint component since the second component is very possible to be the combination of close individual components.



(a) Gene Expression. **(b)** Copy Number.

(c) RPPA. **(d)** Mutation.

Figure 4.9: The scree plots. The components with $\log_{10}$ of singular values above the dashed red line are selected as initial signal components in the first step of JIVE.

Both selections of joint rank, 1 and 2, are tried in the following reconstruction step. When using the joint rank 2, the second singular value of the reconstructed joint matrix of mutation is smaller than the singular value threshold selected in the step 1. This further suggests that one joint variation component should be extracted from these four data types.

The association between the common normalized scores (CNS) of this joint component and genetic subtype differences is visualized in Figure 4.11. For a better understanding of this variation pattern, the dots are a jitter plot of the patients, using colors and symbols to distinguish the subtypes (Blue for Basal-like, cyan for HER2, red for Luminal A and magenta for Luminal B). Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is the height based on data ordering. The curves are Gaussian kernel density estimates i.e. smoothed histograms, which show the distribution of the subtypes.

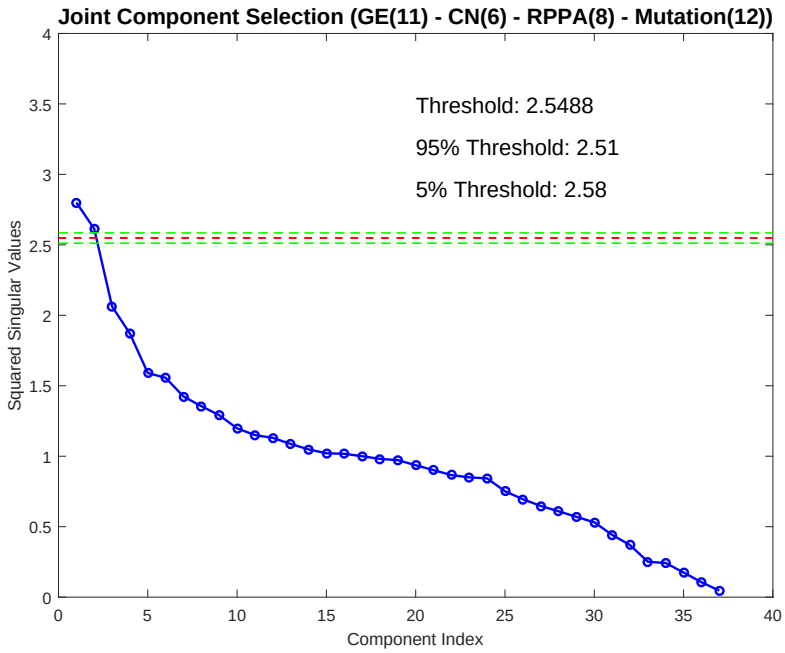


Figure 4.10: The second step of JIVE. The circles indicate the singular values from the second SVD. The dashed red line corresponds to the median estimate of the threshold and the dashed green lines indicate the thresholds from the 5% and 95% one sided confidence intervals. A conservative selection suggests one joint component across the four data types.

The clear separation among density estimates suggest that this joint variation component is strongly connected with the subtype difference between Luminal A versus the other subtypes. To quantify this subtype difference, a test is performed using the CNS

of this joint component evaluated by the DiProPerm hypothesis test (Wei et al., 2015) based on 100 permutations. Strength of the evidence is usually measured by permutation p-values. However, in this context empirical p-values are frequently zero. Thus a more interpretable measure of strength of the evidence is to provide DiProPerm z-score. This is 29.32 for this CNS. An area under the ROC curve (AUC) (Hanley and McNeil, 1982) of 0.915, is also obtained to reflect the classification accuracy. These numbers confirm the strong Luminal A property shared by these four data types.

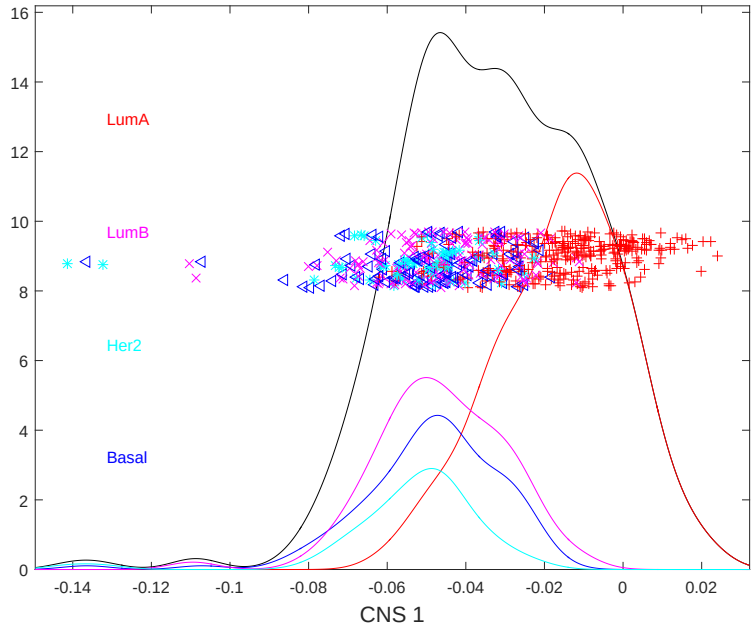
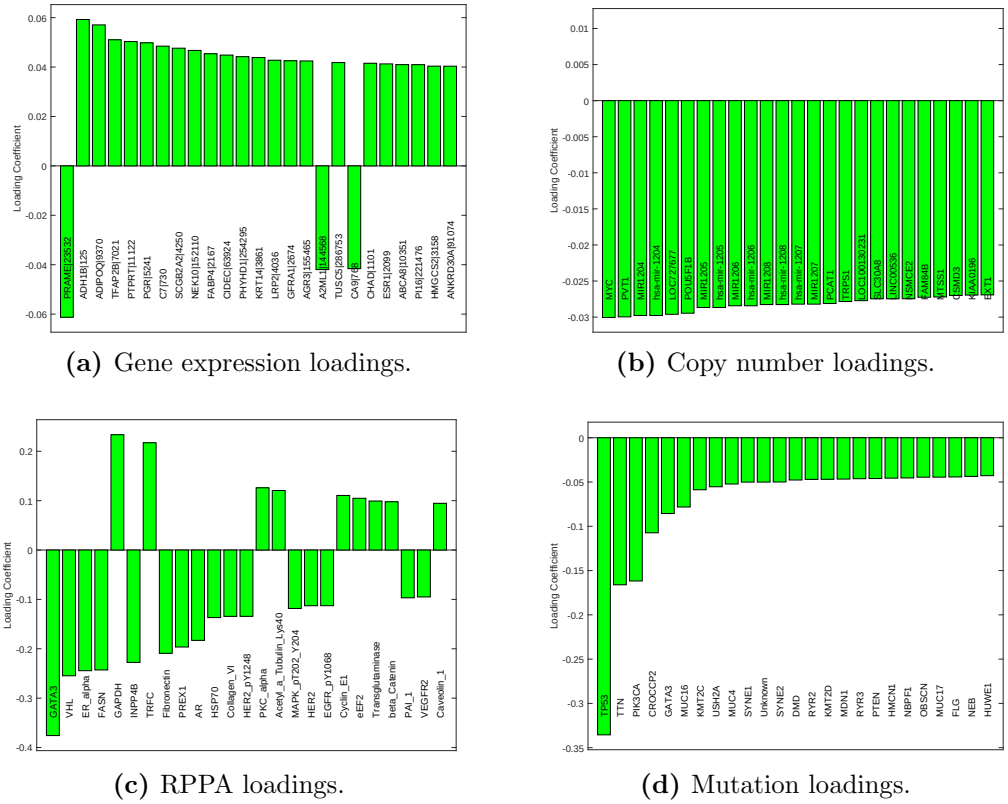


Figure 4.11: The kernel density estimation of the common normalized score (CNS) among gene expression, copy number, RPPA and mutation. The clear separation between Luminal A versus the other subtypes indicates that these four data blocks share a very strong Luminal A property captured in this joint variation component.

A further understanding can be obtained by identifying the feature set of each data type which jointly work with each other in characterizing the Luminal A property. Figure 4.12 presents the top 25 features with largest absolute loading coefficients for each data block. In each panel, each bar represents a feature with its name labeled accordingly. The length of a bar corresponds to the importance in driving this joint variation. A careful look at the sign of CNS vector shows that, a positive loading value indicates that the Luminal A subtype group tends to have a higher level of the corresponding variable than the others, while the negative loading value means the opposite.

The large mutation loading for TP53 is known from previous studies, as our TTN and PIK3CA. Similarly the dominants of GATA3 in RPPA is well known, and is connected with the large GATA3 mutation loading. The copy number loadings are nearly constant, which is related to the strong correlation in this data. A less well known result of this analysis is the genes appearing with large gene expression loadings. Many of these are not dominant in earlier studies, which had focused on subgroup separation, instead of joint behavior.



(a) Gene expression loadings.

(b) Copy number loadings.

(c) RPPA loadings.

(d) Mutation loadings.

Figure 4.12: Loadings plot of the joint common normalized score. Top 25 features with largest absolute loading coefficients are displayed for each data block.

Next step is to study the individual variation of each data block. Figure 4.13, Figure 4.14, Figure 4.15 and Figure 4.16 show the first 3 individual score vectors i.e. individual normalized scores (INSs), colored by subtypes.

Note that the individual variation of gene expression, copy number and RPPA present apparently common and subtype related variation for further analysis. In their figures, the diagonal plots are the INS distribution and the off-diagonal plots are the scatter plots of pairwise INSs. The first INS of gene expression data presents an apparent separation between Basal-like and the other subtypes. Such subtype differences

can also been found in the third INS of copy number and the first INS of RPPA. This suggests that these components may still contains common variation patterns which can be explained by some subtype differences e.g Basal-like versus the others. We explicitly investigate this potential 3–way joint structure using JIVE analysis on the 3 block concatenation of the individual variation matrices of gene expression, copy number and RPPA. The individual variation of Mutation, however, is mainly driven by several influential points and does not seem to contain apparent subtype properties as the other three blocks. Therefore the individual variation of Mutation is not considered for the further JIVE study.

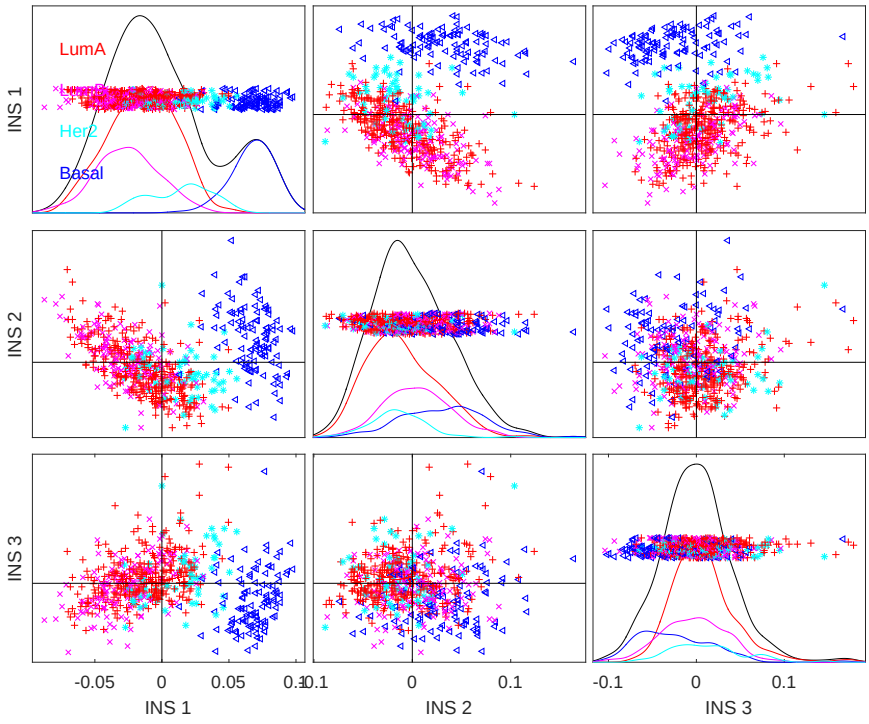


Figure 4.13: The first 3 individual normalized scores (INSs) of gene expression. The first INS presents the subtype differences between Basal-like and the others.

Such observation is also consistent with the joint rank selection shown in Figure fig:tcgamultijive:step2. The individual score spaces of gene expression, copy number and RPPA have components that are very close with each other; however, the individual score space of mutation does not have such components. Therefore, the second squared singular values is slightly above the 5% threshold as three data blocks have one more joint component; and the reconstruction step suggests that this component is not a joint signal for mutation.

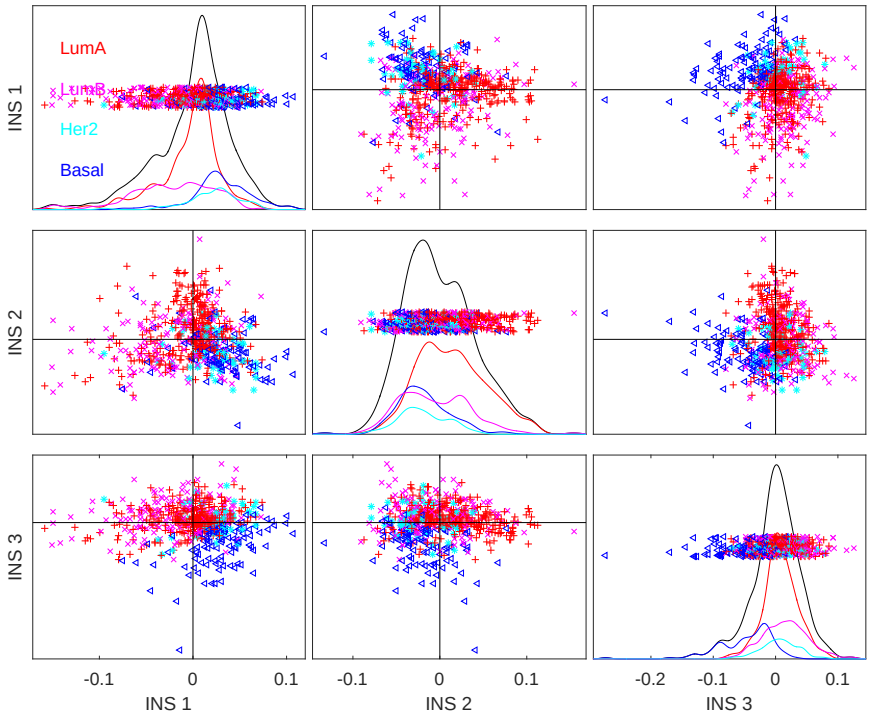


Figure 4.14: The first 3 individual normalized scores (INSs) of copy number. The third INS presents the subtype differences between Basal-like and the others.

