Multivariate Association and Dimension Reduction

by

Ross J Iaci

(Under the direction of Xiangrong Yin and T.N. Sriram)

Abstract

In this thesis, two different nonparametric methods are developed in the statistical field of multivariate association and dimension reduction. While the underlying goal in both methods is to detect both linear and nonlinear relationships between multiple sets and groups of multivariate random vectors, different uses in statistical applications motivate the methods. The primary goal of the information theory based method of Chapter 2 is to provide an overall measure of association between sets of random vectors. In Chapter 3, a method focusing on dimension reduction is developed to extend Canonical Correlation Analysis(CCA), pioneered by Hotelling [6], to identify nonlinear relationships.

Motivated by a problem in morphological integration studies, a field in biological science, a new general index based on Kullback-Leibler(KL) information is proposed to measure the relationships between multiple sets of random vectors. The relationships are detected using a measure of the dependence between multiple sets by calculating the difference between the joint and marginal densities of affine matrix transformations of the random vectors. From this index, we define an overall measure of dependence between multiple sets, initially motivated by a problem in morphometrics. In addition, we develop two methods for dimension reduction for m-sets of random vectors and then extend these to multiple groups of multiple sets.

The second index recovers relationships between sets using a composite L_2 distance measure between linear combinations of one vector and an unknown single index model regression function of the other, interchanging the roles of each respectively. Estimates of the regression functions are calculated using the nonparametric Nadaraya and Watson [19] [27] smoother, thus enabling our index to detect both linear and nonlinear relationships. This method is then extended to identify associations between multiple sets and multiple groups of random vectors. In addition to detecting the nature of the relationships, a bootstrap procedure inspired by Ye and Weiss [32] is developed to determine the number of significant associations. Moreover, this procedure is independent of the measure used to detect the relationships.

Canonical Correlation Analysis is a common measure of the pair-wise linear association between two sets of random vectors and is often used as a benchmark for comparison. In contrast to CCA, both of our methods are shown to determine the existence of both linear and nonlinear relationships, thereby making them useful in many statistical applications.

INDEX WORDS: Information variates; Kernel density estimators; Modules; Permutation test; Dimension reduction, Canonical Correlation Analysis; Projection pursuit; Bootstrapping; Dimension reduction; Single index model.

Multivariate Association and Dimension Reduction

by

Ross J Iaci

B.S., The University of North Carolina, 1994

M.S., University of Nevada Las Vegas, 2000

M.S., The University of Georgia 2003

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2007

© 2007

Ross J Iaci

All Rights Reserved

Multivariate Association and Dimension Reduction

by

Ross J Iaci

Approved:

Major Professors: Xiangrong Yin
T.N. Sriram

Committee: Lynne Seymour
Jaxk Reeves
Nicole Lazar

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2007

Dedication

To my parents and brother.

Acknowledgments

I would first like to express my sincere appreciation for all the advice and guidance that I received from my major professors T.N. Sriram and Xiangrong Yin in the writing of this dissertation. Their instruction and patience in both this work and the classroom were invaluable in my graduate studies at UGA. I would also like to thank Dr. Lynne Seymour for her support and encouragement and the instruction I received in statistical computing, which helped in the completion of this thesis. I would also like to thank Professors Nicole Lazar and Jaxk Reeves for sincerely serving on my committee. Moreover, I would like to especially thank Professor Lazar for her constructive and insightful comments that improved this work.

I would also like to thank Professor Somnath Datta for all his time, effort and instruction throughout my studies. I would like to thank Professor Robert Lund for all that I learned in the courses he taught and all the laughs. Moreover, I would like to express may appreciation to all the faculty for their helpfulness and dedication to teaching, which made attending UGA such a great experience.

I would like to acknowledge and thank Dave Gavisk, Jesse Bowling and Jimmy Cretney for all their technological support and Connie, Loretta, Daphney and Julie for their administrative assistance. I would like to thank my friends Archan Bhattacharya, Ellen Breazel, Guoying Sun, Amy Vaughn, Nicole Ferguson and others who made UGA enjoyable.

I would like to acknowledge and thank my co-author Dr. Peter Christian Klingenberg from the University of Manchester, U.K., for providing me with an interesting application of my research. I would like to express my gratitude to Professor Ashok K. Singh and Dr. Vicky Albert from the University of Nevada Las Vegas for their support and encouragement to further my education.

Finally, I would like to thank my parents and brother for their all their support and encouragement through the years.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

The focus of this thesis is to study the number and nature of the relationships between multiple sets of multivariate random vectors. There is extensive literature on this topic, but most are restricted to identifying linear relationships. In this dissertation, we propose two nonparametric multivariate methods to study the number and nature of associations between multiple sets of random vectors. The common goal of these methods is to determine linear combinations of the random vectors that identify both linear and nonlinear relationships between the sets, often termed projection pursuit methods. Besides proposing methods to identify the nature of the relationships, we also propose procedures to detect the number of significant relationships, thereby achieving dimension reduction in a multivariate setup. The motivation and development for each of the proposed methods is different and together they provide a wide range of use in statistical applications. Before introducing the methods, we discuss related methods from the literature. Detailed descriptions of the methods can be found in the respective chapters.

Canonical Correlation Analysis(CCA), pioneered by Hotelling [2] in the 1930s, is a standard method in multivariate theory for measuring pair-wise linear associations between two sets of random vectors. CCA identifies linear relationships between two sets by determining the linear combinations of the random vectors, termed canonical variates, that have maximum correlation. There may be more than one such linear combination relating the two random vectors, with each representing a different dimension by which the sets are related. Thus, the goal of CCA is to identify the structure or dimensionality of each set that maximizes the linear relationship between them. In the subsequent decades, numerous techniques

have been proposed to generalize CCA to multiple sets. The first attempt was made by Vinograde [11]; however, for two sets of random vectors this procedure does not reduce to Hotelling's [2]. Kettenring [3] [4] proposed extensions to multiple sets, which reduce to CCA with only two sets. His approach is based on maximizing a generalized measure of correlation to detect linear relationships between the random vectors. Van der Burg and De Leeuw [9] first proposed a two-set procedure, termed nonlinear canonical analysis, to find an "optimal" scale for each multivariate variable using an alternating least squares algorithm and then extended this process to several sets (Van de Burg et al [10]). Another method to find nonlinear structures between two sets of random vectors was developed by Shi and Taam [8], who used a conditional mean and nonparametric estimation procedure to find these relationships. More recently, Luijtens et al. [7] developed a linear and nonlinear CCA type method for group-structured data. Neunenschwander and Flury [6] developed the idea of common canonical variates for multiple groups, requiring the canonical variates to have the same coefficients for all sets of variables and thus restricting to the case where all sets of random vectors have the same dimension. Estimates for these common variates are obtained via a maximum likelihood estimate derived from normal theory. Goria and Flury [1] suggested a way to extend common canonical variates to independent but closely related multiple groups of multiple sets of random vectors, with the restriction that the same common canonical variates are applicable across all groups. Their estimation of the variates is based on maximizing a likelihood function of a Wishart distribution. Recently, Yin [12] proposed a different pair-wise nonparametric method to find both linear and nonlinear relationships between two sets of random vectors using Kullback-Leibler (KL) information. Yin and Sriram [13] extended this idea to find common canonical variates for independent groups composed of two sets of random vectors.

The methods above reveal pair-wise associations between sets of random vectors but do not simultaneously measure the overall association. In Chapter 2, we propose an index to measure the overall association between multiple sets of random vectors. The need for such

a measure is motivated by the problem of determining whether sets of random vectors are modules, a problem that arises in morphological integration studies; see e.g., Klingenberg [5]. To this end, we extend Yin's [12] pair-wise method based on information theory to yield an overall joint measurement of association between multiple sets. The general population index based on KL information is

$$\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(m)}\right)=\mathrm{E}\left(\log\frac{p\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}{p\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}\right)\cdots p\left(\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}\right).$$

This index quantifies the amount of association between multiple sets by measuring the dependence of the affine matrix transformed random vectors as the difference between the respective joint and marginal densities. As a measure of dependence, the significance of the detected relationships can be tested using a sequential permutation test. Further, if these affine matrix transformations are nonsingular, the index provides an overall measure of association between the sets. These properties are exploited to develop a test for determining the existence of modules. In addition, constraints on the dimensionality of the matrices and a sequential permutation test are used to construct dimension reduction methods for multiple sets and groups of random vectors.

In Chapter 3, a method is developed with the main focus on dimension reduction, rather than measuring the overall association between sets. The method is constructed as a generalization of CCA to detect both linear and nonlinear relationships and then extended pair-wise to multiple sets and multiple groups. For this purpose we define a two-set population index as,

$$\mathcal{G}(\mathbf{a},\mathbf{b}) \;=\; \mathrm{E}(\mathbf{b}^T\mathbf{Y}-m_1(\mathbf{a}^T\mathbf{X}))^2+\mathrm{E}(\mathbf{a}^T\mathbf{X}-m_2(\mathbf{b}^T\mathbf{Y}))^2,$$

where m_1 and m_2 are unknown single index model regression functions. This index measures the correlation between $\mathbf{b}^T\mathbf{Y}$ and $m_1(\mathbf{a}^T\mathbf{X})$ and $\mathbf{a}^T\mathbf{X}$ and $m_2(\mathbf{b}^T\mathbf{Y})$; in the case where m_1 and m_2 are identity functions this measure is equivalent to CCA. In addition, a bootstrap procedure is proposed, independent of the measure used to detect the relationships, to determine the number of significant relationships.

Each chapter is self-contained in terms of describing and highlighting the performance of the above mentioned methods, but we give a concluding summary of both methods and discuss future work in Chapter 4.

1.1 References

[1] Goria, M.N. and Flury, B.D. (1996). Common canonical variates in k independent groups. *Journal of the American Statistical Association*, Vol. 91, No. 436, 1735-42.

[2] Hotelling, H. (1936). Relations between two sets of variables, *Biometrika*, Vol. 28, No. 3/4., 321-77.

[3] Kettenring, J.R. (1971). Canonical correlation analysis of several sets of variables. *Biometrika*, 58, No. 3, 433-51.

[4] Kettenring, J.R (1985). Canonical correlation analysis. *Encyclopedia of statistical sciences*, 1 ED. S. Kotz and N.L. Johnson, pp. 354-65. New York: John Wiley.

[5] Klingenberg, C.P. (2005). Developmental constraints, modules and evolvability. Pages 219-47 in Variation: A Central Concept in Biology (B. Hallgrimsson and B.K. Hall, eds.). Elsevier, Burlington, MA.

[6] Neuenschwander, B.E. and Flury, B.D. (1995). Common canonical variates. *Biometrika*, 82, No. 3, 553-60.

[7] Luijtens, K., Symons, F. and Vuylsteke-Wauters, M. (1994). Linear and non-linear canonical correlation analysis: and exploratory tool for the analysis of group-structured data. *Journal of Applied Statistics*, V. 21, No. 3, 43-61.

[8] Shi, S.G., Taam, W. (1992). Non-linear canonical correlation analysis with a simulated annealing solution. *Journal of Applied Statistis*, V. 19, No. 1, 155-65.

[9] Van de Burg, E. and De Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54-80.

[10] Van de Burg, E. De Leeuw, J. Verdegaal, R. (1988). Non-linear canonical correlation with m sets of variables. *Psychometrika*, 2, 171-97.

[11] Vinograde, B. (1950). Canonical positive definite matrices under internal linear transformations. *Proceedings of the American Mathematical Society*, Vol. 1, No. 2, 159-61.

[12] Yin, X. (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91, 161-76.

[13] Yin, X. and Sriram, T. N. (2006). Common canonical variates for independent groups using information theory. *Statistica Sinica*, to appear.

Chapter 2

An Informational Measure of Association and Dimension Reduction for Multiple Sets and Groups with Applications in Morphometric Analysis[1]

[1]Iaci, R., Yin, X., Sriram, T. N. and Klingenberg, C.P. Submitted to: *Journal of the American Statistical Association,* 12/06.

Abstract

This paper proposes a new general index which measures relationships between multiple sets of random vectors. The index is based on Kullback-Leibler information, which measures linear or nonlinear dependence between the multiple sets using joint and marginal densities of affine matrix transformations of the random vectors. Estimates of the matrices are obtained by maximizing the Kullback-Leibler information and are shown to be consistent. The motivation for introducing such an index comes from morphological integration studies, a topic in biological science. From this index, we define an overall measure of association and two others for dimension reduction. All these measures are illustrated through data sets and extensive simulations, and compared to those based on canonical correlation analysis. Extensions of the aforementioned measures to multiple groups are also discussed. In contrast to canonical correlation analysis, our general index not only provides an overall measure of association but also determines nonlinear relationships, thereby making it useful in many other applications.

Key Words and Phrases: Information variates; Kernel density estimators; Modules; Permutation test; Dimension reduction.

2.1 Introduction

In this paper, we propose a new general index which measures the association between multiple sets of multivariate random vectors. The need for an overall measure of association arises in morphological integration studies; see e.g., Klingenberg [8]. An important task of morphometric research is to determine whether a structure is a single integrated unit or one that consists of several distinct sets of random vectors, which requires an appropriate overall measure of association; see Section 2.2 for more details. The index is based on KL information, which measures linear or nonlinear dependence between the sets using joint and marginal

densities of affine matrix transformations of the random vectors. Restricting to nonsingular transformations provides an overall measure of association and is used to address the problem in morphometrics. In general, these matrices may not necessarily provide a parsimonious summary of existing relationships if they project onto subspaces of dimension greater than three, where the relationships can not be visualized. Thus, with certain restrictions on the dimensionality of the matrices, we search in succession for the coefficient vectors that identify linear or nonlinear relationships. A permutation test is used to test the significance of these projected random vectors, thereby enabling our index to be used as a dimension reduction method.

As mentioned before, CCA identifies linear relationships between two sets by determining the linear combinations of the random vectors, termed canonical variates, that have maximum correlation. Specifically, suppose $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{q\times 1}$ are multivariate random vectors with zero mean and covariance matrix $\mathbf{\Sigma_{XY}}$. Consider the projections $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ onto a one-dimensional subspace, the goal of CCA is to determine these coefficient vectors such that

$$\rho(\mathbf{a},\mathbf{b}) = \rho(\mathbf{a}^T\mathbf{X},\mathbf{b}^T\mathbf{Y}) = \frac{\mathbf{a}^T\mathbf{\Sigma_{XY}}\mathbf{b}}{\left(\mathbf{a}^T\mathbf{\Sigma_X}\mathbf{a}\right)^{1/2}\left(\mathbf{b}^T\mathbf{\Sigma_Y}\mathbf{b}\right)^{1/2}}$$

is maximum. Finding the coefficient vectors that maximize $\rho(\mathbf{a},\mathbf{b})$, termed the canonical correlation, is equivalent to solving the problem:

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^T\mathbf{\Sigma_{XY}}\mathbf{b} \quad \text{subject to the constraints} \quad \mathbf{a}^T\mathbf{\Sigma_X}\mathbf{a} = \mathbf{b}^T\mathbf{\Sigma_Y}\mathbf{b} = 1.$$

After the first coefficient vectors, $\mathbf{a}_1$ and $\mathbf{b}_1$, have been found the next coefficient vectors are found in an orthogonal or uncorrelated direction. That is, the second coefficient vectors, $\mathbf{a}_2$ and $\mathbf{b}_2$, have the next maximum correlation subject to the constraints: $\mathbf{a}_2^T\mathbf{\Sigma_X}\mathbf{a}_2^T = \mathbf{b}_2\mathbf{\Sigma_Y}\mathbf{b}_2 = 1$(unit variance) and $\mathbf{a}_2^T\mathbf{\Sigma_X}\mathbf{a}_1 = \mathbf{b}_2^T\mathbf{\Sigma_Y}\mathbf{b}_1 = 0$(uncorrelated). This process can be repeated up to the minimum dimension of $\mathbf{X}$ and $\mathbf{Y}$. Large sample tests have been developed to test the significance of the canonical variates; thus, making this a dimension reduction method. Note that, CCA can only detect linear relationships. Below we propose an index to detect

both linear and nonlinear relationships and test the significance of the associations, making our approach a viable alternative to CCA.

In Section 2.2, we give a more detailed motivation for an overall measure of association. In Section 2.3.1, we introduce the general index based on KL information and study basic properties in Section 2.3.2. A computational algorithm for the general index is given in Section 2.3.3 and component-wise calculations are described in Section 2.4.4. In Section 2.3.4, we state the main Theorem on the consistency of the estimated matrices obtained by maximizing the KL information. In Section 2.3.5, we introduce a permutation test based on the general index. In Section 2.4, we introduce an overall measure for detecting modules in the context of morphological integration studies, develop dimension reduction methods for special cases of the general index, and discuss extensions to multiple groups. Simulations and analysis of three data sets (fly wing, mouse mandible and water strider) illustrating our methods are presented in Section 2.5. Concluding remarks are given in Section 2.6 and proofs are given in the Appendix.

2.2 Motivation

Many morphological structures(e.g., Figure 2.1) are composed of multiple key component parts which are integrated to function as a whole. Nevertheless, the component parts of such structures may have a certain degree of independence from one another. In other words, in a morphological structure, there may be multiple sets (called *modules*) consisting of component parts such that each module is integrated internally by strong interactions among its component parts, but has a degree of independence from other modules because interactions between them are fewer or weaker. Modules are therefore units of a system that are defined by the degree of connectivity. For example, in the left panel of Figure 2.1, a fly wing with 15 landmarks partitioned into two sets by a dashed line, termed *subdivisions/partitions*, creates two possible modules. This coordination into subunits has long been known as morphological integration and has become the focus of renewed interest in evolutionary developmental

biology under the heading of modularity(Klingenberg [8]). The study of modularity is not only restricted to analyzing the association between two sets but may also involve multiple sets, created by multiple partitions as in the right panel of Figure 2.1. As explained above, an important task of morphometric research is to determine whether a structure is a single integrated unit or consists of several distinct modules. A formal statistical study of determining whether a structure is a single unit or one that consists of distinct modules would enable researchers to gain more information about a morphological structure.

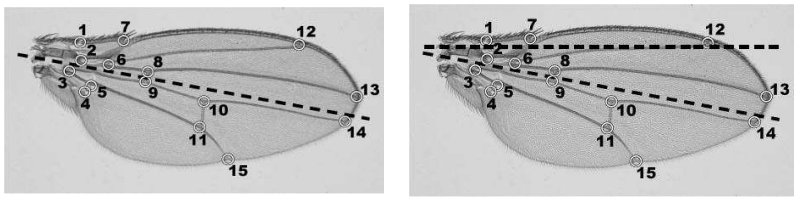


Figure 2.1: Fly wing - 1 partition - two modules; 2 partitions - three modules

The subdivisions create m-sets of vectors defined by $\mathbf{X}^{(1)}_{p_1\times 1}$, $\mathbf{X}^{(2)}_{p_2\times 1}$, $\mathbf{X}^{(3)}_{p_3\times 1}, \ldots, \mathbf{X}^{(m)}_{p_m\times 1}$, with dimensions p_k, $k = 1, \ldots, m$. A target set of modules is usually stated as a hypothesis and tested using an overall measure of association. In the literature, methods based on trace and multi-set trace correlation have been proposed for the purpose of measuring "overall" association between m-sets, which are equivalent to averaging the squared canonical correlations. For example, another trace correlation method(not described here) known as the RV coefficient [12] has been recently suggested as an overall measure of association. However, maximizing this measure under the standard CCA constraints is also equivalent to averaging the squared canonical correlations. The RV coefficient computed without any constraints results in a measure that is not scale invariant, which is not desirable. The following scale invariant overall measure of association has been proposed by Klingenberg

$$R^2_{TM} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} R^2_{T2}\left(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}\right), \tag{2.1}$$

where $R^2_{T2} = \operatorname{tr}\left(\Sigma_{ii}^{-1}\Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}\right) / \min(p_i, p_j) = \sum \lambda_k / \min(p_i, p_j)$, Σ_{ij} is the covariance matrix of $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$, and the λ_k's are eigenvalues of $\Sigma_{ii}^{-1}\Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}$. Note that R^2_{T2} is the

overall measure of association between two sets obtained by averaging the squared canonical correlations.

The above measure may include insignificant canonical correlations, adding bias to the statistic. Furthermore, this measure inherently cannot detect nonlinear relationships. In addition, suppose there are three sets, where the marginal association between $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ is weak and there is a joint relationship, say, $\mathbf{X}^{(1)} = \mathbf{X}^{(2)} + \mathbf{X}^{(3)} + \epsilon$. Then, $R^2_{T2}\left(\mathbf{X}^{(2)}, \mathbf{X}^{(3)}\right)$ only adds more bias to the overall measure R^2_{TM}. In using R^2_{TM}, not only do the relationships need to be linear but also marginal (pair-wise). The same criticism pointed out here extends to the use of the RV measure cited above.

In the next Section, we propose a new general measure of association which overcomes the drawbacks with the above mentioned methods. In Section 5, we make an extensive comparison of our overall measure with R^2_{TM}. Henceforth, dimension of the random vectors will be suppressed in the notation, unless needed for clarity.

2.3 A General index

In this Section, we propose a general index which can measure the overall association between multiple sets of random vectors defined in Section 2.2. We study the properties of this index, give a computational algorithm, state a consistency theorem, and formulate a permutation test of the dependence between affine transformed random vectors.

2.3.1 Definition

We define the following general information index

$$\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) = \mathrm{E}\left(\log \frac{p\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}{p\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}\right) \cdots p\left(\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}\right), \tag{2.2}$$

where $p(\cdot, \ldots, \cdot)$ and $p(\cdot)$ are the joint and marginal densities of the transformed vectors, respectively, and the $\mathbf{A}^{(k)} = \left[a_1^{(k)}, a_2^{(k)} \cdots, a_{p_k^*}^{(k)}\right]$ are $p_k \times p_k^*$ matrices for some $p_k^* \leq p_k$,

$k = 1, \cdots, m$. Here, p_k^* determines the dimension of the subspace onto which $\mathbf{A}^{(k)}$ projects the k^{th} random vector and thus, cannot exceed the dimension p_k. This measure extends the work of Yin [18] to multiple sets with matrix coefficients. The information index $\mathcal{I}^{(\text{G})}$ measures the dependence between the m-sets via the KL divergence of the joint and product of the marginal densities of the transformed vectors. Subject to the constraints $\mathbf{A}^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{A}^{(k)} = \mathbf{I}_{p_k^*}$ for all k, we find the coefficient matrices such that the projected random vectors have the largest dependence by maximizing $\mathcal{I}^{(\text{G})}$ with respect to $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$. If the $\mathbf{A}^{(k)}$ are nonsingular for all k, then $\mathcal{I}^{(\text{G})}$ extracts the full information from the random vectors; thus, providing an overall measure of association between the m-sets. By Proposition 3 stated below, the constraints imposed are not necessary, but are the same as the CCA constraints stated in Section 2.1 and help to reduce the complexity in estimating the coefficient matrices in practice. The ability of our index to measure the amount of association between m-sets in any desired dimension p_k^*, distinguishes our index from others available in the literature.

2.3.2 Properties

For the general index defined in the previous Section, we establish some basic properties, the first of which gives the invariance of the information index under affine transformation.

Proposition 1 *Let* $\mathbf{U}^{(k)} = \mathbf{B}^{(k)^{-1}}\mathbf{X}^{(k)} + b_k$, $k = 1, \ldots, m$, *where* $\mathbf{B}^{(k)}$ *are nonsingular matrices with appropriate dimension, and* b_k *is a fixed* $p_k \times 1$ *vector. Then the following hold:*

1. $\mathcal{I}^{(\text{G})}_{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(m)}}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) = \mathcal{I}^{(\text{G})}_{\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(m)}}\left(\mathbf{B}^{(1)^T}\mathbf{A}^{(1)}, \ldots, \mathbf{B}^{(m)^T}\mathbf{A}^{(m)}\right)$, *where* $\mathcal{I}^{(\text{G})}_{\mathbf{Y}^{(1)},\ldots,\mathbf{Y}^{(m)}}$ *denotes the information index for a m-set* $\{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(m)}\}$.

2. *The information matrices for* $\mathbf{U}^{(k)}$ *are given by* $\mathbf{B}^{(k)^T}\mathbf{A}^{(k)}$, *for all* k.

Proposition 2 *The following hold for the information measure:*

1. $\mathcal{I}^{(\text{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) \geq 0$, *for all* $\mathbf{A}^{(k)}, k = 1, \ldots, m$.

2. $\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(m)}\right)=0$ *if and only if* $\mathbf{A}^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}$ *are mutually independent.*

Proposition 3 *For nonsingular matrices* $\mathbf{C}^{(k)}$, $k=1,\cdots,m$, *we have*

$$\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}\mathbf{C}^{(1)},\ldots,\mathbf{A}^{(m)}\mathbf{C}^{(m)}\right)=\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(m)}\right).$$

Proposition 4 *Suppose* $(\mathbf{X}^{(1)},\mathbf{X}^{(2)})^T$ *is multivariate normal with zero mean, and* $\mathbf{A}^{(1)}_{p_1\times t}$ *and* $\mathbf{A}^{(2)}_{p_2\times s}$ *are matrices with ranks* t *and* s. *Then,* $\mathbf{Y}=(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)},\mathbf{A}^{(2)^T}\mathbf{X}^{(2)})^T\sim N_{(t+s)}\left(\mathbf{0},\Delta\right)$ *and*

$$\begin{aligned}\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)},\mathbf{A}^{(2)}\right) &= -\frac{1}{2}\log\left|\mathbf{I}_{s\times s}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}\Delta_{22}^{-1}\right| \\ &= -\frac{1}{2}\log\left|\mathbf{I}_{t\times t}-\Delta_{12}\Delta_{22}^{-1}\Delta_{21}\Delta_{11}^{-1}\right|.\end{aligned}$$

Proposition 2 gives a necessary and sufficient condition for $\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}$, …, $\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}$ to be mutually independent, which holds without a multivariate normality assumption. Proposition 3 is used for deriving our overall measure of association in Section 2.4.1. Proposition 4 is used to compare our index to those proposed in the literature, specifically Kettenring's [4] methods. This comparison is made in Section 2.4.2. Proofs of the above propositions are given in Appendix 2.7.3.

2.3.3 Computational algorithm

We use the sample version of (2.2) to find the estimated coefficient matrices $\widehat{\mathbf{A}}^{(1)}$, …, $\widehat{\mathbf{A}}^{(m)}$. First we transform the variables to have identity covariance matrices, which simplifies the constraints, lessens the effect of ill-conditioned covariance matrices and rescales the variables to have equivalent magnitude. Note that the scale transformation does not affect or change the relationships between the sets, as shown in Proposition 1.

Step 1 : Construct the following kernel density estimate:

$$\widehat{p}(u_1,u_2,\ldots,u_l)=\frac{1}{nh_1h_2\cdots h_l}\sum_{i=1}^{n}K\left(\frac{u_1-u_{i1}}{h_1}\right)K\left(\frac{u_2-u_{i1}}{h_2}\right)\cdots K\left(\frac{u_l-u_{i1}}{h_l}\right)\text{for},u\in\mathbb{R}^l.$$

As suggested by Scott [15] and Silverman [16], we use Gaussian product kernels to estimate $p(u_1, u_2, \ldots, u_l)$. However, our method holds for any kernel of bounded variation. The terms $h_j = (4/(l+2))^{1/(l+4)} s_j n_j^{-1/(l+4)}$, for $j = 1, 2, \ldots, l$, are calculated with the corresponding standard deviations, $s_1, s_2, \ldots, s_l$ of $u_1, u_2, \ldots, u_l$. This leads to the following sample information index

$$\mathcal{I}_n^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) = \frac{1}{n}\sum_{j=1}^{n} \log\left(\frac{\widehat{p}\left(\mathbf{A}^{(1)^T}\mathbf{x}_j^{(1)}, \ldots, \mathbf{A}^{(m)^T}\mathbf{x}_j^{(m)}\right)}{\widehat{p}\left(\mathbf{A}^{(1)^T}\mathbf{x}_j^{(1)}\right)\cdots\widehat{p}\left(\mathbf{A}^{(m)^T}\mathbf{x}_j^{(m)}\right)}\right), \tag{2.3}$$

where $\mathbf{x}_j^{(k)}$ denotes the j^{th} sample from $\mathbf{X}^{(k)}$.

Step 2 : Under the orthonormal constraint imposed on the columns of $\mathbf{A}^{(k)}$, obtain

$$\left(\widehat{\mathbf{A}}^{(1)}, \ldots, \widehat{\mathbf{A}}^{(m)}\right) = \arg\max_{\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(m)}} \mathcal{I}_n^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right). \tag{2.4}$$

The maximization is carried out by iteration using the constrained minimizer (maximizer) *fmincon* of Matlab, which implements a Sequential Quadratic Programming (SQP) method. The SQP method maximizes the index $\mathcal{I}_n^{(\mathrm{G})}$ while incorporating the constraints simultaneously. Codes in Matlab have been developed for our computations and are available from the authors.

2.3.4 Consistency

For ease of presentation, the consistency result is stated for three sets; however, the result holds for m-sets. Let $\mathbf{X}_{p\times 1}, \mathbf{Y}_{q\times 1}$ and $\mathbf{Z}_{t\times 1}$ denote the three sets and $\mathbf{A}_n, \mathbf{B}_n$ and $\mathbf{C}_n$ denote the estimated coefficient matrices with dimensions $p \times r_1, q \times r_2$ and $t \times r_3$, respectively, with $r_1 \leq p$, $r_2 \leq q$ and $r_3 \leq t$. Let

$$\chi_b = \Big\{ i : p\left(\mathbf{A}^T\mathbf{x}_i\right) > b, p\left(\mathbf{B}^T\mathbf{y}_i\right) > b, p\left(\mathbf{C}^T\mathbf{z}_i\right) > b, \text{ and } p\left(\mathbf{A}^T\mathbf{x}_i, \mathbf{B}^T\mathbf{y}_i, \mathbf{C}^T\mathbf{z}_i\right) > b$$
$$\forall\ \mathbf{A}, \mathbf{B}, \mathbf{C} \text{ subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I}, \mathbf{B}^T\mathbf{B} = \mathbf{I} \text{ and } \mathbf{C}^T\mathbf{C} = \mathbf{I} \Big\},$$

for some b to be chosen later in Appendix 2.7.2. Also, let n_b be the number of observations whose indices are not in χ_b. We have the following result.