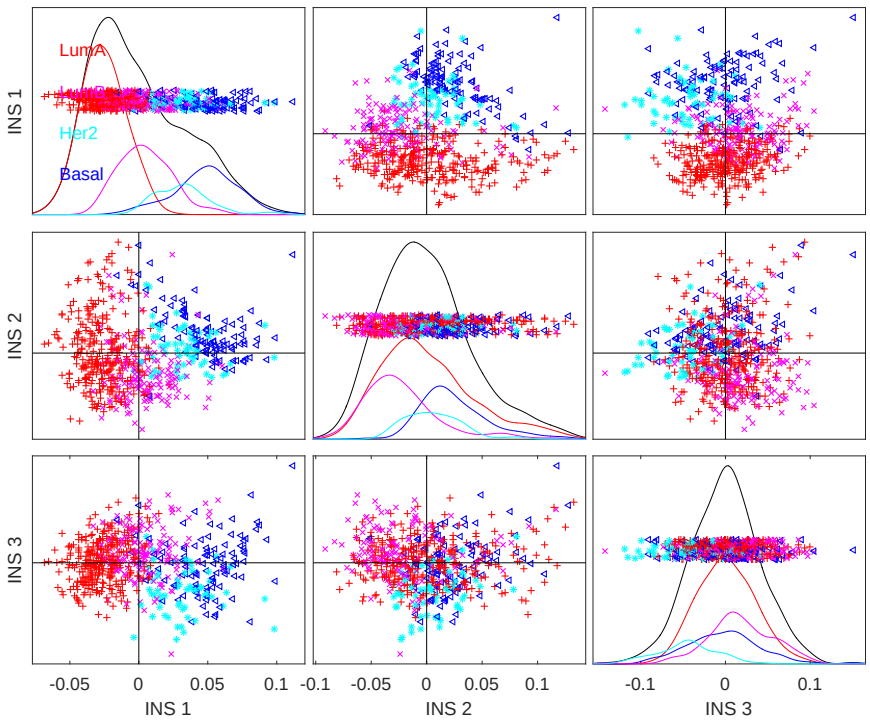


Figure 4.15: The first 3 individual normalized scores (INSs) of RPPA. The first INS presents the subtype differences between Basal-like and the others.

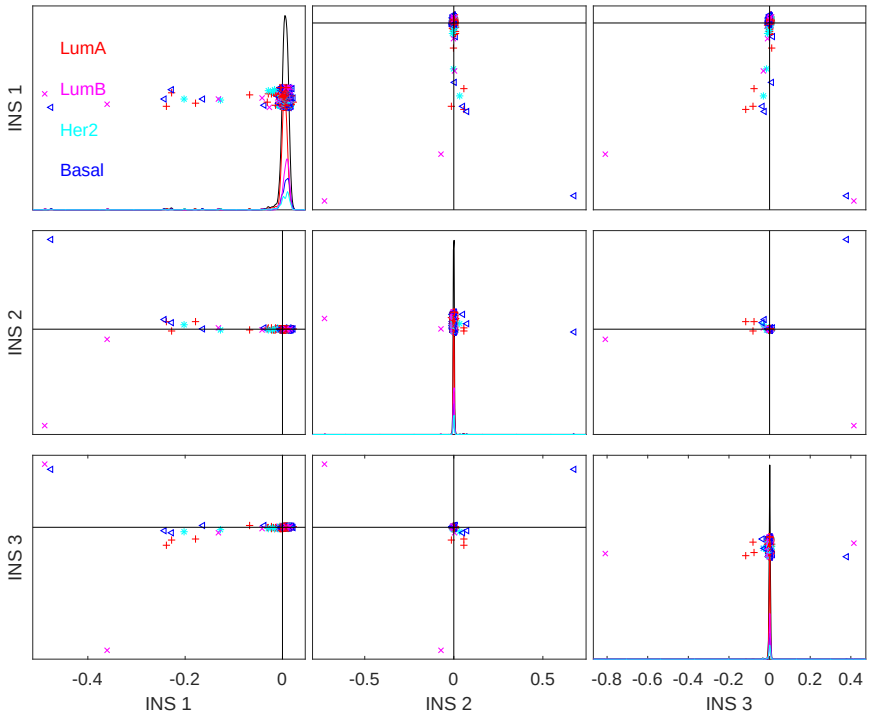


Figure 4.16: The first 3 individual normalized scores (INSs) of Mutation which are mainly driven by several influential points and seem not to be clear subtype–related variation.

Hierarchical JIVE Decomposition

The JIVE analysis on the gene expression, copy number and RPPA individual matrices results in one joint variation component displayed in Figure 4.17. This joint variation component clearly shows the differences among Basal, HER2 and Luminal subtypes. In particular, a subtype difference between Basal-like versus the others is quantified using DiProPerm z-score (29.82) and the AUC (0.998). Considering the fact that the AUC of the classification between Basal-like versus the others using all the original separate GE features is 0.999, this single joint component contains almost all the variation information for separating Basal-like from the others.

This hierarchical application of JIVE reveals an important joint component that is specific to gene expression, copy number and RPPA but not to Mutation. This analysis provides motivation for further extension of JIVE which will be discussed in Section 4.3.

4.2.3 Pairwise TCGA Data

Additional understandings can be obtained by pairwise analyzing these data types of which the analysis of gene expression and copy number data blocks is described here.

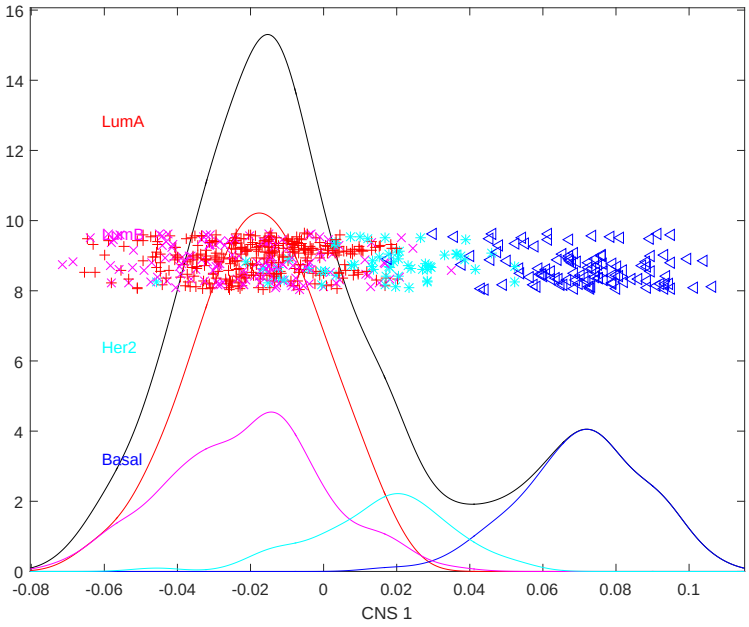


Figure 4.17: The common normalized score (CNS) from applying JIVE to the individual matrices of gene expression, copy number and RPPA. The clear separation between kernel density estimations indicates that the individual matrices of gene expression, copy number and RPPA contains a joint variation component explaining the subtype difference between Basal versus the others.

Genetic subtypes have proven to be fundamental to precision medicine, so insights about these will be used to interpret the variation contained in each data type and also the joint and individual variation extracted by JIVE. Based on the visualizations in Section 4.2.1, we perform JIVE for several selected subsets of the data for gaining more insights about subtype differences, which includes all tumors, HER2 and Luminal, and Luminal alone. Table 4.1 states the variation explained by JIVE decomposition for each subset. As shown in the table, most of the copy number variation (about 80%) is joint with gene expression for all of these subsets. On the other hand, the gene expression data contains a much larger percentage of individual variation (about 60%) that differs from copy number. This observation is consistent with expected biology because copy number variation tends to generate variation in gene expression, while there are many other sources of variation that also drive gene expression.

The classification directions between the studied two subtypes are obtained by Distance Weight Discrimination(DWD) (Marron et al., 2007) which is useful because of the high dimensional nature of these data. Class differences are quantified by DiProPerm hypothesis tests (Wei et al., 2015) based on 100 permutations. Strength of the evidence

Data source	Comparison	Joint	Individual
All Tumors	Gene expression	35%	65%
	Copy number	80%	20%
HER2 & Luminal	Gene expression	34%	66%
	Copy number	73%	27%
Luminal Only	Gene expression	41%	59%
	Copy number	81%	19%

Table 4.1: Percentage of variation explained by joint block specific score (BSS) structure, individual BSS structure for gene expression and copy number data. Shows that copy number variation mainly associates with gene expression, but gene expression is more diverse as expected.

is usually measured by permutation p-values. However, in this context most p-values are zero. Thus a more interpretable measure of strength of evidences is to provide DiProPerm z-scores. We also report the area under the ROC curve (AUC) (Hanley and McNeil, 1982), to show the classification accuracy.

Additional biological insights come from post analysis of these JIVE decompositions. Subtype differences are explored by performing classifications on both joint and individual variation. This was done using both the common normalized score/individual normalized score and the block specific score (BSS) data representations. Results are similar so only common normalized score results are shown here. This gives a straightforward joint analysis because it is based on the common set of joint scores. The classification directions are obtained by Distance Weight Discrimination(DWD) (Marron et al., 2007) which is useful because of the high dimensional nature of these data. Class differences are quantified by DiProPerm hypothesis tests based on 100 permutations. Strength of the evidence is measured by DiProPerm z-scores together with permutation p-values and AUC for showing the classification accuracy.

Figure 4.18 presents the results of classification analysis of joint variation within the three data subsets. Each panel shows a separation of subtypes by projecting the common normalized score of joint structure onto the DWD discrimination direction. The dots are a jitter plot of the data, using colors and symbols to distinguish the subtypes. Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is the height based on data ordering. The curves are Gaussian kernel density estimates i.e. smoothed histograms, which show the distribution of the subtypes.

The left plot of Figure 4.18 presents a clear visual separation between Basal-like and other tumor subtypes. The high value z-score of 37.6 and AUC also suggest a strongly

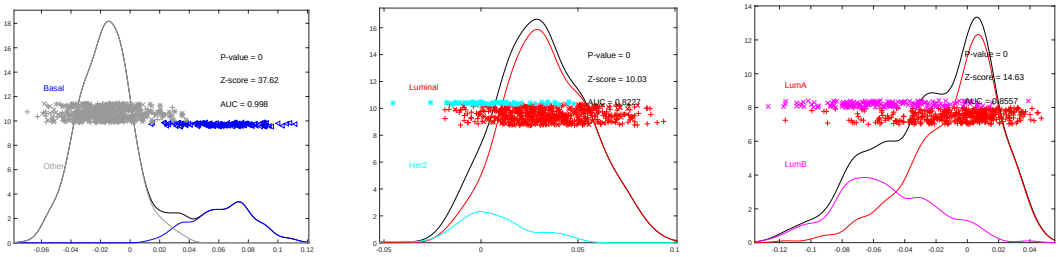


Figure 4.18: One dimensional projection of joint structures onto the DWD discriminant direction. Basal-like vs. the other tumor subtypes (left), HER2 vs. Luminal(center) and Luminal A vs. B (right). A strong separation is apparent between Basal and the other tumor subtypes, while there is more overlap for the other two classifications. This contrast indicates different discriminatory power of joint variations between these different subsets of gene expression and copy number.

significant class difference. The middle plot visualizes the discrimination between HER2 and Luminal. Although the z-score of 10.0 from DiProPerm indicates a significant difference, the visual separation is not as large. The separation between Luminal A and B, depicted in the right plot, is still not as strong as the Basal-like vs. the other tumor subtypes but has stronger evidence than HER2 and Luminal suggested by the DiProPerm z-score of 14.6. The contrast of separations in Figure 4.18 indicates the distinct discriminant powers of the joint signals within different data subsets. The joint signal between gene expression and copy number shows strong power for distinguishing Basal-like from the other tumor subtypes but is not quite as powerful for the other two class comparisons. This contrast is consistent with the known biological fact that the Basal-like subtype has much stronger copy number variations than the Her2 subtype.

A similar study is conducted for the individual variation within gene expression and copy number, which reveals a contrast with the joint variation. Table 4.2 gives the DiProPerm z-scores and AUCs for the individual normalized scores of each individual variation. Differing from the joint variation, the individual variations within copy number do not have power for distinguishing Basal-like from Other tumors, and Luminal A from B. The table shows that the DiProPerm z-scores are not significant and the AUCs are almost equivalent to random guessing(around 0.5). For these two class comparisons, the individual variation within gene expression still present substantial discriminant power but much weaker than the joint variation. The insignificant separation of individual variation within copy number and the dramatic decrease in discriminant power of individual variation within gene expression suggest that the class differences

are mostly explained by the joint variation between gene expression and copy number. Besides, in view of the fact shown in Table 4.1 that gene expression contains a large proportion of individual variation, this is a strong indicator that the individual structure of gene expression may be driven by some additional biological components. A further investigation could be a clustering analysis of these individual variations to identity new subtypes which might lead to better treatments.

The discrimination between HER2 and Luminal tells a different story. The individual variations within both gene expression and copy number present significant discriminatory power; in particular, the individual gene expression has an even better classification than its joint variation. This suggests that copy number features may not work jointly with gene expression features to distinguish HER2 and Luminal.

Data source	Data Type	Z-score (P-value)	AUC
All Tumors	Gene expression	9.61 (0)	0.7829
	Copy number	1.1 (0.145)	0.5663
HER2 & Luminal	Gene expression	20.64 (0)	0.9643
	Copy number	9.37 (0)	0.7551
Luminal Only	Gene expression	11.77 (0)	0.8052
	Copy number	0.67 (0.267)	0.5704

Table 4.2: Z-scores and AUC of individual structure in classifying different pairwise classes. Except HER2 versus Luminal, the other two comparisons indicate a less significant discrimination in the individual variation.

A further understanding of these genomic sources can be obtained by looking at the loading plots given by each classification. In particular, we have identified a set of gene expression features associated with a set of copy number features that work together to separate the compared classes.