Theorem 1 (Consistency) *Assume the conditions of Lemma 1 in Appendix 2.7.2 and that* $n_b/n \xrightarrow{p} 0$ *as* $n \to \infty$. *Let* $(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) = \operatorname{argmax} \mathcal{I}_n^b \left(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}\right)$ *and* $(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \operatorname{argmax} \mathcal{I} \left(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}\right)$, *where*

$$\mathcal{I}_n^b (\mathbf{A}, \mathbf{B}, \mathbf{C}) = \frac{1}{n} \sum_{i=1}^{n} J (i \in \chi_b) \log\left(\frac{p_n \left(\mathbf{A}^T \mathbf{x}_i, \mathbf{B}^T \mathbf{y}_i, \mathbf{C}^T \mathbf{z}_i\right)}{p_n (\mathbf{A}^T \mathbf{x}_i)\, p_n (\mathbf{B}^T \mathbf{y}_i)\, p_n (\mathbf{C}^T \mathbf{z}_i)} \right),$$

with p_n *as defined in Appendix 2.7.2 and* $J (i \in \chi_b)$ *is the indicator function for* χ_b. *Then,*

$$(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) \to (\mathbf{A}, \mathbf{B}, \mathbf{C}) \;\; \textit{as } n \to \infty, \quad \textit{with probability 1.}$$

2.3.5 Permutation test

An advantage in using the general index $\mathcal{I}^{(\mathrm{G})}$ is that we can test the independence of the affine transformed random vectors in any desired dimension without a parametric assumption because by Proposition 2

$$\mathcal{I}^{(\mathrm{G})} = 0 \text{ if and only if } \mathbf{A}^{(1)^T} \mathbf{X}^{(1)}, \ldots, \mathbf{A}^{(m)^T} \mathbf{X}^{(m)} \text{ are mutually independent.}$$

As a result, we formulate a permutation test of

$$H_0 : \; \mathcal{I}^{(\mathrm{G})} = 0 \;\; \text{vs} \;\; H_1 : \; \mathcal{I}^{(\mathrm{G})} > 0.$$

Under H_0, a permutation of the rows of the data matrices corresponding to $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots,$ $\mathbf{X}^{(m)}$ should preserve the independence. If the null is false, then a permutation of the rows will destroy the dependence relationship and produce a smaller value of the information index. Let $\mathcal{I}^{(\mathrm{G})*}$ be the information index for a permutation of the rows of the data matrices. Then the p-value for the permutation test is given by

$$\text{p-value} = \left(\frac{\#\; \mathcal{I}^{(\mathrm{G})*} > \mathcal{I}^{(\mathrm{G})}}{\#\text{ of permutations}} \right).$$

In our numerical studies, we set the number of permutations to be 1000 and use the estimated information index to calculate the p-value. Any bias introduced by density estimation is nullified by using the same kernel and bandwidth during each permutation.

2.4 Modularity and Dimension Reduction

In this Section, we show how the general measure defined in Section 2.3.1 can be used to test for the presence of modules in the study of morphological structures described in Section 2.2. Furthermore, we exploit the measure to visualize relationships and discuss extensions to multiple groups.

2.4.1 Modularity

As described in Section 2.2, the subdivisions of landmarks in a morphological structure denote m-sets of random vectors. Also, as mentioned in Section 2.3.1, if the $\mathbf{A}^{(k)}$ are non-singular for all k, then $\mathcal{I}^{(\mathrm{G})}$ provides an overall measure of association between the m-sets. Letting $\mathbf{C}^{(k)} = \mathbf{A}^{(k)^{-1}}$ in Proposition 3 gives $\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) = \mathcal{I}^{(\mathrm{G})}\left(\mathbf{I}_{p_1}, \ldots, \mathbf{I}_{p_k}\right)$. This means that the maximum value of the information index obtained by any rotation in the full dimension is equal in value to the index evaluated at the original non-transformed vectors. This provides a significant reduction in computation since the matrix maximization step is eliminated. As a result of this identity, we will denote the overall measure of association between the m-sets as

$$\mathcal{I}^{(\mathrm{O})} = \mathcal{I}^{(\mathrm{G})}\left(\mathbf{I}_{p_1}, \ldots, \mathbf{I}_{p_k}\right). \tag{2.5}$$

For discrete random vectors, Joe [3] discussed $\mathcal{I}^{(\mathrm{O})}$ and referred to it as a relative entropy measure. We state the following hypotheses to test whether the m-sets form modules.

$$\begin{aligned} H_0 &: \ \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)} \ \text{ are modules } \ \text{vs} \\ H_1 &: \ \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)} \ \text{are not modules.} \end{aligned}$$

Equivalently, interpreting modules as independent sets, we test

$$H_0 : \ \mathcal{I}^{(\mathrm{O})} = 0 \ \ \text{vs} \ \ H_1 : \ \mathcal{I}^{(\mathrm{O})} > 0.$$

The above hypotheses is tested using the permutation test defined in Section 2.3.5 with $\mathcal{I}^{(\mathrm{O})}$ replacing $\mathcal{I}^{(\mathrm{G})}$.

A different testing procedure for modularity has been suggested that involves finding every possible subdivision/partition that create modules of the same dimension. This procedure is referred to as the Partition test. For simplicity, we describe this testing method for two target modules: $\mathbf{X}^{(1)}_{p_1\times 1}$ and $\mathbf{X}^{(2)}_{p_2\times 1}$($p_1$ and p_2 are the number of landmarks). There are $w = \binom{p_1+p_2}{\min(p_1,p_2)}$ possible partitions that create sets of the same dimension. A new partition is equivalent to interchanging an equal number of elements of $\mathbf{X}^{(1)}_{p_1\times 1}$ and $\mathbf{X}^{(2)}_{p_2\times 1}$. For each partitioned set calculate the test statistic $\mathcal{I}^{(\mathrm{O})*}$ as above, and define the percentile rank or "partitioned" p-value: p-value $= \left(\frac{\#\ \mathcal{I}^{(\mathrm{O})*}>\mathcal{I}^{(\mathrm{O})}}{w-1}\right)$, where $\mathcal{I}^{(\mathrm{O})}$ is the value of the test statistic of the target modules. Note that, at the significance level α, the null hypothesis is accepted if the p-value exceeds $1-\alpha$.

In our opinion this procedure is not a test of modularity, but is better used as a data mining tool to identify the partitions that are most likely to correspond to modules. That is, the method tests whether a specific subdivision is among the least dependent of all possible subdivisions but does not test whether these partitioned sets are in fact modules. For example, assume all sets corresponding to all possible subdivisions are dependent(not modules), then the partition test can identify which is the least dependent, but to conclude that these sets are in fact modules would be incorrect. The partition test was performed in all the simulation studies and real data analysis performed in section 2.5 and, consistent with these comments, produced unreliable results. For these reasons the results using the partition test have been omitted in the analysis.

2.4.2 Dimension reduction measures

Suppose we conclude based on the test of hypotheses in Section 2.3.5 that the m-sets are not independent. Then, the general index provides an overall measure of dependence but the coefficient matrices do not necessarily provide a parsimonious summary describing all existing relationships between the sets. To this end, we will search for the orthogonal coefficient vectors $(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)})$ successively such that $(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{a}^{(m)^T}\mathbf{X}^{(m)})$ have the largest

amount of information. For this purpose, we propose two measures, $\mathcal{I}^{(\mathrm{J})}$ and $\mathcal{I}^{(\mathrm{P})}$, by setting the matrices $\mathbf{A}^{(k)}$ in (2.2) to $p_k \times 1$ vectors denoted by $\mathbf{a}^{(k)}$. Let

$$\begin{aligned}\mathcal{I}^{(\mathrm{J})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}\right) &= \mathcal{I}^{(\mathrm{G})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}\right)\\ \mathcal{I}^{(\mathrm{P})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}\right) &= \sum_{1\le i<j\le m}\mathcal{I}^{(\mathrm{G})}\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right).\end{aligned}$$

We find the coefficient vectors $\mathbf{a}_i^{(1)},\ldots,\mathbf{a}_i^{(m)}$ by maximizing the above measures, subject to the constraints,

$$\mathbf{a}_i^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{a}_i^{(k)} = 1 \text{ for all } i = 1,\ldots,\min(p_k) \text{ and } \mathbf{a}_i^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{a}_j^{(k)} = 0 \text{ for all } j = 1,\ldots,i-1.$$

That is, we constrain the information variates to have unit variance and be uncorrelated for all sets. $\mathcal{I}^{(\mathrm{J})}$ is a joint measure of association between m-sets, while $\mathcal{I}^{(\mathrm{P})}$ is the sum of all possible two-set associations. Naturally, when the relationships between the sets are joint, e.g. in the case of three sets, $y_1 = x_1 + z_1^2 + \epsilon$, then $\mathcal{I}^{(\mathrm{J})}$ is a preferred measure over $\mathcal{I}^{(\mathrm{P})}$. If the relationships are marginal, e.g., $y_1 = x_1 + \epsilon$ and $z_1 = x_2^2 + \epsilon$, then $\mathcal{I}^{(\mathrm{P})}$ is preferred. However, it is shown in simulation 2.5.3 that the difference between the measures is negligible.

Let $\mathcal{I}_i$ (either $\mathcal{I}_i^{(\mathrm{J})}$ or $\mathcal{I}_i^{(\mathrm{P})}$) denote the index corresponding to the i^{th} coefficient vector $(\mathbf{a}_i^{(1)},\ldots,\mathbf{a}_i^{(m)})$. It can be shown using results in Yin [18] that $\mathcal{I}_i \ge \mathcal{I}_j$ for $i \ge j$. Also, by Proposition 2, $\mathcal{I}_i = 0$ if and only if $\mathbf{a}_i^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{a}_i^{(m)^T}\mathbf{X}^{(m)}$ are independent or pairwise independent, according to whether $\mathcal{I}_i$ is $\mathcal{I}_i^{(\mathrm{J})}$ or $\mathcal{I}_i^{(\mathrm{P})}$. Therefore, we can successively test $H_0 : \mathcal{I}_i = 0$ vs $H_1 : \mathcal{I}_i > 0$ using the permutation test described in Section 2.3.5 to determine the number of significant coefficient vectors, and this is referred to as a dimension reduction method. In this case, $\mathbf{A}^{(k)}$ is a vector, hence the maximization is made more simple. An alternative, subjective approach, is to plot the variate pairs to visually determine the number of significant variates.

In general, the information values are much harder to interpret than the canonical correlations because the former are more general measures of dependency. One way to look at this new measure is to consider the simple case of multivariate normal $(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\mathbf{X}^{(3)},\ldots,\mathbf{X}^{(m)})$

vectors with zero mean. Then it can be shown using Proposition 4 that,

$$\begin{aligned}\mathcal{I}^{(\mathrm{P})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}\right) &= -\frac{1}{2}\sum_{1\le i<j\le m}\log\left[1-\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)\right]\\ &= -\frac{1}{2}\log\prod_{1\le i<j\le m}\left[1-\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)\right],\end{aligned}$$

where $\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)$ is the correlation coefficient between ${\mathbf{a}^{(i)}}^T\mathbf{X}^{(i)}$ and ${\mathbf{a}^{(j)}}^T\mathbf{X}^{(j)}$. Maximizing our index $\mathcal{I}^{(\mathrm{P})}$ is then equivalent to minimizing $\log\prod_{1\le i<j\le m}\left[1-\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)\right]$. If we define Σ_m to be the correlation coefficient matrix of $({\mathbf{a}^{(1)}}^T\mathbf{X}^{(1)},\ldots,{\mathbf{a}^{(m)}}^T\mathbf{X}^{(m)})^T$ and assume Σ_m is positive definite, then $|\Sigma_m|\le\left[1-\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)\right]$ for any $1\le i<j\le m$; see Appendix 2.7.1. Thus,

$$|\Sigma_m|^{m(m-1)/2}\le\prod_{1\le i<j\le m}\left[1-\rho_{ij}^2\left(\mathbf{a}^{(i)},\mathbf{a}^{(j)}\right)\right].$$

One of the methods (generalized variance method) in Kettenring [4] is equivalent to maximizing the left hand side in the above inequality. In fact, our method is not equivalent to any of the five methods studied by Kettenring [4], even under a multivariate normal assumption.

2.4.3 Multiple groups

Suppose the data on m-sets are collected from g independent but closely related groups and one is interested in measuring the relationship between m-sets within each group. Let $\mathbf{X}^{(1)w},\mathbf{X}^{(2)w},\ldots,\mathbf{X}^{(m)w}$ denote the m-sets in group w, $w=1,\ldots,g$. The general information index for g groups is then defined as $\mathcal{I}^{(\mathrm{K})\mathrm{C}}=\sum_{w=1}^{g}\mathcal{I}^{(\mathrm{K})w}\mathrm{P}\,(W=w)$, where $\mathcal{I}^{(\mathrm{K})w}$ is $\mathcal{I}^{(\mathrm{K})}$ for group w with K = G, J or P. We weight the information for each group by $\mathrm{P}\,(W=w)$, which is commonly taken to be $n_w/(n_1+n_2+\cdots+n_g)$, where n_w is the number of observations in the w^{th} group. As before, we maximize the index $\mathcal{I}^{(\mathrm{K})\mathrm{C}}$ with respect to the matrices or vectors; however, the constraints may be imposed using the pooled covariance matrices or group-specific covariance matrices. Once again, it is possible to establish the properties of $\mathcal{I}^{(\mathrm{K})\mathrm{C}}$. Furthermore, we can determine whether the common information variates are significant by extending the permutation test in Section 2.3.5 (or Section 2.4.2), where permutations are carried out in each group separately. The case $m=2$, is discussed in Yin and Sriram [19].

2.4.4 Dimension reduction measures computational algorithm

For m-sets of random vectors, the dimension reduction measure $\mathcal{I}^{(\mathrm{J})}\left(\mathbf{a}^{(i)},\ldots,\mathbf{a}^{(m)}\right)$ of Section 2.4.2, is simply the general index $\mathcal{I}^{(\mathrm{G})}$, with the matrices replaced by vectors. The same product kernel density estimators are used in the computational algorithm in Section 2.3.3 to obtain the sample version,

$$\mathcal{I}_n^{(\mathrm{J})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}\right) = \frac{1}{n}\sum_{k=1}^{n}\log\left(\frac{\widehat{p}\left(\mathbf{a}^{(1)^T}\mathbf{x}_k^{(1)},\ldots,\mathbf{a}^{(m)^T}\mathbf{x}_k^{(m)}\right)}{\widehat{p}\left(\mathbf{a}^{(1)^T}\mathbf{x}_k^{(1)}\right)\cdots\widehat{p}\left(\mathbf{a}^{(m)^T}\mathbf{x}_k^{(m)}\right)}\right).$$

The first coefficient vectors are found using the general index computational algorithm; however, finding successive coefficient vectors in orthogonal directions cannot be accomplished simultaneously as with the general index. One possible algorithm to find successive coefficient vectors is discussed below.

For the dimension reduction measures of Section 2.4.2, $\mathcal{I}^{(\mathrm{P})}$ is the sum of all possible two-set combinations of $\mathcal{I}^{(\mathrm{J})}$ and the common group index , $\mathcal{I}^{(\mathrm{K})\mathrm{C}}$ of Section 2.4.3, is a weighted sum over all groups using $\mathcal{I}^{(\mathrm{J})}$ or $\mathcal{I}^{(\mathrm{P})}$. Therefore, we detail the computational algorithm for the m-set population version $\mathcal{I}^{(\mathrm{J})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(k)}\right)$, noting that any of the measures are simply score functions to be maximized. In addition, since our measures are invariant under nonsingular matrix transformation, we develop the algorithm assuming that the random vectors have been "whitened" to have unit covariance matrices,

$$\mathbf{Z}^{(k)} = \Sigma_{\mathbf{X}^{(k)}}^{-1/2}\left(\mathbf{X}^{(k)}\right) \Longrightarrow \mathrm{cov}\left(\mathbf{Z}^{(k)}\right) = \mathbf{I}.$$

This transformation of the vectors simplifies the constraints and lessens the effects of ill-conditioned covariance matrices and variables of differing magnitudes. The algorithm is detailed for one iteration, finding the 1st and 2nd coefficient vectors, $\mathbf{a}_1^{(k)}$ and $\mathbf{a}_2^{(k)}$.

Step 0: Find the 1st coefficient vectors $\mathbf{a}_1^{(k)}$ that maximize $\mathcal{I}^{(\mathrm{J})}\left(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(k)}\right)$ for all k, subject to the constraint, $\mathbf{a}_1^{(k)^T}\mathbf{a}_1^{(k)} = 1$ for all k.

The following steps outline an algorithm to find $\mathbf{a}_2^{(k)}$ orthogonal to $\mathbf{a}_1^{(k)}$. For simplicity, we will fix k and denote $\mathbf{a}_1 = \mathbf{a}_1^{(k)}$, $\mathbf{a}_2 = \mathbf{a}_2^{(k)}$ and $p = p_k$, with the understanding that the algorithm is carried out simultaneously for all $k = 1, \ldots, m$. Let $\mathbf{D}_{n\times p} = \mathbf{D}_{\mathbf{x}^{(k)}}$ be the data matrix for $\mathbf{X}^{(k)}$.

Step 1: Define $\mathbf{A} = \mathbf{a}_1\mathbf{a}_1^T$ and singular value decompose $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{\Sigma}$ is the diagonal matrix of singular values. Using the left singular vectors, let

$\mathbf{B}_{p\times p} = [\mathbf{a}_1\ \mathbf{u}_2 \ldots \mathbf{u}_p] = [\mathbf{u}_1\ \ldots \mathbf{u}_p]$.

Step 2: Define $\mathbf{B}^*_{p\times(p-1)} = [\mathbf{u}_2 \ldots \mathbf{u}_p]$ and project the data matrix $\mathbf{D}$ onto this $\perp$ subspace:

$$\mathbf{D}^*_{n\times(p-1)} = \mathbf{D}_{n\times p}\ \mathbf{B}^*_{p\times(p-1)}$$

Step 3: Repeat Step 0, find the coefficient vector $\mathbf{a}_1^*$ that maximizes the index for the new dataset determined by $\mathbf{D}^*$:

$\mathbf{a}^*_{1_{(p-1)\times 1}} = (a^*_{11}, a^*_{12}, \ldots, a^*_{1(p-1)})^T$ such that $\mathbf{a}_1^{*^T}\mathbf{a}_1^* = 1$.

Step 4: Define $\mathbf{a}_2$, corresponding to the original data $\mathbf{D}$, as:

$$\mathbf{a}_{2_{p\times 1}} = a^*_{11}\mathbf{u}_2 + a^*_{12}\mathbf{u}_3 + \cdots a^*_{1(p-1)}\mathbf{u}_p = [\mathbf{u}_2\ \mathbf{u}_3 \ldots \mathbf{u}_p]\ \mathbf{a}_1^* = \mathbf{B}^*\mathbf{a}_1^*$$

The above steps are repeated until the desired number of coefficient vectors $\mathbf{a}_i^{(k)}$, $i = 1, \ldots, \min(p_k)$, have been found. In practice the sample version $\mathcal{I}_n^{(\mathrm{J})}$ is used in place of the population version. Codes, available from the authors, have been developed in Matlab for our computations.

Some notes on the steps in the above algorithm. In Step 1, for finding the successive coefficient vectors $\mathbf{a}_i$, $2 < i < p$, define $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2 \cdots\ \mathbf{a}_{i-1}][\mathbf{a}_1\ \mathbf{a}_2 \cdots\ \mathbf{a}_{i-1}]^T$ and repeat the next steps. In step 2, note that the column space of $\mathbf{B}^*$ is orthogonal to $\mathbf{a}_1$. The constraints are satisfied since:

$$\begin{aligned}
\mathbf{a}_1^T\mathbf{a}_2 &= a^*_{11}\mathbf{a}_1^T\mathbf{u}_2 + a^*_{12}\mathbf{a}_1^T\mathbf{u}_3 + \cdots a^*_{1(p-1)}\mathbf{a}_1^T\mathbf{u}_p = 0 \\
\mathbf{a}_2^T\mathbf{a}_2 &= \mathbf{a}_1^{*^T}\mathbf{B}^{*^T}\mathbf{B}^*\ \mathbf{a}_1^* = \mathbf{a}_1^{*^T}\mathbf{I}_{(p-1)}\ \mathbf{a}_1^* = \mathbf{a}_1^{*^T}\mathbf{a}_1^* = 1.
\end{aligned}$$

Additional details and justifications for each step can be found in Chapter 3, Section 3.2.3.

2.5 Simulations and real data sets

Before we analyze the morphometric data that motivated our method, we carry out two comprehensive simulation studies in Sections 2.5.1 and 2.5.2 to illustrate the use of our overall index, $\mathcal{I}^{(\mathrm{O})} = \mathcal{I}^{(\mathrm{G})}(\mathbf{I}_{p_1}, \ldots, \mathbf{I}_{p_k})$ (see Section 2.4.1), and compare its performance with that of R^2_{TM} defined in Section 2.2. Simulations are performed on two sets of landmarks, $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{q\times 1}$. From visual inspection, it appears that the variables in the fly wing and mouse mandible data sets are marginally normal (see Section 2.5.5). Unless otherwise specified, our simulated data sets are generated from a multivariate normal distribution. As in the fly wing data, we set $p = 14$, $q = 16$ and the sample size $n = 109$. In Section 2.5.3, we perform a simulation to illustrate the ability of our information measures to handle multiple sets and, in the following Section 2.5.4, we show that the same performance is achieved for multiple groups of multiple sets. In Section 2.5.5 we analyze the fly wing and mouse mandible datasets. Finally, in Section 2.5.6 we compare our multiple set measure $\mathcal{I}^{(\mathrm{P})}$ to Neuenschwander and Flury's [11] common CCA method to show our method does not lose power when linear relationships and normality holds.

We refer to our methods as Kullback-Leibler Information Canonical Analysis(KLICA). Since the KLICA measures are invariant to nonsingular scale transformations, to be consistent with other dimension reduction methods the following real data analysis and simulations are reported for the whitened random vectors in Sections 2.5.3, 2.5.4 and 2.5.6. That is, we transform the vectors to have identity covariance matrices. However, for the morphometric data analysis the modularity simulations and real data analysis in Sections 2.5.1, 2.5.2 and 2.5.5 are done in the original scale. However, for ease and clarity the notation $\mathbf{x}, \mathbf{y}$ etc. is maintained.

2.5.1 Modularity simulation

Assume that the mean of $\mathbf{X}$ and $\mathbf{Y}$ is zero and the correlation matrix is given by

$$\rho = \begin{pmatrix} \mathbf{U}_X & \mathbf{S} \\ \mathbf{S} & \mathbf{U}_Y \end{pmatrix},$$

where u is the equi-correlation in the (equi-correlation) matrices $\mathbf{U}_X$ and $\mathbf{U}_Y$, and s is the equi-correlation in the (equi-correlation) matrix $\mathbf{S}$. Here $u \in [0, 1]$ defines the linear correlation within each module. The value s controls the amount of linear association between $\mathbf{X}$ and $\mathbf{Y}$. For a simulated dataset, the sample correlation $\mathbf{S}_n$ naturally does not equal $\mathbf{S}$. As a result, a tolerance of 0.04 is set and datasets are repeatedly generated until the $\min(\mathbf{S}_n)$ and the $\max(\mathbf{S}_n)$ are within this tolerance. For each simulation the value u is large and the corresponding sample correlation matrices, $\mathbf{U}_n$, are allowed to vary randomly.

Having set the within correlation to be high, we will investigate the accuracy of $\mathcal{I}^{(\mathrm{O})}$ and R^2_{TM} by varying the level of association between modules with different values of s. In addition, for each level of s one of the simulations will contain a nonlinear relationship across the two sets, e.g. $y_1 = x_1^2 + \epsilon$. The nonlinear relationship will create a strong dependence between the sets but will not increase the correlation within the sets. The values and relationships in each of our simulations are given below:

Simulation 1: $u = 0.9$ and $s = 0.3$, all linear associations

Simulation 2: $u = 0.9$ and $s = 0.3$, with nonlinear associations $y_1 = x_1^2 + 0.3\epsilon$

Simulation 3: $u = 0.9$ and $s = 0.6$, all linear associations

Simulation 4: $u = 0.9$ and $s = 0.6$, with nonlinear associations $y_1 = x_1^2 + 0.3\epsilon$

For each simulation, we generate 1000 datasets and for each dataset calculate p-values based on 1000 permutations using the sample information test statistics $\mathcal{I}_n^{(\mathrm{O})}$ and R^2_{TM}. Table 2.1 shows the number of rejected null hypothesis at the significance levels: $\alpha = 0.10, 0.05, 0.01$ and 0.001.

# of rejected H_0 - Simulation 2.5.1																
	$\alpha = .10$				$\alpha = .05$				$\alpha = .01$				$\alpha = .001$			
Sim	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$\mathcal{I}^{(O)}$	309	1000	1000	1000	203	1000	1000	1000	53	1000	992	1000	11	998	809	1000
R^2_{TM}	132	179	566	612	64	104	398	479	12	27	170	200	0	4	33	45

Table 2.1: Modularity simulation study 1 (Simulation 2.5.1)

We believe that permutation test is meaningful and should be used to determine modularity. In Simulation 1, the correlation between both sets is low, $s = 0.3$, and both methods, in general, support the null hypothesis. However, note that the index $\mathcal{I}^{(O)}$ is more sensitive to the correlation between the sets as evidenced by more rejected null hypotheses at the reported significance levels. In Simulation 2, a nonlinear association is added and the correct conclusion is that the sets are not modules. The number of rejections show that the correct conclusion is made using our index, while the R^2_{TM} statistic seems unable to recognize the nonlinear dependence. Simulations 3 and 4 mimic simulations 1 and 2 with the exception that the correlation between the sets is increased to 0.6. With this increase, the ability of our test statistic to arrive at the correct conclusion is reinforced. In sharp contrast, the p-values based on the R^2_{TM} index consistently provide incorrect conclusions, further showing the difference between the two indices.

In conclusion, we believe the $\mathcal{I}^{(O)}$ index outperforms R^2_{TM} whether there are linear or nonlinear relationships. The difference between the methods is substantial as evidenced by the difference in the number of rejected null hypothesis. Also, these simulations were performed on two sets where CCA is defined and one could only expect less satisfactory results in a multiple set scenario.

2.5.2 Simulation modularity

Unlike Simulation 1, the within sample correlations will be allowed to vary at random and in some simulations the variables will be non-normal. We choose the simulation parameters in such a way as not to favor a specific method. The procedure followed in Simulation 1 will be followed in Simulation 2. Data sets were generated in the following way:

independent: generate $\mathbf{X} \sim MVN_p(\mathbf{0}, \rho_x)$,$\mathbf{Y} \sim MVN(\mathbf{0}, \rho_y)$

dependent: generate Data=$[\mathbf{X}\ \mathbf{Y}]^T \sim MVN_{p+q}(\mathbf{0}, \rho_{xy})$,

where $\rho_x, \rho_y,$ and ρ_{xy} are equi-correlation matrices with correlations generated from a uniform distribution $U \sim (b, 1)$ for some $b \geq 0$. Combinations of linear, nonlinear and non-normal transformations of variables were used to create some of the final simulated data sets. The simulations are defined as follow:

Independent simulations

sim 1 $\mathbf{X}$ and $\mathbf{Y}$ independent, correlation within each set, u_1 and u_2, generated at random from a $\mathbf{U}(0.9, 1)$ distribution.

sim 2 $\mathbf{X}$ and $\mathbf{Y}$ independent, correlation within each set, u_1 and u_2, generated at random from a $\mathbf{U}(0.9, 1)$ distribution. Next, the following nonlinear transformations within each set are made: $x_3 = x_1^2 + 0.3\epsilon$, $x_4 = x_2^2 + 0.4\epsilon$, $y_9 = y_{13}^2 + 0.5\epsilon$ and $y_{10} = y_{14}^2 + 0.2\epsilon$.

sim 3 $\mathbf{X}$ and $\mathbf{Y}$ all generated at random from a $N(0, 1)$ distribution.

sim 4 $\mathbf{X}$ and $\mathbf{Y}$ independent, correlation within each set, u_1 and u_2, generated at random from a $\mathbf{U}(0.5, 0.7)$ distribution. Next, the following non-normal random variables are defined: $x_1 \sim \chi^2(7)$, $x_2 \sim F(14, 10)$, $y_5 \sim \Gamma(3, 6)$ and $y_6 \sim \beta(10, 5)$.

Dependent simulations

sim 5 $\mathbf{X}$ and $\mathbf{Y}$ generated from a multivariate normal distribution with correlation, u, generated at random from a $\mathbf{U}(0.85, 1)$.

sim 6 $\mathbf{X}$ and $\mathbf{Y}$ generated from a multivariate normal distribution with correlation, u, generated at random from a $\mathbf{U}(0.85, 1)$. Next, the following nonlinear transformations within each set are made: $x_3 = x_1^2 + 0.3\epsilon$ and $y_2 = y_4^2 + 0.2\epsilon$.

sim 7 $\mathbf{X}$ and $\mathbf{Y}$ initially independent, correlation within each set, u_1 and u_2, generated at random from a $\mathbf{U}(0.9, 1)$ distribution. Next, the following nonlinear dependence relationship between $\mathbf{X}$ and $\mathbf{Y}$ is made: $x_3 = x_1^2 + 0.3\epsilon$ and $y_1 = x_{14}^2 + 0.4\epsilon$.

sim 8 $\mathbf{X}$ and $\mathbf{Y}$ generated from a multivariate normal distribution with correlation, u, generated at random from a $\mathbf{U}(0.8, 1)$. Next, the following non-normal variables are defined: $x_1 \sim \chi^2(7)$, $x_2 \sim F(14, 10)$, $y_1 \sim \Gamma(3, 6)$ and $y_{16} \sim \beta(10, 5)$.

For each model, we generate 1000 simulated datasets. Table 2.2 reports the number of rejections of the null hypothesis for both methods at various significance levels, α.