4.3 JIVE Discussion

Our proposed Non-iterative JIVE method targets decompositions for each data block result in two types of variation structure, joint over all data blocks and individual, as defined in Section 3.3.1. Besides extracting the joint structure across all data blocks together, additional insights may come from another type of decomposition based on the joint variation structure within different subsets of the data blocks. For instance, in the TCGA analysis, after applying JIVE on the four data blocks and extracting the joint

variation, the individual variation matrices of gene expression, copy number and RPPA still contain joint variation within that subset, which revealed by another hierarchical JIVE decomposition. Figure 4.19 illustrates this for a three block data set.

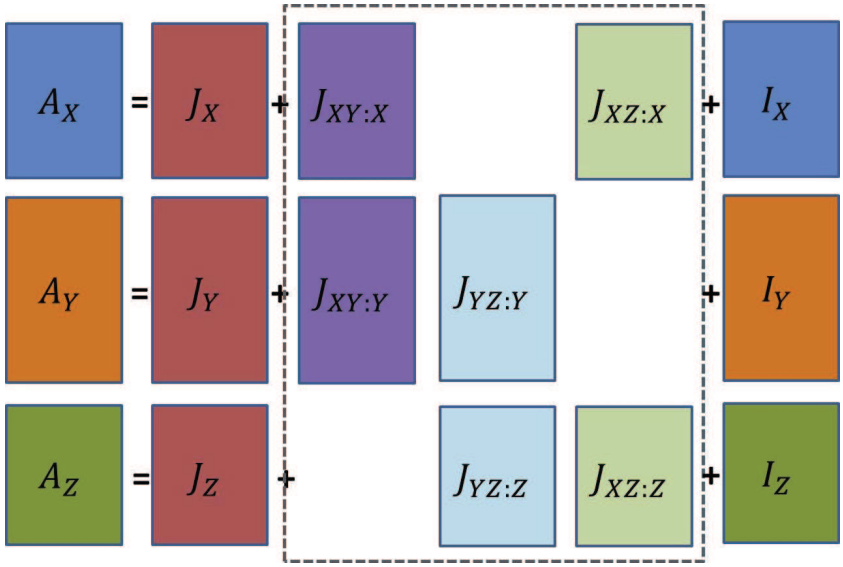


Figure 4.19: A three-block data set example shows a generalized Non-iterative JIVE decomposition. Besides the joint variation structure across all the three data blocks, the joint variation structures of each pair of the blocks are taken into consideration, depicted within the dashed rectangle.

Figure 4.19 starts with the same signal matrices A_X, A_Y, A_Z. As before there are joint components J_X, J_Y, J_Z and fully individual components I_X, I_Y, I_Z. In addition there are 2–block pairwise joint components. In the current JIVE, these are all treated as part of the individual components. This more generalized signal decomposition model gives more insights into the data in the spirit of that discovered in the three block TCGA analysis above.

CHAPTER 5: FUSION LEARNING FOR INTERLABORATORY COMPARISON

5.1 Introduction

As noted in Hannig et al. (2015b). Interlaboratory trials are often conducted by leading metrology laboratories in the world to compare each others' capabilities for measuring various fundamental properties of substances. Such a trial typically involves two or more participants each of whom measures the (nominally) same unknown value (called *measurand*) and provides the result along with an assessment of the uncertainty in the result. The results are meant to be the best estimates of the measurand the participating laboratories are able to provide. Often the same or very similar protocols are used by the participating laboratories. In some cases different subsets of participants use different methods for measuring the same unknown quantity. This is particularly so when specific laboratories have special expertise in particular measurement methods. The results from such experiments are used to gauge how comparable the measurement capabilities are across the participating laboratories. In some cases such experiments lead to the creation of certified reference materials (CRMs) and a consensus value for the measurand is arrived at by combining the results from the participating laboratories. This consensus value is used as the certified value for the CRM. The uncertainty associated with this certified value is used to provide an interval estimate of the value for the CRM.

Key Comparisons

There is a particular class of interlaboratory trials which takes on international significance. With the signing of the Mutual Recognition Arrangement (MRA) (CIPM, 1999) in 1999, National Metrology Institutes (NMI's) and Regional Metrology Organizations (RMO's) around the world have undertaken the task of examining the *degree of equivalence* of their measurement standards. The CIPM (*Comité international des poids et mesures* – The International Committee on Weights and Measures), an en-

tity whose principal task is to promote world-wide uniformity in units of measurement, works with member countries on issues related to the creation of measurement standards and comparisons of measurement capabilities of various national metrological laboratories (such as the National Institute of Standards and Technology (NIST) in the U.S, the National Physical Laboratory (NPL) in Great Britain, and Physikalisch-Technische Bundesanstalt (PTB) in Germany), and oversees the conduct of interlaboratory experiments by participating NMIs to evaluate the relative measurement capabilities of each other and also to establish standard reference values (called Key Comparison Reference Value(s) or KCRV) for many important fundamental measurements and standards. The results obtained by the different laboratories are combined to arrive at the consensus KCRV value. Such comparisons *provide for the mutual recognition of calibration and measurement certificates issued by NMIs and thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.*

During any interlaboratory trial it is generally the case that the results from one or a few laboratories differ noticeably from the rest even though all participating laboratories are considered to be more or less equally competent. It is natural to think that these apparently nonconforming values should perhaps be excluded from the calculation of a consensus value. There are at least two problems with this thinking. First, since the true value of the measurand is not known, one cannot say, based on any objective evidence, that one result is more believable than another. Second, there are political overtones associated with leaving out measured results of a laboratory since all participating laboratories are considered to be competent in their own right. Although discrepant results are subjected to further scrutiny to make sure such discrepancies are not the result of identifiable errors, when no errors are identified, each laboratory stands behind its result and the associated uncertainty. Hence the problem of arriving at a consensus value takes on a greater level of significance when it comes to International Key Comparison Studies.

Gauge Blocks

A gauge block (Thalmann, 2002) is a length standard having flat and parallel opposing surfaces. The cross-sectional shape is not very important, although the stan-

dard does give suggested dimensions for rectangular, square and circular cross-sections. Gauge blocks have nominal lengths defined in either the metric system (millimeters) or in the English system (1 inch = 25.4 mm). The length of the gauge block is defined at standard reference conditions:

temperature = 20 °C (68 °F)

barometric pressure = 101,325 Pa (1 atmosphere)

water vapor pressure = 1,333 Pa (10 mm of mercury)

CO2 content of air = 0.03%.

The length of a gauge block is defined as the perpendicular distance from a gauging point on one end of the block to an auxiliary true plane wrung to the other end of the block, as shown in Figure 5.1.

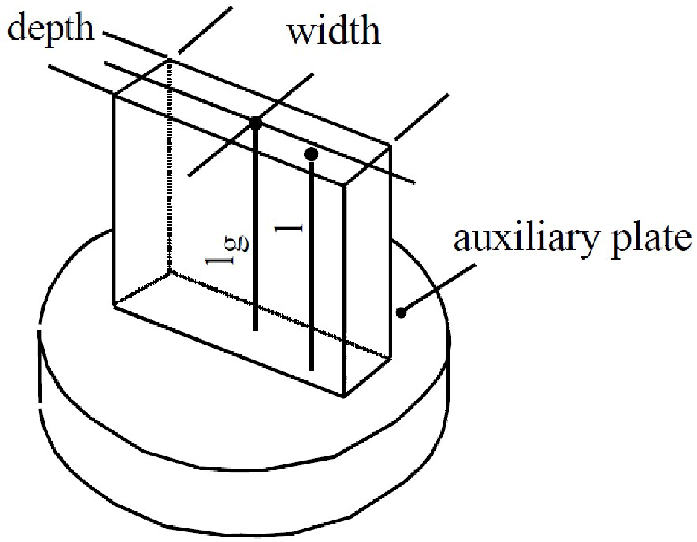


Figure 5.1: The length of a gauge block is the distance from the gauging point on the top surface to the plane of the platen adjacent to the wrung gauge block.

Figure 5.2 shows a portion of the results from an international key comparison study (CCL-K1) involving the measurement of the central length of steel gauge blocks (nominal length 1.01 mm) using interferometry according to ISO 3650. Detailed results are available from the website of the International Bureau of Weights & Measures (BIPM). The URL for the website is `http://kcdb.bipm.org/`. For instance, one can see, given the reported uncertainties, VNIIM (D.I. Mendeleev All-Russian Institute for Metrology) appears to deviate the most from the rest of the measurements.

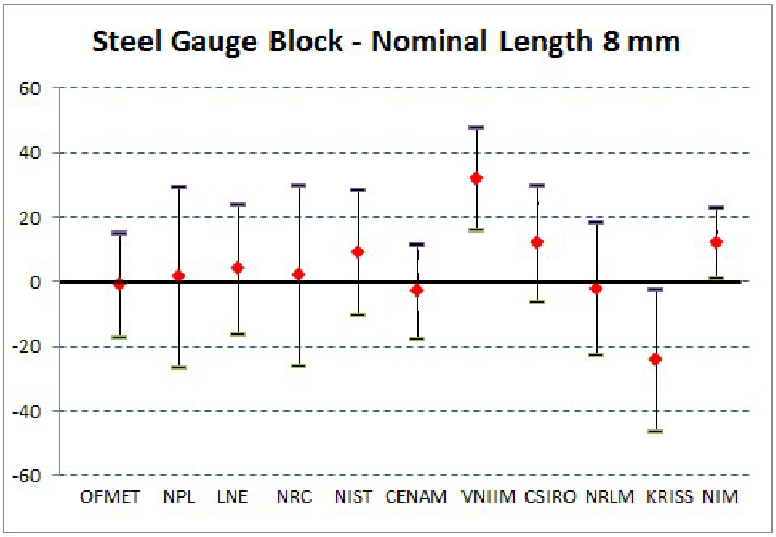


Figure 5.2: Gauge Block Measurements by 11 National Metrological Laboratories. Nominal length is 8 mm. The horizontal axis shows deviations (in nm) from the nominal value.

One of the issues that needs to be resolved is "how to treat this apparent outlier?" Alternatively, how much weight should be given to this particular measurement if one were using a weighted average approach to arrive at the KCRV?

The Key Comparison Study also considered gauge blocks of other nominal lengths besides the 8 mm gauge block. The estimates and uncertainties for the full set of steel gauge blocks for the 11 NMIs is given in Table 5.1. The entire array of issues related to this problem is more involved than what we are able to present here.

Although the potential of Fiducial methods in this area has been investigated in the literature (Iyer et al., 2004b,a), a systematic and thorough treatment of Robust Fiducial Methods has not been carried out. In this Chapter we propose the use of a generalized fiducial model averaging approach to finding a robust consensus value.

When combining information in the labs together we use *fusion learning* techniques (CD combination techniques) based on the Generalized Fiducial Inference Ideas of (Hannig and Xie, 2012). This is described in Section 5.3. A highly computationally efficient algorithm for model averaging is presented in Section 5.3.1. We show good small sample properties of the proposed method in Section 5.4. Finally, we demonstrate the new technique on the steel gauge block data and measurements of Newton's constant of gravitation (G) in Section 5.6.

Lab	Nominal Lengths (in mm)								
	0.5	1.01	6	7	8	15	80	90	100
OFMET	17 ± 9	34 ± 9	52 ± 8	31 ± 8	-1 ± 8	16 ± 8	22 ± 11	-21 ± 12	-96 ± 13
NPL	20 ± 14	25.5 ± 14	54.5 ± 14	33.5 ± 14	1.5 ± 14	22.5 ± 15	38.5 ± 28	-14 ± 31	-140 ± 33
LNE	15 ± 10	25 ± 10	54 ± 10	35 ± 10	4 ± 10	20 ± 10	28 ± 14	-24 ± 15	-110 ± 16
NRC	29 ± 13	28 ± 13	36 ± 14	30 ± 14	2 ± 14	14 ± 14	9 ± 21	-37 ± 22	-126 ± 24
NIST	26 ± 8.9	42 ± 9	57 ± 9.4	34 ± 9.5	9 ± 9.6	30 ± 10.3	33 ± 16.1	-23 ± 17	-117 ± 17.9
CENAM	15 ± 7	20 ± 7	47 ± 7.1	26 ± 7.1	-3 ± 7.2	13 ± 7.4	21 ± 15.6	-19 ± 17.3	-119 ± 18.7
VNIIM	*	60 ± 8	68 ± 8	25 ± 8	32 ± 8	36 ± 12	25 ± 14	-32 ± 15 -	104 ±
CSIRO	28 ± 9	46 ± 9	53 ± 9	37 ± 9	12 ± 9	51 ± 9	27 ± 14	-20 ± 15	-114 ± 16
NRLM	23.9 ± 8.6	17.7 ± 10.3	44.1 ± 10.3	27 ± 8.7	-2.2 ± 10.3	15.1 ± 10.9	47.3 ± 13.5	9.1 ± 14.3	-89.4 ± 16.3
KRISS	18.7 ± 13.1	20.3 ± 12.2	22.1 ± 13.6	12.8 ± 11	-24.2 ± 11	8.1 ± 13.2	30.4 ± 17	-18.4 ± 18.9	-104.3 ± 20.6
NIM	30 ± 5.4	48 ± 5.4	56 ± 5.5	42 ± 5.5	12 ± 5.5	28 ± 5.6	44 ± 8.9	18 ± 9.6	-90 ± 10.3

Table 5.1: CCL-K1 Measured results by 11 NMIs and combined standard uncertainties for steel gauge blocks for 9 different nominal lengths. The nominal lengths are in millimeters (mm). The values shown in the table are deviations from the nominal values (in nm) plus or minus the combined standard uncertainty (also in nm)

5.2 Background

5.2.1 Generalized Fiducial Inference

Fiducial inference was originally proposed by R.A. Fisher in 1930 to address the need to select a prior when none is available. Concepts of fiducial inference was never fully accepted by other statisticians. Since 2000, there has been a resurgence of fiducial inspired approaches (Berger and Bernardo, 1992; Martin and Liu, 2013; Xie and Singh, 2013; Hannig et al., 2015c). One of these approaches Hannig et al. (2015c) was termed *generalized fiducial inference* (GFI).

The key idea of GFI is to define a data dependent measure on the parameter space without the use of Bayes theorem; this data dependent distribution is called the *generalized fiducial distribution* (GFD). GFD can be used to derive approximate confidence intervals for parameters of interest. The transference of randomness from the model space to the parameter space is done by an inverse of a deterministic *data generating equation* (DGE), also known as the *structural equation*. Take a simple example for illustration. Define a random variable $Y \sim N(\mu, 1)$ i.e. a normal distribution with unknown mean parameter μ and standard deviation 1. The random variable W can be written as $Y = \mu + Z$ where Z follows a standard normal distribution. This equation is the DGE in this context. If a realization of Y say y is given e.g. $y = 5$, we can solve the equation to get μ expressed as $\mu = y - Z$ which defines a distribution called the *fiducial distribution* of the parameter μ. The known distribution of Z is used to deduce the distribution of mu through the inverse of the DGE.