# of rejected H_0 - Simulation 2.5.2																
Sim	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
	$\alpha = .10$								$\alpha = .05$							
$\mathcal{I}^{(\mathrm{O})}$	89	94	94	98	1000	1000	976	1000	48	41	58	48	1000	1000	946	1000
R^2_{TM}	100	125	99	98	1000	1000	927	1000	55	64	54	50	999	1000	866	1000
	$\alpha = .01$								$\alpha = .001$							
$\mathcal{I}^{(\mathrm{O})}$	8	8	13	7	1000	1000	853	1000	1	2	1	1	1000	1000	586	1000
R^2_{TM}	8	12	9	4	987	1000	679	997	0	1	1	0	841	999	350	946

Table 2.2: Modularity simulation study 2 (Simulation 2.5.2)

For simulations 1-4 both methods arrive, in general, at the same correct conclusion that the sets of variables are independent, as the way generated. However, the number of incorrectly rejected hypotheses for the $\mathcal{I}^{(\mathrm{O})}$ based statistic is nearly always less than the R^2_{TM} based method. The dependent simulations 5,6 and 8, do not show much difference for both methods, except that our measure detects the dependence relationships more often when there is a difference. Simulation 7 significantly indicates the advantage of $\mathcal{I}^{(\mathrm{O})}$ for detecting nonlinearity, as evidenced by the higher number of rejected null hypotheses.

2.5.3 Multiple set simulation

In this simulation we test the accuracy of our methods with a moderate sample size of three sets of random vectors in high dimension composed of variables with a wide range of distributions. This scenario is made more complex with complicated linear and nonlinear relationships between the sets. Using the information measures $\mathcal{I}^{(\mathrm{J})}$ and $\mathcal{I}^{(\mathrm{P})}$ we perform a separate simulation for each under the following scenario for three sets. For a sample size of $n = 75$, we define the variables $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)^T$, $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$ and $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$ where $X_1 \sim U(0,1)$, $X_2 \sim \chi^2_{(7)}$, $X_3 \sim N(0,1)$, $X_4 \sim t(8)$, $X_5 \sim F(3,12)$, $X_6 \sim t(5)$, $X_7 \sim N(0,1)$, $X_8 \sim t(12)$, $Y_3 \sim t(9)$, $Z_3 \sim \chi^2_{(14)}$ and $\epsilon_j \sim N(0,1), j = 1, \ldots, 4$. The remaining variables are defined as:

$$\begin{aligned} Y_1 &= (2X_1 + X_2 + 3X_3)^2 + 0.5\epsilon_1 \text{ and } Z_1 = 2X_1 + X_2 + 3X_3 + 0.4\epsilon_2 \\ Y_2 &= \cos(X_5 + X_6) + 0.2\epsilon_3 \quad \text{and } Z_2 = X_5 + X_6 + 0.3\epsilon_4. \end{aligned}$$

The true coefficient vectors are:

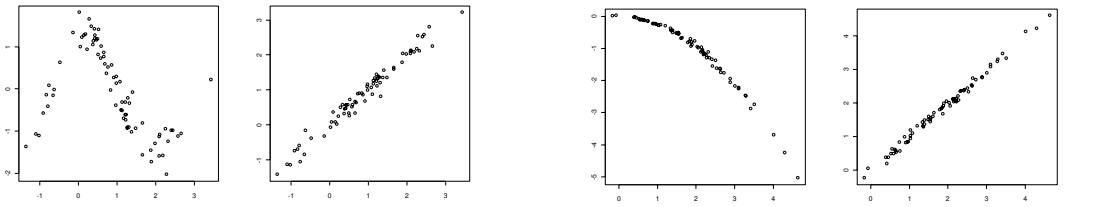
$$\begin{aligned} \mathbf{a} &= (2,1,3,0,0,0,0,0)^T, \mathbf{b} = (1,0,0)^T \text{ and } \mathbf{c} = (1,0,0)^T \text{ and} \\ \widetilde{\mathbf{a}} &= (0,0,0,0,1,1,0,0)^T, \widetilde{\mathbf{b}} = (0,1,0)^T \text{ and } \widetilde{\mathbf{c}} = (0,1,0)^T. \end{aligned}$$

For a sample dataset drawn according to the above specifications, we estimate the information coefficient vectors using $\mathcal{I}^{(\mathrm{P})}$ and $\mathcal{I}^{(\mathrm{J})}$ and repeat the process for each 1000 times. We calculate estimates of the 1st, 2nd and 3rd information variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}$, $\widehat{\mathbf{b}}_j^T\mathbf{y}$ and $\widehat{\mathbf{c}}_j^T\mathbf{z}$, $j = 1,2,3$, where $\mathbf{x}$,$\mathbf{y}$ and $\mathbf{z}$ are samples from $\mathbf{X}$,$\mathbf{Y}$ and $\mathbf{Z}$ respectively. Depending on the simulation, $\widehat{\mathbf{a}}_1$ or $\widehat{\mathbf{a}}_2$ may be the estimated coefficient vector for either the linear or nonlinear relationship. If $\widehat{\mathbf{a}}_1$ is the estimated coefficient for the linear relationship then it should equal the true coefficient vector $\mathbf{a}$, defined above. To assess the accuracy of the estimates we calculate the absolute value of the correlation between these variates as $|\rho\left(\widehat{\mathbf{a}}_1^T\mathbf{x}, \mathbf{a}^T\mathbf{x}\right)|$, denoted $|\bar{\rho}_1|$. For both the information measures, the means and estimated standard errors of the absolute value of the correlations are given in Table 2.3.

Simulation 2.5.3				
	$\mathcal{I}^{(\mathrm{P})}$		$\mathcal{I}^{(\mathrm{J})}$	
	1^{st} variates	2^{nd} variates	1^{st} variates	2^{nd} variates
	$\lvert\bar{\rho}_1\rvert$(se)	$\lvert\bar{\rho}_2\rvert$(se)	$\lvert\bar{\rho}_1\rvert$(se)	$\lvert\bar{\rho}_2\rvert$(se)
$\widehat{\mathbf{a}}^T\mathbf{x}$	.99324(.001444)	.97213(.003180)	.99000(.001378)	.95993(.004058)
$\widehat{\mathbf{b}}^T\mathbf{y}$	.99247(.001127)	.98204(.000997)	.98813(.001438)	.97436(.002004)
$\widehat{\mathbf{c}}^T\mathbf{z}$	.99565(.000710)	.98826(.000766)	.99179(.001089)	.98066(.001637)

Table 2.3: Correlation means(standard errors) (Simulation 2.5.3)

The absolute average correlations for all the variates are high with small standard errors, showing that our methods are able to detect the known relations accurately and consistently even under extreme assumptions. To demonstrate this further, a generated dataset was selected and the 1^{st} and 2^{nd} $\mathcal{I}^{(\mathrm{P})}$ information variate pairs of $\widehat{\mathbf{a}}^T\mathbf{x}$ and $\widehat{\mathbf{b}}^T\mathbf{y}$ were plotted; see Figure 2.2. These plots show that the first information variates(left panel) correspond to the quadratic and linear relationships, $Y_2 = \cos(X_5 + X_6) + 0.2\epsilon_3$ and $Z_2 = X_5 + X_6 + 0.3\epsilon_4$, and the second information variates identify the associations, $Y_1 = (2X_1 + X_2 + 3X_3)^2 + 0.5\epsilon_1$ and $Z_1 = 2X_1 + X_2 + 3X_3 + 0.4\epsilon_2$.



1^{st} variates: $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{b}}^T\mathbf{y}$ and $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{c}}^T\mathbf{z}$ 2^{nd} variates: $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{b}}^T\mathbf{y}$ and $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{c}}^T\mathbf{z}$

Figure 2.2: KLICA - Variate plots (Simulation 2.5.3)

A permutation test (1000 permutations) to determine significant numbers of variates was also performed on each, producing p-values of 0.0, 0.0 and 0.29 for the 1^{st}, 2^{nd} and 3^{rd}

information variates, respectively. As expected, the p-value for the 3rd information variate is not significant, since there are only two defined relationships. The correlations between the true and estimated information variate pairs for each are: $\rho_{\widehat{\mathbf{a}}_1} = \rho(\widehat{\mathbf{a}}_1^T\mathbf{x}, \widetilde{\mathbf{a}}^T\mathbf{x}) = .9983$, $\rho_{\widehat{\mathbf{b}}_1} = .9925$, $\rho_{\widehat{\mathbf{c}}_1} = .9973$ and $\rho_{\widehat{\mathbf{a}}_2} = .9981$, $\rho_{\widehat{\mathbf{b}}} = -.999$ and $\rho_{\widehat{\mathbf{c}}_2} = .9978$. The directions of the coefficient vectors are unique up to the sign, therefore, the sign of the correlations are not meaningful.

Next, a few datasets were generated at random and the permutation test was performed on the one with the worst correlations between the true and estimated information variate pairs. This is done since the correlations between the true and estimated variates were strong for the above dataset on which the permutation test was performed. Computing the permutation test on this dataset is meant to reaffirm the above results using this procedure. The correlations for each are: $\rho_{\widehat{\mathbf{a}}_1} = \rho(\widehat{\mathbf{a}}_1^T\mathbf{x}, \widetilde{\mathbf{a}}^T\mathbf{x}) = .9987$, $\rho_{\widehat{\mathbf{b}}_1} = .9984$, $\rho_{\widehat{\mathbf{c}}_1} = .999$ and $\rho_{\widehat{\mathbf{a}}_2} = .9672$, $\rho_{\widehat{\mathbf{b}}} = -.9830$ and $\rho_{\widehat{\mathbf{c}}_2} = -.9726$. The p-values for the 1st, 2nd and 3rd information variates were, 0.0, 0.0 and 0.5740, respectively, further illustrating the accuracy of the test.

Finally, it is important to note the relationships were defined marginally, which favors $\mathcal{I}^{(\mathrm{P})}$. However, the difference in the performance of these measures is negligible. We also simulated a joint relationship which favored $\mathcal{I}^{(\mathrm{J})}$ but the difference in the performance of these measures was once again negligible.

2.5.4 Multiple group simulation

Simulations were performed to assess the performance of the proposed common information method $\mathcal{I}^{(\mathrm{K})\mathrm{C}}$ of Section 2.4.3 for multiple groups. Here, relationships are defined marginally, so the sample information of $\mathcal{I}^{(\mathrm{P})\mathrm{C}}$ is used. Let $\mathrm{P}(W = w) = 1/2$ for $w = 1, 2$. Suppose $X_1^w, \ldots, X_8^w$, $Y_2^w, Y_3^w, Z_2^w, Z_3^w$ and ϵ are all $N(0,1)$ random variables variables, for each w with

$$Y_1 = X_1 + 0.4\epsilon_1 \quad \text{and} \quad Z_1 = X_1^2 + 0.3\epsilon_2 \text{ when } W = 1$$

$$Y_1 = \cos(X_1) + 0.3\epsilon_3 \quad \text{and} \quad Z_1 = X_1 + 0.3\epsilon_4 \text{ when } W = 2.$$

There is only one true common pair for each group given by $\mathbf{a}^T\mathbf{X}^w$, $\mathbf{b}^T\mathbf{Y}^w$ and $\mathbf{c}^T\mathbf{Z}^w$ with $\mathbf{a} = (1, 0, \ldots, 0)^T$, $\mathbf{b} = (1, 0, 0)^T$ and $\mathbf{c} = (1, 0, 0)^T$. When $W = 1$ the random vectors Y_1 and Z_1 are linearly and nonlinearly related to X_1, respectively. However, for $W = 2$ the type of association between X_1, linear and nonlinear, is reversed.

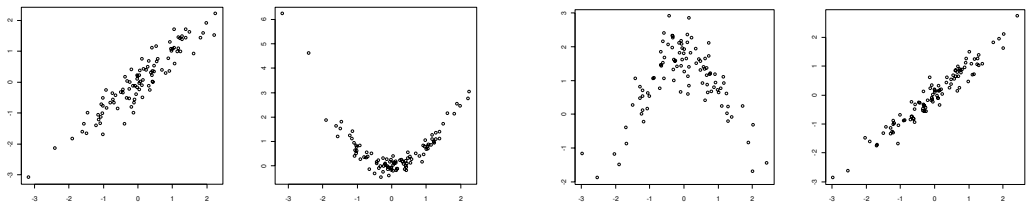
Based on a random sample of size $n = 100$, the common information coefficients are computed. The calculation was repeated 500 times. For each group, we calculate estimates of the 1st common information variates as $\widehat{\mathbf{a}}^T\mathbf{x}^w$, $\widehat{\mathbf{b}}^T\mathbf{y}^w$ and $\widehat{\mathbf{c}}^T\mathbf{z}^w$, where $\mathbf{x}^w$, $\mathbf{y}^w$ and $\mathbf{z}^w$ denote the sample from the w^{th} group. We then calculate the absolute value of the correlation between the information variates $\left(\widehat{\mathbf{a}}^T\mathbf{x}^w, \mathbf{a}^T\mathbf{x}^w\right)$, $(\widehat{\mathbf{b}}^T\mathbf{y}^w, \mathbf{b}^T\mathbf{y}^w)$ and $\left(\widehat{\mathbf{c}}^T\mathbf{z}^w, \mathbf{c}^T\mathbf{z}^w\right)$, for $w = 1, 2$. In addition, we calculate the average distance between the true and estimated common coefficient vectors using $||(\mathbf{I}-\mathbf{d}\mathbf{d}^T)\widehat{\mathbf{d}}^T||_2$ (see Xia et al. [17] and Li, Zha and Chiaromonte [10] for instance). Here the standard vector 2-norm is used, $||u||_2 = \sqrt{u^T u}$. For an estimated coefficient vector, say $\widehat{\mathbf{a}}$, $(\mathbf{I} - \mathbf{a}\mathbf{a}^T)\widehat{\mathbf{a}}^T$ is the projection of $\widehat{\mathbf{a}}$ into the orthogonal subspace spanned by the vector $\mathbf{a}$, and hence should be close to zero if $\widehat{\mathbf{a}}$ is an estimate of $\mathbf{a}$. Since the analysis and results are performed in the standardized scale, the distance is calculated in this scale by transforming the true coefficient as $\mathbf{\Sigma}_{\mathbf{X}}^{1/2}\mathbf{a}$, where $\mathbf{\Sigma}_{\mathbf{X}}^{1/2}$ is the pooled covariance matrix across the two groups. This distance is calculated for each estimated common coefficient vector in all our simulations and averaged; we denote this $||\cdot||_2$. The means and estimated standard errors of the absolute correlations and distances are given in Table 2.4.

Simulation 2.5.4				
Group W $= 1$		**Group W** $= 2$		$||(\mathbf{I} - \mathbf{d}\mathbf{d}^T)\widehat{\mathbf{d}}^T||_2$
Relation	$\lvert\bar{\rho}_1\rvert$(se)	Relation	$\lvert\bar{\rho}_1\rvert$(se)	$||\cdot||_2$(se)
$\widehat{\mathbf{a}}^T\mathbf{x}, \mathbf{a}^T\mathbf{x}$	.9929(.00101)	$\widehat{\mathbf{a}}^T\mathbf{x}, \mathbf{a}^T\mathbf{x}$	.9919(.00094)	.0548(.00217)
$\widehat{\mathbf{b}}^T\mathbf{y}, \mathbf{b}^T\mathbf{y}$	.9918(.00320)	$\widehat{\mathbf{b}}^T\mathbf{y}, \mathbf{b}^T\mathbf{y}$	.9887(.00316)	.0615(.00364)
$\widehat{\mathbf{c}}^T\mathbf{z}, \mathbf{c}^T\mathbf{z}$	.9962(.00026)	$\widehat{\mathbf{c}}^T\mathbf{z}, \mathbf{c}^T\mathbf{z}$	.9981(.00024)	.0542(.00146)

Table 2.4: Absolute average correlations(standard errors) (Simulation 2.5.4)

The average correlations show that our estimation method accurately estimates each component of the true common pair. For a generated dataset, graphs of the 1st common

information variates for each of the individual groups are given in Figure 2.3. The graphs show that our method recovers the relationships between the three sets for both groups.



Group 1: $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{b}}^T\mathbf{y}$ and $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{c}}^T\mathbf{z}$ Group 2: $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{b}}^T\mathbf{y}$ and $\widehat{\mathbf{a}}^T\mathbf{x}$ vs $\widehat{\mathbf{c}}^T\mathbf{z}$

Figure 2.3: KLICA - Variate plots by group (Simulation 2.5.4)

The correlations between the true and estimated common information variate pairs for each group are: $\rho_{\widehat{\mathbf{a}}} = \rho(\widehat{\mathbf{a}}^T\mathbf{x}, \mathbf{a}^T\mathbf{x}) = 0.996$, $\rho_{\widehat{\mathbf{b}}} = 0.999$ and $\rho_{\widehat{\mathbf{c}}} = 0.998$ (Group 1) and $\rho_{\widehat{\mathbf{a}}} = 0.998$, $\rho_{\widehat{\mathbf{b}}} = 0.996$ and $\rho_{\widehat{\mathbf{c}}} = 0.999$ (Group 2).

2.5.5 Modularity real data example

Here we consider data on fly wing and mouse mandible, each consisting of 15 landmarks represented by coordinate measures x_1 and x_2 for a total of 30 variables. These landmarks are grouped into sets and tested to determine whether the configurations of the landmarks are modules. Note that pictured examples of the fly wing with landmarks grouped into sets corresponding to possible modules are given in Section 2.2. For the data of the fly wing, *Drosophila melanogaster*, the two sets are the anterior and posterior compartments consisting of 7 and 8 landmarks, respectively. The mouse mandible is also subdivided into subsets of 7 and 8 landmarks corresponding to the alveolar and ascending ramus regions, respectively. Measurements are taken on both the left and right sides of the mandible, and the left and right wing. To extract the information on shape, we used a full generalized Procrustes fit and projection to tangent space (Dryden [1]), which remove variation in size, position and orientation. As a result, the dimensionality of the data is reduced by four. The data are then combined either by taking the average or difference of the corresponding left and right

configurations, referred to below as the symmetric and asymmetric data, respectively. Specific details of these methods can be found in Klingenberg [7] and [6]. Of importance is that the procrustes fit methods create linear relationships that need to be ignored when determining whether the sets are modules. That is, we need to measure the variation in the sets ignoring these linear relationships. A method to do so is given below.

For the analysis that follows, $\mathbf{X}_{14\times 1}$ denotes the first 14 measurements on 7 landmarks, and $\mathbf{Y}_{16\times 1}$ denotes the remaining 16 measurements on 8 landmarks.. There are $n = 109$ observations on the fly wing and $n = 90$ on the mouse mandible. Initially an analysis using the information measure $\mathcal{I}^{(\mathrm{P})}(= \mathcal{I}^{(\mathrm{J})})$ and CCA is performed on each dataset to recover these four perfect linear relationships. Our method (and the CCA method) recovers the four perfect linear relations among 30 variables in the small sample. For example, plots of the $1^{\mathrm{st}} - 4^{\mathrm{th}}$ variates for the symmetric fly wing dataset using the multiple set index $\mathcal{I}^{(\mathrm{J})}$ is shown in Figure 2.4. As previously mentioned, in order to test whether the sets $\mathbf{X}$ and $\mathbf{Y}$ are modules

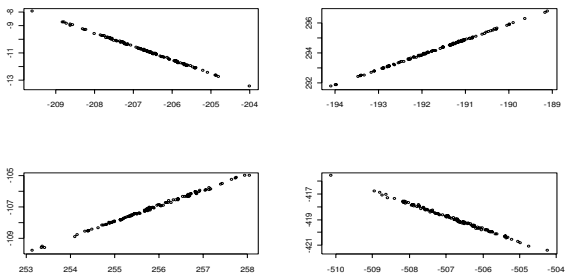


Figure 2.4: Variate plots - symmetric fly wing dataset

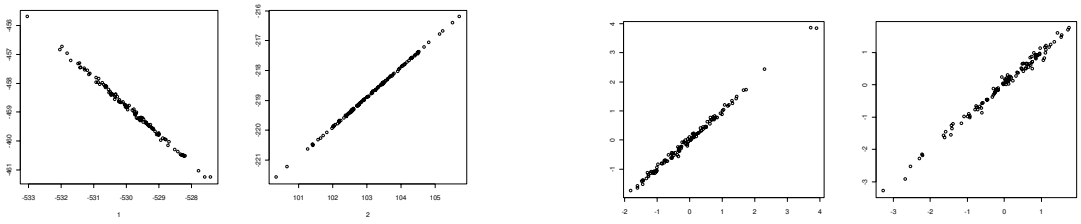
we need to measure the variation in the sets, ignoring the perfect linear relationships. This is done by first defining the matrices $\mathbf{A} = [\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \dots \widehat{\mathbf{a}}_{14}]$ and $\mathbf{B} = [\widehat{\mathbf{b}}_1\ \widehat{\mathbf{b}}_2 \dots \widehat{\mathbf{b}}_{14}]$, where the columns are the estimated information coefficient vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Next, obtain the matrices $\mathbf{E} = [\mathbf{e}_1 \dots \mathbf{e}_{14}]$ and $\mathbf{F} = [\mathbf{f}_1 \dots \mathbf{f}_{16}]$, where the columns are the left singular vectors(eigen-vectors) of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{B}\mathbf{B}^T$. Then project the data for $\mathbf{X}$ and $\mathbf{Y}$, denoted by $\mathbf{D}_{\mathbf{x}}$ and $\mathbf{D}_{\mathbf{y}}$, onto the subspace orthogonal to the first four to obtain,

$$\mathbf{D}_{\mathbf{x}^*} = \mathbf{D}_{\mathbf{x}} \times [\mathbf{e}_5 \dots \mathbf{e}_{14}] \text{ and } \mathbf{D}_{\mathbf{y}^*} = \mathbf{D}_{\mathbf{y}} \times [\mathbf{f}_5 \dots \mathbf{f}_{16}].$$

Since the statistic R^2_{TM} is equivalent to summing the squared canonical correlations, we repeat the above using the coefficient vectors produced by CCA.

We use the transformed datasets to test the null hypothesis that the two sets constitute modules. For these two new databases, we perform permutation tests using $\mathcal{I}^{(\mathrm{O})}$ and R^2_{TM} detailed in Section 2.4.1. For the permutation test, 1000 random permutations were performed to get the p-values.

For the symmetric fly data, the permutation p-values based on $\mathcal{I}^{(\mathrm{O})}$ and R^2_{TM} are .017 and 0.0, respectively. Note that the tests based on both measures reject the null at $\alpha = 0.05$ and do not detect two modules. To see whether there is any relationships between the two sets, we find the first two pairs of KLICA and CCA variates. The variate pairs and their associated correlations are plotted in Figure 2.5 for the symmetric fly data. From the plots in Figure 2.5, both methods seem to indicate that the two sets do not constitute modules. For the asymmetric fly data, the permutation p-value based on $\mathcal{I}^{(\mathrm{O})}$ and R^2_{TM} are .032 and 0.0, respectively. Again, from the plots (not given here) both methods arrive at the same decision, which appear to be consistent with the plots.



a. $\mathcal{I}^{(\mathrm{P})}$ variates; $\rho_1 = -0.9985$, $\rho_2 = 0.9999$ b. CCA variates; $\rho_1 = 0.9975$, $\rho_2 = 0.9949$

Figure 2.5: KLICA - Variate plots: Symmetric fly transformed data (Example 2.5.5)

Again, the conclusions between the p-values and the variate plots in the analysis of both fly wing datasets is observed in the analysis of the symmetric and asymmetric mouse mandible datasets. For the symmetric mouse data the permutation p-value based on $\mathcal{I}^{(\mathrm{O})}$ and R^2_{TM} are .02 and 0.0, respectively. For the asymmetric mouse mandible data the p-values are .006 and 0. At the $\alpha = 0.05$ significance level, both methods do not detect two modules,

which is in agreement with the plots (not given here) of the variates and their respective correlations($\rho_1 = -0.9998$, $\rho_2 = 0.9999$ and $\rho_1 = 0.9976$, $\rho_2 = 0.9798$).

In conclusion, the previous simulations and the real data analysis show that our proposed method performs as well as or outperforms the methods based on trace correlations defined in Section 2.2. Moreover, all the examples were performed in a two-set configuration, for which CCA is defined. The methods presented based on information theory extend naturally to multiple sets unlike the CCA based methods, which are often extended in an ad hoc fashion based on pairwise measures.

As an aside, initially all datasets were analyzed using both KLICA and CCA and produced nearly identical results. For example, the correlations for the first four information variates, using $\mathcal{I}^{(\mathrm{P})}(\mathcal{I}^{(\mathrm{J})})$, for the fly wing dataset were: $\rho(\widehat{\mathbf{a}}_1^T\mathbf{x}, \widehat{\mathbf{b}}_1^T\mathbf{y}) = -0.9997$, $\rho(\widehat{\mathbf{a}}_2^T\mathbf{x}, \widehat{\mathbf{b}}_2^T\mathbf{y}) = 0.9996$, $\rho(\widehat{\mathbf{a}}_3^T\mathbf{x}, \widehat{\mathbf{b}}_3^T\mathbf{y}) = 0.9998$ and $\rho(\widehat{\mathbf{a}}_4^T\mathbf{x}, \widehat{\mathbf{b}}_4^T\mathbf{y}) = -0.9987$. These relationships were identified among 30 variables with low variation and a sample size of only $n = 109$. In addition, the estimated coefficient vectors for $\mathbf{X}$ and $\mathbf{Y}$ are complex, take for example the first estimated coefficient vector for $\mathbf{X}$:

$$\widehat{\mathbf{a}}_1 = (.255, .337, .163, .249, .1679, .200, .330, .445, .135, .145, .340, .163, .385, .155)^T.$$

This illustrates that our method detects relationships accurately for random vectors of high dimension and does not lose power to CCA in the presence of linear relationships.

2.5.6 Water strider data

This dataset, analyzed by Neuenschwander and Flury [11], consists of $n = 88$ female water striders of the species *Limnoporus canaliculatus*. The lengths of the femur and tibia were collected for the first three of six stages of growth and log-transformed. The three sets are the stages of growth with two variables within each set. Flury et al. [11] analyzed this dataset assuming normality and common canonical variates, that is, the coefficient vectors are the same across all sets. In our analysis, we do not use common coefficient vectors, rather the information based analysis is performed treating the data as three sets of random vectors.

If the assumptions of common variates and multivariate normality are valid, then the variates should be similar in both methods; hence, the correlations between the variates should be high. Our results show that the correlation coefficient between the first common canonical variate and our $\mathcal{I}^{(\mathrm{P})}$ variates are 0.985, 0.994 and 0.990 for the three sets. Similarly, for the second variates, the correlations are 0.996, 0.999 and 0.968.

This example demonstrates that the information based method is able to produce equivalent results to Neuenschwander and Flury's [11] method, which shows that our method does not lose power to CCA type methods, when multivariate normality holds or linear relationships are of primary concern.

2.6 Discussion

Motivated by data sets that arise in morphological integration studies, we suggest a general measure ($\mathcal{I}^{(\mathrm{G})}$) which measures the association between m-sets of random vectors. Based on this general index, we define three useful measures, the first of which gives an index of overall association ($\mathcal{I}^{(\mathrm{O})}$). This index together with a permutation test is used to test for the presence of modules. The other two measures ($\mathcal{I}^{(\mathrm{J})}$ and $\mathcal{I}^{(\mathrm{P})}$) provide powerful dimension reduction methods to recover (joint and/or marginal) linear and nonlinear relationships between the sets. Extension of these measures to multiple groups with multiple sets is also discussed.

Alternative methods using the CCA principle along with the permutation test have been proposed to analyze our motivational data. Through extensive simulation studies, we establish the superiority of $\mathcal{I}^{(\mathrm{O})}$ to the CCA based methods. Of course, a certain amount of dependence may be acceptable between modules. However, in order to ascertain what amount is too much or too little, one needs first to be able to detect it. A further topic of research is to obtain a p-value correction that accounts for a certain amount of acceptable dependence as determined by the researcher. In this study, we also discover that the partition method is better viewed as a data mining tool to find the partition with the smallest association among all possible subdivisions.

As for visualizing the relationships, the simulations show that the two measures $\mathcal{I}^{(\mathrm{J})}$ and $\mathcal{I}^{(\mathrm{P})}$ recover linear and nonlinear associations whether there are multiple groups or multiple sets. When linear relationships are of primary concern or normality holds, as in the water strider data, our methods do not lose power to the common CCA method.

The overall association index $\mathcal{I}^{(\mathrm{O})}$ needs no maximization, whereas the dimension reduction measures $\mathcal{I}^{(\mathrm{J})}$ and $\mathcal{I}^{(\mathrm{P})}$ do. In addition to the choice of kernel density estimators and selection of bandwidths, the maximization requires initial guesses. Generally, the choice of kernel may not be crucial (Härdle [2]) but the bandwidth selection may be important. However, our choices for these seem reasonable, as evidenced by our simulations. Initial guesses can speed up the convergence but determining them need not be complicated. A basic random grid search was used for the initial guesses, but in only rare instances did this make any significant difference in the final answer. For the stability of the algorithm, it is better to standardize the variables, which yield identity covariance matrices. Note that standardization only changes the scale but not the relationship. Finally, the simulations show that our methods perform well for moderate sample sizes and improve upon this performance as the sample size increases.

2.7 Appendix

2.7.1 Correlation matrix inequality

Proof of correlation determinant inequality Section 2.4.2:

Denote the correlation entries of $\mathbf{\Sigma}_m$ as $\rho(\mathbf{a}^{(i)}, \mathbf{a}^{(j)}) = \rho_{ij}$ and note the following theorem from Schott [14] pg. 250 for the partition of any square matrix.

Theorem. Let $\mathbf{A}_{m\times m}$ be partitioned as follows:

$$\mathbf{A} = \begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix}.$$

Then for $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ nonsingular,

(a) $|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|$.