The DGE, expressing the relationship between data Y and the parameters $\boldsymbol{\theta}$, is generally represented as

$$Y = G(U, \boldsymbol{\theta}), \tag{5.1}$$

where G is a deterministic function, and U represents the random component with completely known and independent of parameter distributions. Examples include the random variable Z in the simple example above.

After observing a realization y, the inverse of the data generating equation can be defined in the parameter space as

$$Q_y(U) = \{\boldsymbol{\theta} : \ G(U, \boldsymbol{\theta}) = y\}. \tag{5.2}$$

The inverse image $Q_y(U)$ induces to a distribution on $\boldsymbol{\theta}$ from the randomness of U. However, the distribution might be ill-defined for two possible reasons (Hannig, 2013). One is that there are multiple solutions of $\boldsymbol{\theta}$ which might be caused by discrete distributions. The other is that there may be no solution satisfying the equation. One remedy for this is to remove the realizations $U^\star$ resulting in non-existence of a solution. A conditional distribution

$$Q_y(U^\star) \mid \{Q_y(U^\star) \neq \emptyset\}$$

is introduced as a result of this adjustment. However, when the condition has a probability zero, the conditional probability is not well defined due to the non-uniqueness known as the Borel paradox. In response to this, Hannig et al. (2015c) proposes an attractive definition of generalized fiducial distribution via the limit of discretization of the conditional distribution.

Definition 5.1. *A probability measure on the parameter space* Θ *is called a* generalized fiducial distribution *(GFD) if it can be obtained as a weak limit*

$$\lim_{\epsilon \to 0} \left\{ \underset{\boldsymbol{\theta}^\star}{\operatorname{argmin}} \|y - G(U^\star, \boldsymbol{\theta}^\star)\| \ \middle| \ \min_{\boldsymbol{\theta}^\star} \|y - G(U^\star, \boldsymbol{\theta}^\star)\| \leq \epsilon \right\} \tag{5.3}$$

If there are multiple choices of $\underset{\boldsymbol{\theta}^\star}{\operatorname{argmin}} \|y - G(U^\star, \boldsymbol{\theta}^\star)\|$*, one of them is potentially selected at random.*

With this definition, a closed form of the limit in 5.3 was derived by Hannig (2013) under the l^∞ norm and a generalized result is provided in Hannig et al. (2015c), which is applicable to many practical situations.

Theorem 5.2. *Assume that the parameter $\boldsymbol{\theta} \in \Theta$ is p-dimensional, the data y are n-dimensional. Suppose the assumptions A.1 to A.3 in Appendix A of Hannig et al. (2015c), the limiting distribution in 5.3 has a density*

$$r(\boldsymbol{\theta}|y) = \frac{f(y,\boldsymbol{\theta})J(y,\boldsymbol{\theta})}{\int_\Theta f(y,\boldsymbol{\theta}')J(y,\boldsymbol{\theta}')d\boldsymbol{\theta}'} \tag{5.4}$$

where $f(y,\theta)$ is the likelihood and the function

$$J(y,\boldsymbol{\theta}) = D\big(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}G(U,\boldsymbol{\theta})|_{\mathbf{u}=G^{-1}(y,\boldsymbol{\theta})}\big) \tag{5.5}$$

1. *If $n = p$ then $D(A) = |\det A|$. Otherwise the function $D(.)$ depends on the norm that is used;*
2. *The l^∞ norm makes $D(A) = \sum_{i=(i_1,\ldots,i_p)} |\det(A)_i|$ where the sum spans over $\binom{n}{p}$ p-tuples of indices $i = (1 \le i_1, < \cdots < i_p \le n)$;*
3. *Under an additional assumption A.4, the l^2 norm gives $D(A) = (\det A^T A)^{1/2}$*

GFI has increasingly attracted interest and has been demonstrated to be inferentially meaningful in many practical applications without the need for subjective prior information. See Hannig et al. (2003), Iyer et al. (2004c), McNally et al. (2003), Wang et al. (2012), Wang and Iyer (2005) for examples. This chapter focuses on one practical usage of GFI; the fusion learning of key comparisons.

5.2.2 Confidence Distributions

A Confidence Distribution (CD) is a way to summarize information about a parameter contained in the data. It is similar to a Bayes posterior distribution but is grounded in frequentist methodology. Heuristically speaking, the CD function is obtained by stacking up one-sided confidence intervals of all levels. Schweder and Hjort (2002); Singh et al. (2005) provide the following formal definition of a CD function

Definition 5.3. *A function $H(\cdot|\mathbf{x})$ on $\Theta \times \mathcal{X} \to [0,1]$ is called a confidence distribution (CD) for a parameter θ, if it follows two requirements:*

R1) For each given $\mathbf{X} \in \mathcal{X}$, $H(\cdot)$ is a continuous cumulative distribution function on Θ;

R2) At the true parameter value $\theta = \theta_0$, $H(\theta_0) \equiv H(\mathbf{X}, \theta_0)$, as a function of the sample $\mathbf{X}$, follows the uniform distribution $U[0,1]$.

Also, the function $H(\cdot)$ is an asymptotic CD (aCD), if the $U[0,1]$ requirement is true only asymptotically (as sample size goes to infinity) and the continuity requirement on $H(\cdot)$ is dropped.

In general, fiducial distributions, objective Bayes distributions, inversions of one sided confidence intervals are all examples of CDs. As a particular example consider a sample of size n from a $N(\theta, \sigma^2)$ distribution with sample mean $\bar{x}$ and sample standard deviation s. The corresponding CD is the location-scale t distribution with distribution function

$$H(\theta|\mathbf{x}) = F_{n-1}^t \left(\frac{\theta - \bar{x}}{s/\sqrt{n}} \right),$$

where F_{n-1}^t is the distribution function of the Student's t distribution with $n-1$ degrees of freedom.

A useful graphical tool for visualizing a confidence distribution is a confidence curve (Birnbaum, 1961). For a given confidence distribution $H(\theta, \mathbf{x})$, its corresponding confidence curve is defined as $CV(\theta) = 2|H(\theta, \mathbf{x}) - 0.5|$. On a plot of $CV(\theta)$ versus θ, a line across the height (y-axis) of α, for any $0 < \alpha < 1$, intersects with the confidence curve at two points, and these two points correspond (on the x-axis) to a α level, equal tailed, two sided confidence interval for θ. Thus, a confidence curve is a graphical device that shows confidence intervals of all levels; see, e.g. Birnbaum (1961); Bender et al. (2005). The minimum of a confidence curve is the median of the confidence distribution. It provides a point estimator which is typically median unbiased (Birnbaum, 1961). Figure 5.3 shows an example of a confidence curve.

5.3 Method

Let us assume that there are K labs and lab i measures the object n_i times, $i = 1, \ldots, K$, and reports the mean, X_i, of these n measurements. We assume that the data

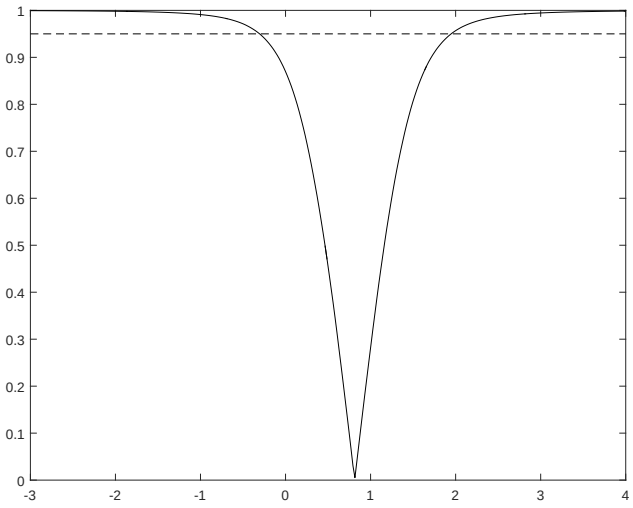


Figure 5.3: Confidence curve for the mean of the normal distribution based on a sample of size 8 with mean $\bar{x} = 1.1$ and sample standard deviation $s = 2.1$. The interval between the two points where the dotted line intersects the CD is the 95% confidence interval.

generating equation for these measurements is

$$X_i = \mu + B_i + \frac{\sigma_{A_i}}{\sqrt{n_i}} Z_i, \quad i = 1, \ldots, K \tag{5.6}$$

Here μ is the true value of the measurand, ${n_i}^{-1/2}\sigma_{A_i} Z_i$ are measurement errors assumed to have $N(0, \sigma^2_{A_i}/n_i)$ distribution and B_i are lab specific unknown systematic errors. The B_i cannot be measured directly. However it is assumed that there is some prior information available for it. This prior information often differs significantly from lab to lab, so modeling it as a random effect with a common distribution across labs is not appropriate.

Typically, B_i is modeled as a random variable with zero mean and known standard deviation $\sigma_{B,i}$ often referred to as type-B uncertainty. Hence, the variance of X_i, denoted as $\sigma^2_{C_i}$, is given by

$$\sigma^2_{C_i} = \sigma^2_{B_i} + \frac{\sigma^2_{A_i}}{n_i}.$$

The inferences for μ are performed separately by each lab and reported as the triple (x_i, u_i, d_i) where x_i is the realized value of X_i, u_i is an estimate of σ_{C_i} and d_i is an *effective degrees of freedom* associated with u_i. The quantity u_i is called the *combined standard uncertainty* (GUM, 1995). The value of d_i is generally determined

using the Satterthwaite approximation (Satterthwaite, 1946) for a linear combination of independent χ^2 random variables.

In particular, the estimate of the combined variance $\sigma^2_{C_i}$ is

$$u_i^2 = \sigma^2_{B_i} + \frac{s_i^2}{n_i}$$

where s_i is the sample standard deviation of the n_i observations from lab i whose mean is X_i. It is assumed that $d_i u_i^2/\sigma^2_{C_i}$ is distributed (approximately) as a χ^2 random variable with d_i degrees of freedom. The approximate combined degrees of freedom

$$d_i = (n_i - 1)\frac{u_i^4}{s_i^4/n_i^2} \tag{5.7}$$

is based on the Satterthwaite (1946) approximation using the fact that we assume the type B error has a known variance. The labs therefore report what is essentially a conservative Confidence Distribution given by the location-scale t distribution with distribution function

$$H_i(\mu_i \mid \mathbf{X}_i) = F^t_{d_i}\left(\frac{\mu_i - x_i}{u_i}\right).$$

We take these lab reported CDs as a starting point for our model averaging. Trying to improve the lab reported CDs goes beyond the scope of this work and will be subject of future work.

Because the labs are measuring the same quantity, it is reasonable to assume that most, if not all, of the labs are actually providing unbiased estimates of μ. However, it is not uncommon for a handful of labs to provide discrepant results. This may be the consequence of incorrect adjustments by the labs to account for systematic errors or incorrect specification of σ_{B_i}. Our goal is to provide a combined confidence distribution for the common value μ that is robust to discrepant results. We first provide a formula assuming that $E(X_i) = \mu$ for all labs.

Hannig and Xie (2012) provide a simple formula based on Dempster's rule of recombination (Dempster, 2008) and generalized fiducial distribution (Hannig et al., 2015a). The density of the combined CD for μ is

$$h(\mu \mid \mathbf{x}) = \frac{\sum_{i=1}^K \frac{\partial}{\partial\mu} H_i(\mu \mid \mathbf{x}_i) \prod_{j\neq i} D_{\mathbf{x}_j} H_j(\mu \mid \mathbf{x}_j)}{\int_{-\infty}^{\infty} \sum_{i=1}^K \frac{\partial}{\partial\mu} H_i(\bar{\mu} \mid \mathbf{x}_i) \prod_{j\neq i} D_{\mathbf{x}_j} H_j(\bar{\mu} \mid \mathbf{x}_j)\, d\bar{\mu}} \tag{5.8}$$

where $D_{\mathbf{x}_j} H_j(\theta|\mathbf{x}_j) = \|\nabla_{\mathbf{x}_j} H_j(\theta|\mathbf{x}_j)\|_2$ is the norm of the gradient of the $H_j(\theta|\mathbf{x}_j)$ computed with respect to the observed measurements $\mathbf{x}_j$.

Calculations similar to Hannig and Xie (2012) show that

$$D_{\mathbf{x}_j} H_j(\mu|\mathbf{x}_j) = t_{d_j}\left(\frac{\mu - x_j}{u_j}\right)\frac{1}{n_j^{1/2}u_j}\left(1 + \frac{(\mu - x_j)^2}{((n_j-1)d_j)^{1/2}u_j^2}\right)^{1/2}, \tag{5.9}$$

where $t_{d_j}(s)$ is the density of the T distribution with d_j degrees of freedom.

To numerically compute the confidence interval based on the combined generalized fiducial distribution (5.8) we can use the following importance sampling algorithm from Robert and Casella (2004, Section 3.3).

1. Generate $R_{i,l}$, a sample of size m from each of the generalized fiducial distributions $H_i(\mu|\mathbf{x}_i)$, using $R_{i,l} = x_i - u_i T_{i,l}$ where $T_{i,l}, l = 1, \ldots, m$ are independent standard T random variables with d_i degrees of freedom.

2. For each $R_{i,l}$, $i = 1, \ldots K$, $l = 1 \ldots, m$, compute unnormalized weights $W_{i,l} = \prod_{j \neq i} D_{\mathbf{x}} H_j(R_{i,l}|\mathbf{x}_j)$ using (5.9).

3. Compute the importance sampling estimate of the distribution function of (5.8) by
$$\hat{H}(\mu) = \frac{\sum_{i=1}^{K} m^{-1} \sum_{l=1}^{m} W_{i,l} I_{[R_{i,l},\infty)}(\mu)}{\sum_{i=1}^{K} m^{-1} \sum_{l=1}^{m} W_{i,l}}, \tag{5.10}$$
where the indicator $I_{[R_{i,j},\infty)}(\mu) = 1$ if $R_{i,j} \leq \mu$ and $I_{[R_{i,j},\infty)}(\mu) = 0$ otherwise. To form approximate confidence intervals use the appropriate quantiles of $\hat{H}(\mu)$.

Finally notice that the normalizing constant in (5.10) is an estimate of the normalizing constant in (5.8).

5.3.1 Model Selection

Let us now consider the situation where *most* of the labs are measuring the same correct value μ while each remaining lab is measuring some incorrect value. We are interested in making inferences about the true value μ without making any a priori assumptions about which labs are correct. There are $2^K - 1$ possible such models ranging from only a single lab measuring the true value to all the labs measuring the correct value.

Hannig and Lee (2009) have introduced model selection into the generalized fiducial paradigm. Their results have been used for a multivariate normal model by Wandler and Hannig (2011, 2012). The idea is to include the various models as a parameter in the setup of the problem and has been formalized in Theorem 3.1 of Hannig et al. (2015a) where a formula for fiducial probability of each model is described.

In our situation we consider as a model $\mathbf{i} \subset \{1, \ldots, K\}$, with $i \in \mathbf{i}$ if the lab i was measuring μ and $s \notin \mathbf{i}$ if the lab measured some other value. The joint fiducial density of the common mean μ and the discrepant means $\mu_s,\ s \notin \mathbf{i}$ is proportional to

$$\left(\sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{x}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{x}_j} H_j(\mu \mid \mathbf{x}_j) \right) \prod_{s \notin \mathbf{i}} \frac{\partial}{\partial \mu} H_s(\mu_s \mid \mathbf{x}_s).$$

Notice that $\int_{-\infty}^{\infty} \frac{\partial}{\partial \mu} H_s(\mu_s \mid \mathbf{x}_s)\, d\mu_s = 1$ and therefore the marginal combined density for the common parameter μ and model $\mathbf{i}$ is

$$h_{\mathbf{i}}(\mu|\mathbf{x}) = \frac{\sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{x}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{x}_j} H_j(\mu \mid \mathbf{x}_j)}{\int_{-\infty}^{\infty} \sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{x}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{x}_j} H_j(\bar{\mu} \mid \mathbf{x}_j)\, d\bar{\mu}}.$$

Theorem 3.1 of Hannig et al. (2015a) gives a generalized fiducial probability of each model. After integrating out the nuisance parameters this simplifies to

$$h(\mathbf{i}|\mathbf{x}) = \frac{q^{K-|\mathbf{i}|+1} \int_{-\infty}^{\infty} \sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{x}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{x}_j} H_j(\bar{\mu} \mid \mathbf{x}_j)\, d\bar{\mu}}{\sum_{\bar{\mathbf{i}} \in 2^{1,\ldots,K}} q^{K-|\bar{\mathbf{i}}|+1} \int_{-\infty}^{\infty} \sum_{i \in \bar{\mathbf{i}}} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{x}_i) \prod_{j \in \bar{\mathbf{i}},\ j \neq i} D_{\mathbf{x}_j} H_j(\bar{\mu} \mid \mathbf{x}_j)\, d\bar{\mu}},$$

where q is a penalty term to be specified below. The combined confidence distribution for μ obtained by model averaging based on the fiducial probabilities of the model is given by

$$h(\mu \mid \mathbf{x}) = \frac{\sum_{\bar{\mathbf{i}} \in 2^{1,\ldots,K}} q^{K-|\bar{\mathbf{i}}|+1} h_{\bar{\mathbf{i}}}(\mu|\mathbf{x}) h(\bar{\mathbf{i}}|\mathbf{x})}{\int_{-\infty}^{\infty} \sum_{\bar{\mathbf{i}} \in 2^{1,\ldots,K}} q^{K-|\bar{\mathbf{i}}|+1} h_{\bar{\mathbf{i}}}(\bar{\mu}|\mathbf{x}) h(\bar{\mathbf{i}}|\mathbf{x})\, d\bar{\mu}}.$$

The sum above is over $2^K - 1$ summands which would be prohibitively large even for medium values of K. However, by rearranging the terms and combining them into a product we get the following computationally friendly version of the combined density

$$h(\mu \mid \mathbf{x}) = \frac{\sum_{i=1}^{K} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{x}_i) \prod_{j \neq i} \left(1 + q^{-1} D_{\mathbf{x}_j} H_j(\mu \mid \mathbf{x}_j)\right)}{\int_{-\infty}^{\infty} \sum_{i=1}^{K} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{x}_i) \prod_{j \neq i} \left(1 + q^{-1} D_{\mathbf{x}_j} H_j(\bar{\mu} \mid \mathbf{x}_j)\right)\, d\bar{\mu}}. \tag{5.11}$$

The penalty term q is required to offset the propensity of the generalized fiducial distribution to select models with a larger number of parameters. We propose to use the following penalty

$$q = \text{MSE} \left(\sum_{i=1}^{k} u_i^{-2} \right)^{-1/2} \left(\sum_{i=1}^{k} n_i \right)^{-1/2} \tag{5.12}$$

where the type A mean square error $\text{MSE} = K^{-1} \sum_{i=1}^{K} n_i u_i^2 \sqrt{(n_i - 1)/d_i}$. This penalty is inspired by the Minimum Description Length principle (Lee, 2001) with the addition of the MSE factor that is meant to make the method scale invariant.

Based on (5.11) we propose the following importance sampling algorithm that is usable for practical computations:

1. Generate $R_{i,l}$, a sample of size m from each of the generalized fiducial distribution $H_i(\mu|\mathbf{x}_i)$, using $R_{i,l} = \bar{x}_i - u_i T_{i,l}$ where $T_{i,l}, l = 1, \ldots, m$ are independent standard T random variables with d_i degrees of freedom.

2. For each $R_{i,l}$, $i = 1, \ldots K$, $l = 1 \ldots, m$, compute unnormalized weights

$$\tilde{W}_{i,l} = \prod_{j \neq i} \left[1 + D_{\mathbf{x}_j} H_j(R_{i,l}|\mathbf{x}_j)\, q^{-1} \right],$$

   where $D_{\mathbf{x}_j} H_j$ is given in (5.9) and q is in (5.12).

3. Compute the importance sampling estimate of the distribution function of (5.11) by

$$\hat{H}(\mu) = \frac{\sum_{i=1}^{K} \sum_{l=1}^{M} \tilde{W}_{i,l} I_{[R_{i,l},\infty)}(\mu)}{\sum_{i=1}^{K} \sum_{j=1}^{M} \tilde{W}_{i,l}}.$$

   To form approximate confidence intervals use the appropriate quantiles of $\tilde{H}(\mu)$.

Remark 5.3.1. The combined confidence distribution in (5.11) treats all the labs equally. However in some situations we want to combine results that are similar to a particular lab. This is achieved by making sure that this lab is included in all the models considered. If the lab r is preferred, this exhibits itself in (5.11) and the corresponding part of the importance sampling algorithm by replacing "1+" in the formula with "$I_{\{j \neq r\}}+$".

5.4 Simulation Study

To demonstrate the small sample performance of our proposed algorithm, we conducted a simulation study consisting of measurements from 7 labs generated from each of three different scenarios listed below.

- *Scenario 0*: All 7 labs provide unbiased estimates of the true value μ. We take $\mu = 45$ for concreteness.

- *Scenario 1*: Six labs provide unbiased estimates of the true value $\mu = 45$ and while one lab provides a biased estimate whose expectation is $\mu + 3 = 48$. This mimics the situation where one lab may incorrectly estimate the lab bias B_k and/or the standard deviation of lab bias $\sigma_{B,k}$.

- *Scenario 2*: Two clusters of labs. One cluster of size 4 make measurements with expected value equal to 45 and the other cluster of size 3 make measurements with expected value equal to 48. This setting simulates the situation where labs use fundamentally different methods for measurement and it is impossible to know which of the labs, if any, are providing unbiased estimates of the true value μ. Thus, there is no answer to which value is the truth.

For each scenario, we assume each lab makes the same number of measurements n_i and thus same type A degrees of freedom $n_i - 1$. Two values $n_i = 5, 15$ are used in the simulation study. To model the heterogeneity among the labs, different standard deviations of type A error and type B error, $\sigma_{A,i}$ and $\sigma_{B,i}$ respectively, are generated from a Gamma distribution for each lab, i.e.

$$\sigma_{A,i} \sim \Gamma(n_i, \frac{1}{n_i}), \quad \sigma_{B,i} \sim \Gamma(n_i, \frac{R}{n_i})$$

in which R is the ratio of the mean of $\sigma_{B,i}$s over the mean of $\sigma_{A,i}$s. Four different ratios are considered ($R = 0$, 1/3, 1, 2) for generating the data sets. Note that $R = 0$ implies type B error is not present. For each collection of $\sigma_{B,i}$s, the type B errors B_i are simulated, one per lab, from normal distribution $B_i \sim N(0, \sigma^2_{B,i})$.

One hundred parameter sets of $\{\sigma_{A,i}, \sigma_{B,i}, B_i, i = 1, \cdots, 7\}$ are simulated following this procedure. For each fixed parameter set, 1000 repetitions of the laboratory

measurements, type A and combined standard errors are generated using

$$X_i = \mu_i + B_i + \frac{\sigma_{A,i}}{\sqrt{n_i}} Z_i, \quad s_i = \sigma_{A,i}\sqrt{\frac{W_i}{(n_i - 1)}}, \quad u_i = \sqrt{\frac{s_i^2}{n_i} + \sigma_{B,i}^2},$$

where $Z_i \sim N(0, 1)$ and $W_i \sim \chi^2_{n_i - 1}$ are independent.

In addition to the proposed method we also used classical methods, the arithmetic mean and the variance weighted mean, for calculating a consensus value for the simulated data. These two classical methods are commonly used in metrology (GUM, 1995). Detailed results for each scenario are discussed in Section 5.4.1, Section 5.4.2 and Section 5.4.3, respectively.

5.4.1 Scenario 0

The expected values for the measurements by the 7 labs are all equal to $\mu = 45$. For illustration, Figure 5.4 provides an example of the fiducial distribution of the consensus value for one of the datasets generated for $d.f. = 4$ and $R = 0$. The blue curves in the left panel are kernel density estimates for the fiducial density for each lab. It can be seen that the expected result for each lab deviates slightly from the true value of 45 with different amounts of dispersion. The top black kernel density estimator shows that the center of the consensus value distribution is around the true value 45. The top black confidence curve in the right panel depicts the median estimate as 44.9 and 95% fiducial confidence interval as $[44.4,\ 45.4]$ which successfully covers the truth.

For each of the 100 parameter sets, we compute the coverage and lengths of the 95% confidence intervals based on the 1000 simulated data sets. Box-plots shown in Figure 5.5 summarize the results for the 100 parameter sets under different ratios R and $d.f.$. The blue boxes display the coverages for fiducial method, while the green and yellow boxes, respectively, show the coverages of arithmetic mean and weighted mean. Results are grouped by ratios for $d.f. = 4$ (left) and $d.f. = 14$ (right). The average coverages are given underneath each box. When only type A error is present ($R = 0$), the coverage of fiducial estimates and arithmetic mean are around 95%, while the weighted mean has a much lower coverage, especially when $d.f. = 4$ (with median coverage being around 80%). When type B error exists and increases (larger R), all three methods tend to get

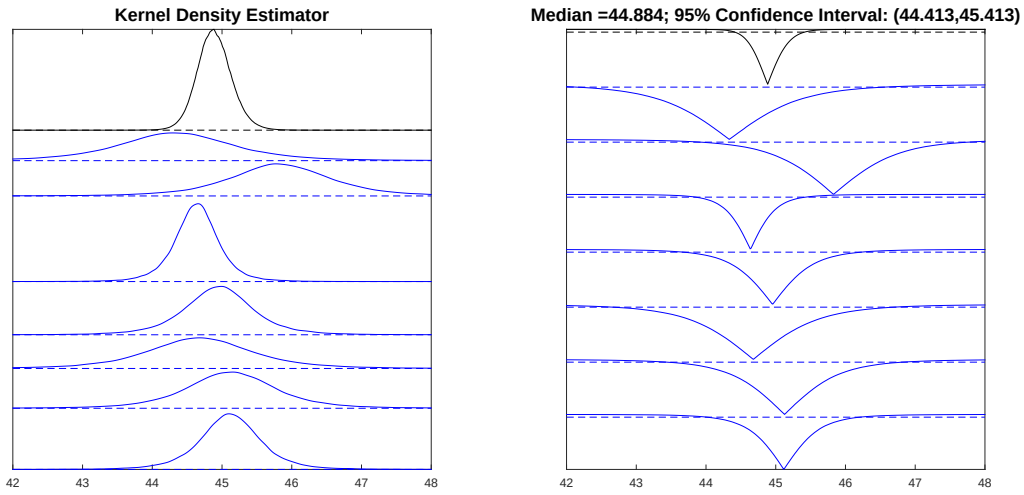


Figure 5.4: Fiducial estimate of one simulated data with $\sigma_{A,i}$, $\sigma_{B,i}$, B_i generated from $d.f. = 4$ and $R = 0$ of scenario 0. The top left black curve shows the kernel density estimate for the consensus value with mode around 45. The top right black curve shows that the confidence curve covers the true value 45 at 95% confidence level.

100% coverage. However, the arithmetic mean and the weighted mean are less robust in the sense that they might get 0 coverage for certain parameter sets.

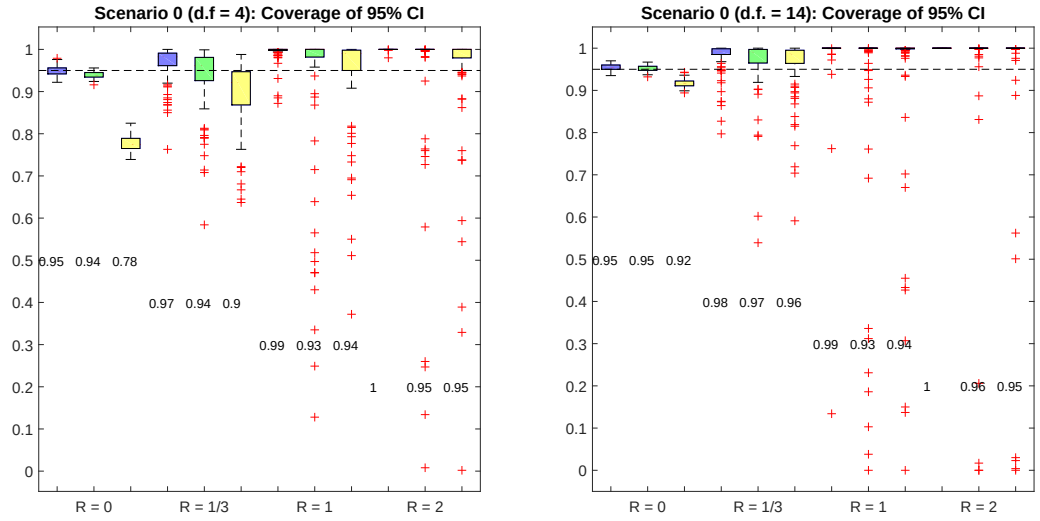


Figure 5.5: Coverage Comparison for Scenario 0 grouped by ratios R for $d.f. = 4$ (left) and $d.f. = 14$ (right). The results of fiducial method are given in the blue boxes with average coverage written underneath. The green and yellow boxes present the results of arithmetic and weighted mean, respectively.