Let $\mathbf{\Sigma}_{m-1}$ be the $m-1 \times m-1$ leading sub-matrix of the partition of $\mathbf{\Sigma}_m$,

$$\mathbf{\Sigma}_m = \begin{pmatrix} 1 & \cdots & \rho_{1(m-1)} & \rho_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ \rho_{(m-1)1} & \cdots & 1 & \rho_{(m-1)m} \\ \rho_{m1} & \cdots & \rho_{m(m-1)} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{m-1} & \mathbf{\Sigma}_{1m} \\ \mathbf{\Sigma}_{m1} & 1 \end{pmatrix}$$

then,

$$|\mathbf{\Sigma}_m| = |\mathbf{\Sigma}_{m-1}||1 - \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m}| \leq |\mathbf{\Sigma}_{m-1}|.$$

The inequality results from noting that since both $|\mathbf{\Sigma}_m|$ and $|\mathbf{\Sigma}_{m-1}|$ are positive definite correlation matrices, then $|\mathbf{\Sigma}_m|, |\mathbf{\Sigma}_{m-1}| \geq 0$ implies $|1 - \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m}| \geq 0$. In addition, $|1 - \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m}| = 1 - \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m} \geq 0$, which implies $\mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m} \leq 1$. Finally, $\mathbf{\Sigma}_{1m} = \mathbf{\Sigma}_{m1}^T$ are vectors and $\mathbf{\Sigma}_{m-1}$ positive definite implies $\mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m} \geq 0$, thus $0 \leq \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m} \leq 1$. Therefore, $0 \leq |1 - \mathbf{\Sigma}_{m1}\mathbf{\Sigma}_{m-1}^{-1}\mathbf{\Sigma}_{1m}| \leq 1$ and hence the inequality. Using the same argument and partition as above on $\mathbf{\Sigma}_{m-1}$ and each successive correlation matrix yields the following inequality,

$$|\mathbf{\Sigma}_m| \leq |\mathbf{\Sigma}_{m-1}| \leq |\mathbf{\Sigma}_{m-2}| \leq \cdots \leq |\mathbf{\Sigma}_3| \leq |\mathbf{\Sigma}_2| = (1 - \rho_{12}^2).$$

Since $|\mathbf{\Sigma}_m|$ remains unchanged for any arrangement of the correlations in $\mathbf{\Sigma}_m$, letting the correlations in the leading 2×2 sub-matrix of $\mathbf{\Sigma}_m$ be any of the correlations ρ_{ij}, gives $|\mathbf{\Sigma}_m| \leq (1 - \rho_{ij}^2)$ for all $1 \leq i < j \leq m$. Therefore,

$$|\Sigma_m|^{m(m-1)/2} \leq \prod_{1 \leq i < j \leq m} \left[1 - \rho_{ij}^2\right].$$

2.7.2 Consistency

Proof of Consistency, Theorem 1 Section 2.3.4:

Let $\mathbf{U}_i$ be a sequence of k-dimensional random variables with distribution function F and Lebesgue measurable density p. Define the kernel density estimate of p as:

$$p_n(\mathbf{u}) = \frac{1}{na_n^k} \sum_{j=1}^{n} K\left(\frac{\mathbf{u} - \mathbf{U}_j}{a_n^k}\right), \qquad \text{for } \mathbf{u} \in \mathbb{R}^k,$$

where $K : \mathbb{R}^k \to \mathbb{R}^+$ is a probability density on $\mathbb{R}^k$, uniformly for $\|\mathbf{u}\| \to \infty$ and where $a_n > 0$ and $\lim_{n\to\infty} a_n = 0$.

Noting that the theorem 1-m of Kiefer [5] holds for all F, a direct application of Theorem 1 of Ruschendorf [13] yields the following lemma.

Lemma 1 *Let* $\{(\mathbf{Y}_i^T, \mathbf{X}_i^T, \mathbf{Z}_i^T)\}$, $i = 1, \ldots, n$, *be iid, and*

$$\sum_{n=1}^{\infty} \epsilon^{-\gamma n a_n^{2k_r}} < \infty, \text{ for all } \gamma > 0. \text{ where } k_r \text{ depends on } r.$$

Let K *be of bounded variation and let,*

$$p\left(\mathbf{A}^T\mathbf{x}\right), p\left(\mathbf{B}^T\mathbf{y}\right), p\left(\mathbf{C}^T\mathbf{z}\right) \quad \text{be uniformly continuous in } \mathbf{A} \text{ and } \mathbf{x}, \mathbf{B} \text{ and } \mathbf{y}, \mathbf{C} \text{ and } \mathbf{z}$$

$$p\left(\mathbf{A}^T\mathbf{x}, \mathbf{B}^T\mathbf{y}, \mathbf{C}^T\mathbf{z}\right) \quad \text{be uniformly continuous in } \mathbf{A}, \mathbf{B}, \mathbf{C} \text{ and } \mathbf{x}, \mathbf{y}, \mathbf{z}$$

Under these conditions we have:

$$\sup_{\mathbf{A}_{p\times r_1}, \mathbf{x}\in\mathbb{R}^p} \left|p_n\left(\mathbf{A}^T\mathbf{x}\right) - p\left(\mathbf{A}^T\mathbf{x}\right)\right| \to 0 \text{ a.s.} \qquad \text{where } k_r = r_1$$

$$\sup_{\mathbf{B}_{q\times r_2}, \mathbf{y}\in\mathbb{R}^q} \left|p_n\left(\mathbf{B}^T\mathbf{y}\right) - p\left(\mathbf{B}^T\mathbf{y}\right)\right| \to 0 \text{ a.s.} \qquad \text{where } k_r = r_2$$

$$\sup_{\mathbf{C}_{t\times r_3}, \mathbf{z}\in\mathbb{R}^t} \left|p_n\left(\mathbf{C}^T\mathbf{z}\right) - p\left(\mathbf{C}^T\mathbf{z}\right)\right| \to 0 \text{ a.s.} \qquad \text{where } k_r = r_3$$

$$\sup_{\mathbf{A}_{p\times r_1}\mathbf{B}_{q\times r_2}\mathbf{C}_{t\times r_3}, \mathbf{x}\in\mathbb{R}^p, \mathbf{y}\in\mathbb{R}^q, \mathbf{z}\in\mathbb{R}^t} \left|p_n\left(\mathbf{A}^T\mathbf{x}, \mathbf{B}^T\mathbf{y}, \mathbf{C}^T\mathbf{z}\right) - p\left(\mathbf{A}^T\mathbf{x}, \mathbf{B}^T\mathbf{y}, \mathbf{C}^T\mathbf{z}\right)\right| \to 0 \text{ a.s.}$$

$$\text{where } k_r = r_1 + r_2 + r_3$$

Proof of **Theorem 1:**

Assume the conditions of Lemma 1 and let $\epsilon > 0, b > 0 \to 0$ as $n \to \infty$ such that $\epsilon/b \to 0$. In addition, assume $n_b/n \xrightarrow{p} 0$ as $n \to \infty$, where n_b is the number of observations whose indices are not in χ_b.

Let $\mathbf{A}_{p\times r_1}$ be a matrix such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. Note that this matrix is not unique. Also, since for any $r_1 \times r_1$ orthonormal matrix $\mathbf{D}$, $S(\mathbf{A}) = S(\mathbf{AD})$, where $S(\mathbf{A})$ is the subspace spanned by the columns of $\mathbf{A}$; however, $\mathbf{A} \neq \mathbf{AD}$. Therefore, we may use the unique orthogonal projections matrices, $P_{\mathbf{A}} = \mathbf{A}\mathbf{A}^T = P_{\mathbf{D}} = \mathbf{D}\mathbf{D}^T$ instead. That is, the orthogonal matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are not unique but their columns span unique subspaces. For simplicity and ease in notation we assume that these matrices are unique.

Using proof by contradiction, suppose $(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n)$ fails to converge to the unique matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ *with probability* 1. This implies that there exists a subsequence (still denoted by n) and matrices $\mathbf{A}_{0_{p\times r_1}}, \mathbf{B}_{0_{q\times r_2}}$ and $\mathbf{C}_{0_{t\times r_3}}$ satisfying $\mathbf{A}_0^T\mathbf{A}_0 = \mathbf{I}$, $\mathbf{B}_0^T\mathbf{B}_0 = \mathbf{I}$ and $\mathbf{C}_0^T\mathbf{C}_0 = \mathbf{I}$, such that $(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) \to (\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)$ and $(\mathbf{A}, \mathbf{B}, \mathbf{C}) \neq (\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)$.

Thus for any $\epsilon > 0$ and n large enough, from Lemma 1

$$p_n\left(\mathbf{A}_n^T\mathbf{x}_i\right) = p\left(\mathbf{A}_n^T\mathbf{x}_i\right) + \Delta_{1i} = p\left(\mathbf{A}_0^T\mathbf{x}_i\right) + \delta_{1i}$$

$$p_n\left(\mathbf{B}_n^T\mathbf{y}_i\right) = p\left(\mathbf{B}_n^T\mathbf{y}_i\right) + \Delta_{2i} = p\left(\mathbf{B}_0^T\mathbf{y}_i\right) + \delta_{2i}$$

$$p_n\left(\mathbf{C}_n^T\mathbf{z}_i\right) = p\left(\mathbf{C}_n^T\mathbf{z}_i\right) + \Delta_{3i} = p\left(\mathbf{C}_0^T\mathbf{z}_i\right) + \delta_{3i}$$

$$p_n\left(\mathbf{A}_n^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i\right) = p\left(\mathbf{A}_n^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i\right) + \Delta_{4i} = p\left(\mathbf{A}_0^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i\right) + \delta_{4i},$$

such that $\Delta_{ki}, \delta_{ki} < \epsilon$ for all i and $k = 1, \ldots, 4$.

The first equalities above follow from the conclusion of Lemma 1 and the second equalities above from the uniform continuity conditions. Therefore, taking the natural logarithm of each equality above,

$$\begin{aligned}
\log\left(p_n(\mathbf{A}_n^T\mathbf{x}_i)\right) &= \log\left(p(\mathbf{A}_0^T\mathbf{x}_i)\right) + \log\left(1 + \frac{\delta_{1i}}{p(\mathbf{A}_0^T\mathbf{x}_i)}\right) \\
\log(p_n\left(\mathbf{B}_n^T\mathbf{y}_i\right)) &= \log\left(p(\mathbf{B}_0^T\mathbf{y}_i)\right) + \log\left(1 + \frac{\delta_{2i}}{p(\mathbf{B}_0^T\mathbf{y}_i)}\right) \\
\log(p_n\left(\mathbf{C}_n^T\mathbf{z}_i\right)) &= \log\left(p(\mathbf{C}_0^T\mathbf{z}_i)\right) + \log\left(1 + \frac{\delta_{3i}}{p(\mathbf{C}_0^T\mathbf{z}_i)}\right) \\
\log\left(p_n(\mathbf{A}_n^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i)\right) &= \log\left(p(\mathbf{A}_0^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i)\right) \\
&+ \log\left(1 + \frac{\delta_{4i}}{p(\mathbf{A}_0^T\mathbf{x}_i, \mathbf{B}_n^T\mathbf{y}_i, \mathbf{C}_n^T\mathbf{z}_i)}\right).
\end{aligned}$$

Note that, since by definition $\delta_{ki} < \epsilon$ and $\epsilon/b \to 0$, then the last terms in the each equation are $o(1)$. For example, under these conditions and restriction to χ_b (see Theorem 1 Section 2.3.4), $p(\mathbf{A}_0^T \mathbf{x}_i) > b$ implies $0 \leq \frac{\delta_{1i}}{p(\mathbf{A}_0^T \mathbf{x}_i)} < \frac{\epsilon}{b} \to 0$, which implies $\log\left(1 + \frac{\delta_{1i}}{p(\mathbf{A}_0^T \mathbf{x}_i)}\right) \to 0$ as $n \to \infty$.

Next, by the definition of χ_b and subtracting the first three terms from the last in the above,

$$\begin{aligned} \mathcal{I}_n^b(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) &= \frac{1}{n}\sum_{i=1}^{n} J(i \in \chi_b) \log\left(\frac{p\left(\mathbf{A}_0^T \mathbf{x}_i, \mathbf{B}_0^T \mathbf{y}_i, \mathbf{C}_0^T \mathbf{z}_i\right)}{p\left(\mathbf{A}_0^T \mathbf{x}_i\right) p\left(\mathbf{B}_0^T \mathbf{y}_i\right) p\left(\mathbf{C}_0^T \mathbf{z}_i\right)}\right) + o(1) \\ &= \mathcal{I}_n^b(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) + o(1). \end{aligned}$$

Letting χ_b^c denote the complement of the set χ_b, we have the following:

$$\begin{aligned} \mathcal{I}_n^b(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) &= \mathcal{I}_n^b(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) + o(1) \\ \mathcal{I}_n^b(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) &= \mathcal{I}_n(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) - \mathcal{I}_n^{b^c}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0). \end{aligned}$$

Combining these two terms,

$$\mathcal{I}_n^b(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) = \mathcal{I}_n(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) - \mathcal{I}_n^{b^c}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) + o(1).$$

Therefore,

$$\begin{aligned} \mathcal{I}_n^b(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) - \mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) &= \mathcal{I}_n(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) - \mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) \\ &- \mathcal{I}_n^{b^c}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) + o(1). \end{aligned}$$

That is, expanding for clarity,

$$\begin{aligned} \mathcal{I}_n(\mathbf{A}_0\mathbf{B}_0, \mathbf{C}_0) - \mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) &= \left[\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p\left(\mathbf{A}_0^T \mathbf{x}_i, \mathbf{B}_0^T \mathbf{y}_i, \mathbf{C}_0^T \mathbf{z}_i\right)}{p\left(\mathbf{A}_0^T \mathbf{x}_i\right) p\left(\mathbf{B}_0^T \mathbf{y}_i\right) p\left(\mathbf{C}_0^T \mathbf{z}_i\right)}\right)\right. \\ &- \left.\mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)\right] \\ &- \frac{1}{n}\sum_{i=1}^{n} J(i \in \chi_b^c) \log\left(\frac{p\left(\mathbf{A}_0^T \mathbf{x}_i, \mathbf{B}_0^T \mathbf{y}_i, \mathbf{C}_0^T \mathbf{z}_i\right)}{p\left(\mathbf{A}_0^T \mathbf{x}_i\right) p\left(\mathbf{B}_0^T \mathbf{y}_i\right) p\left(\mathbf{C}_0^T \mathbf{z}_i\right)}\right). \end{aligned}$$

By the Law of Large Numbers $\mathcal{I}_n(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) \to \mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)$ and, since $n_b/n \xrightarrow{p} 0$ as $n \to \infty$, $\mathcal{I}_n^{b^c}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) \to 0$. Thus,

$$\lim_{n\to\infty} \mathcal{I}_n^b(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) = \mathcal{I}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0).$$

Note that, by assumption,

$$(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) = \text{argmax}\, \mathcal{I}_n\left(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*\right) \text{ and } (\mathbf{A}, \mathbf{B}, \mathbf{C}) = \text{argmax}\, \mathcal{I}\left(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*\right).$$

Therefore, $\mathcal{I}_n^b\left(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n\right) \geq \mathcal{I}_n^b\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right)$, which implies

$$\mathcal{I}\left(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0\right) = \lim_{n\to\infty} \mathcal{I}_n^b\left(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n\right) \geq \lim_{n\to\infty} \mathcal{I}_n^b\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right) = \mathcal{I}\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right)$$

and also by the assumption $\mathcal{I}\left(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0\right) \leq \mathcal{I}\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right)$. Therefore,

$$\mathcal{I}\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right) \leq \mathcal{I}\left(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0\right) \leq \mathcal{I}\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right), \text{ which implies } \mathcal{I}\left(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0\right) = \mathcal{I}\left(\mathbf{A}, \mathbf{B}, \mathbf{C}\right).$$

But we assumed that the matrices $\mathbf{A}, \mathbf{B},$ and $\mathbf{C}$ are unique, which contradicts the above arguments. Therefore,

$$(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n) \to (\mathbf{A}, \mathbf{B}, \mathbf{C}) \quad \textit{with probability } 1.$$

2.7.3 Properties

Proof of Proposition 1, Section 2.3.2:

As for the first assertion let the number of sets be $m = 3$. Note that, $\mathbf{U}^{(k)} = \mathbf{B}^{(k)^{-1}}\mathbf{X}^{(k)} + \mathbf{b}_k \to \mathbf{X}^{(k)} = \mathbf{B}^{(k)}\mathbf{U}^{(k)} - \mathbf{B}^{(k)}\mathbf{b}_k$ and

$$\begin{aligned} p(\mathbf{A}^{(i)^T}\mathbf{X}^{(i)}, \mathbf{A}^{(j)^T}\mathbf{X}^{(j)} | \mathbf{A}^{(k)^T}\mathbf{X}^{(k)}) &= p(\mathbf{A}^{(i)^T}\mathbf{X}^{(i)}, \mathbf{A}^{(j)^T}\mathbf{X}^{(j)} | \mathbf{A}^{(k)^T}\mathbf{B}^{(k)}\mathbf{U}^{(k)} - \mathbf{A}^{(k)^T}\mathbf{B}^{(k)}\mathbf{b}_k) \\ &= p(\mathbf{A}^{(i)^T}\mathbf{X}^{(i)}, \mathbf{A}^{(j)^T}\mathbf{X}^{(j)} | \mathbf{A}^{(k)^T}\mathbf{B}^{(k)}\mathbf{U}^{(k)}), \end{aligned}$$

where $i \neq j \neq k$. Let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)})$ and $\mathbf{U} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)})$. Use these notations to write,

$$\begin{aligned} \mathcal{I}_{\mathbf{X}}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \mathbf{A}^{(3)}, \mathbf{A}^{(3)}\right) &= \mathrm{E}\left(\log \frac{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \mathbf{A}^{(2)^T}\mathbf{X}^{(2)} | \mathbf{A}^{(3)^T}\mathbf{X}^{(3)})}{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}) p(\mathbf{A}^{(2)^T}\mathbf{X}^{(2)})}\right) \\ &= \mathrm{E}\left(\log \frac{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \mathbf{A}^{(2)^T}\mathbf{X}^{(2)} | \mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}) p(\mathbf{A}^{(2)^T}\mathbf{X}^{(2)})}\right) \\ &= \mathrm{E}\left(\log \frac{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \mathbf{A}^{(2)^T}\mathbf{X}^{(2)}, \mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}) p(\mathbf{A}^{(2)^T}\mathbf{X}^{(2)}) p(\mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left(\log\frac{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)}|\mathbf{A}^{(2)^T}\mathbf{X}^{(2)})}{p(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)})p(\mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}\right)\\
&= \cdots\\
&= \mathrm{E}\left(\log\frac{p(\mathbf{A}^{(1)^T}\mathbf{B}^{(1)}\mathbf{U}^{(1)}, \mathbf{A}^{(2)^T}\mathbf{B}^{(2)}\mathbf{U}^{(2)}, \mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}{p(\mathbf{A}^{(1)^T}\mathbf{B}^{(1)}\mathbf{U}^{(1)})p(\mathbf{A}^{(2)^T}\mathbf{B}^{(2)}\mathbf{U}^{(2)})p(\mathbf{A}^{(3)^T}\mathbf{B}^{(3)}\mathbf{U}^{(3)})}\right)\\
&= \mathcal{I}_{\mathbf{U}}^{(3)}\left(\mathbf{B}^{(1)^T}\mathbf{A}^{(1)}, \mathbf{B}^{(2)^T}\mathbf{A}^{(2)}, \mathbf{B}^{(3)^T}\mathbf{A}^{(3)}\right).
\end{aligned}
$$

The same argument extended to m-sets proves the assertion.

Proof of Proposition 2, Section 2.3.2:

By the definition of KL information (Kullback [9]) pages 14 and 15, $\mathrm{E}\left(\log\frac{p(\mathbf{X})}{q(\mathbf{X})}\right) \geq 0$, where p and q are probability density functions. The result of proposition 2 is immediate since, the numerator is a probability density function(joint) and the denominator is a product of density functions, which is also density function.

Proof of Proposition 3, Section 2.3.2:

Assuming $\mathbf{C}^{(1)}, \ldots, \mathbf{C}^{(m)}$ are all nonsingular, hence $\mathbf{C}^{(k)^{-1}}$ exists for all k, $k = 1, \ldots, m$, define the following transformations,

$$
\begin{gathered}
\mathbf{U}_1 = \mathbf{C}^{(1)^T}\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{U}_m = \mathbf{C}^{(m)^T}\mathbf{A}^{(m)^T}\mathbf{X}^{(m)} \rightarrow\\
\mathbf{C}^{(1)^{T^{-1}}}\mathbf{U}_1 = \mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{C}^{(m)^{T^{-1}}}\mathbf{U}_m = \mathbf{A}^{(m)^T}\mathbf{X}^{(m)}.
\end{gathered}
$$

By variable transformation, letting $\frac{\partial_j}{\partial \mathbf{U}_j} = \frac{\partial_j \mathbf{A}^{(j)^T}\mathbf{X}^{(j)}}{\partial \mathbf{U}_j}, j = 1, \ldots, m$, the following is true,

$$
\begin{aligned}
\frac{p_{\mathbf{U}_1,\ldots,\mathbf{U}_m}(\mathbf{U}_1, \ldots, \mathbf{U}_m)}{p_{\mathbf{U}_1}(\mathbf{U}_1)\cdots p_{\mathbf{U}_m}(\mathbf{U}_m)} &= \frac{p_{\mathbf{A}^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}}\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}{p_{\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}}\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}\right)\cdots p_{\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}}\left(\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}\\
&\times \frac{\left|\frac{\partial_1}{\partial \mathbf{U}_1}\mathbf{C}^{(1)^{T^{-1}}}\mathbf{U}_1\right| \times \cdots \times \left|\frac{\partial_m}{\partial \mathbf{U}_m}\mathbf{C}^{(m)^{T^{-1}}}\mathbf{U}_m\right|}{\left|\frac{\partial_1}{\partial \mathbf{U}_1}\mathbf{C}^{(1)^{T^{-1}}}\mathbf{U}_1\right| \times \cdots \times \left|\frac{\partial_m}{\partial \mathbf{U}_m}\mathbf{C}^{(m)^{T^{-1}}}\mathbf{U}_m\right|}\\
&= \frac{p_{\mathbf{A}^{(1)^T}\mathbf{X}^{(1)},\ldots,\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}}\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}{p_{\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}}\left(\mathbf{A}^{(1)^T}\mathbf{X}^{(1)}\right)\cdots p_{\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}}\left(\mathbf{A}^{(m)^T}\mathbf{X}^{(m)}\right)}.
\end{aligned}
$$

Therefore,

$$
\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}\mathbf{C}^{(1)}, \ldots, \mathbf{A}^{(m)}\mathbf{C}^{(m)}\right) = \mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right).
$$

Proof of Proposition 4, Section 2.3.2:

Since Δ_{11} and Δ_{22} are non-singular, the following is true,
$|\Delta| = |\Delta_{11}| \left|\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right| = |\Delta_{22}| \left|\Delta_{11} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21}\right|$; see Appendix 2.7.1. Then,

$$\begin{aligned}
\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}\right) &= \mathrm{E}\log\left(\frac{p(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})}{p(\mathbf{Y}^{(1)})p(\mathbf{Y}^{(2)})}\right) = \mathrm{E}\log\frac{|\Delta_{11}|^{1/2}|\Delta_{22}|^{1/2}}{|\Delta|^{1/2}} \\
&+ \mathrm{E}\left[\frac{1}{2}\mathbf{Y}^{(1)^T}\Delta_{11}^{-1}\mathbf{Y}^{(1)} + \frac{1}{2}\mathbf{Y}^{(2)^T}\Delta_{22}^{-1}\mathbf{Y}^{(2)} - \frac{1}{2}\mathbf{Y}^T\Delta^{-1}\mathbf{Y}\right] \\
&= -\frac{1}{2}\log\frac{|\Delta|}{|\Delta_{11}|\,|\Delta_{22}|} + \frac{1}{2}\mathrm{tr}\left(\Delta_{11}^{-1}\Delta_{11}\right) \\
&+ \frac{1}{2}\mathrm{tr}\left(\Delta_{22}^{-1}\Delta_{22}\right) - \frac{1}{2}\mathrm{tr}\left(\Delta^{-1}\Delta\right) \\
&= -\frac{1}{2}\log\left(|\Delta|\left|\Delta_{11}^{-1}\right|\left|\Delta_{22}^{-1}\right|\right) + \frac{1}{2}\left[\mathrm{t}+\mathrm{s}-(\mathrm{t}+\mathrm{s})\right] \\
&= -\frac{1}{2}\log\left[|\Delta_{11}|\left|\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right|\left|\Delta_{11}^{-1}\right|\left|\Delta_{22}^{-1}\right|\right] \\
&= -\frac{1}{2}\log\left|\mathbf{I}_{s\times s} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\Delta_{22}^{-1}\right| \\
&= -\frac{1}{2}\log\left|\mathbf{I}_{t\times t} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21}\Delta_{11}^{-1}\right|.
\end{aligned}$$

Now, let $Z = \begin{pmatrix} \Delta_{11}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \Delta_{22}^{-1/2} \end{pmatrix} \mathbf{Y} \sim N_{(t+s)}\left(\mathbf{0}, \Delta_{(\mathbf{z})}\right)$ where,

$$\Delta_{(\mathbf{z})} = \begin{pmatrix} \mathbf{I} & \Delta_{11}^{-1/2}\Delta_{12}\Delta_{22}^{-1/2} \\ \Delta_{22}^{-1}\Delta_{21}\Delta_{11} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \rho_{\mathbf{Y}^{(1)},\mathbf{Y}^{(2)}} \\ \rho_{\mathbf{Y}^{(2)},\mathbf{Y}^{(1)}} & \mathbf{I} \end{pmatrix}.$$

Thus,

$$\mathcal{I}^{(\mathrm{G})}\left(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}\right) = -\frac{1}{2}\log\left|\mathbf{I}_{t\times t} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21}\Delta_{11}^{-1}\right| = -\frac{1}{2}\log\left|\mathbf{I}_{t\times t} - \rho_{\mathbf{Y}^{(1)},\mathbf{Y}^{(2)}}\rho^T_{\mathbf{Y}^{(1)},\mathbf{Y}^{(2)}}\right|.$$

2.8 References

[1] Dryden, I.L., and Mardia, K.V. (1998). Statistical Shape Analysis. Wiley, Chichester.

[2] Härdle, W. (1990). Applied nonparametric regression. Cambridge University Press.

[3] Joe, H. (1989). Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Society*, Vol. 84, No. 405. (Mar., 1989) 157-64.

[4] Kettenring, J.R. (1971). Canonical correlation analysis of several sets of variables. In *Biometrika*, 58, No. 3, 433-51.

[5] Kiefer, J (1961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, *Pacific Journal of Mathematics*, 11, 649-59.

[6] Klingenberg, C.P., McIntyre, G. S. (1998). Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with procrustes methods. Evolution, 52, 1363-75.

[7] Klingenberg, C.P., Barluenga, M. and Meyer, A. (2002). Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. Evolution, 56, 1909-20.

[8] Klingenberg, C.P. (2005). Developmental constraints, modules and evolvability. Pages 219-47 in Variation: A Central Concept in Biology (B. Hallgrimsson and B.K. Hall, eds.). Elsevier, Burlington, MA.

[9] Kullback S.(1959). Information Theory and Statistics. John Wiley & Sons.

[10] Li, B. Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33, 1580-1616.

[11] Neuenschwander, B.E. and Flury, B.D. (1995). Common canonical variates. *Biometrika*, 82, No. 3, 553-60.

[12] Robert, P. and Escoufier, Y.(1976). A unifying tool for multivariate statistical methods: The RV Coefficient. *Applied Statistics*, Vol. 25, No. 3, 257-65.

[13] Ruschendorf, L. (1977). Consistency of estimators for multivariate density functions and for the mode, *Sankhyā*, Series A 39 243-50.

[14] Schott, J.R. (1997). Matrix Analysis for Statistics. John Wiley & Sons, Inc.

[15] Scott, D.W. (1992). Multivariate density estimation: Theory, Practice and Visualization. John Wiley & Sons.

[16] Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman & Hall.

[17] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), An adaptive estimation of dimension reduction. *Journal of the Royal Statistical Society*, Ser. B, Vol. 64, No. 3, 363-410.

[18] Yin, X (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91, 161-76.

[19] Yin, X and Sriram, T. N. (2006). Common canonical variates for independent groups using information theory. *Statistica Sinica*, to appear.

Chapter 3

A Multivariate Dimension Reduction Method for Multiple Sets: An Extension of Canonical Correlation Analysis[2]

[2]Iaci, R., Yin, X. and Sriram, T. N. To be submitted to: *Journal of Computational and Graphical Statistics.*

Abstract

This paper proposes a new index to measure relationships between two sets of multivariate random vectors adopting a projection pursuit regression approach in association studies. The index is based on the idea of finding the projected random vectors which minimize an overall L_2 distance between a linear combination of one and an unknown function of the other. The unknown functions are estimated using the Nadaraya-Watson [9] [12] estimator, as in the nonparametric approach to estimating a regression curve. Extensions to multiple sets of random vectors and multiple groups of multiple sets are also discussed. All these methods to detect relationships are illustrated through extensive simulations and used to analyze a real dataset. In contrast to canonical correlation analysis, our method detects the existence of both linear and nonlinear relationships, thereby making it useful in many other applications.

Key Words and Phrases: Single index model; Nadaraya-Watson estimator; Dimension reduction; Projection pursuit; Bootstrapping.

3.1 Introduction

In this paper, we first propose a new index to detect relationships between two sets of random vectors and then extend this index to multiple sets and multiple groups. The index can be initially described as a projection pursuit method with a single index model step where the usual multivariate response is projected along with the predictor variables. That is, we find the linear combinations of each of the random vectors that minimize a composite distance between the projected vectors. To this end, we interchange the roles of the vectors as response and predictor and simultaneously minimize a composite least squares criterion for each of the roles. In the regression step we use the nonparametric Nadaraya-Watson [9] [12] estimator, thus enabling our index to detect both linear and nonlinear relationships. After the first coefficient vectors that minimize our index have been found we search in the orthogonal

direction for the next set of minimizing coefficient vectors. These coefficient vectors project the random vectors onto a subspace of maximum correlation making our method a projection pursuit method and a generalization of Canonical Correlation Analysis(CCA). Although pairwise plots of these projected random vectors, termed variates as in canonical correlation analysis, can be used to determine visually the number of significant pairs we formally propose a bootstrap procedure in Section 3.2.4 to determine numerically the number of significant relationships. The bootstrap methodology is an extension of that used by Ye and Weiss [14] to select between classes of dimension reduction methods in a regression setting. Different distance measures between subspaces, that of Ye and Weiss [14] and 2-norm based measures are used to determine the number of significant relationships between the sets of vectors, thereby making it a dimension reduction method.

In Section 3.2 we introduce our new method, termed Generalized Canonical Analysis(GCA), in detail. Section 3.2.1 extends this index to multiple sets and multiple groups of random vectors. A computational algorithm for the generalized index is given in Section 3.2.3. The bootstrap procedure to determine the number of significant associations is described in Section 3.2.4. Extensive simulations are performed in Sections 3.3.1-3.3.4 and a real dataset is analyzed in Section 3.3.5.