Additionally, we compute the average length of 95% confidence intervals of 1000 simulated data sets for comparing the three methods. As before, Figure 5.6 displays the box-plots of fiducial method (blue), arithmetic mean (green) and weighted mean

(yellow) for different choices of $d.f.$ and R. The confidence intervals gets wider with an increase in the ratio R and gets shorter when the degree of freedom increases for all three methods. In general, the fiducial intervals are wider than the others which is consistent with the coverage comparison in Figure 5.5.

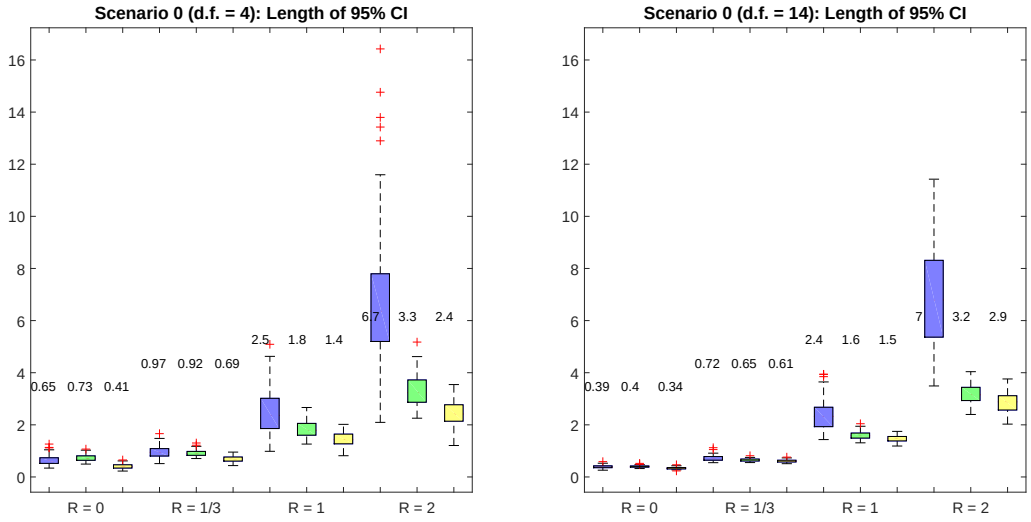


Figure 5.6: 95% CI length comparison, under Scenario 0, for $d.f.= 4$ (left) and $d.f. = 14$ (right) with different ratios R: fiducial method (blue), arithmetic mean (green), weighted mean (yellow). The CI gets wider with the increase in the ratio and gets narrower with the increase in degree of freedom for all three methods.

5.4.2 Scenario 1

In this scenario, we try to mimic the consensus value estimation with a single apparently discrepant lab. Again, Figure 5.7 provides an illustration of the GFD of the consensus value for one simulated data set with $d.f. = 4$ and $R = 0$. The bottom blue curves in both panels indicates the presence of a discrepant lab. The black curves on the top show that the consensus value estimate from the fiducial approach is around 45 and appears to be uninfluenced by the apparently discrepant lab. The 95% confidence interval is $[44.40,\ 45.18]$ which covers the true value of the six *nondiscrepant* labs. It can be seen that the fiducial consensus estimate is robust against an outlier measurement from a discrepant lab.

We similarly compute the coverage of the 95% confidence interval for the true value $\mu = 45$. Box-plots for different choices of $d.f.$ and R are shown in Figure 5.8. The fiducial method stays robust against the discrepant lab measurement and obtains similar

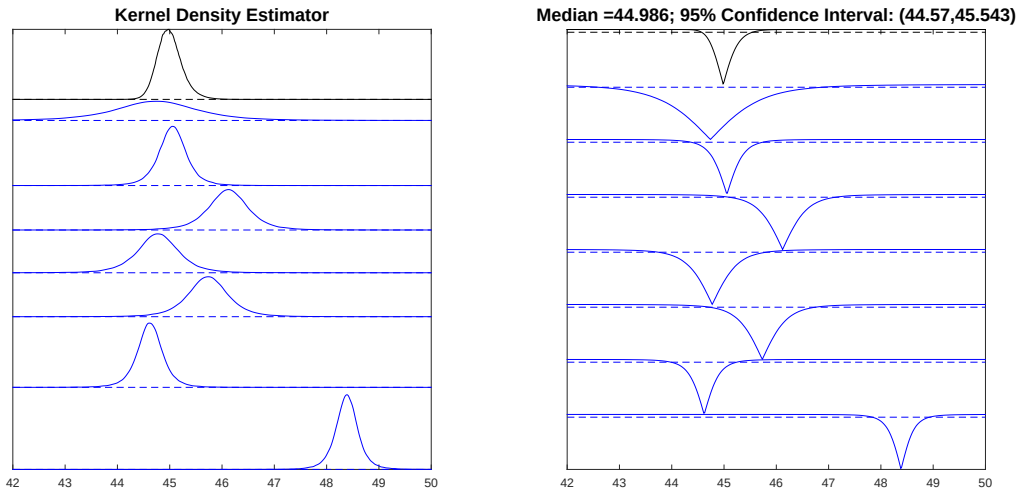


Figure 5.7: One simulated data with $\sigma_{A,i}$ and $\sigma_{B,i}$ generated from $d.f. = 4$ and $R = 0$. The kernel density estimates (left) indicate an apparently discrepant measurement from the last lab. The black kernel density estimate and confidence curves for the consensus value demonstrate the robustness of our proposed method against discrepant measurements.

coverages as Scenario 0. Both arithmetic mean and weighted mean are adversly influenced by the discrepant lab. When $d.f. = 4$, the coverages are only around 40% with no type B error or a small ratio of type B error. Differing from Scenario 0, the coverages get worse with an increase in degree of freedom since the evidence of the outlier lab gets stronger. The median coverages even drop to near 0 when $R = 0$ and $d.f. = 14$. When type B error dominates, both arithmetic and weighted mean are unstable with some zero coverages as in the previous scenario.

5.4.3 Scenario 2

We consider two clusters of labs in this scenario for simulating the situation where labs might use different measuring methods. Recall that the true value of 4 labs is equal to 45 and the true value of the other 3 labs is equal to 48. In this situation, it is not clear which of the two, 45 or 48, should be the consensus value. This is illustrated in Figure 5.9 where GFD of the consensus value for two simulated data sets generated with $d.f. = 4$ and $R = 0$ are shown. The first four blue curves are centered around 45 representing the first cluster and the last three blue curves are centered around 48 from the second cluster. In the first row, the top black density estimate curve suggests that the estimate of the consensus value is about 45 which is dominated by the cluster whose true value is 45. A small peak can be seen around 48 indicating the impact from

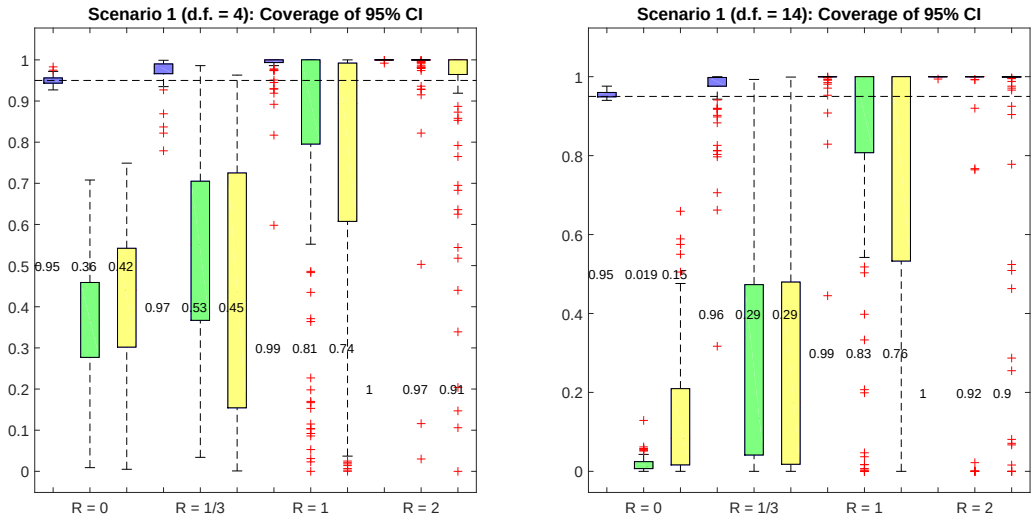


Figure 5.8: Coverage Comparison for Scenario 1: fiducial estimate (blue), arithmetic mean (green), weighted mean (yellow). Fiducial estimate is robust against the apparent discrepancy of one of the labs, while the other methods are strongly influenced, especially in the case without type B error ($R = 0$).

the other cluster. Besides, the black confidence curve shows the 95% confidence interval is [44.6, 47.9] which stretches towards the true value of the other cluster of labs. The second row shows an example of a situation where the value of 48 dominates. One of the labs in the second cluster has much smaller uncertainty compared with the other labs. This is enough to move the mode of the fiducial distribution of the consensus value to 48. The confidence curve suggests the 95% confidence interval is [44.6, 48.4], successfully covering both true values.

The assessment of coverage is tricky in the current situation since there is no single correct value. We evaluate two different coverage probabilities – (a) probability that at least one of the two values (45 or 48) will be covered, and (b) the probability of covering both 45 and 48. Results are shown in Figure 5.10 and Figure 5.11. The fiducial confidence intervals (blue boxes) cover at least one of the two values nearly 100% of the time under the different parameter settings. However, both arithmetic mean and weighted mean fail to capture any of the true values when ratio $R = 0$ and $R = 1/3$.

When it comes to simultaneous coverage of both values, 75% of the fiducial confidence intervals have a coverage around or above 60% for $R = 0$ and $R = \frac{1}{3}$. This should not be surprising because our method was designed to capture the most dominant value,

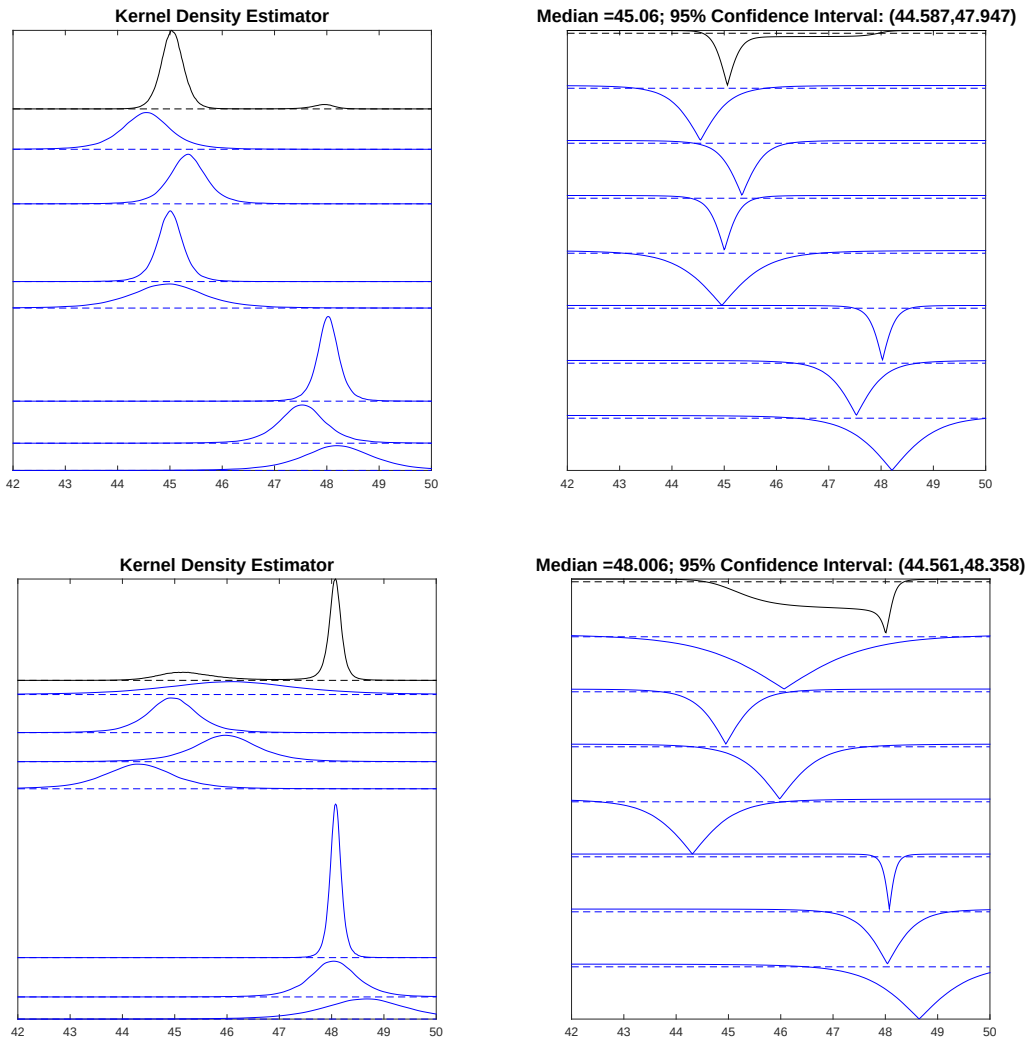


Figure 5.9: Two simulated data sets under Scenario 2 with $\sigma_{A,i}$ and $\sigma_{B,i}$ generated from $d.f. = 4$ and $R = 0$. The first example shows that the first cluster with 4 labs dominate the consensus value estimation. The second example presents the strong record from one lab in the cluster with true value being 48. Therefore, the consensus estimate is shifted towards 48.

not both values. The other two methods are unable to simultaneously cover both of the true values for any of the cases.