3.2 A Generalized index

Consider the model $\mathbf{b}^T\mathbf{Y} = \mathbf{a}^T\mathbf{X} + \epsilon$, where $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{q\times 1}$ are multivariate random vectors, $\mathbf{a}_{p\times 1}$ and $\mathbf{b}_{q\times 1}$ coefficient vectors and ϵ an error term with some unknown distribution F. For a fixed $\mathbf{b}$, finding the coefficient vector $\mathbf{a}$ that minimizes the error function $\mathbf{Q}(\mathbf{a}) = \mathrm{E}(\mathbf{b}^T\mathbf{Y} - \mathbf{a}^T\mathbf{X})^2$ is the method of classical linear regression for the model $\mathbf{b}^T\mathbf{Y} = m(\mathbf{a}^T\mathbf{X}) + \epsilon$ with regression function $m(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{X} = \mathrm{E}(\mathbf{Y}|\mathbf{X})$.

The goal of Canonical Correlation Analysis(CCA) is to find the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ that project $\mathbf{X}$ and $\mathbf{Y}$ into a one-dimensional subspace such that the correlation between $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ is maximum. These coefficient vectors maximize $\mathbf{C}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{\Sigma_{XY}}\mathbf{b}$, where $\mathbf{\Sigma_{XY}}$

is the joint covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, with the constraint $\mathbf{a}^T\mathbf{\Sigma_X}\mathbf{a} = \mathbf{b}^T\mathbf{\Sigma_Y}\mathbf{b} = \mathbf{I}$. However, this is equivalent to finding the $\mathbf{a}$ and $\mathbf{b}$ that minimize $\mathbf{Q}^*(\mathbf{a}, \mathbf{b}) = \mathrm{E}(m_2(\mathbf{b}^T\mathbf{Y}) - m_1(\mathbf{a}^T\mathbf{X}))^2$ where $m_i(u) = u,\ i = 1, 2$; see Appendix 3.5.2. To relate CCA to linear regression one interchanges the roles of the predictor and response for both vectors and defines the two regression models; $\mathbf{b}^T\mathbf{Y} = m_1(\mathbf{a}^T\mathbf{X}) + \epsilon$ and $\mathbf{a}^T\mathbf{X} = m_2(\mathbf{b}^T\mathbf{Y}) + \epsilon$, where the $m_i,\ i = 1, 2$ are viewed as regression functions. Now, for $\mathbf{b}$ fixed, minimize $\mathbf{Q}(\mathbf{a}) = \mathrm{E}(\mathbf{b}^T\mathbf{Y} - m_1(\mathbf{a}^T\mathbf{X}))^2$ with respect to $\mathbf{a}$ and, for $\mathbf{a}$ fixed, minimize $\mathbf{Q}(\mathbf{b}) = \mathrm{E}(\mathbf{a}^T\mathbf{X} - m_2(\mathbf{b}^T\mathbf{Y}))^2$ with respect to $\mathbf{b}$. Observe that simultaneously minimizing $\mathbf{Q}^*(\mathbf{a}, \mathbf{b}) = \mathbf{Q}(\mathbf{a}) + \mathbf{Q}(\mathbf{b})$, where $m_i(u) = u,\ i = 1, 2$, with respect to $\mathbf{a}$ and $\mathbf{b}$ is equivalent to maximizing $\mathbf{C}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{\Sigma_{XY}}\mathbf{b}$, as shown in Appendix 3.5.2.

For a two-set configuration we extend the above arguments by assuming a more general single index regression model $\mathbf{Y} = m(\mathbf{a}^T\mathbf{X}) + \epsilon$, where m is a linear or nonlinear function of $\mathbf{X}$. That is, the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ that project $\mathbf{X}$ and $\mathbf{Y}$ into a one-dimensional subspace with maximum correlation are found by minimizing the following population index with respect to $\mathbf{a}$ and $\mathbf{b}$,

$$\mathcal{G}(\mathbf{a}, \mathbf{b}) \quad = \quad \mathrm{E}(\mathbf{b}^T\mathbf{Y} - m_1(\mathbf{a}^T\mathbf{X}))^2 + \mathrm{E}(\mathbf{a}^T\mathbf{X} - m_2(\mathbf{b}^T\mathbf{Y}))^2. \tag{3.1}$$

Note that, minimizing this index obtains estimates of $\mathbf{a}$ and $\mathbf{b}$ by simultaneously utilizing the functional relationships between both $\mathbf{X}$ and $\mathbf{Y}$. Because of the above discussion, when $m_1(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{X}$ and $m_2(\mathbf{b}^T\mathbf{Y}) = \mathbf{b}^T\mathbf{Y}$, our method is equivalent to CCA(Appendix 3.5.2) and therefore can be thought of as a generalization of CCA. We refer to our method as Generalized Canonical Analysis (GCA). If m_1 and m_2 are known then our index is a measure of the correlation between $\mathbf{b}^T\mathbf{Y}$ and $m_1(\mathbf{a}^T\mathbf{X})$ and $\mathbf{a}^T\mathbf{X}$ and $m_2(\mathbf{b}^T\mathbf{Y})$. However, the functions m_1 and m_2 are generally unknown in practice, and hence need to be estimated. Our main goal is to identify the linear combinations $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$, which also involves choosing an appropriate estimator of the regression functions m_1 and m_2. For our calculations we use the Nadarya-Watson [9] [12] smoother to estimate the regression functions and thereby our method is shown to detect both linear and nonlinear relationships.

We summarize the pair-wise associations between $\mathbf{X}_{q\times 1}$ and $\mathbf{Y}_{p\times 1}$ by finding the first pair of coefficient vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ that minimize $\mathcal{G}(\mathbf{a},\mathbf{b})$. These coefficient vectors project the random vectors, $\mathbf{a}_1^T\mathbf{X}$ and $\mathbf{b}_1^T\mathbf{Y}$, termed variates, onto a subspace such that they have maximum correlation. As in CCA, we constrain these variates to have unit variance. Next, we search in an orthogonal direction for the next set of coefficient vectors, $\mathbf{a}_2$ and $\mathbf{b}_2$, such that they project into a subspace that is uncorrelated with the first projection. That is, we constrain the variates to have unit variance, $\text{Var}(\mathbf{a}_i^T\mathbf{X}) = \mathbf{a}_i^T\Sigma_{\mathbf{X}}\mathbf{a}_i = 1$ and, be uncorrelated, $\rho(\mathbf{a}_i^T\mathbf{X}, \mathbf{a}_j^T\mathbf{X}) = \mathbf{a}_i^T\Sigma_{\mathbf{X}}\mathbf{a}_j = 0$, $i \neq j$. More succinctly:

$$\begin{aligned}\mathbf{a}_i^T\Sigma_{\mathbf{X}}\mathbf{a}_i &= \mathbf{b}_i^T\Sigma_{\mathbf{Y}}\mathbf{b}_i = 1 \text{ for all } i = 1,\ldots,\min(q,p)\\ \mathbf{a}_i^T\Sigma_{\mathbf{X}}\mathbf{a}_j &= \mathbf{b}_i^T\Sigma_{\mathbf{Y}}\mathbf{b}_j = 0 \text{ for all } j = 1,\ldots,i-1.\end{aligned}$$

It is possible that when the estimated variates, $\widehat{\mathbf{a}}^T\mathbf{X}$ and $\widehat{\mathbf{b}}^T\mathbf{Y}$, are plotted against each other, one of the plots could reveal a non-functional relationship. For example, if the true relationship is $Y_1 = X_1^2 + \epsilon$, for the true coefficient vectors $\mathbf{a} = (1,0,\ldots)^T$ and $\mathbf{b} = (1,0,\ldots)^T$, regressing $\mathbf{a}^T\mathbf{X}(=\mathbf{X}_1)$ on $\mathbf{b}^T\mathbf{Y}(=\mathbf{Y}_1)$ results in a regression on the non-functional relationship $\mathbf{X}_1 = \pm\sqrt{\mathbf{Y}_1} + \epsilon$. The best fit, in a regression sense, is an estimate estimate of the mean of $\mathbf{b}^T\mathbf{Y} = \mathbf{Y}_1$, which would be a constant function. Therefore, in this situation the second term in (3.1) creates unnecessary noise that may affect the accuracy of the estimates. To overcome this, we suggest another index, termed the reduced index:

$$\begin{aligned}\mathcal{G}_r(\mathbf{a},\mathbf{b}) &= \text{E}(\mathbf{b}^T\mathbf{Y} - m_1(\mathbf{a}^T\mathbf{X}))^2\\ &\text{or}\\ &= \text{E}(\mathbf{a}^T\mathbf{X} - m_1(\mathbf{b}^T\mathbf{Y}))^2. \qquad (3.2)\end{aligned}$$

The performance of $\mathcal{G}_r(\mathbf{a},\mathbf{b})$ is investigated in Simulation 3.3.1 below.

3.2.1 Extensions of the general index

Multiple sets

For m-sets of random vectors $\mathbf{X}^{(1)}_{p_1\times 1}, \mathbf{X}^{(2)}_{p_2\times 1}, \mathbf{X}^{(3)}_{p_3\times 1}, \ldots, \mathbf{X}^{(m)}_{p_m\times 1}$ with associated coefficient vectors denoted $\mathbf{a}^{(k)}$, $k = 1, \ldots, m$, we extend our measure in (3.1) as,

$$\mathcal{G}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}\right) = \sum_{1\le i<j\le m} \mathcal{G}\left(\mathbf{a}^{(i)}, \mathbf{a}^{(j)}\right). \tag{3.3}$$

That is, we sum our general index over all possible two-set combinations and minimize subject to the constraints:

$$\begin{aligned} \mathbf{a}_i^{(k)T} \Sigma_{\mathbf{X}^{(k)}} \mathbf{a}_i^{(k)} &= 1 \text{ for all } i = 1, \ldots, \min(p_k) \\ \mathbf{a}_i^{(k)T} \Sigma_{\mathbf{X}^{(k)}} \mathbf{a}_j^{(k)} &= 0 \text{ for all } j = 1, \ldots, i-1. \end{aligned}$$

That is, the variates are constrained to have unit variance and be uncorrelated for all sets.

Multiple groups

Suppose the data on m-sets of random vectors are collected from g independent but closely related groups and one is interested in measuring the relationship between the m-sets within each group. Let $\mathbf{X}^{(1)w}, \mathbf{X}^{(2)w}, \ldots, \mathbf{X}^{(m)w}$ denote the m-sets in group w, $w = 1, \ldots, g$. We define the common GCA as the weighted average of the multiple set index over each group,

$$\mathcal{G}^{(\mathrm{C})}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}\right) = \sum_{w=1}^{g} \mathcal{G}^{w}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}\right) \mathrm{P}\left(W = w\right). \tag{3.4}$$

As can be seen in these measures we are assuming common coefficient vectors across the g groups. We weight the GCA index for each group by $\mathrm{P}\left(W = w\right)$, which is commonly taken to be $n_w/(n_1 + n_2 + \cdots + n_g)$, where n_w is the number of observations in the w^{th} group. The coefficient vectors are constrained as follows:

common coefficient constraints

$$\mathbf{a}_i^{(k)T} \Sigma_{\mathbf{X}^{(k)}} \mathbf{a}_i^{(k)} = 1 \quad \text{for all } i = 1, \ldots, \min(p_k)$$

$$\mathbf{a}_i^{(k)^T} \Sigma_{\mathbf{X}^{(k)}} \mathbf{a}_j^{(k)} = 0 \quad \text{for all } j = 1, \ldots, i-1$$

where $\Sigma_{\mathbf{X}^{(k)}}$ is the pooled covariance matrix of the k^{th} set of variables across the g groups.

Note that if the vectors are "whitened" for each set of variables within every group,

$$\mathbf{Z}_{p_k}^{(k)w} = \Sigma_{x^{(k)w}}^{-1/2} \mathbf{X}_{p_k}^{(k)w} \Longrightarrow \text{cov}\left(\mathbf{Z}_{p_k}^{(k)w}\right) = \mathbf{I}_{p_k \times p_k},$$

then the common coefficient vectors satisfy the constraints $\mathbf{a}_j^{(k)^T} \mathbf{a}_i^{(k)} = 0$ and $\mathbf{a}_j^{(k)^T} \mathbf{a}_j^{(k)} = 1,\ i \neq j$. This allows two options when transforming back to the original scale. First, the common coefficient vectors in the original scale are satisfied by the transformation $\mathbf{a}_j^{(k)w*} = \Sigma_{\mathbf{X}}^{1/2} \mathbf{a}_j^{(k)}$ where $\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}^{(k)}}$ is the pooled covariance matrix. Second, is to let $\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}^{(k)w}}$ be the covariance matrix for the k^{th} set in the w^{th} group. We refer to these constraints as the "individual" common coefficient constraints, which satisfy:

"individual" common coefficient constraints

$$\mathbf{a}_i^{(k)w^T} \Sigma_{\mathbf{X}^{(k)w}} \mathbf{a}_i^{(k)w} = 1 \quad \text{for all } i = 1, \ldots, \min(p_k)$$

$$\mathbf{a}_i^{(k)w^T} \Sigma_{\mathbf{X}^{(k)w}} \mathbf{a}_j^{(k)w} = 0 \quad \text{for all } i = 1, \ldots, j-1, \text{ and } w = 1, \ldots, g$$

where $\mathbf{a}_i^{(k)w}$ and $\Sigma_{\mathbf{X}^{(k)w}}$ are the i^{th} coefficient vector and covariance matrix for the k^{th} set of variables in the w^{th} group.

These two different methods for translating back to the original scale depend on whether an equal covariance structure for both groups is assumed. The common coefficient vector constraints are used throughout the analysis, since this is an appropriate choice for the datasets analyzed.

3.2.2 Sample version

We estimate the population version $\mathcal{G}(\mathbf{a}, \mathbf{b})$ using a composite least squares criterion defined as follows:

$$\mathcal{G}_n(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n} \left[(\mathbf{b}^T \mathbf{y}_j - \widehat{m}_1(\mathbf{a}^T \mathbf{x}_j))^2 + (\mathbf{a}^T \mathbf{x}_j - \widehat{m}_2(\mathbf{b}^T \mathbf{y}_j))^2 \right], \tag{3.5}$$

where the functions $\widehat{m}_1$ and $\widehat{m}_2$ are estimated as

$$\widehat{m}_1\left(\mathbf{a}^T\mathbf{x}_j\right) = \frac{1}{n}\sum_{i=1}^{n} W_{ni}\left(\mathbf{a}^T\mathbf{x}_j\right)\mathbf{b}^T\mathbf{y}_i \quad \text{and} \quad \widehat{m}_2(\mathbf{b}^T\mathbf{y}_j) = \frac{1}{n}\sum_{i=1}^{n} W_{ni}(\mathbf{b}^T\mathbf{y}_j)\mathbf{a}^T\mathbf{x}_i,$$

for a sequence of weights $W_{ni}(z)$ such that $\sum_{i=1}^{n} W_{ni}(z) = 1$. In our analysis we use the kernel estimated weight sequence proposed by Nadaraya [9] and Watson [12],

$$W_{ni}\left(z\right) = \frac{1}{h}K\left(\frac{z - z_i}{h}\right) \Big/ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{z - z_i}{h}\right).$$

The estimated coefficient vectors for $\mathbf{a}$ and $\mathbf{b}$, denoted $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$, are found by minimizing the sample version $\mathcal{G}_n\left(\mathbf{a}, \mathbf{b}\right)$, that is,

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}) = \arg\min_{\mathbf{a},\mathbf{b}} \mathcal{G}_n\left(\mathbf{a}, \mathbf{b}\right).$$

The choice of kernel and associated bandwidth is always a concern in methods dealing with kernel density estimation. We choose two commonly selected kernels and three bandwidths used in univariate density estimation and investigate their performance for our method.

Generally the choice of kernels may not be crucial in projection pursuit methods, Härdle [4], therefore we investigate the choice of kernel based on a criterion developed for univariate density estimation. Specifically we choose the Epanechnikov kernel, which is the most efficient in terms of minimizing the approximate mean integrated square error(AMISE) and the commonly used Gaussian kernel; see Silverman [10]. The MISE and AMISE are defined as:

$$\begin{aligned} \mathrm{MISE}(\widehat{f}) &= \mathrm{E}\int\left(\widehat{f}(x) - f(x)\right)^2 dx \\ \mathrm{MISE}(\widehat{f}) \approx \mathrm{AMISE}(\widehat{f}) &= \frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx + \frac{1}{nh}\int \mathrm{K}(t)^2 dt, \end{aligned}$$

where $k_2 = \int t^2 \mathrm{K}(t)dt \neq 0$. The Gaussian kernel, $\mathrm{K}_1(t)$, and Epanechnikov kernel, $\mathrm{K}_2(t)$, are defined as follow:

$$\mathrm{K}_1(t) = (1/\sqrt{2\pi})\mathrm{e}^{-(1/2)t^2} \text{ and } \mathrm{K}_2(t) = (3/4)\left(1 - t/5\right)/\sqrt{5} \text{ for } |t| \leq \sqrt{5}\ ; 0 \text{ otherwise}$$

While the choice of kernel may not be crucial, bandwidth selection may be; see Härdle [4]. We choose three bandwidths, relative to the criterion used in the choice of the Gaussian kernel, to investigate our methods sensitivity to the choice of bandwidth. The following bandwidths, suggested by Silverman [10], are used in each of the simulations and are defined as follow:

$$\mathrm{h}_1 = (4/3)^{1/5}\, s n^{-1/5},\ \mathrm{h}_2 = 0.79\mathrm{R}n^{-1/5} \text{ and } \mathrm{h}_3 = 0.9\mathrm{A}n^{-1/5}$$

Here s is the sample standard deviation, R = IQR(Interquartile Quartile Range) and A = min (s, IQR/1.34). In terms of minimizing the AMISE for the Gaussian kernel, h_1 is optimal if the unknown density is normally distributed, h_2 is an improvement in the presence of skewed data and h_3 is an adaptive measure of the first two; see Silverman [10].

For projection pursuit regression, Hall [3] showed that the $\widehat{\mathbf{a}}$ that minimizes $\mathbf{Q}(\mathbf{a}) = \Sigma_{i=1}^{n}(\mathbf{y}_i - \widehat{m}(\mathbf{a}^T\mathbf{x}_i))^2$ is $\sqrt{n}$-consistent for an appropriately chosen bandwidth h, but this h may not be optimal for the estimation of the regression function m. Ichimura [7] gives a range of bandwidths for which $\widehat{\mathbf{a}}$ is $\sqrt{n}$-consistent for $\mathbf{a}$, but this range excludes the optimal bandwidth for estimating m. Härdle, Hall and Ichimura [5] provide an empirical way of selecting the best bandwidth, which is optimal for the estimation of both $\mathbf{a}$ and the regression function m. In future work we propose to show that our estimates of $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ are consistent for $\mathbf{a}$ and $\mathbf{b}$, respectively.

3.2.3 Computational algorithm

We detail the computational algorithm for the m-set population version $\mathcal{G}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}\right)$ of Section 3.2.1, replaced by the sample version in practice.

Since our measures are invariant to nonsingular matrix transformation, we develop the algorithm assuming the random vectors have been centered and "whitened" to have unit covariance matrices,

$$\mathbf{Z}^{(k)} = \Sigma_{\mathbf{X}^{(k)}}^{-1/2}\left(\mathbf{X}^{(k)} - \mathrm{E}\mathbf{X}^{(k)}\right) \Longrightarrow \mathrm{cov}\left(\mathbf{Z}^{(k)}\right) = \mathbf{I}.$$

This transformation of the vectors simplifies the constraints of Section 3.2 and lessens the effects of ill-conditioned covariance matrices and variables of differing magnitudes. The algorithm is detailed for one iteration, finding the 1st and 2nd coefficient vectors $\mathbf{a}_1^{(k)}$ and $\mathbf{a}_2^{(k)}$. Note, the bullets in each step are additional comments and details pertaining to that step.

Step 0: Find the 1st coefficient vectors $\mathbf{a}_1^{(k)}$ that minimize $\mathcal{G}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}\right)$ for all k, subject to the constraint $\mathbf{a}_1^{(k)^T}\mathbf{a}_1^{(k)} = 1$ for all k.

- The minimization is carried out iteratively using a Sequential Quadratic Programming (SQP) method, specifically the *fmincon* function available in Matlab. The SQP method minimizes the index $\mathcal{G}$ while incorporating the constraints $\mathbf{a}_1^{(k)^T}\mathbf{a}_1^{(k)} = 1$ simultaneously.
- Termination tolerance on the function value and the constraints is 1e-006, the default values in optimset for the *fmincon* function. Convergence is measured when both of these tolerances are satisfied. Since both tolerances are necessary for convergence, the actual function value and constraint tolerances at convergence are smaller than these defaults.

The following steps outline an algorithm to find $\mathbf{a}_2^{(k)}$ orthogonal to $\mathbf{a}_1^{(k)}$. For simplicity, we will fix k and denote $\mathbf{a}_1 = \mathbf{a}_1^{(k)}$, $\mathbf{a}_2 = \mathbf{a}_2^{(k)}$ and $p = p_k$, with the understanding that the algorithm is carried out simultaneously for all $k = 1, \ldots, m$. Let $\mathbf{D}_{n\times p} = \mathbf{D}_{\mathbf{x}^{(k)}}$ be the data matrix for $\mathbf{X}^{(k)}$.

Step 1: Define $\mathbf{A} = \mathbf{a}_1\mathbf{a}_1^T$ and singular value decompose $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{\Sigma}$ is the diagonal matrix of singular values. Using the left singular vectors, let $\mathbf{B}_{p\times p} = [\mathbf{a}_1\ \mathbf{u}_2 \ldots \mathbf{u}_p] = [\mathbf{u}_1\ \ldots \mathbf{u}_p]$.

- For finding $\mathbf{a}_i$, $2 < i < p$, define $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2 \cdots\ \mathbf{a}_{i-1}][\mathbf{a}_1\ \mathbf{a}_2 \cdots\ \mathbf{a}_{i-1}]^T$
- $\mathbf{A}$ is an orthonormal projection matrix with eigenvalues $\lambda = 0$ or 1 with multiplicities, $g_{\mathbf{A}}(\lambda = 1) = \text{rank}(\mathbf{A})$ and $g_{\mathbf{A}}(\lambda = 0) = n - \text{rank}(\mathbf{A})$. The singular

values, σ, are the eigenvalues of $\mathbf{A}$ and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ is a spectral decomposition of $\mathbf{A}$; see Appendix 3.5.3.

- Using the SVD decomposition ensures that $\mathbf{U} = [\mathbf{u}_1 \ \ldots \mathbf{u}_p]$ are an orthonormal basis for $\mathbb{R}^p$.

Step 2: Define $\mathbf{B}^*_{p\times(p-1)} = [\mathbf{u}_2 \ldots \mathbf{u}_p]$ and project the data matrix $\mathbf{D}$ onto this $\perp$ subspace:

$$\mathbf{D}^*_{n\times(p-1)} = \mathbf{D}_{n\times p}\ \mathbf{B}^*_{p\times(p-1)}$$

- The column space of $\mathbf{B}^*$ is orthogonal to $\mathbf{a}_1$.

Step 3: Repeat Step 0, find the coefficient vector $\mathbf{a}^*_1$ that minimizes the index for the new dataset determined by $\mathbf{D}^*$:

$$\mathbf{a}^*_{1_{(p-1)\times 1}} = (a^*_{11}, a^*_{12}, \ldots, a^*_{1(p-1)})^T \text{ such that } \mathbf{a}^{*^T}_1 \mathbf{a}^*_1 = 1.$$

Step 4: Define $\mathbf{a}_2$, corresponding to the original data $\mathbf{D}$, as:

$$\mathbf{a}_{2_{p\times 1}} = a^*_{11}\mathbf{u}_2 + a^*_{12}\mathbf{u}_3 + \cdots a^*_{1(p-1)}\mathbf{u}_p = [\mathbf{u}_2\ \mathbf{u}_3 \ldots \mathbf{u}_p]\ \mathbf{a}^*_1 = \mathbf{B}^*\mathbf{a}^*_1$$

- The constraints are satisfied since:

$$\begin{aligned}
\mathbf{a}_1^T\mathbf{a}_2 &= a^*_{11}\mathbf{a}_1^T\mathbf{u}_2 + a^*_{12}\mathbf{a}_1^T\mathbf{u}_3 + \cdots a^*_{1(p-1)}\mathbf{a}_1^T\mathbf{u}_p = 0 \\
\mathbf{a}_2^T\mathbf{a}_2 &= \mathbf{a}_1^{*^T}\mathbf{B}^{*^T}\mathbf{B}^*\ \mathbf{a}^*_1 = \mathbf{a}_1^{*^T}\mathbf{I}_{(p-1)}\ \mathbf{a}^*_1 = \mathbf{a}_1^{*^T}\mathbf{a}^*_1 = 1.
\end{aligned}$$

Repeat the above steps until the desired number of coefficient vectors $\mathbf{a}_i^{(k)}$, $i = 1, \ldots, \min(p_k)$, have been found. Codes in Matlab have been developed for our computations and are available from the authors. The code requires the user to only input the data and select any of the kernel/bandwidth combinations detailed in Section 3.2.2.

3.2.4 Bootstrap dimension detection method

Suppose there are m-sets of multivariate random vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$, each with dimension p_k and let $p = \min(p_k)$, $k = 1, \ldots, m$. A goal of this method is to determine the number of $d \leq p$ linear combinations $\mathbf{A}_d^{(k)^T}\mathbf{X}^{(k)}$, where $\mathbf{A}_d^{(k)}$ has dimension $p_k \times d$, that detect the

relationships between the m-sets of random vectors such that the loss of information in using the p possible linear combinations is minimal. For each k, let $\mathcal{S}(\mathbf{A}_d^{(k)})$ be the d-dimensional subspace spanned by the columns of $\mathbf{A}_d^{(k)}$. We formulate the problem of determining d as finding the subspaces $\mathcal{S}(\mathbf{A}_d^{(k)})$ that have the lowest variability. To estimate the variability of $\mathcal{S}(\mathbf{A}_d^{(k)})$, we bootstrap the data and calculate a distance between the original and bootstrapped subspaces, for each d, and choose the value $d = d_0$ which yields the smallest distance for all sets. This is an extension of the idea used in Ye and Weiss's [14] bootstrap procedure to choose between different methods for dimension reduction in a regression setting.

We estimate the subspaces using projection matrices. Specifically, for each k, let $\widehat{\mathbf{A}}_d^{(k)} = [\widehat{\mathbf{a}}_1^{(k)} \dots \widehat{\mathbf{a}}_d^{(k)}]$ be the estimated coefficient matrix of $\mathbf{A}_d^{(k)}$, where the columns are the estimated coefficient vectors $\widehat{\mathbf{a}}_j^{(k)^T}$, $j = 1 \dots d \leq p_k - 1$, obtained from GCA (or any method to detect relationships between sets, such as CCA for that matter). Let $\mathcal{S}(\widehat{\mathbf{A}}_d^{(k)})$ denote the estimated subspace of $\mathcal{S}(\mathbf{A}_d^{(k)})$. If the true number of significant relationships is $l < p_k - 1$ then for $d > l$ the remaining $d - l$ coefficient vectors span the null space of $\mathcal{S}(\mathbf{A}_d^{(k)})$ at random. Thus we can measure the variability in $\mathcal{S}(\mathbf{A}_d^{(k)})$ by calculating bootstrapped estimates of a distance, $||\mathcal{S}(\widehat{\mathbf{A}}_d^{(k)}) - \mathcal{S}(\widehat{\mathbf{A}}_d^{(k)b})||$ to be defined below, where $b = 1, \dots, B$ is the number of bootstrap iterations. Now, for $d < l$, assuming not all the columns of the bootstrapped estimated coefficient matrices correspond to the same detected relationships in $\widehat{\mathbf{A}}_d^{(k)}$ for at least one of the m-sets, then the variability between the subspaces will be larger than the variability of the subspaces when $d = l$. That is, even if just one of the columns of one bootstrapped estimated coefficient matrix $\widehat{\mathbf{A}}_d^{(k)b}$, for any $k = 1, \dots, m$, corresponds to a relationship not detected in $\widehat{\mathbf{A}}_d^{(k)}$, the variability should be larger then when $d = l$. Allowing for the very unlikely possibility of all bootstrapped estimates detecting the relationships in the same order as those detected in the original for all sets, then the variability of the subspaces for $d \leq l$ will be of the same order as $d = l$ and plots of the variate pairs can be used to determine visually the correct d.

We propose three different distance measures for measuring the variability between subspaces. The first method uses Hotelling's [6] vector correlation coefficient, suggested in Ye and Weiss [14], extended to m-sets. For a fixed k, this actually measures the correlation between the estimated and bootstrapped subspaces. This method will be compared to Methods 2 and 3 which use the matrix 2-norm distance. The dimension d that minimizes these measures for all k is taken to be the number of significant relationships between the sets of random vectors.

In the first method, we extend Ye and Weiss's [14] measure based on Hotelling's [6] vector correlation coefficient. For a fixed k, b and d, the squared vector correlation coefficient for the orthonormal bases, $\mathbf{A} = \widehat{\mathbf{A}}_d^{(k)} (\mathbf{A}^T\mathbf{A} = \mathbf{I})$ and $\mathbf{B} = \widehat{\mathbf{A}}_d^{(k)b} (\mathbf{B}^T\mathbf{B} = \mathbf{I})$ is,

$$\begin{aligned} q^2_{(k,b,d)} &= \frac{(-1)^d \begin{vmatrix} \mathbf{0} & \mathbf{B}^T\mathbf{A} \\ \mathbf{A}^T\mathbf{B} & \mathbf{A}^T\mathbf{A} \end{vmatrix}}{|\mathbf{B}^T\mathbf{B}||\mathbf{A}^T\mathbf{A}|} \\ &= |\mathbf{B}^T\mathbf{B}|^{-1}|\mathbf{A}^T\mathbf{A}|^{-1}|\mathbf{B}^T\mathbf{A}||\mathbf{A}^T\mathbf{B}| \\ &= |\mathbf{B}^T\mathbf{A}||\mathbf{A}^T\mathbf{B}| = |\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}| = \prod_{i=1}^{d} \lambda_i \end{aligned}$$

where the λ_i are the eigenvalues of $\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}$. The squared vector correlation is positive since, $|\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}| = |\mathbf{B}^T\mathbf{A}||\mathbf{A}^T\mathbf{B}| = |\mathbf{B}^T\mathbf{A}|^2 = \prod_{i=1}^{d} \tau_i^2 \geq 0$, where the τ_i are the eigenvalues of $\mathbf{B}^T\mathbf{A}$. As a measure of the correlation between the subspaces $\mathcal{S}(\widehat{\mathbf{A}}_d^{(k)})$ and $\mathcal{S}(\widehat{\mathbf{A}}_d^{(k)b})$, $0 \leq q^2_{(k,b,d)} \leq 1$, a value of $q = 1$ implies the subspaces are equal and $q = 0$ means the subspaces are orthogonal. Therefore, as in Ye and Weiss [14], we take $(1 - q_{(k,b,d)})$ as our measure of distance between the subspaces so that smaller distances correspond to smaller values. We estimate the variability of the subspace of dimension d, by averaging first over all k and then over the B bootstrapped iterations:

$$\text{Method 1} : \bar{q}_d = \frac{1}{b}\sum_{b=1}^{B} \frac{1}{m}\sum_{k=1}^{m} \left(1 - q_{(k,b,d)}\right). \tag{3.6}$$

Next, we propose a measure of the distance between the subspaces using the matrix 2-norm. We measure the distance between the subspaces by multiplying the orthogonal

projection matrices corresponding to the column space of $\widehat{\mathbf{A}}_d^{(k)}$ and the null space of $\widehat{\mathbf{A}}_d^{(k)b}$. Specifically, we calculate $||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}})||_2$, where the projection matrices are defined as $\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}} = \widehat{\mathbf{A}}_d^{(k)}\widehat{\mathbf{A}}_d^{(k)^T}$ and $\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}} = \widehat{\mathbf{A}}_d^{(k)b}\widehat{\mathbf{A}}_d^{(k)b^T}$. Here, we use the standard matrix 2-norm $||\mathbf{D}_{t\times t}||_2 = \max_{1\leq i\leq t}\sigma_i$, where the $\sigma_i,\ i = 1,\ldots t$, are the singular values of $\mathbf{D}$. Note that the orthonormal constraints imposed on the coefficient vectors provide a unique orthonormal projection matrix onto a d-dimensional subspace of $\mathbb{R}^{p_k}$ and $||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}})||_2 \leq 1$; see Appendix 3.5.3 Result 2. If this distance is close to zero for some d then the estimated and bootstrapped estimated coefficient matrices span the same subspace and therefore correspond to significant relationships. As a measure of variability of the subspaces for a fixed d over the m-sets, we look at both the product and the average of the 2-norm distance over all k and average over the number of bootstrap iterations:

$$\text{Method 2} \ : \ \mathrm{L}_{2_d} = \frac{1}{b}\sum_{b=1}^{B}\prod_{k=1}^{m}||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}})||_2 \tag{3.7}$$

$$\text{Method 3} \ : \ \bar{\mathrm{L}}_{2_d} = \frac{1}{b}\sum_{b=1}^{B}\frac{1}{m}\sum_{k=1}^{m}||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}})||_2. \tag{3.8}$$

For both methods, smaller values imply the projection matrices of the original and bootstrap estimated coefficient matrices estimate the same subspace, $\mathcal{S}(\mathbf{A}_d^{(k)})$.

Note that for all sets where $d = p_k$, the orthonormal constraints imposed on the coefficient vectors mean $\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}}$ is a unique orthonormal projection matrix onto $\mathbb{R}^{p_k}$, where $\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}} = \mathbf{I}$. In this case, the vector correlation coefficient $(1 - q_{(k,b,d)}) = 0$ since for $\mathbf{A} = \widehat{\mathbf{A}}_d^{(k)}$ and $\mathbf{B} = \widehat{\mathbf{A}}_d^{(k)b}$,

$$q^2_{(k,b,d)} \ = \ |\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}| = |\mathbf{A}\mathbf{A}^T||\mathbf{B}\mathbf{B}^T| = |\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}}| = |\mathbf{I}_d||\mathbf{I}_d| = 1$$

and the 2-norm measure $||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)b}})||_2 = ||\mathbf{P}_{\widehat{\mathbf{A}}_d^{(k)}}(\mathbf{0})||_2 = 0$. Thus, for all sets where $d = p_k$, the measures in Methods 1, 2 and 3 are all zero; and all misrepresent the estimated variability of the subspaces $\mathcal{S}(\mathbf{A}_d^{(k)})$.

In practice, we calculate the above measures for $d = 1,\ldots,p-1$, $p = \min(p_k)$ for all k, and create a bar-plot of each measure and choose the d corresponding to the smallest bar. Plots of the 1^{st} through the p^{th} variate pairs can also be compared to the bar-plots

for confirmation that d is determined correctly. In cases where there are relationships up to dimension $d = (p-1)$, that is if $d = (p-1)$ corresponds to the smallest value (or bar), pair-wise plots of the p^{th} variate pairs can be used to determine visually whether there are $d = p$ or $(p-1)$ significant relationships. That is, if the p^{th} variate pairs show no relationship between the k sets then $d = (p-1)$. However, if the plots of the p^{th} variate pairs show a relationship then we choose $d = p$.

3.3 Simulations and real data sets

In the simulations that follow various scenarios involving sample sizes, mixtures of linear and nonlinear relationships, initial guess strategies and random vectors with varying distributions are investigated. In addition, the GCA index is calculated for each simulated dataset using all the kernel/bandwidth combinations given in Section 3.2.2.

To assess our method's ability to detect the dependence relationships we use plots of the variates and two point estimates of association. For one measure we calculate the absolute value of the correlation between the estimated and true variates. For example, if the 1^{st} estimated variate is $\widehat{\mathbf{a}}_1^T\mathbf{x}$, corresponding to the true variate $\mathbf{a}^T\mathbf{x}$, we calculate the absolute value of the correlation between these variates as $|\rho\left(\widehat{\mathbf{a}}_1^T\mathbf{x}, \mathbf{a}^T\mathbf{x}\right)|$ and report the average over all simulations, denoted $|\bar{\rho}_1|$. Here $\mathbf{x}$ denotes a sample from the random vector $\mathbf{X}$. The absolute value is taken since the coefficient vectors can vary in sign, that is $-\widehat{\mathbf{a}}$ detects the same relationship as $\widehat{\mathbf{a}}$. The second is a distance measure, similar to that used by Xia et al. [13] and Li, Zha and Chiaromonte [8], between the true and estimated coefficient vectors, defined as $||(\mathbf{I} - \mathbf{a}\mathbf{a}^T)\widehat{\mathbf{a}}^T||_2$. This distance is calculated in all our simulations and the average reported in the tables, denoted $||\cdot||_2$. Here the standard vector 2-norm is used, $||u||_2 = \sqrt{u^T u}$. $(\mathbf{I} - \mathbf{a}\mathbf{a}^T)\widehat{\mathbf{a}}^T$ is the projection of $\widehat{\mathbf{a}}$ into the orthogonal subspace spanned by the true coefficient vector $\mathbf{a}$, and hence should be close to zero if $\widehat{\mathbf{a}}$ is an estimate of $\mathbf{a}$. Again these measures are averaged over all simulations. For simulations involving more than one relationship between the random vectors, we also report the average distance taking into account the order in

which the true association is detected. That is, if $\mathbf{a}$ corresponds to a relationship and $\widetilde{\mathbf{a}}$ to another, the first estimated coefficient vector $\widehat{\mathbf{a}_1}$ could identify either of the two relationships. So we report the average distance between the relationships in the order of detection.

To show that the method is robust in the choice of initial guess, we randomly generate vectors consisting only of zeros and ones, a restricted random grid search only in the positive direction. For any numerical minimization method this would be best if for example in two dimensions, $\mathbf{X}_{2\times 1}$ and $\mathbf{Y}_{2\times 1}$ have a strictly positive relationship, e.g. $Y_1 = X_1 + X_2 + \epsilon$, but would be less favorable if the relationship was $Y_1 = X_1 - X_2 + \epsilon$. Of course in high dimensions with a mixture of positive and negative associations a less restrictive random search in each of the quadrants, generating random vectors consisting of ± 1 and 0's, or a full random grid search, generating vectors with $U(-1, 1)$ components, would likely be best for methods sensitive to initial guess selection but usually come at the cost of increased function evaluations. We show in the following simulations that our method is not sensitive to initial guess selection by using the restricted random grid search in the positive direction for all simulations. We take this idea one step further in Simulation 3.3.1 by using initial guesses consisting only of ones.

Since our methods are invariant to nonsingular scale transformations the following real data analysis and simulation results are reported for the whitened random vectors, transformed to have zero mean and identity covariance matrices. However, for ease and clarity the notation $\mathbf{x}$, $\mathbf{y}$ etc. is maintained. The distance measure is calculated in the transformed scale by transforming the true coefficient vectors as, $\mathbf{\Sigma}^{1/2}\mathbf{a}$.

Simulation 3.3.1 compares the reduced model and the full model when a non-functional dependence relationship exists between two sets, another two-set comparison with different sample sizes is done in Simulation 3.3.2, a three-set analysis is performed in Simulation 3.3.3 and Simulation 3.3.4 is a multiple group two-set simulation. An analysis of a real dataset is performed in Section 3.3.5.

3.3.1 Simulation

In this simulation we test $\mathcal{G}(\mathbf{a}, \mathbf{b})$ and $\mathcal{G}_r(\mathbf{a}, \mathbf{b})$ in the presence of a non-functional dependence relationship, when $\mathbf{a}^T\mathbf{X}$ is regressed on $\mathbf{b}^T\mathbf{Y}$. For both the full and reduced model we use an initial guess of ones and the restricted random initial guess strategy detailed at the start of this Section. The Gaussian kernel, K_1, with bandwidth h_1 is used throughout.

For a sample of $n = 125$, we define the random vectors $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ and $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^T$, where:

$$X_1 \sim U(-2, 2),\ X_2 \sim t(20),\ X_3 \sim N(0,1),\ X_4 \sim t(12)\ Y_3 \sim N(0,1),\ Y_4 \sim N(0,1),$$
$$Y_5 \sim \Gamma(2,3) \text{ and } \epsilon_1, \epsilon_2 \sim N(0,1).$$

The remaining variables are defined as:

$$Y_1 = \begin{cases} 2 - X_1^2 + 0.2\epsilon_1 & \text{if } X_1 \leq 1 \\ 4X_1 - 3 + 0.2\epsilon_1 & \text{if } X_1 > 1 \end{cases} \quad \text{and} \quad Y_2 = X_2^2 + 0.4\epsilon_2$$

The true coefficient vectors are:

$$\mathbf{a} = (1,0,0,0)^T, \mathbf{b} = (1,0,0,0,0)^T \quad \text{and} \quad \widetilde{\mathbf{a}} = (0,1,0,0)^T, \widetilde{\mathbf{b}} = (0,1,0,0,0)^T.$$

For a dataset drawn according to the above specifications we estimate the generalized coefficient vectors using, $\mathcal{G}(\mathbf{a}, \mathbf{b})$ and $\mathcal{G}_r(\mathbf{a}, \mathbf{b}) = \mathrm{E}(\mathbf{b}^T\mathbf{Y} - m_1(\mathbf{a}^T\mathbf{X}))^2$, and repeat the process 500 times. We calculate estimates of the 1st and 2nd variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}$ and $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$, where $\mathbf{x}$ and $\mathbf{y}$ denote the sample from the random vectors $\mathbf{X}$ and $\mathbf{Y}$. The means and estimated standard errors of the absolute correlations and distances are given in Table 3.1 for each model/initial guess combination. The distance measures are also decomposed and averaged in the order of detection and reported in Table 3.2, along with the frequency of detection.

The results show that the full and reduced methods using random initial guesses perform similarly on average, with near one correlations and small distance measures, in detecting the first relationship. However, there is an improvement in using the reduced model. First note, the ordered distances of Table 3.2 show that the reduced index detects the quadratic

relationship first 394 times compared to just 35 times for the full measure. Next, the absolute mean correlation for the second variates corresponding to $\widehat{\mathbf{b}}_2^T\mathbf{y}$ is 0.9942 and for the full model is 0.9755. In addition, the standard errors for the mean correlations for the second variates are an order of magnitude larger for the full model; see Table 3.1. The reason for the difference in performance is due to the full model regressing $\mathbf{a}^T\mathbf{X}$ on $\mathbf{Y}$, corresponding to a non-functional relationship when the true directions $\mathbf{a}$ and $\mathbf{b}$ are estimated correctly, which results in a constant estimate for the conditional mean function. This estimate of the mean adds unnecessary noise to the full model resulting in the piecewise function usually being the stronger of the two relationships, which is not the case for the reduced model. Since in practice the true relationships are unknown, one could use the full model and plot the variates to reveal any non-functional dependence relationships. If such a relationship exists the reduced model can be used to see if more accurate results can be obtained.

For the simulations using initial guesses of ones, again the second dependence relationship is detected more accurately using the reduced index compared to the full model. However, there is a loss of power on average in detecting the second relationship compared to the simulations involving random initial guesses. This is due to a few of the simulations using the initial guess of ones not converging. Therefore, in practice for ease of implementation an initial guess of ones can be used and if convergence becomes an issue, then use a random initial guess strategy.

Next, using the reduced model with random initial guesses a generated dataset was selected and another generated at random and the 1^{st}, 2^{nd} and 3^{rd} variates pairs, $\widehat{\mathbf{a}}_j^T\mathbf{x}$ versus $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2, 3$, were plotted in Figures 3.1 and 3.2. We see from both plots that the purely quadratic relationship is detected first and the piece-wise second but the relationships are stronger in Figure 3.1.

The bootstrap method for detecting the number of dimensions, detailed in Section 3.2.4, is run on both datasets using $\mathcal{G}_r(\mathbf{a}, \mathbf{b})$ and CCA. For the GCA method, the estimated subspace with the lowest variability across the m-sets occurs when $d = 2$, correctly determining the

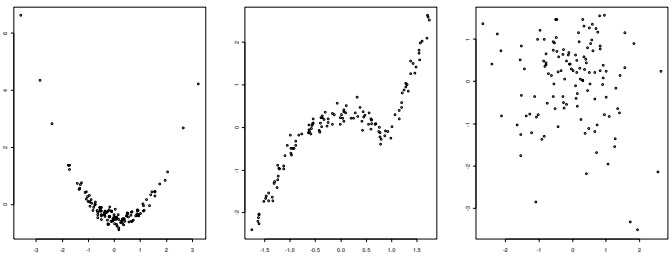
true number of coefficient vectors that identify the defined relationships. In contrast, the least variable estimated subspace using CCA is obtained for $d = 1$. This is shown in the barplots in Figures 3.3 and 3.4, where the smallest bar corresponds to the estimated subspace with the lowest variability. The CCA method is able to detect one dependence relationship since a sample from the piecewise relationship is expected to be linear half the time.

Simulation 3.3.1						
$\mathcal{G}(\mathbf{a},\mathbf{b})$	variate	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	variate	$\lvert\bar{\rho}_2\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$
random	$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9986(.0001)	.0489(.0010)	$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9919(.0012)	.1061(.0030)
	$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9959(.0002)	.0843(.0018)	$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9755(.0022)	.1923(.0043)
ones	$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9986(.0002)	.0538(.0015)	$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9173(.0104)	.1898(.0117)
	$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9930(.0007)	.0985(.0031)	$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9041(.0101)	.2673(.0110)
$\mathcal{G}_r(\mathbf{a},\mathbf{b})$	variate	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	variate	$\lvert\bar{\rho}_2\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$
random	$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9992(.0001)	.0380(.0007)	$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9944(.0004)	.0884(.0025)
	$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9989(.0001)	.0507(.0008)	$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9942(.0003)	.1008(.0023)
ones	$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9978(.0007)	.0424(.0022)	$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9167(.0103)	.1844(.0121)
	$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9972(.0011)	.0557(.0029)	$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9319(.0091)	.1790(.0111)

Table 3.1: Absolute average correlation(standard errors) (Simulation 3.3.1)

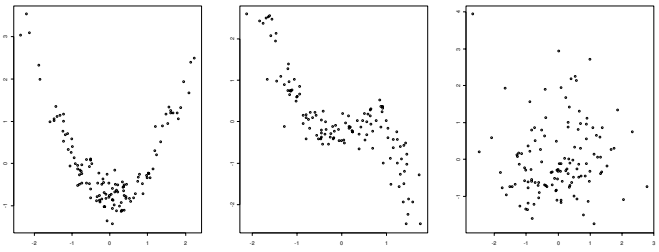
Simulation 3.3.1					
$\mathcal{G}(\mathbf{a},\mathbf{b})$	variate	$\mathbf{a}$	$\mathbf{b}$	$\widetilde{\mathbf{a}}$	$\widetilde{\mathbf{b}}$
random	1^{st}	.0478(465)	.0776(465)	.0900(035)	.1730(035)
	2^{nd}	.0635(035)	.1457(035)	.1073(465)	.1958(465)
ones	1^{st}	.0497(409)	.0799(409)	.0720(091)	.1824(091)
	2^{nd}	.0942(091)	.1360(091)	.2110(409)	.2965(409)
$\mathcal{G}_r(\mathbf{a},\mathbf{b})$	variate	$\mathbf{a}$	$\mathbf{b}$	$\widetilde{\mathbf{a}}$	$\widetilde{\mathbf{b}}$
random	1^{st}	.0361(106)	.0536(106)	.0384(394)	.0499(394)
	2^{nd}	.0866(394)	.1003(394)	.0952(106)	.1028(106)
ones	1^{st}	.0415(341)	.0559(341)	.0443(159)	.0554(159)
	2^{nd}	.0914(159)	.1055(159)	.2277(341)	.2133(341)

Table 3.2: Ordered distance calculations(frequency) (Simulation 3.3.1)



Plots from left to right - $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2, 3$ (Selected dataset)

Figure 3.1: GCA - Variate plots (Simulation 3.3.1)



Plots from left to right - $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2, 3$ (Random dataset)

Figure 3.2: GCA - Variate plots (Simulation 3.3.1)

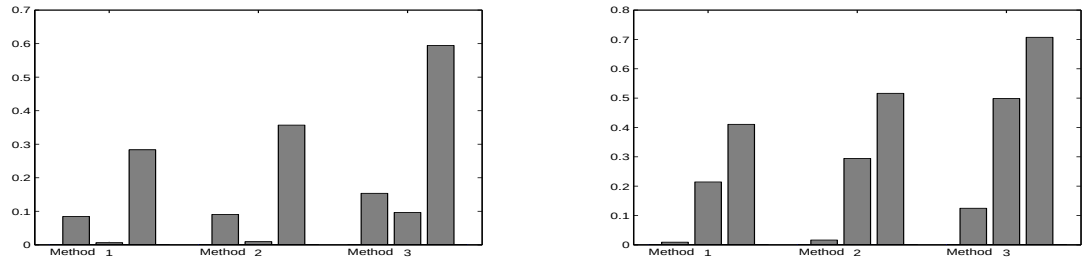


Figure 3.3: Bootstrap bar-plot GCA (Left panel), CCA (Right panel) (Simulation 3.3.1) (Selected dataset)

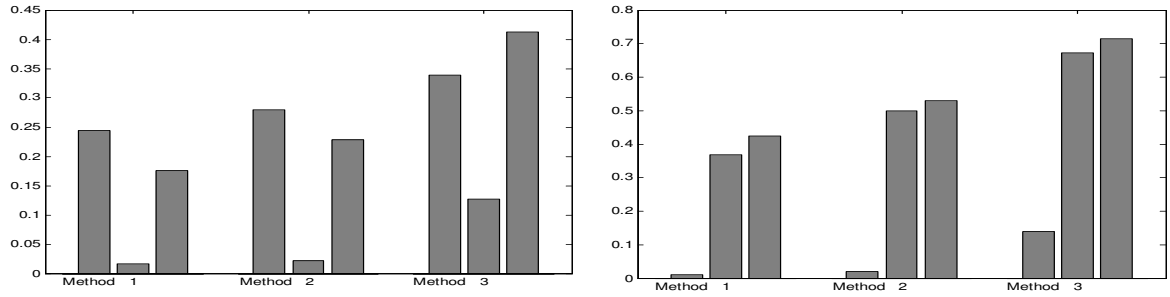


Figure 3.4: Bootstrap bar-plot GCA (Left panel), CCA (Right panel) (Simulation 3.3.1) (Random dataset)

From this example, we see that regardless of the chosen model the GCA method is able to detect the known relationships and does not require extensive or complicated generation of initial guesses. In addition, the bootstrap procedure using our method detects the correct the number of significant relationships, despite the strength of these associations.

3.3.2 Simulation

In this simulation we keep the same piece-wise relationship as in Simulation 3.3.1, and change the quadratic association to an exponential relationship. For two different samples of size $n = 75$ and 125, we define the multivariate random vectors $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ and $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^T$, where:

$X_1 \sim U(-2, 2)$, $X_2 \sim N(0, 1)$, $X_3 \sim N(0, 1)$, $X_4 \sim t(12)$ $Y_3 \sim N(0, 1)$, $Y_4 \sim N(0, 1)$, $Y_5 \sim \Gamma(2, 3)$ and $\epsilon_1, \epsilon_2 \sim N(0, 1)$.

The remaining variables are defined as:

$$Y_1 = \begin{cases} 2 - X_1^2 + 0.2\epsilon_1 & \text{if } X_1 \leq 1 \\ 4X_1 - 3 + 0.2\epsilon_1 & \text{if } X_1 > 1 \end{cases} \quad \text{and} \quad Y_2 = \mathrm{e}^{X_2} + 0.4\epsilon_2$$

The true coefficient vectors are:

$$\mathbf{a} = (1,0,0,0)^T, \mathbf{b} = (1,0,0,0,0)^T \quad \text{and} \quad \widetilde{\mathbf{a}} = (0,1,0,0)^T, \widetilde{\mathbf{b}} = (0,1,0,0,0)^T.$$

For a dataset drawn according to the above specifications we estimate the generalized coefficient vectors using $\mathcal{G}(\mathbf{a}, \mathbf{b})$ and repeat the process 500 times for both the Gaussian, K_1, and the Epanechnikov kernel, K_2, at each of the three bandwidths defined in Section 3.2.2. We compute estimates of the 1st and 2nd variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}$ and $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$. The means and estimated standard errors of the absolute correlations and distances are given in Table 3.3 for the Gaussian kernel and Table 3.4 for the Epanechnikov kernel. The distance measures are decomposed and averaged in the order of detection and reported in Tables 3.5 and 3.6, along with the frequency.

The results show that for a sample of size $n = 75$ the correlations between the estimated and true variates are high on average with small standard errors for both of the detected relationships. The accuracy in detecting the known associations is further seen in the small average distance between the true and estimated coefficient vectors, again with small standard errors. Moreover, choosing between a Gaussian or Epanechnikov kernel does not seem important as evidenced by the near identical results using each. In a similar vein, bandwidth selection does not appear to be crucial with bandwidths h_2 and h_3 showing similar performance with only a negligible improvement over the bandwidth h_1.

When the sample size is increased to $n = 125$ the difference is negligible between the choice of kernel and bandwidth and both relationships are detected with near one absolute average correlation. For all levels the average distance between the estimated and true coefficient vectors is also small and as expected the increased sample size increased the accuracy of our method.

Gaussian Kernel K_1 **- Simulation 3.3.2**						
	h_1		h_2		h_3	
$n = 75$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9878(.0012)	.1220(.0041)	.9935(.0004)	.1001(.0024)	.9952(.0003)	.0860(.0021)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9896(.0007)	.1316(.0037)	.9951(.0003)	.0945(.0021)	.9953(.0003)	.0935(.0020)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9739(.0020)	.1795(.0058)	.9845(.0009)	.1525(.0037)	.9857(.0010)	.1397(.0040)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9742(.0017)	.1980(.0055)	.9841(.0010)	.1639(.0039)	.9842(.0010)	.1600(.0041)
$n = 125$						
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9958(.0004)	.0757(.0022)	.9968(.0002)	.0714(.0016)	.9978(.0001)	.0575(.0015)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9960(.0002)	.0851(.0019)	.9977(.0001)	.0673(.0013)	.9978(.0001)	.0672(.0013)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9888(.0009)	.1204(.0037)	.9920(.0004)	.1098(.0026)	.9929(.0005)	.0986(.0028)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9889(.0007)	.1358(.0033)	.9923(.0004)	.1188(.0026)	.9926(.0004)	.1147(.0026)

Table 3.3: Gaussian kernel - Absolute average correlation(standard errors) (Simulation 3.3.2)

Epanechnikov Kernel K_2 **- Simulation 3.3.2**						
	h_1		h_2		h_3	
$n = 75$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$	$\|\bar{\rho}_1\|$	$\|\|\cdot\|\|_2$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9849(.0012)	.1416(.0042)	.9937(.0003)	.1005(.0023)	.9949(.0003)	.0905(.0020)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9867(.0008)	.1506(.0040)	.9953(.0002)	.0922(.0020)	.9955(.0003)	.0930(.0019)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9704(.0021)	.1962(.0059)	.9806(.0018)	.1685(.0039)	.9846(.0016)	.1429(.0041)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9704(.0017)	.2143(.0057)	.9820(.0021)	.1686(.0041)	.9838(.0012)	.1605(.0042)
$n = 125$						
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9940(.0005)	.0901(.0026)	.9965(.0002)	.0743(.0018)	.9968(.0006)	.0650(.0020)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9942(.0004)	.0997(.0024)	.9975(.0001)	.0689(.0014)	.9972(.0006)	.0696(.0018)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9862(.0011)	.1336(.0041)	.9901(.0005)	.1244(.0027)	.9920(.0006)	.1038(.0031)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9866(.0009)	.1475(.0036)	.9916(.0004)	.1239(.0026)	.9920(.0005)	.1177(.0028)

Table 3.4: Epanechnikov kernel - Absolute average correlation(standard errors) (Simulation 3.3.2)

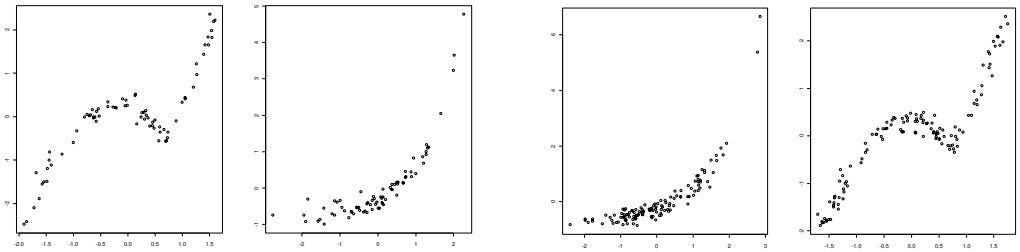
Distances - Gaussian Kernel K_1 - Simulation 3.3.2						
$n = 75$	h_1		h_2		h_3	
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.1034(148)	.1297(352)	.1242(064)	.0965(436)	.0728(114)	.0899(386)
2^{nd}	.1870(352)	.1618(148)	.1505(436)	.1657(064)	.1373(386)	.1480(114)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1371(148)	.1293(352)	.1202(064)	.0907(436)	.1062(114)	.0898(386)
2^{nd}	.2047(352)	.1822(148)	.1612(436)	.1828(064)	.1609(386)	.1568(114)
$n = 125$						
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.0540(133)	.0836(367)	.0760(051)	.0708(449)	.0389(105)	.0624(395)
2^{nd}	.1260(367)	.1047(133)	.1109(449)	.1001(051)	.0991(395)	.0967(105)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.0865(133)	.0846(367)	.0863(051)	.0651(449)	.0751(105)	.0651(395)
2^{nd}	.1453(367)	.1096(133)	.1209(449)	.1002(051)	.1186(395)	.1001(105)

Table 3.5: Gaussian kernel - Ordered average distances(frequency) (Simulation 3.3.2)

Distances - Epanechnikov Kernel K_2 - Simulation 3.3.2						
$n = 75$	h_1		h_2		h_3	
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.1354(134)	.1439(366)	.1306(048)	.0973(452)	.0859(095)	.0916(405)
2^{nd}	.2038(366)	.1753(134)	.1700(452)	.1549(048)	.1395(405)	.1574(095)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1581(134)	.1478(366)	.1171(048)	.0895(452)	.1043(095)	.0903(405)
2^{nd}	.2206(366)	.1969(134)	.1698(452)	.1572(048)	.1609(405)	.1586(095)
$n = 125$						
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.0793(128)	.0939(372)	.1022(028)	.0727(472)	.0587(086)	.0662(414)
2^{nd}	.1394(372)	.1167(128)	.1250(472)	.1138(028)	.1028(414)	.1085(086)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1038(128)	.0983(372)	.0973(028)	.0672(472)	.0854(086)	.0663(414)
2^{nd}	.1554(372)	.1246(128)	.1249(472)	.1072(028)	.1201(414)	.1062(086)

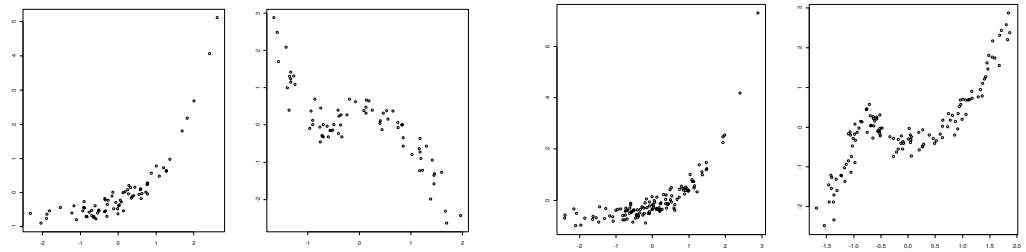
Table 3.6: Epanechnikov kernel - Ordered average distances(frequency) (Simulation 3.3.2)

Next a generated dataset was selected with high correlations and another generated at random for each of the two samples sizes and the 1st and 2nd variate pairs, $\widehat{\mathbf{a}}_j^T\mathbf{x}$ versus $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$, were plotted. The estimated coefficient vectors are found using $\mathcal{G}(\mathbf{a}, \mathbf{b})$ with a Epanechnikov kernel and bandwidth h_3. For the selected dataset, the plots in Figure 3.5 show that both relationships are detected accurately and the order of detection is reversed depending on the sample size. The same relationships are seen in the plots for the randomly generated dataset, see Figure 3.6, except the relationships for this dataset are not as strong.



a. 1st and 2nd variates $n = 75$ b. 1st and 2nd variates $n = 125$

Figure 3.5: GCA - Variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$ (Simulation 3.3.2) (Selected dataset)



a. 1st and 2nd variates $n = 75$ b. 1st and 2nd variates $n = 125$

Figure 3.6: GCA - Variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$ (Simulation 3.3.2) (Random dataset)

The bootstrap based method for detecting the number of dimensions, detailed in Section 3.2.4, is run on these same datasets and compared to CCA. For both methods, the estimated subspace with the lowest variability across the m-sets occurs when $d = 2$, correctly determining the true number of coefficient vectors. This is shown in the bar-plots in Figures 3.7

and 3.8 for the selected dataset and 3.9 and 3.10 for the random dataset. In all plots the smallest bars occur for $d = 2$. The CCA based method is able to detect the first defined relationship since it has a linear piecewise component. The second relationship is an exponentiated standard normal which has enough linear trend to be identified by CCA. Note that in the previous Simulation 3.3.1 when the second relationship was quadratic, instead of exponential, CCA was unable to determine the correct number of relationships. However, determining the number of significant relationships is different from detecting the nature of the associations. That is, even when CCA detects the correct number of relationships, $d = 2$ in this example, it incorrectly defines these relationships as linear. In contrast, the simulations show that the GCA method determines both the number and nature of the relationships correctly, indicating that our method is more comprehensive than CCA in this regard.