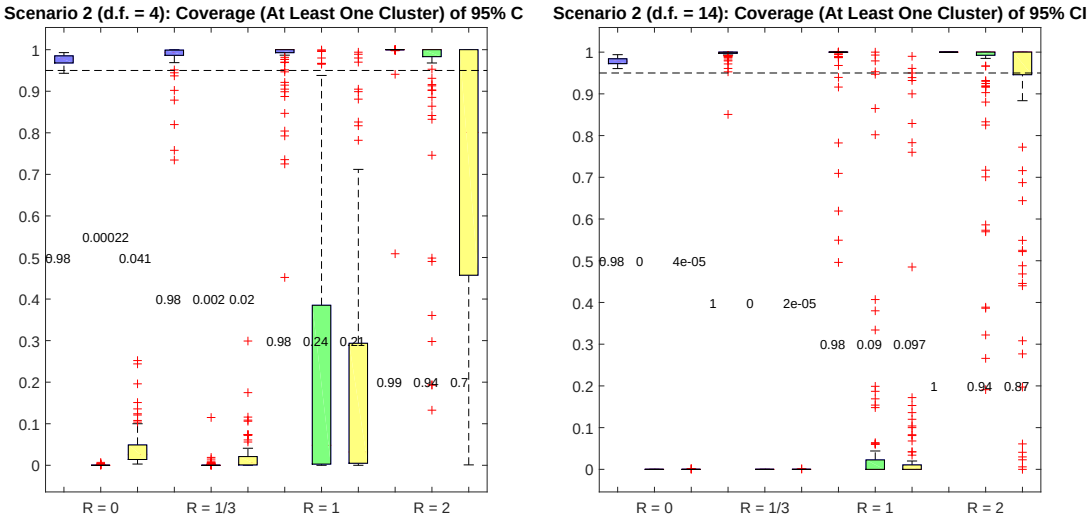


Figure 5.10: Coverage Comparison (At Least One Cluster) for Scenario 2 for different ratios and degrees of freedom.

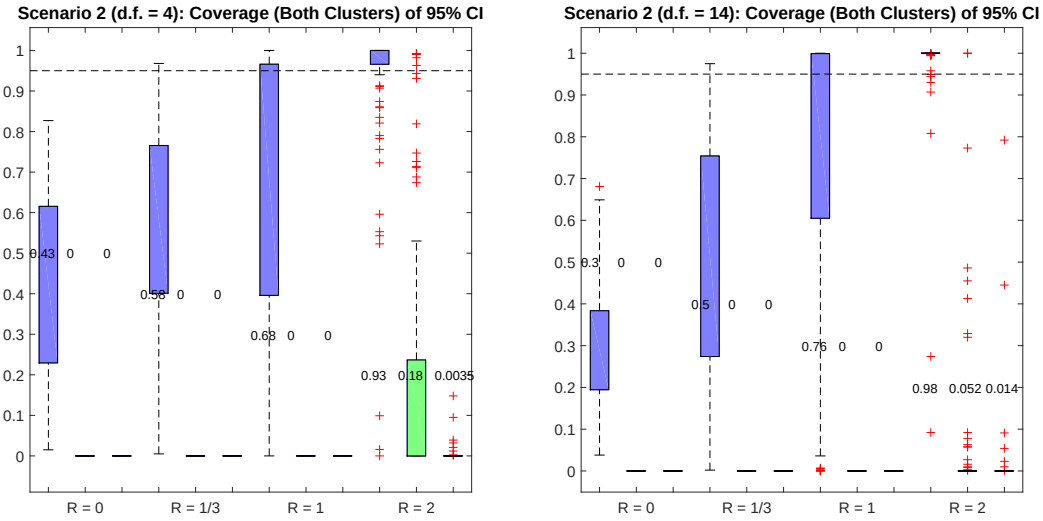


Figure 5.11: Coverage Comparison (Both Clusters) for Scenario 2 for different ratios and degrees of freedom.

5.5 Discussion of Simulation Results

Below is a brief summary of some main observations from the simulation study.

- In Scenario-0 and Scenario-1 the fiducial intervals cover the true μ with confidence level greater than or equal to the nominal value of 95%. For Scenario-2 the coverage probability for covering at least one of the two cluster means is greater than or equal to the nominal value.

- For arithmetic mean and weighted mean approaches the coverage is nowhere near nominal in most situations examined. For scenarios 1 and 2 the coverage is particularly bad. These methods are unsuitable for situations when there may be discrepant labs.

- Although the arithmetic mean and the weighted mean provide 95% confidence intervals of shorter expected length this is meaningless since their coverages are highly inadequate.

5.6 Data examples

We illustrate our method with two real data examples. The first example is taken from a key comparison study called CCL-K1 which involved length measurements of steel gauge blocks. The second example involves measurements, by many different labs, of the Newton's constant of gravitation, called *big G* (to distinguish it from g, the acceleration due to gravity). The details follow.

5.6.1 Steel Gauge Blocks

In order to establish the metrological equivalence of national measurement standards and of calibration certificates issued by national metrology institutes a set of key comparisons are chosen and organized by the Consultative Committees of the CIPM or by the regional metrology organizations in collaboration with the Consultative Committees (Thalmann, 2002). In September 1997, the Consultative Committee for Length, CCL, decided upon a key comparison on gauge block measurements by interferometry, named CCL-K1, starting in spring 1998, with the Swiss Federal Office of Metrology (OFMET) as the pilot laboratory. The results of this international comparison contribute to the mutual recognition arrangement (MRA) between the national metrology institutes of the Metre Convention.

Ten gauge blocks of steel and 10 gauge blocks of tungsten carbide, of varying nominal lengths, were circulated to 11 different NMIs. For the purpose of illustration we considered one particular set of gauge block measurements corresponding to the nominal value of 8 mm. The results along with their associated uncertainties are shown in Figure 5.2. What is actually reported by each participating lab is the *deviation (in nm)* of the measured length from the nominal value.

The published reports did not clearly spell out the degrees of freedom. In order to apply the proposed method we selected the total degrees of freedom $d = 60$ which correspond to the usual multiplier of 2. The type B to type A standard deviation ratio is typically 1.5 in these problems and (5.7) gives the corresponding type A degrees of freedom as $n - 1 = 6$. Figure 5.12 presents the estimates of kernel density curve (left) and confidence curves (right). The 95% confidence interval is $[-31.6,\ 37.3]$ nm with the median estimate being 4.08 nm. The consensus value estimate mainly picks up the measurements of the labs with mode around 0. The confidence interval takes the uncertainty caused by two discrepant labs into consideration.

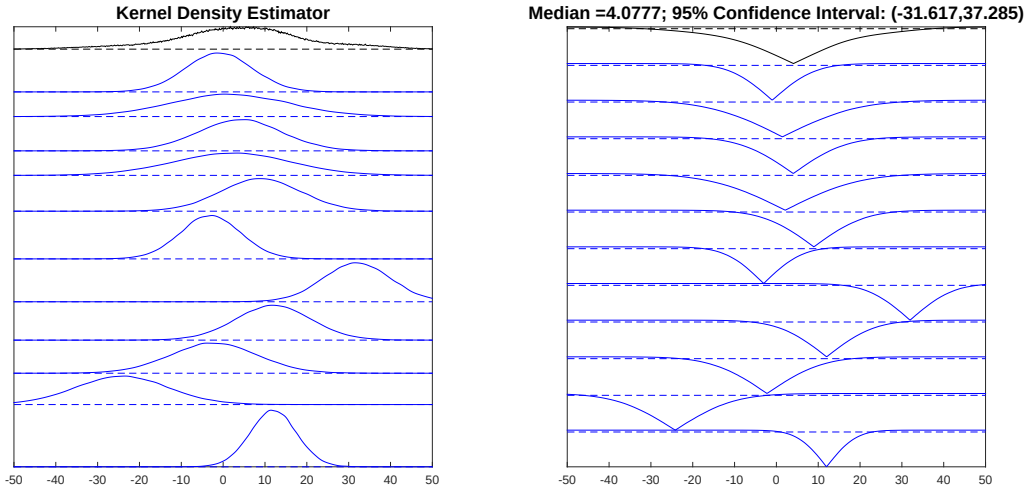


Figure 5.12: Results of CCL data set with total degrees of freedom equal to 60 and type A degrees of freedom equal to 6.

The arithmetic mean -0.2 ± 3.5 nm and the weighted mean 0.1 ± 3.2 nm are given in Thalmann (2002) as the reference value. These results exclude the values of VNIIM and NIM based on the decision of the CCL Working Group Dimensional Metrology (WGDM). Hence their confidence intervals are narrower as VNIIM is the one most different from the others.

5.6.2 Newton's Constant of Gravitation, G

Newton's constant of gravitation G is a key constant that is needed for much fundamental research in physics. Many advanced scientific labs measure G and report a value and an uncertainty. The data set contains the values from 11 labs shown in Table 5.2. See Mohr et al. (2012) for details.

Organization	Result	Combined Standard Uncertainty
NIST-82	6.67248	0.00043
TR&D-96	6.6729	0.00050
LANL-97	6.67398	0.00070
UWash-00	6.674255	0.000092
BIPM-01	6.67559	0.00027
UWup-02	6.67422	0.00098
MSL-03	6.67387	0.00027
HUST-05	6.67228	0.00087
UZur-06	6.67425	0.00012
HUST-09	6.67349	0.00018
JILA-10	6.67234	0.00014

Table 5.2: Summary of the results of measurements of the Newton's constant of gravitation G. The units are $10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$.

It turns out that the confidence interval for G from some labs exclude values from other labs, so there is some inconsistency. This is perhaps due to severe underestimation by some or all the labs of uncertainties in their results. The community seeks a consensus value that uses all available information. We applied the proposed method and obtained an estimate depicted in Figure 5.13 computed using the default values of $d = 60$ and $n - 1 = 6$.

The blue curves show that two labs, with small uncertainties, perhaps coincidently, have nearly the same mode around $6.674 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$. Besides, there are several labs whose results are near this value with varying levels of uncertainties. The consensus estimate is therefore pulled towards this number with 95% confidence interval being $[6.6740,\ 6.6743] \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$.

The value of G given by Mohr et al. (2012) is $6.67384 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$. The uncertainty is $0.00080 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$ which is the weighted mean of the 11 values in Table 5.2 multiplied the factor 14. The multiplication is intended to cover all the 11 values of G as none of them has an apparent issue besides the disagreement. Hence,

although the confidence interval in the report better covers all the measurements, it might not be robust due an arbitrary magnification of uncertainty.

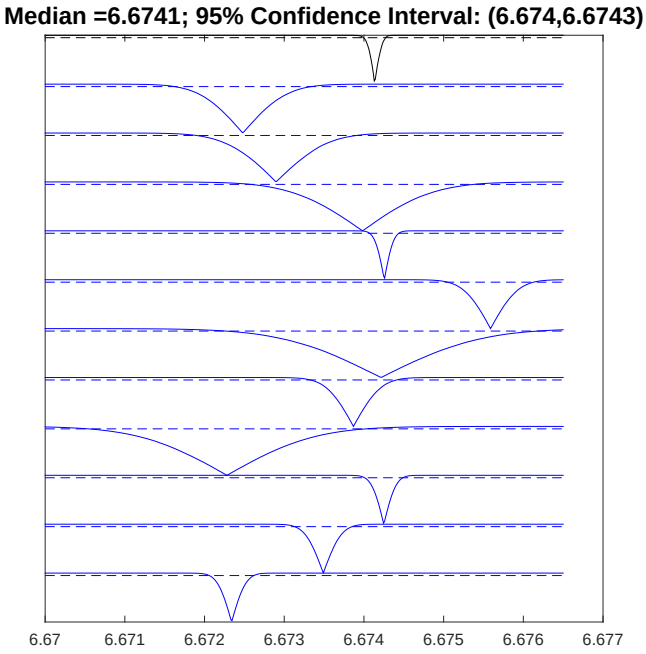


Figure 5.13: Results of Big-G data set with total degrees of freedom equal to 60 and type A degrees of freedom equal to 6.

APPENDIX A: NON-ITERATIVE JIVE PROOF

Proof of Theorem 3.1. Define the row subspaces respectively for each matrix A_k as $\text{row}(A_k) \subseteq \mathbb{R}^n$. For each row subspace, there exists a corresponding projection matrix P_k $(n \times n)$ which is idempotent and symmetric. For non-trivial cases, define a subspace $\text{row}(J) \neq \{\mathbf{0}\}$ as the intersection of row spaces of $\{\text{row}(A_k),\ k = 1, \ldots, K\}$ i.e.

$$\text{row}(J) \triangleq \bigcap_{k=1}^{K} \text{row}(A_k).$$

The projection matrix of subspace $\text{row}(J)$, P_J, can thus be represented as $P_J = \prod_{k=1}^{K} P_k$. Then for each matrix A_k, two matrices J_k, I_k can be obtained using projection matrix P_J and its orthogonal complement $P_{I_k} \triangleq P_k - P_J$ i.e. $J_k = A_k P_J$ and $I_k = A_k P_{I_k}$. The two matrices satisfy $J_k + I_k = A_k$ and their row subspaces are orthogonal with each other $\text{row}(J) \perp \text{row}(I_k)$, $k = 1, \ldots, K$.

Moreover, the intersections of row subspaces $\{\text{row}(I_k),\ k = 1, \cdots, K\}$, $\bigcap_{k=1}^{K} \text{row}(I_k)$, has a projection matrix written as

$$\prod_{k=1}^{K} P_{I_k} = \prod_{k=1}^{K} P_k(I - P_J) = \prod_{k=1}^{K} P_k \prod_{k=1}^{K} (I - P_J) = 0$$

Therefore, we have $\bigcap_{k=1}^{K} \text{row}(I_k) = \{\mathbf{0}\}$ satisfied and obtain a set of matrices simultaneously satisfying the stated constraints.