For this example, the GCA method detects the known relationships with increasing accuracy as the sample size becomes larger and, more importantly the choice of kernel and bandwidth is not critical. Finally, the bootstrap procedure using our method detects the correct number of significant relationships despite the strength of these associations.

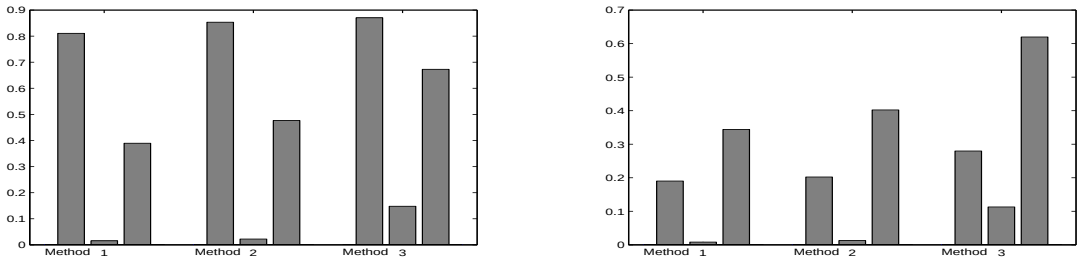


Figure 3.7: GCA - Bootstrap bar-plot $n = 75$ - Left, $n = 125$ - Right (Simulation 3.3.2) (Selected dataset)

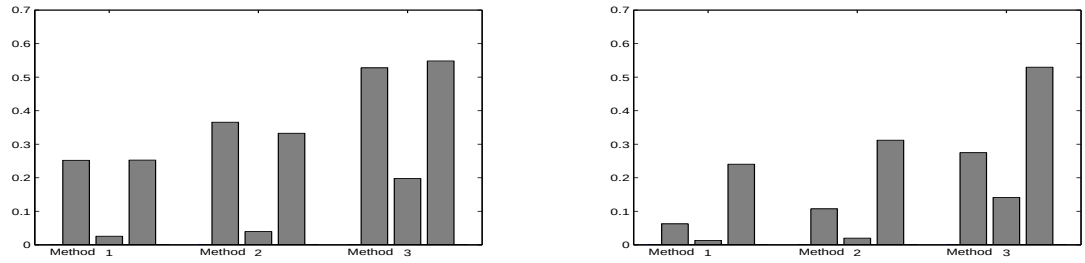


Figure 3.8: CCA - Bootstrap bar-plot $n = 75$ - Left, $n = 125$ - Right (Simulation 3.3.2) (Selected dataset)

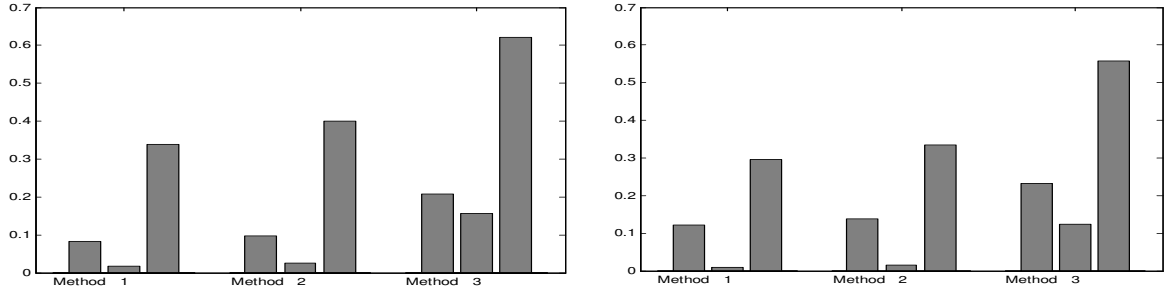


Figure 3.9: GCA - Bootstrap bar-plot $n = 75$ - Left, $n = 125$ - Right (Simulation 3.3.2) (Random dataset)

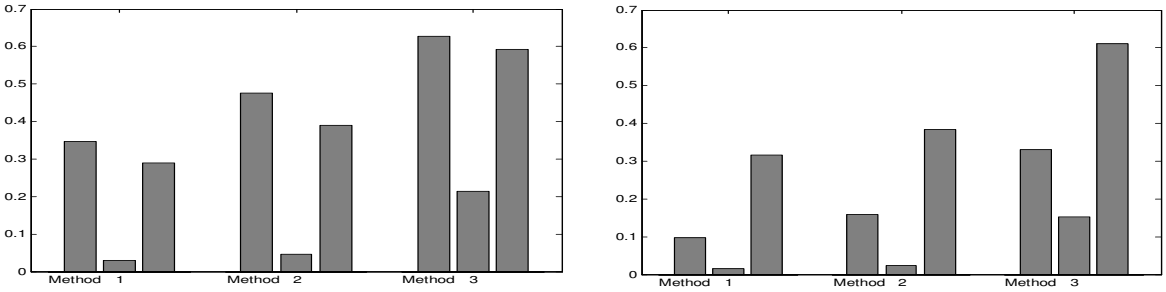


Figure 3.10: CCA - Bootstrap bar-plot $n = 75$ - Left, $n = 125$ - Right (Simulation 3.3.2) (Random dataset)

3.3.3 Simulation

In this simulation we test the accuracy of our method with a moderate sample size of three sets of random vectors in high dimension composed of variables with a wide range of distributions. This scenario is made more complex with complicated linear and nonlinear relationships between the sets. For a sample size of $n = 100$, we define the multivariate random

vectors $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)^T$, $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$ and $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$, where

$$X_1, X_2, X_3 \sim N(0,1),\ X_4 \sim \chi^2_{(7)},\ X_5 \sim t(5),\ X_6 \sim F(3,12) \text{ and } X_7, X_8 \sim N(0,1).$$

$$Y_3 \sim t(9),\ Z_3 \sim \chi^2_{(14)} \text{ and } \epsilon_j \sim N(0,1), j = 1, \ldots, 4.$$

The remaining variables are defined as:

$$Y_1 = (2X_1 + X_2 + X_3)^2 + 0.5\epsilon_1 \text{ and } Z_1 = 2X_1 + X_2 + X_3 + 0.4\epsilon_2.$$

$$Y_2 = \cos(X_2 - X_3) + 0.2\epsilon_3 \qquad \text{and } Z_2 = X_2 - X_3 + 0.3\epsilon_4.$$

The true coefficient vectors are:

$$\mathbf{a} = (2,1,1,0,0,0,0,0)^T, \mathbf{b} = (1,0,0)^T \text{ and } \mathbf{c} = (1,0,0)^T$$

$$\widetilde{\mathbf{a}} = (0,1,-1,0,0,0,0,0)^T, \widetilde{\mathbf{b}} = (0,1,0)^T \text{ and } \widetilde{\mathbf{c}} = (0,1,0)^T.$$

For a dataset drawn according to the above specifications we estimate the generalized coefficient vectors using $\mathcal{G}(\mathbf{a}, \mathbf{b})$ and repeat the process 500 times for the Gaussian and Epanechnikov kernels at each of the three bandwidths defined in Section 3.2.2. We compute estimates of the 1st and 2nd variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}$, $\widehat{\mathbf{b}}_j^T\mathbf{y}$ and $\widehat{\mathbf{c}}_j^T\mathbf{z}$, $j = 1, 2$. The means and estimated standard errors of the absolute correlations and distances are given in Table 3.7 for the Gaussian kernel and Table 3.8 for the Epanechnikov kernel. The distance measures are also decomposed and averaged in the order of detection and reported in Tables 3.9 and 3.10, along with the frequency.

The results show that the Gaussian kernel with bandwidth h_1 detects the relationships accurately with the highest absolute average correlation and smallest standard error. The accuracy using the Epanechnikov and other bandwidths is slightly less in varying degrees. In addition, the distance measures are consistent with these inferences. The varying levels of performance can be attributed to the complexities of this example, where not only is the dimension high and the relationships a complex mixture of linear and nonlinear associations, the true variates lie in orthogonal directions and are composed of two of the same variables.

Gaussian Kernel K_1 **- Simulation 3.3.3**						
band	h_1		h_2		h_3	
	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9980(.0001)	.0670(.0011)	.9677(.0066)	.1257(.0074)	.9583(.0077)	.1222(.0089)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9874(.0008)	.1340(.0040)	.9728(.0042)	.1677(.0062)	.9612(.0063)	.1761(.0072)
$\widehat{\mathbf{c}}_1^T\mathbf{z}$	.9990(.0001)	.0529(.0014)	.9952(.0008)	.0756(.0029)	.9958(.0010)	.0642(.0030)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9900(.0020)	.1165(.0034)	.9738(.0051)	.1490(.0060)	.9765(.0048)	.1382(.0059)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9787(.0027)	.1749(.0049)	.9631(.0050)	.2012(.0069)	.9584(.0054)	.2166(.0077)
$\widehat{\mathbf{c}}_2^T\mathbf{z}$	.9930(.0005)	.1087(.0032)	.9900(.0012)	.1216(.0041)	.9881(.0022)	.1160(.0044)

Table 3.7: Gaussian kernel - Absolute average correlation(standard errors) (Simulation 3.3.3)

Epanechnikov Kernel K_2 **- Simulation 3.3.3**						
band	h_1		h_2		h_3	
	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$	$\lvert\bar{\rho}_1\rvert$	$\lvert\lvert\cdot\rvert\rvert_2$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9917(.0029)	.0780(.0038)	.9769(.0054)	.1183(.0061)	.9508(.0085)	.1302(.0094)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9856(.0013)	.1451(.0048)	.9720(.0043)	.1695(.0069)	.9620(.0058)	.1800(.0081)
$\widehat{\mathbf{c}}_1^T\mathbf{z}$	.9975(.0005)	.0569(.0021)	.9958(.0028)	.0772(.0025)	.9965(.0008)	.0630(.0026)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9883(.0020)	.1202(.0040)	.9738(.0050)	.1529(.0058)	.9671(.0061)	.1481(.0071)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9613(.0059)	.1907(.0072)	.9499(.0066)	.2171(.0085)	.9442(.0073)	.2290(.0089)
$\widehat{\mathbf{c}}_2^T\mathbf{z}$	.9895(.0018)	.1120(.0042)	.9903(.0028)	.1207(.0036)	.9891(.0020)	.1162(.0042)

Table 3.8: Epanechnikov kernel - Absolute average correlation(standard errors) (Simulation 3.3.3)

Distances - Gaussian Kernel K_1 **- Simulation 3.3.3**						
band	h_1		h_2		h_3	
	$\mathbf{a}$	$\widetilde{\mathbf{a}}$	$\mathbf{a}$	$\widetilde{\mathbf{a}}$	$\mathbf{a}$	$\widetilde{\mathbf{a}}$
1^{st}	.0673(337)	.0664(163)	.1302(348)	.1154(152)	.1267(342)	.1124(158)
2^{nd}	.1155(163)	.1170(337)	.1401(152)	.1528(348)	.1416(158)	.1367(342)
	$\mathbf{b}$	$\widetilde{\mathbf{b}}$	$\mathbf{b}$	$\widetilde{\mathbf{b}}$	$\mathbf{b}$	$\widetilde{\mathbf{b}}$
1^{st}	.1362(337)	.1468(163)	.1638(348)	.1766(152)	.1699(342)	.1897(158)
2^{nd}	.1618(163)	.1813(337)	.2034(152)	.2003(348)	.2048(158)	.2220(342)
	$\mathbf{c}$	$\widetilde{\mathbf{c}}$	$\mathbf{c}$	$\widetilde{\mathbf{c}}$	$\mathbf{c}$	$\widetilde{\mathbf{c}}$
1^{st}	.0519(337)	.0550(163)	.0720(348)	.0840(152)	.0650(342)	.0625(158)
2^{nd}	.0903(163)	.1177(337)	.1107(152)	.1264(348)	.1046(158)	.1213(342)

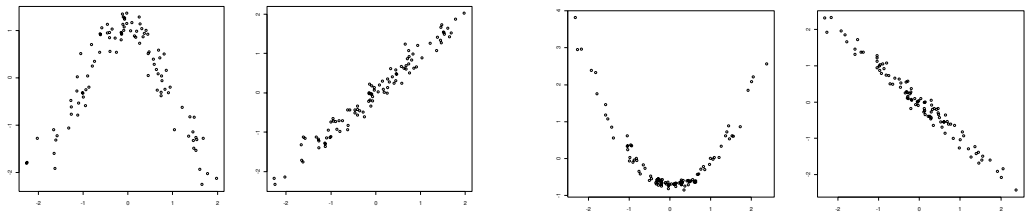
Table 3.9: Gaussian kernel - Ordered average distance(frequency) (Simulation 3.3.3)

Distances - Epanechnikov Kernel K_2 - Simulation 3.3.3						
band	h_1		h_2		h_3	
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.0832(333)	.0676(167)	.1242(346)	.1051(154)	.1413(352)	.1039(148)
2^{nd}	.1113(167)	.1247(333)	.1638(154)	.1481(346)	.1535(148)	.1458(352)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1472(333)	.1411(167)	.1675(346)	.1742(154)	.1815(352)	.1765(148)
2^{nd}	.1861(167)	.1930(333)	.2285(154)	.2120(346)	.2386(148)	.2243(352)
	c	$\widetilde{\mathbf{c}}$	**c**	$\widetilde{\mathbf{c}}$	**c**	$\widetilde{\mathbf{c}}$
1^{st}	.0574(333)	.0558(167)	.0748(346)	.0825(154)	.0641(352)	.0603(148)
2^{nd}	.0883(167)	.1239(333)	.1053(154)	.1275(346)	.0953(148)	.1250(352)

Table 3.10: Epanechnikov kernel - Ordered average distance(frequency) (Simulation 3.3.3)

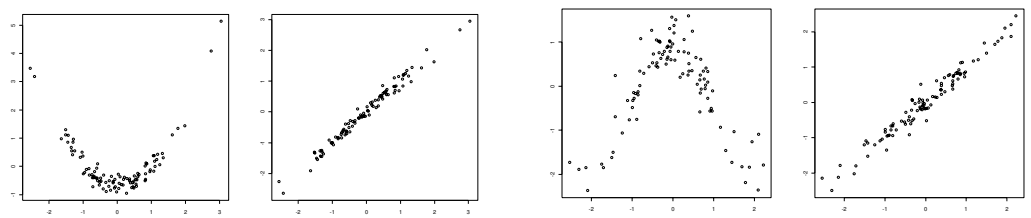
Next a generated dataset was selected and another generated at random and the 1^{st} and 2^{nd} variate pairs, $\widehat{\mathbf{a}}_j^T\mathbf{x}$ versus $\widehat{\mathbf{b}}_j^T\mathbf{y}$ and $\widehat{\mathbf{a}}_j^T\mathbf{x}$ versus $\widehat{\mathbf{c}}_j^T\mathbf{y}$, $j = 1, 2$, were plotted for each; see Figures 3.11 and 3.12. The estimated coefficient vectors are found using $\mathcal{G}(\mathbf{a}, \mathbf{b})$ with a Gaussian kernel and bandwidth h_1. For both datasets, the plots show that both relationships are detected accurately.

Using GCA, the bootstrap method for detecting the number of dimensions is then run on these datasets. The minimum dimension here is three with $d = 2$ true relationships, thus we can only compare the bar-plots for the first two dimensions($d = 1$ and $d = 2$). In Figures 3.13 and 3.14 the smallest bars corresponding to the estimated subspaces with the smallest variability occur when $d = 2$ for all three methods. The plots of the 3^{rd} variates in the left panel of Figures 3.13 and 3.14 show that there is not a third significant relationship, thus we infer that $d = 2$.



a. GCA - 1st variate plots b. GCA - 2nd variate plots

Figure 3.11: GCA - Variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$ and $\widehat{\mathbf{c}}_j^T\mathbf{z}$ $j = 1, 2$ (Simulation 3.3.3) (Selected dataset)



a. GCA - 1st variate plots b. GCA - 2nd variate plots

Figure 3.12: GCA - Variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$ and $\widehat{\mathbf{c}}_j^T\mathbf{z}$ $j = 1, 2$ (Simulation 3.3.3) (Random dataset)

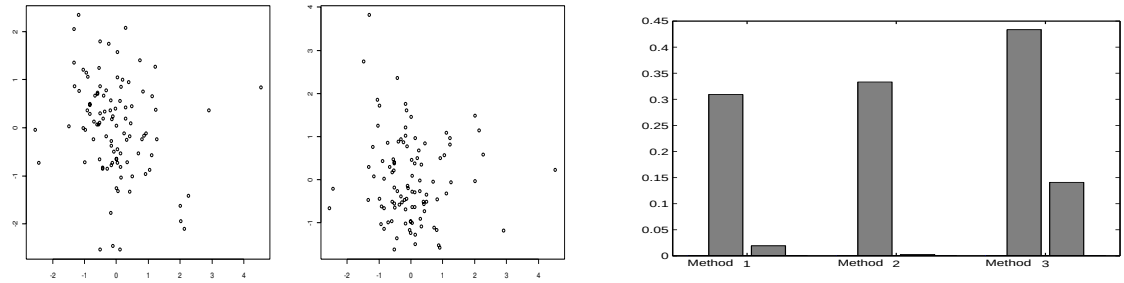


Figure 3.13: GCA - 3rd variate plots & Bootstrap bar-plot (Simulation 3.3.3) (Selected dataset)

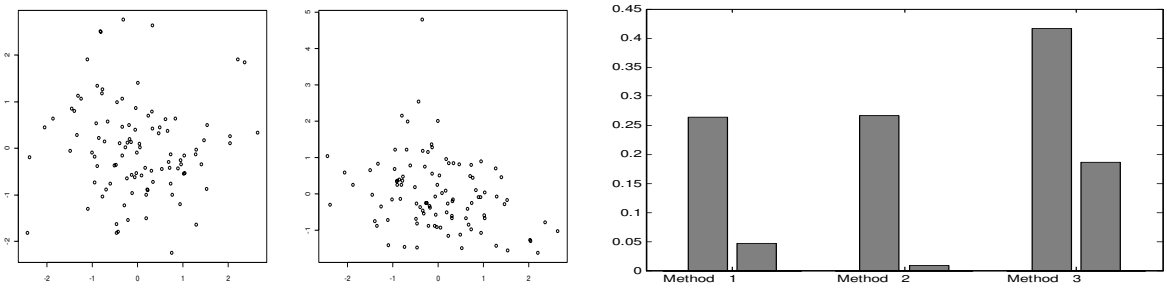


Figure 3.14: GCA - 3rd variate plots & Bootstrap bar-plot (Simulation 3.3.3) (Random dataset)

This example shows that even under a complicated scenario our methods still perform well. While the selection of different kernels and bandwidths has an effect on performance, the overall accuracy is favorable for any of the bandwidth kernel combinations.

3.3.4 Simulation

In this simulation two groups consisting of two sets of random vectors using the proposed common GCA method, $\mathcal{G}^{(\mathrm{C})}(\mathbf{a}, \mathbf{b})$. Suppose $\mathbf{X}^W = (X_1^w, X_2^w, X_3^w, X_4^w, X_5^w)^T$ and $\mathbf{Y}^W = (Y_1^w, Y_2^w, Y_3^w)^T$ are random vectors for each w and let $\mathrm{P}(W = w) = 1/2$, $w = 1, 2$. For each group define the variables, $X_1^w \sim N(0,1)$, $X_2^w \sim t(12)$, $X_3^w \sim F(3,12)$, $X_4^w \sim \chi^2_{(2)}$, $X_5^w \sim N(0,1)$, $Y_3^w \sim N(0,1)$ and $\epsilon_j \sim N(0,1), j = 1, \ldots, 4$. The remaining variables,

defined by group, are as follow:

$$Y_1 = 2Y_3 + X_1 + 0.4\epsilon_1 \quad \text{and} \quad Y_2 = \mathrm{e}^{X_5} + 0.3\epsilon_2 \text{ when } W = 1$$
$$Y_1 = 2Y_3 + X_1^3 + 0.3\epsilon_3 \quad \text{and} \quad Y_2 = X_5 + 0.4\epsilon_4 \text{ when } W = 2.$$

The common general coefficient vectors for each group, given by $\mathbf{a}^T\mathbf{X}^w$ and $\mathbf{b}^T\mathbf{Y}^w$, are:

$$\mathbf{a} = (1,0,0,0,0)^T,\ \mathbf{b} = (1,0,-2)^T \quad \text{and} \quad \widetilde{\mathbf{a}} = (0,0,0,0,1)^T,\ \widetilde{\mathbf{b}} = (0,1,0)^T.$$

Note that the relationships between the pairs are a mixture of linear and nonlinear associations depending on whether $W = 1$ or 2 and there is a further within set dependence relationship between Y_1 and Y_3 in each group.

For sample sizes of $n = 100$, datasets are drawn according to the above specifications and estimates of the common coefficient vectors using $\mathcal{G}^{(\mathrm{C})}(\mathbf{a}, \mathbf{b})$ for 500 repetitions are calculated. The Gaussian and Epanechnikov kernels at each of the three bandwidths. For each group we compute estimates of the 1^{st} and 2^{nd} variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}^w$ and $\widehat{\mathbf{b}}_j^T\mathbf{y}^w$, $j = 1, 2$, where $\mathbf{x}^w$ and $\mathbf{y}^w$ denote the samples from $\mathbf{X}^w$ and $\mathbf{Y}^w$ in the w^{th} group. For each group the means and estimated standard errors of the absolute correlations are given in Table 3.11 for the Gaussian kernel and Table 3.12 for the Epanechnikov kernel. The distance measures for each group are decomposed and averaged in the order of detection and reported in Tables 3.13 and 3.14, with the frequency of detection. From the tables, it's clear that the common GCA method accurately identifies the relationships between the two groups. Moreover, the choice of kernel and bandwidth seems unimportant since they each produced nearly identical results.

Gaussian Kernel K_1 **- Simulation 3.3.4**						
	Group 1			Group 2		
band	h_1	h_2	h_3	h_1	h_2	h_3
	$\lvert\bar{\rho}_1\rvert$					
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9921(.0005)	.9948(.0002)	.9948(.0002)	.9911(.0006)	.9939(.0002)	.9940(.0002)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9950(.0005)	.9973(.0003)	.9972(.0003)	.9944(.0004)	.9965(.0002)	.9963(.0002)
	$\lvert\bar{\rho}_2\rvert$					
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9852(.0008)	.9904(.0004)	.9897(.0004)	.9862(.0008)	.9911(.0004)	.9906(.0004)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9765(.0008)	.9700(.0010)	.9695(.0010)	.9810(.0007)	.9920(.0004)	.9912(.0005)

Table 3.11: Gaussian kernel - Absolute average correlation(standard errors) (Simulation 3.3.4)

Epanechnikov Kernel K_2 **- Simulation 3.3.4**						
	Group 1			Group 2		
band	h_1	h_2	h_3	h_1	h_2	h_3
	$\lvert\bar{\rho}_1\rvert$					
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	.9898(.0006)	.9947(.0002)	.9946(.0002)	.9890(.0007)	.9939(.0002)	.9940(.0002)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	.9926(.0006)	.9973(.0003)	.9972(.0003)	.9925(.0005)	.9965(.0002)	.9964(.0002)
	$\lvert\bar{\rho}_2\rvert$					
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	.9825(.0010)	.9904(.0004)	.9899(.0004)	.9833(.0010)	.9912(.0004)	.9905(.0005)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	.9751(.0010)	.9700(.0010)	.9697(.0010)	.9771(.0009)	.9921(.0004)	.9913(.0004)

Table 3.12: Epanechnikov kernel - Absolute average correlation(standard errors) (Simulation 3.3.4)

Distances - Gaussian Kernel K_1 - Simulation 3.3.4						
band	h_1		h_2		h_3	
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.1626(013)	.0910(487)	.0708(014)	.0733(486)	.0717(014)	.0728(486)
2^{nd}	.1285(487)	.2070(013)	.1042(486)	.1020(014)	.1072(486)	.1052(014)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1924(013)	.0699(487)	.1785(014)	.0478(486)	.1860(014)	.0494(486)
2^{nd}	.1547(487)	.1690(013)	.1815(486)	.0710(014)	.1838(486)	.0858(014)

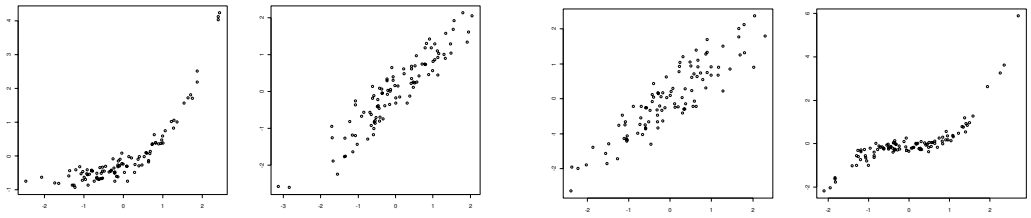
Table 3.13: Gaussian kernel - Ordered average distance(frequency) (Simulation 3.3.4)

Distances - Epanechnikov Kernel K_2 - Simulation 3.3.4						
band	h_1		h_2		h_3	
	a	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$	**a**	$\widetilde{\mathbf{a}}$
1^{st}	.1661(013)	.1065(487)	.0801(014)	.0739(486)	.0694(014)	.0738(486)
2^{nd}	.1415(487)	.2147(013)	.1041(486)	.0988(014)	.1063(486)	.1078(014)
	b	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$	**b**	$\widetilde{\mathbf{b}}$
1^{st}	.1960(013)	.0879(487)	.1814(014)	.0478(486)	.1828(014)	.0494(486)
2^{nd}	.1622(487)	.1803(013)	.1812(486)	.0712(014)	.1829(486)	.0846(014)

Table 3.14: Epanechnikov kernel - Ordered average distance(frequency) (Simulation 3.3.4)

Next a generated dataset was selected and the 1^{st} and 2^{nd} variates pairs, $\widehat{\mathbf{a}}_j^T\mathbf{x}^w$ versus $\widehat{\mathbf{b}}_j^T\mathbf{y}^w$, $j = 1, 2$, for each group were plotted; see Figure 3.15. The estimated coefficient vectors are found using $\mathcal{G}^{(C)}(\mathbf{a}, \mathbf{b})$ with a Gaussian kernel and bandwidth h_2. These plots show that each of the relationships in the groups is detected accurately.

The GCA based measure and the bootstrap method for detecting the number of dimensions corresponding to the significant relationships is then run on this dataset. The minimum dimension here is three with $d = 2$ true relationships, thus we compare the bar-plots of the first two dimensions to the 3^{rd} variate plots in each group. The bar plots for all three methods in Figure 3.16 are smallest when $d = 2$, indicating that this is the less variable dimension for the estimated subspaces. We conclude that $d = 2$ by plotting the 3^{rd} variates for each group(see the left panel of Figure 3.16) which shows no dependence relationships beyond the first two.



a. Group 1-1st and 2nd variates b. Group 2-1st and 2nd variates

Figure 3.15: GCA - Group variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$ by group (Simulation 3.3.4)

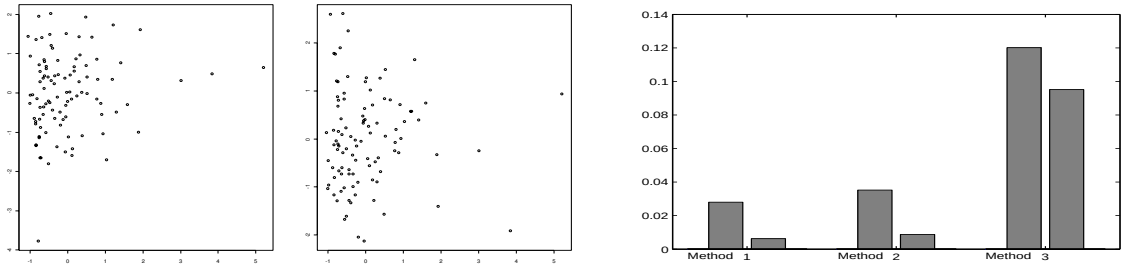


Figure 3.16: GCA - 3rd variate plots & Bootstrap bar-plot (Simulation 3.3.4)

From this example we see that our common index accurately detects the relationships for each group and does so for any kernel/bandwidth combination.

3.3.5 Machine data

This is an example based on electrode data first analyzed by Flury and Riedwyl [1] and later analyzed by Yin [15]. There were five measurements taken on fifty electrodes produced by two different machines. The first set of variables, denoted $\mathbf{X}^W = (X_1^w, X_2^w, X_3^w)^T$, are widths and the other two, denoted $\mathbf{Y}^W = (\mathbf{Y}_1^w, \mathbf{Y}_2^w)^T$, are lengths. Here each machine is a group consisting of two sets of random vectors, hence $W = 1, 2$ for Machines 1 and 2 respectively.

We first consider machine 2 only, analyzing it as a separate dataset consisting of two multivariate random vectors using $\mathcal{G}(\mathbf{a}, \mathbf{b})$ with a Gaussian kernel and bandwidth h_1. Yin [15] showed that a nonlinear relationship exists between the random vectors $\mathbf{X}$ and $\mathbf{Y}$ using his method termed Information Canonical Correlation Analysis(ICCA); which is equivalent to the KLICA measure $\mathcal{I}^{(P)}$ for two sets. The estimated 1st variates are reported in Table 3.15 using each method. The correlation between both estimated variates, $\rho(\widehat{\mathbf{u}}_1^T\mathbf{x}, \widehat{\mathbf{a}}_1^T\mathbf{x})$ and $\rho(\widehat{\mathbf{v}}_1^T\mathbf{y}, \widehat{\mathbf{b}}_1^T\mathbf{y})$, are reported in the last column. Note, our coefficients are reported in the transformed scale which does not effect the correlation.