Next we show the sets of matrices $\{J_k\ (d_k \times n),\ k = 1, \ldots, K\}$ and $\{I_k\ (d_k \times n),\ k = 1, \ldots, K\}$ are uniquely defined. Assume the row subspace of matrices J_k, $\text{row}(J_k) = \text{row}(J)$, is spanned by a set of bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ and the row subspaces of I_k, $\text{row}(I_k)$, is spanned by a set of bases $\{\mathbf{w}_1, \ldots, \mathbf{w}_{Ik}\}$. The row subspace $\text{row}(A_k)$ is thus spanned by their union i.e. $\{\mathbf{v}_1, \ldots, \mathbf{v}_J, \mathbf{w}_1, \ldots, \mathbf{w}_{Ik}\}$, since $\text{row}(J_k) = \text{row}(J) \perp \text{row}(A_k)$ for all k. Hence, given an arbitrary vector $\mathbf{v} \in \text{row}(J)$, we always has $\mathbf{v} \in \text{row}(A_k)$ for all k, which indicates

$$\text{row}(J_k) = \text{row}(J) \nsubseteq \text{row}(A_k), \quad k = 1, \ldots, K,$$

and therefore

$$\text{row}(J) \subseteq \bigcap_{k=1}^{K} \text{row}(A_k).$$

Furthermore, suppose there exist a non-zero vector $\mathbf{a} \in \bigcap_{k=1}^{K} \text{row}(A_k)$ but $\mathbf{a} \notin \text{row}(J)$ and $\mathbf{a} \perp \text{row}(J)$. This vector should have $\mathbf{a} \in \text{row}(I_k),\ k = 1, \ldots, K$ and thus $\mathbf{a} \in \bigcap_{k=1}^{K} \text{row}(I_k)$ which contradicts the constraint $\bigcap_{k=1}^{K} \text{row}(I_k) = \{\mathbf{0}\}$. This implies that the row subspace $\text{row}(J)$ is uniquely defined as

$$\text{row}(J) = \bigcap_{k=1}^{K} \text{row}(A_k).$$

Accordingly, the matrices J_k and I_k for $k = 1, \ldots, K$ are also uniquely defined. Otherwise assume there have another set of matrices $A_k = \tilde{J}_k + \tilde{A}_k$ and P_J is the projection matrices of $\text{row}(J)$, we have $J_k = A_k P_J = \tilde{J}_k$. □

Proof of Lemma 1. Let P_1 and P_2 be the projection matrices onto the individually perturbed joint row spaces. And let P be the projection matrices onto the common joint row space J. Thus, we have

$$\begin{aligned}
\sin\theta &= \|(I - P_1)P_2\| && \text{(A.1)}\\
&\leq \|(I - P_1)(I - P)P_2\| + \|(I - P_1)PP_2\| && \text{(A.2)}\\
&\leq \|(I - P_1)(I - P)\|\|(I - P)P_2\| + \|(I - P_1)P\|\|PP_2\| && \text{(A.3)}
\end{aligned}$$

in which $\|(I - P_1)P\| = \sin\phi_1,\ \|(I - P_1)(I - P)\| = \cos\phi_1,\ \|(I - P_2)P\| = \sin\phi_2$ and $\|(I - P_2)(I - P)\| = \cos\phi_2$. Therefore,

$$\sin\theta \leq \cos\phi_1 \sin\phi_2 + \sin\phi_1 \cos\phi_2 = \sin(\phi_1 + \phi_2).$$

□

Proof of Lemma 2. Denote the spanning basis for the estimates of each signal score spaces $\text{row}(\tilde{A}_k)$ as $\{\tilde{V}_k, k = 1, \cdots, K\}$ and M as the vertical concatenation of right singular vector matrices $\{\tilde{V}_k^T, k = 1, \cdots, K\}$ (denoted as M) for SVD.

$$M \triangleq \begin{bmatrix} \tilde{V}_1^T \\ \vdots \\ \tilde{V}_K^T \end{bmatrix} = U_M \Sigma_M V_M^T.$$

For each singular value, it can be formulated as a sequential optimization problem i.e

$$\sigma_i^2 = \max\|MQ\|_F^2 = \max\sum_{k=1}^{K}\|\tilde{V}_1^T Q\|_F^2,$$

in which Q is a rank 1 projection matrix that is orthogonal to the previous $i-1$ optima i.e. $Q_1, \ldots, Q_{i-1}$. For the one that maximizing the Frobneius norm of M projected onto it i.e. σ_i, we denote as Q_i.

For an arbitary component in the theoretical joint score subspace row(J), write its projection matrix as P_J. The Frobneius norm of M projected onto P_J is

$$\|MP_J\|_F^2 = \begin{bmatrix} \tilde{V}_1^T P_J \\ \vdots \\ \tilde{V}_K^T P_J \end{bmatrix}_F^2 \geq \begin{bmatrix} \cos\phi_1 \\ \vdots \\ \cos\phi_K \end{bmatrix}_F^2 = \sum_{k=1}^{K}\cos^2\phi_k \tag{A.4}$$

Considering the mechanism of SVD, σ_1^2 is the maximal norm obtained from the optimal projection matrix $Q_1 \subseteq \bigcup_{k=1}^{K}\text{row}(\tilde{A}_k) \subseteq \mathbb{R}^n$. Assuming the low rank approximations $\tilde{A}_k$ are correctly given for each data, we have $\text{row}(J) \subseteq \bigcup_{k=1}^{K}\text{row}(\tilde{A}_k)$ and therefore

$$\sigma_1^2 \geq \|MP_J\|_F^2 \geq \sum_{k=1}^{K}\cos^2\phi_k$$

to be considered as a component of joint score subspace.

Sequentially such argument can be applied to the following projection matrices Q_i. For the $Q_2 \in Q_1^{\perp} \cap \{\bigcup_{k=1}^{K}\text{row}(\tilde{A}_k)\}$, there exist a non-empty joint subspace ($\subseteq \text{row}(J)$) and therefore an joint component with projection matrix $P_J^{(2)}$ such that

$$\sigma_2^2 \geq \|MP_J^{(2)}\|_F^2 \geq \sum_{k=1}^{K}\cos^2\phi_k$$

This will continue until the r_J joint components are extracted and no such component for optimization in SVD. Therefore, the right singular values should larger than $\sum_{k=1}^{K}\cos^2\phi_k$ to be considered as the estimates of joint components. □

BIBLIOGRAPHY

Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

R. Bender, G. Berg, and H. Zeeb. Tutorial: Using Confidence Curves in Medical Research. *Biometrical Journal*, 47:237–247, 2005.

James O Berger and José M Bernardo. On the development of reference priors. *Bayesian Statistics*, 4(4):35–60, 1992.

Hans Binder, Lydia Hopp, Kathrin Lembcke, and Henry Wirth. Personalized disease phenotypes from massive omics data. *Big Data Analytics in Bioinformatics and Healthcare*, pages 359–378, 2014.

A. Birnbaum. Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of American Statistical Association*, 56:246–249, 1961.

Åke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

Raymond J. Carroll. A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(1):71–78, 1980.

Lu Charboneau, Heather Scott, Tina Chen, Mary Winters, Emanuel F Petricoin, Lance A Liotta, and Cloud P Paweletz. Utility of reverse phase protein arrays: applications to signalling pathways and human body arrays. *Briefings in functional genomics & proteomics*, 1(3):305–315, 2002.

CIPM. Text of the CIPM MRA. *http://www.bipm.org/en/cipm-mra/cipm-mra-text/*, 1999.

Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer, 2007.

A. P. Dempster. The Dempster-Shafer Calculus for Statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, June 2008.

Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press, 1928.

BV Gnedenko. Sur la distribution limité du terme dkune série aléatoire. *The Annals of Mathematics*, 44:195–225, 1943.

GUM. *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization (ISO), Geneva, Switzerland, 1995.

Aaron K Han. A non-parametric analysis of transformations. *Journal of Econometrics*, 35(2):191–209, 1987.

Mohamed Hanafi, Achim Kohler, and El-Mostafa Qannari. Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and intelligent laboratory systems*, 106(1):37–40, 2011.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

J Hannig, Hari K. Iyer, Randy C. S. Lai, and Thomas C. M. Lee. Generalized Fiducial Inference: A Review. Submitted for publication, 2015a.

Jan Hannig. Generalized fiducial inference via discretization. *Statist. Sinica*, 23(2): 489–514, 2013.

Jan Hannig and Thomas C. M. Lee. Generalized Fiducial Inference for Wavelet Regression. *Biometrika*, 96(4):847–860, 2009.

Jan Hannig and Minge Xie. A note on Dempster-Shafer Recombinations of Confidence Distributions. *Electronic Journal of Statistics*, 6(1943-1966), 2012.

Jan Hannig, CM Wang, and Hari K Iyer. Uncertainty calculation for the ratio of dependent measurements. *Metrologia*, 40(4):177, 2003.

Jan Hannig, Qing Feng, Hari Iyer, Jack Wang, and Xuhua Liu. Fusion learning for key comparisons. Manuscript submitted for publication, 2015b.

Jan Hannig, Hari Iyer, Randy CS Lai, and Thomas CM Lee. Generalized fiducial inference: A review. *Unpublished manuscript, http://www. unc. edu/˜ hannig/publications/HannigIyerLaiLee2015. pdf*, 2015c.

Fabian Hernandez and Richard A Johnson. The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75(372):855–861, 1980.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Hari K. Iyer, C. M. Jack Wang, and Thomas Mathew. Models and Confidence Intervals for True Values in Interlaboratory Trials. *Journal of the American Statistical Association*, 99(468):1060–1071, 2004a.

Hari K Iyer, CM Wang, and DF Vecchia. Consistency tests for key comparison data. *Metrologia*, 41(4):223, 2004b.

Hari K Iyer, CM Jack Wang, and Thomas Mathew. Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99(468):1060–1071, 2004c.

Shashank Jere, Justin Dauwels, Muhammad Tayyab Asif, Nikola Mitro Vie, Andrzej Cichocki, and Patrick Jaillet. Extracting commuting patterns in railway networks through matrix decompositions. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 541–546. IEEE, 2014.

Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174, 1875.

Samuel Kotz and Saralees Nadarajah. *Multivariate t–distributions and their applications*. Cambridge University Press, 2004.

Oliver Kühnle. *Integration of multiple high-throughput data-types in cancer research*. PhD thesis, Ludwig Maximilian University of Munich, 2011.

Julia Kuligowski, David Pérez-Guaita, Ángel Sánchez-Illana, Zacarías León-González, Miguel de la Guardia, Máximo Vento, Eric F Lock, and Guillermo Quintás. Analysis of multi-source metabolomic data using joint and individual variation explained (jive). *Analyst*, 2015.

R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer New York, 2011. ISBN 9781461254515.

Thomas C. M. Lee. An introduction to coding theory and the two–part minimum description length principle. *International Statistical Review*, 69:169–183, 2001.

Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, page btt425, 2013.

Eric F Lock, Katherine A Hoadley, JS Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

J Steve Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

JS Marron, Michael J Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

Ryan Martin and Chuanhai Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313, 2013.

Gregoria Mateos-Aparicio. Partial least squares (pls) methods: Origins, evolution, and application to social sciences. *Communications in Statistics-Theory and Methods*, 40 (13):2305–2317, 2011.

Richard J McNally, Hari Iyer, and Thomas Mathew. Tests for individual and population bioequivalence based on generalized p-values. *Statistics in medicine*, 22(1):31–53, 2003.

Jianming Miao and Adi Ben-Israel. On principal angles between subspaces in R^n. *Linear algebra and its applications*, 171:81–98, 1992.

Jayson Miedema, James Stephen Marron, Marc Niethammer, David Borland, John Woosley, Jason Coposky, Susan Wei, Howard Reisner, and Nancy E Thomas. Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–444, 2012.

Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

P.J. Mohr, B.N. Taylor, and D.B. Newell. Codata recommended values of the fundamental physical constants: 2010. *Reviews of Modern Physics*, 84(4):1527–1605, 2012.

Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-21239-6.

R. M. Sakia. The box–cox transformation technique: a review. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(2):169–178, 1992.

Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, pages 110–114, 1946.

Tore Schweder and Nils Lid Hjort. Confidence and Likelihood. *Scandinavian Journal of Statistics. Theory and Applications*, 29(2):309–332, 2002.

Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, et al. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

Kesar Singh, Minge Xie, and William E. Strawderman. Combining Information from Independent Sources Through Confidence Distributions. *The Annals of Statistics*, 33 (1):159–183, 2005.

Age K Smilde, Johan A Westerhuis, and Sijmen de Jong. A framework for sequential multiblock component methods. *Journal of chemometrics*, 17(6):323–337, 2003.

Michael A Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

G.W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309.

R Thalmann. Ccl key comparison: calibration of gauge blocks by interferometry. *Metrologia*, 39(2):165, 2002.

Raoul Tibes, YiHua Qiu, Yiling Lu, Bryan Hennessy, Michael Andreeff, Gordon B Mills, and Steven M Kornblau. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics*, 5(10):2512–2521, 2006.

Johan Trygg and Svante Wold. O2-pls, a two-block (x±y) latent variable regression (lvr) method with an integral osc® lter^2. *J. chemometrics*, 17:53–64, 2003.

Nadine Tung, Chiara Battelli, Brian Allen, Rajesh Kaldate, Satish Bhatnagar, Karla Bowles, Kirsten Timms, Judy E Garber, Christina Herold, Leif Ellisen, et al. Frequency of mutations in individuals with breast cancer referred for brca1 and brca2 testing using next-generation sequencing with a 25-gene panel. *Cancer*, 121(1):25–33, 2015.

Willem Rutger van Zwet. *Convex transformations of random variables*. Mathematisch centrum, 1964.

Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

D. V. Wandler and J. Hannig. Fiducial Inference on the Maximum Mean of a Multivariate Normal Distribution. *Journal of Multivariate Analysis*, 102(1):87–104, 2011.

Damian V. Wandler and Jan Hannig. A Fiducial Approach to Multiple Comparisons. *Journal of Statistical Planning and Inference*, 142(4):878–895, 2012.

CM Wang and Hari K Iyer. Propagation of uncertainties in measurements using generalized inference. *Metrologia*, 42(2):145, 2005.

CM Wang, Jan Hannig, and Hari K Iyer. Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142(7):1980–1990, 2012.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Susan Wei, Chihoon Lee, Lindsay Wichers, and JS Marron. Direction–projection–permutation for high dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 0(just-accepted):0–0, 2015.

Johan A Westerhuis, Theodora Kourti, and John F MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301–321, 1998.

Herman Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.

Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

Svante Wold, Paul Geladi, Kim Esbensen, and Jerker Öhman. Multi-way principal components-and pls-analysis. *Journal of chemometrics*, 1(1):41–56, 1987.

Svante Wold, Nouna Kettaneh, and Kjell Tjessem. Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics*, 10(5-6):463–482, 1996.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

Bin Xu, J Louw Roos, Shawn Levy, EJ Van Rensburg, Joseph A Gogos, and Maria Karayiorgou. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature genetics*, 40(7):880–885, 2008.

L. Yang. *Transformation–Density Estimation*. PhD thesis, University of North Carolina, Chapel Hill, 1995.

Lingsong Zhang, JS Marron, Haipeng Shen, and Zhengyuan Zhu. Singular value decomposition and its visualization. *Journal of Computational and Graphical Statistics*, 16 (4):833–854, 2007.

Yu Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Ssvep recognition using common feature analysis in brain–computer interface. *Journal of neuroscience methods*, 244:8–15, 2015.

Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 2015.