1st Variates-Machine 2		
ICCA	**GCA**	ρ
$\widehat{\mathbf{u}}_1^T\mathbf{x} = 0.747\mathbf{x}_1 + 0.322\mathbf{x}_2 + 0.581\mathbf{x}_3$	$\widehat{\mathbf{a}}_1^T\mathbf{x} = 0.668\mathbf{x}_1 + 0.443\mathbf{x}_2 + 0.600\mathbf{x}_3$	0.9989
$\widehat{\mathbf{v}}_1^T\mathbf{y} = 0.726\mathbf{y}_1 + 0.687\mathbf{y}_2$	$\widehat{\mathbf{b}}_1^T\mathbf{y} = 0.720\mathbf{y}_1 + 0.695\mathbf{y}_2$	0.9999

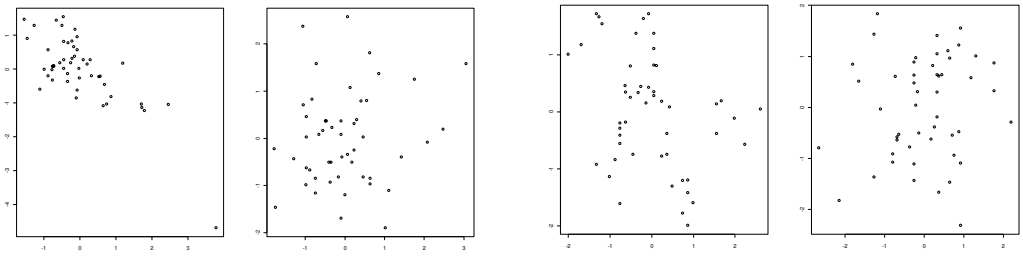
Table 3.15: Machine 2 - 1st variates comparison (Example 3.3.5)

Yin showed the regression function,

$$\mathrm{E}\left(\mathbf{v}_1^T\mathbf{y}|\mathbf{u}_1^T\mathbf{x}\right) = 1420.9 - 3.1\mathbf{u}_1 + 65\sin\left(2\pi\mathbf{u}_1/8.5 - 88.8\pi\right)$$

fit the variate plots of $\mathbf{u}_1^T\mathbf{x}$ and $\mathbf{v}_1^T\mathbf{y}$ well, except in the tails. The high correlation between both methods variate pairs show that our method does not lose power to the ICCA method in detecting this nonlinear relationship.

Machine 1 was also analyzed separately and plots of the 1st and 2nd variates for each machine are shown in Figure 3.17. The graph in the far left panel indicates an existence of a linear relationship between $\mathbf{X}$ and $\mathbf{Y}$ for Machine 1. The graph in the left panel for Machine 2 indicates a nonlinear trend as shown by Yin and discussed above. Note, this graph is in the whitened scale while the plot given in Yin [15] is translated back to the original scale, where the sine relationship given above is much more discernible. However, due to the invariance of our method the same nonlinear relationship involving the sine function can be seen in the plot using either the regression function above or the plot given in Yin [15] as a reference. In



a. Machine 1 - 1st and 2nd variates b. Machine 2 - 1st and 2nd variates

Figure 3.17: GCA - Individual machine variate plots $\widehat{\mathbf{a}}_j^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$ (Example 3.3.5)

addition, the correlation between our estimates and Yin's given in Table 3.15 further shows that this relationship exists and is detected using our method.

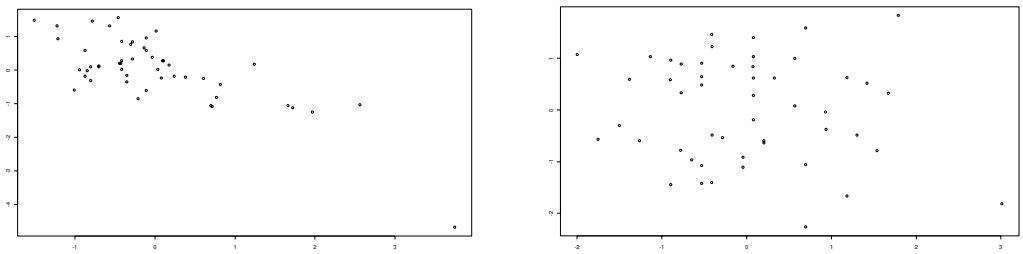
To determine whether a common analysis is appropriate we use the results of the machines analyzed individually and compare them to the results using our common index $\mathcal{G}^{(\mathrm{C})}(\mathbf{a}, \mathbf{b})$. Each machine is weighted equally, $P(W = w) = 1/2$, and a Gaussian kernel and bandwidth h_1 is used as in the separate analysis performed on Machine 2.

The variates from the individual analysis are compared to the variates obtained using our common model, which if appropriate should have the same coefficients. Note that, all coefficient vectors are the directions corresponding to the whitened mean centered data. The results are presented in Table 3.16.

	Individual Analysis		Common Analysis
	Machine 1	Machine 2	
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	$0.503\mathbf{x}_1 + 0.864\mathbf{x}_2 - 0.030\mathbf{x}_3$	$0.668\mathbf{x}_1 + 0.443\mathbf{x}_2 + 0.600\mathbf{x}_3$	$0.526\mathbf{x}_1 + 0.848\mathbf{x}_2 + 0.056\mathbf{x}_3$
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	$-0.702\mathbf{y}_1 + 0.713\mathbf{y}_2$	$0.720\mathbf{y}_1 + 0.695\mathbf{y}_2$	$-0.710\mathbf{y}_1 + 0.704\mathbf{y}_2$

Table 3.16: Common model comparison (Example 3.3.5)

We see from the individual analysis the variates corresponding to the width variables, $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$, are a weighted average of all three variables for Machine 2. In contrast, for Machine 1 the variates are a weighted average of the widths, $\mathbf{x}_1$ and $\mathbf{x}_2$. The length variables, $\mathbf{y}_1$ and $\mathbf{y}_2$, are weighted equally for both machines but in orthogonal directions. Since the individual variables are weighted differently a common analysis would not be expected to be appropriate. This is evidenced by the common variates being nearly identical to the individual variates for Machine 1 and the correlations between the individual and common variates for the width and length variables being $\rho(\widehat{\mathbf{a}}_1^T\mathbf{x}, 0.526\mathbf{x}_1 + 0.848\mathbf{x}_2 + 0.056\mathbf{x}_3) = 0.996$ and $\rho(\widehat{\mathbf{b}}_1^T\mathbf{y}, -0.710\mathbf{y}_1 + 0.704\mathbf{y}_2) = 0.999$. The corresponding correlations for Machine 2 are 0.759 and -0.0224. This is further supported by the plots of the 1st variates for each group, see Figure 3.18, where the common variates detect the linear relationship in Machine 1 but fail to detect any relationship for Machine 2.



a. Machine 1 - 1st common variates b. Machine 2 - 1st common variates

Figure 3.18: GCA - Common variate plots $\widehat{\mathbf{a}}_1^T\mathbf{x}$ vs $\widehat{\mathbf{b}}_1^T\mathbf{y}$, by Machine (Example 3.3.5)

As mentioned before, if a true common analysis is appropriate the 1st common coefficient vectors say, $\mathbf{a}_1^c$ and $\mathbf{b}_1^c$, that minimize $\mathcal{G}^{(\mathrm{C})}(\mathbf{a}, \mathbf{b})$ are the same for each group and thus $\mathbf{a}_1^{(1)} = \mathbf{a}_1^c$ and $\mathbf{b}_1^{(1)} = \mathbf{b}_1^c$ should minimize the individual index $\mathcal{G}^1(\mathbf{a}, \mathbf{b})$ for group 1 and $\mathbf{a}_1^{(2)} = \mathbf{a}_1^c$ and $\mathbf{b}_1^{(2)} = \mathbf{b}_1^c$ the index $\mathcal{G}^2(\mathbf{a}, \mathbf{b})$ for group 2. That is, the minimum value of the common index can be decomposed as follows:

$$\mathcal{G}^{(\mathrm{C})}(\mathbf{a}_1^c, \mathbf{b}_1^c) = \sum_{w=1}^{2} \mathcal{G}^w(\mathbf{a}_1^c, \mathbf{b}_1^c)\,\mathrm{P}(W = w) = \frac{1}{2}\mathcal{G}^1(\mathbf{a}_1^c, \mathbf{b}_1^c) + \frac{1}{2}\mathcal{G}^2(\mathbf{a}_1^c, \mathbf{b}_1^c)$$

$$= \frac{1}{2}[\mathcal{G}^1\left(\mathbf{a}_1^{(1)}, \mathbf{b}_2^{(1)}\right) + \mathcal{G}^2\left(\mathbf{a}_1^{(2)}, \mathbf{b}_1^{(2)}\right)],$$

which is just the average of the minimum individual index values for each group. The decomposed sample versions for the 1st variates are,

$$\begin{aligned} \frac{1}{2}[\mathcal{G}_n^1(\widehat{\mathbf{a}}_1, \widehat{\mathbf{b}}_1) + \mathcal{G}_n^2(\widehat{\mathbf{a}}_1, \widehat{\mathbf{b}}_1)] &= \mathcal{G}_n^{(\mathrm{C})}(\widehat{\mathbf{a}}_1, \widehat{\mathbf{b}}_1) \\ \frac{1}{2}[34.212 + 72.742] = 53.47 &\neq 61.03. \end{aligned}$$

While we cannot easily quantify whether this difference is statistically significant, this difference and the above arguments support our view that a common analysis is not appropriate for this dataset.

Based on a large sample test using a log-likelihood ratio statistic, Goria and Flury [2] concluded that a common model fit well for the machine data belonging to Machines 1 and 2. However, in analyzing this same dataset with our indexes we believe the correct conclusion is that a common analysis is not appropriate. Our conclusions confirm those of Yin and Sriram [16], who also argued that a common canonical analysis was not appropriate. A plausible explanation for the conflicting results is that the methods used by Goria and Flury only detect linear relationships, while our methods are able to identify both linear and nonlinear associations.

3.4 Discussion

In this paper, we introduce a new nonparametric index $\mathcal{G}(\mathbf{a}, \mathbf{b})$, which is shown to recover both linear and nonlinear relationships between two sets of multivariate random vectors. When the relationships are linear, we have shown analytically that minimizing our index is equivalent to CCA; hence, our index is a generalization of CCA. The index measures an L_2 distance between linear combinations of one vector and an unknown regression function of the other, interchanging the roles of each respectively, thus making it a least squares type method. Assuming the form of the regression function to be unknown as in single index regression models, enables us to measure both linear and nonlinear relationships between the

sets. In addition, a bootstrap procedure, inspired by Ye and Weiss [14], has been developed to determine the number of significant relationships based on measuring the variability of d-dimensional subspaces. Three measures of variability have been proposed, the first of which is based on Hotelling's [6] vector correlation coefficient and the other two use the matrix 2-norm. It should be noted that the bootstrap procedures are independent of the measure used to detect the relationships, thereby allowing for a wide range of statistical applications.

Inherent in our methods is minimizing the sample version of $\mathcal{G}(\mathbf{a}, \mathbf{b})$ and estimating the unknown regression functions. The Nadaraya-Watson smoother is used to estimate these functions, which requires choosing kernel density estimators, bandwidths and initial guesses. Generally, in projection pursuit methods the choice of kernel may not be crucial (Härdle [4]) but the choice of bandwidth may be important. Of most consequence to the stability of the algorithm and ease of implementation, is the standardizing of the random vectors to have identity covariance matrices. Computationally, this is beneficial because the covariance matrices are set to the identity in the constraints and in each iteration of the algorithm. Moreover, this allows datasets with low variation to be dealt with before implementation of the algorithm. Note that the standardization only changes the scale and it is easy to translate back to the original scale. In summary, the algorithm described in Section 3.2.3 is very stable when the constrained minimizer *fmincon* of Matlab, which implements a Sequential Quadratic Programming algorithm, is used. The *fmincon* function worked well for our calculations, but alternative minimizers can be used.

Since our method is a projection pursuit method with the goal of estimating the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ and the estimation of m_1 and m_2 is viewed as an intermediate step, we believe an argument for consistency can be developed using the results of Hall [3]. That is, a possible proof may be outlined as follows: for a given $\mathbf{b}$ the $\widehat{\mathbf{a}}$ that minimizes $\mathbf{Q}(\mathbf{a}) = \Sigma_{i=1}^{n}(\mathbf{b}^T\mathbf{y}_i - \widehat{m}(\mathbf{a}^T\mathbf{x}_i))^2$ gives $\widehat{\mathbf{a}} \to \mathbf{a}$, by Hall [3]. Also, for a given $\mathbf{a}$ then we have $\widehat{\mathbf{b}} \to \mathbf{b}$ where $\widehat{\mathbf{b}}$ minimizes $\mathbf{Q}(\mathbf{b}) = \Sigma_{i=1}^{n}(\mathbf{a}^T\mathbf{x}_i - \widehat{m}(\mathbf{b}^T\mathbf{y}_i))^2$. Next, we believe an argument combining

these results would show that $\widehat{\mathbf{a}} \rightarrow \mathbf{a}$ and $\widehat{\mathbf{b}} \rightarrow \mathbf{b}$, where $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ minimize $\mathcal{G}(\mathbf{a}, \mathbf{b}) = \Sigma_{i=1}^{n}[(\mathbf{b}^T\mathbf{y}_i - m_1(\mathbf{a}^T\mathbf{x}_i)^2 + (\mathbf{a}^T\mathbf{x}_i - m_2(\mathbf{b}^T\mathbf{y}_i))^2]$.

As for visualizing the relationships, the simulations show that our measures recover both linear and nonlinear associations for multiple groups and/or multiple sets. In calculating the sample version of $\mathcal{G}(\mathbf{a}, \mathbf{b})$, we investigated two commonly used kernels, namely the Gaussian and Epanechnikov, for each combination of three bandwidths. The simulation studies illustrate that neither the choice of kernel nor the selection of bandwidth was crucial to the performance of our method. Moreover, the performance seemed unaffected by our simple strategy of generating initial guesses. In order to detect the number of relationships between sets using our bootstrap procedure, we compared the performance of the method using both GCA and CCA. We show that our method, unlike CCA, consistently detects the true number of associations in addition to identifying the true nature of the relationships. For Simulation 3.3.2, while the CCA method correctly determines the true number of relationships for one of the datasets, it does not identify the true nature of the associations. Therefore, in correctly detecting the nature and the number of relationships, our method clearly outperforms CCA in this regard. Finally, the simulations show that our methods perform well for moderate sample sizes and gain accuracy as the sample size increases.

3.5 Appendix

3.5.1 Invariance

Proof of Invariance:

Let $\mathbf{X}$ and $\mathbf{Y}$ be mean centered multivariate random vectors and define the following non-singular matrix transformations, $\mathbf{U}_1 = \mathbf{B}_1\mathbf{X}$ and $\mathbf{U}_2 = \mathbf{B}_2\mathbf{Y} \rightarrow \mathbf{X} = \mathbf{B}_1^{-1}\mathbf{U}_1$ and $\mathbf{Y} = \mathbf{B}_2^{-1}\mathbf{U}_2$ then,

$$\begin{aligned}
\mathcal{G}_{\mathbf{X},\mathbf{Y}}(\mathbf{a}, \mathbf{b}) &= \mathrm{E}(\mathbf{b}^T\mathbf{Y} - m_1(\mathbf{a}^T\mathbf{X}))^2 + \mathrm{E}(\mathbf{a}^T\mathbf{X} - m_1(\mathbf{b}^T\mathbf{Y}))^2 \\
&= \mathrm{E}(\mathbf{b}^T\mathbf{B}_2^{-1}\mathbf{U}_2 - m_1(\mathbf{a}^T\mathbf{B}_1^{-1}\mathbf{U}_1))^2 + \mathrm{E}(\mathbf{a}^T\mathbf{B}_1^{-1}\mathbf{U}_1 - m_1(\mathbf{b}^T\mathbf{B}_2^{-1}\mathbf{U}_2))^2
\end{aligned}$$

$$= \mathcal{G}_{\mathbf{U}_1,\mathbf{U}_1}\left(\mathbf{B}_1^{-T}\mathbf{a}, \mathbf{B}_2^{-T}\mathbf{b}\right)$$

Therefore, $\mathcal{G}(\mathbf{a},\mathbf{b})$ is invariant under nonsingular transformation.

3.5.2 CCA equivalence

Proof of equivalence to CCA:

Let $m_1(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{X}$ and $m_2(\mathbf{b}^T\mathbf{Y}) = \mathbf{b}^T\mathbf{Y}$ and w.o.l.g. assume $\mathbf{a}^T\mathbf{\Sigma_X}\mathbf{a} = \mathbf{b}^T\mathbf{\Sigma_Y}\mathbf{b} = \mathbf{I}$. Then,

$$\begin{aligned}
\mathrm{E}\left(\mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y}\right)^2 &= \mathrm{E}\left(\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a} + \mathbf{b}^T\mathbf{Y}\mathbf{Y}^T\mathbf{b} - \mathbf{a}^T\mathbf{X}\mathbf{Y}^T\mathbf{b} - \mathbf{b}^T\mathbf{Y}\mathbf{X}^T\mathbf{a}\right) \\
&= \mathrm{E}\left(\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a} + \mathbf{b}^T\mathbf{Y}\mathbf{Y}^T\mathbf{b} - 2\mathbf{a}^T\mathbf{X}\mathbf{Y}^T\mathbf{b}\right) \\
&= \mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a} + \mathbf{b}^T\Sigma_{\mathbf{Y}}\mathbf{b} - 2\mathbf{a}^T\Sigma_{\mathbf{XY}}\mathbf{b} \\
&= 2 - 2\mathbf{a}^T\Sigma_{\mathbf{XY}}\mathbf{b}
\end{aligned}$$

Therefore, the $\min_{\mathbf{a},\mathbf{b}} \mathrm{E}\left(\mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y}\right)^2 \equiv \min_{\mathbf{a},\mathbf{b}}\left(-\mathbf{a}^T\Sigma_{\mathbf{xy}}\mathbf{b}\right) \equiv \max_{\mathbf{a},\mathbf{b}}\left(\mathbf{a}^T\Sigma_{\mathbf{xy}}\mathbf{b}\right)$. In addition, since $\mathcal{G}(\mathbf{a},\mathbf{b}) = 2\mathrm{E}\left(\mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y}\right)^2$ the $\min_{\mathbf{a},\mathbf{b}} \mathcal{G}(\mathbf{a},\mathbf{b}) \equiv \max_{\mathbf{a},\mathbf{b}}\left(\mathbf{a}^T\Sigma_{\mathbf{xy}}\mathbf{b}\right)$, thus both are equivalent to CCA.

3.5.3

Result 1: If $\mathbf{A}_{n\times n}$ is a projection matrix, then $\mathbf{A}^T = \mathbf{A}$ and $\mathbf{A}^2 = \mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{A}$ and the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is a spectral decomposition of $\mathbf{A}$.

proof:

$$\mathbf{A} = \mathbf{A}\mathbf{A}^T = \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\right)\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\right)^T = \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\right)\left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\right) = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

The matrix of singular values $\mathbf{\Sigma}^2 = \mathbf{\Sigma}$ is the matrix of eigenvalues of a projection matrix($\lambda =$ 0 or 1) and $\mathbf{U}$ is an orthonormal basis.

Result 2: If $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ are projection matrices then $||\mathbf{A}\mathbf{B}||_2 \leq 1$.

proof: Let $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_2^T$ and $\mathbf{B} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$ be the singular values decompositions of $\mathbf{A}$ and $\mathbf{B}$ respectively and note the following,

a) If $\mathbf{Q}$ is unitary, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ or $\mathbf{Q}^{-1} = \mathbf{Q}^T$, then $||\mathbf{QA}||_2 = ||\mathbf{A}||_2$.

b) The 2-norm of $\mathbf{A}(\mathbf{B})$ is:

$$\begin{aligned} ||\mathbf{A}||_2 = ||\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T||_2 &= ||\mathbf{U}_1(\mathbf{\Sigma}_1\mathbf{V}_1^T)||_2 && (\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I} \text{ apply a)} \\ &= ||\mathbf{\Sigma}_1\mathbf{V}_1^T||_2 = ||V_1^T\mathbf{\Sigma}_1||_2 && (\mathbf{V}_1^T\mathbf{V}_1 = \mathbf{I} \text{ apply a)} \\ &= ||\mathbf{\Sigma}_1||_2 = \max_{1\le i\le n} |\sigma_{1_i}| \end{aligned}$$

Since $\mathbf{A}$ and $\mathbf{B}$ are projection matrices with maximum eigenvalue $\lambda = 1$ and the singular values are equal to the eigenvalues,

$$\begin{aligned} ||\mathbf{AB}||_2 \le ||\mathbf{A}||_2||\mathbf{B}||_2 &= ||\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T||_2||\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T||_2 \\ &= ||\mathbf{\Sigma}_1||_2||\mathbf{\Sigma}_2||_2 = \max_{1\le i\le n} |\sigma_{1_i}| \max_{1\le i\le n} |\sigma_{2_i}| \le 1 \end{aligned}$$

3.6 References

[1] Flury, B. and Riedwyl, H. (1988). Multivariate Statistics: A Practical Approach. Chapman & Hall, London.

[2] Goria, M.N. and Flury, B.D. (1996). Common canonical variates in k independent groups. *Journal of the American Statistical Association*, Vol. 91, No. 436, 1735-42.

[3] Hall, P. (1989). On Projection Pursuit Regression. *The Annals of Statistics*, Vol. 17, No. 2, 573-88.

[4] Härdle, W. (1990). Applied nonparametric regression. Cambridge University Press.

[5] Härdle, W., Hall P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, Vol. 21, No. 1., 157-78.

[6] Hotelling, H. (1936). Relations between two sets of variables, *Biometrika*, Vol. 28, No. 3/4., 321-77.

[7] Ichimura, H. (1987). Estimation of single index models. Ph.D. dissertation, Dept. Economics, MIT.

[8] Li, B. Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33, 1580-1616.

[9] Nadaraya, E. A. (1964) On Estimating Regression. *Theory of Probability and Its Applications*, 9, 141-42.

[10] Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman & Hall.

[11] Van de Burg, E. and De Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54-80.

[12] Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā*, Series A, 26, 359-72.

[13] Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002), An adaptive estimation of dimension reduction. *Journal of the Royal Statistical Society*, Ser. B, Vol. 64, No. 3, 363-410.

[14] Ye, Z. and Weiss, R. E. (2003). Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association*, Vol. 98, No. 464, 968-79.

[15] Yin, X (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91, 161-76.

[16] Yin, X and Sriram, T. N. (2006). Common canonical variates for independent groups using information theory. *Statistica Sinica*, to appear.

Chapter 4

Conclusions

In this chapter we give some brief ending remarks regarding each of the two nonparametric dimension reduction methods developed in this thesis. In Chapter 2, we developed a method, termed Kullback-Leibler Information Canonical Analysis(KLICA), to measure the relationships between multiple sets of random vectors using information theory. In contrast to other dimension reduction methods, our index provides an overall measure of association between multiple sets. The need for such a measure of association arises in morphological integration studies where an important task is to determine whether a structure is a single integrated unit or one that consists of several distinct modules. Through extensive simulation studies, we established the superiority of this overall measure of association to the CCA based methods suggested in the literature. Next, from this index, two powerful measures for dimension reduction were developed to recover both (joint and/or marginal) linear and nonlinear relationships between multiple sets. As for visualizing the relationships, the simulations show that these two measures identify both linear and nonlinear associations whether there are multiple sets or multiple groups. Moreover, when linear relationships are of primary concern or normality holds, as in the water-strider example, our methods are as powerful as the common CCA method. For dimension reduction, the p-values of the sequential permutation test provide an analytic measure of the significance of the detected associations. Although this research is initially motivated by a problem in biological science, we believe that these methods have a wide range of application; for example, in the behavioral and ecological sciences, where CCA is often used.

In reference to the problem of determining the existence of modules, alternative measures (similar to CCA) using a partition or permutation test have been proposed. A certain amount of dependence may be acceptable between modules, but in order to determine whether or not this level is exceeded, one needs to first measure the strength of dependence accurately. A future topic of research is to obtain a p-value correction that accounts for a certain amount of acceptable dependence, as determined by the researcher.

In Chapter 3, the Generalized Canonical Analysis(GCA) method was proposed as a nonparametric extension of CCA to detect both linear and nonlinear relationships. This index is a composite L_2 distance involving the use of single index model regression functions to measure relationships between sets. In comparison, this approach is completely different from the KLICA index, which measures the distance between joint and marginal densities, respectively, of linear combinations of the random vectors. While the motivation for the KLICA methodology was to provide an overall measure of association between multiple sets of random vectors, here our primary focus is on dimension reduction and we show that the new index $\mathcal{G}(\mathbf{a}, \mathbf{b})$ meets this objective. Note that the KLICA methodology can also be used for this purpose. Extensions of the $\mathcal{G}(\mathbf{a}, \mathbf{b})$ index to multiple sets and multiple groups are also given.

As for identifying relationships, the extensive simulations studies show that the GCA method recovers associations between multiple sets and/or multiple groups accurately. Of importance to any nonparametric method involving density estimation, these simulations reveal the robustness of our method to the choice of kernel, bandwidth and initial guess. In addition, the bootstrap procedure provides another analytic technique for detecting the number of significant relationships. Of note here is that this bootstrap methodology is independent of the measure used to detect the relationships between sets. Finally, in light of the performance of the GCA method in simulation, we propose to establish the consistency of the estimated coefficient vectors using the results of Hall [3].

In conclusion, the different methodologies and their respective properties make them powerful tools for dimension reduction in a wide range of statistical applications.

Bibliography

[1] Dryden, I.L., and Mardia, K.V. (1998). Statistical Shape Analysis. Wiley, Chichester.

[2] Flury, B. and Riedwyl, H. (1988). Multivariate Statistics: A Practical Approach. Chapman & Hall, London.

[3] Goria, M.N. and Flury, B.D. (1996). Common canonical variates in k independent groups. *Journal of the American Statistical Association*, Vol. 91, No. 436, 1735-42.

[4] Hall, P. (1989). On Projection Pursuit Regression. *The Annals of Statistics*, Vol. 17, No. 2, 573-88.

[5] Härdle, W. (1990). Applied nonparametric regression. Cambridge University Press. *Journal of the Royal Statistical Society*, Ser. B, 64, 363-410.

[6] Härdle, W., Hall P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, Vol. 21, No. 1., 157-78.

[7] Hotelling, H. (1936). Relations between two sets of variables, *Biometrika*, Vol. 28, No. 3/4., 321-77.

[8] Ichimura, H. (1987). Estimation of single index models. Ph.D. dissertation, Dept. Economics, MIT.

[9] Joe, H. (1989). Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Society*, Vol. 84, No. 405. (Mar., 1989) 157-64.

[10] Kettenring, J.R. (1971). Canonical correlation analysis of several sets of variables. In *Biometrika*, 58, No. 3, 433-51.

[11] Kettenring, J.R (1985). Canonical correlation analysis. *Encyclopedia of statistical sciences*, 1 ED. S. Kotz and N.L. Johnson, pp. 354-65. New York: John Wiley.

[12] Kiefer, J (1961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, *Pacific Journal of Mathematics*, 11, 649-59.

[13] Klingenberg, C.P., McIntyre, G. S. (1998). Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with procrustes methods. Evolution, 52, 1363-75.

[14] Klingenberg, C.P., Barluenga, M. and Meyer, A. (2002). Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. Evolution, 56, 1909-20.

[15] Klingenberg, C.P. (2005). Developmental constraints, modules and evolvability. Pages 219-47 in Variation: A Central Concept in Biology (B. Hallgrimsson and B.K. Hall, eds.). Elsevier, Burlington, MA.

[16] Kullback S.(1959). Information Theory and Statistics. John Wiley & Sons.

[17] Li, B. Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33, 1580-1616.

[18] Luijtens, K., Symons, F. and Vuylsteke-Wauters, M. (1994). Linear and non-linear canonical correlation analysis: and exploratory tool for the analysis of group-structured data. *Journal of Applied Statistics*, V. 21, No. 3, 43-61.

[19] Nadaraya, E. A. (1964) On Estimating Regression. *Theory of Probability and Its Applications*, 9, 141-42.

[20] Neuenschwander, B.E. and Flury, B.D. (1995). Common canonical variates. *Biometrika*, 82, No. 3, 553-60. *The Annals of Statistics*, 33, 1580-1616.

[21] Robert, P. and Escoufier, Y.(1976). A unifying tool for multivariate statistical methods: The RV Coefficient. *Applied Statistics*, Vol. 25, No. 3, 257-65.

[22] Ruschendorf, L. (1977). Consistency of estimators for multivariate density functions and for the mode, *Sankhyā*, Series A 39 243-50.

[23] Schott, J.R. (1997). Matrix Analysis for Statistics. John Wiley & Sons, Inc.

[24] Scott, D.W. (1992). Multivariate density estimation: Theory, Practice and Visualization. John Wiley & Sons.

[25] Shi, S.G., Taam, W. (1992). Non-linear canonical correlation analysis with a simulated annealing solution. *Journal of Applied Statistis*, V. 19, No. 1, 155-65.

[26] Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman & Hall.

[27] Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā*, Series A, 26, 359-72.

[28] Van de Burg, E. and De Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54-80.

[29] Van de Burg, E. De Leeuw, J. Verdegaal, R. (1988). Non-linear canonical correlation with m sets of variables. *Psychometrika*, 2, 171-97.

[30] Vinograde, B. (1950). Canonical positive definite matrices under internal linear transformations. *Proceedings of the American Mathematical Society*, Vol. 1, No. 2, 159-61.

[31] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), An adaptive estimation of dimension reduction. *Journal of the Royal Statistical Society*, Ser. B, Vol. 64, No. 3, 363-410.

[32] Ye, Z. and Weiss, R. E. (2003). Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association*, Vol. 98, No. 464, 968-79.

[33] Yin, X (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91, 161-76.

[34] Yin, X and Sriram, T. N. (2006). Common canonical variates for independent groups using information theory. *Statistica Sinica*, to appear.