Random Subspace Analysis on Canonical Correlation of High Dimensional Data

By Ryo Yamazaki

Department of Statistics
Uppsala University

Supervisor: Måns Thulin

Year 2016
Abstract

High dimensional, low sample, data have singular sample covariance matrices, rendering them impossible to analyse by regular canonical correlation (CC). By using random subspace method (RSM) calculation of canonical correlation becomes possible, and a Monte Carlo analysis shows resulting maximal CC can reliably distinguish between data with true correlation (above 0.5) and without. Statistics gathered from RSMCCA can be used to model true population correlation by beta regression, given certain characteristic of data set. RSM-CCA applied on real biological data however show that the method can be sensitive to deviation from normality and high degrees of multi-collinearity.
Contents

1 Introduction 3

2 Explanation of method 4
  2.1 Canonical correlation ........................................... 4
  2.2 Random subspace method ....................................... 4
  2.3 Overfitting, population CCA, sample CCA .................. 5

3 Pre-analysis 7
  3.1 Potentially relevant parameters ............................... 7
  3.2 RSMCC for \( n \) and \( k \) ....................................... 7
  3.3 RSMCC and \( p \) ............................................... 8
  3.4 RSMCC and \( t \) ............................................... 8
  3.5 Summary ...................................................... 10

4 Main study 11
  4.1 Hypothesis and research question ............................ 11
  4.2 Experiment design ........................................... 12
    4.2.1 Material ................................................. 12
    4.2.2 Significance check ...................................... 13
  4.3 Result .......................................................... 14
    4.3.1 Critical limits for alpha = 0.05 ......................... 14
    4.3.2 Maximal CC by true population correlation ............. 14
    4.3.3 Exploring other statistics ............................... 20
    4.3.4 Estimating true population ............................... 24

5 Discussion 30
  5.1 Limitation in generalizability of result .................... 30
  5.2 Can RSM CC be used to detect real correlation in real chaotic data? 31
  5.3 What does a significant correlation tell us? ................. 31
  5.4 Effect of choice \( k \) .......................................... 32

6 RSM CC on real data 32
  6.1 Introduction .................................................. 32
  6.2 Material ...................................................... 33
  6.3 Result ........................................................ 34
  6.4 Discussion ................................................... 36

7 References 36
1 Introduction

The use of multivariate techniques presuppose sufficient size of observations in data. The minimum number of observations required is influenced by characteristic of the model, and often higher number of variables included will increase the number of observations necessary. For methods where covariance matrix inverts are required for the calculation of the multivariate statistic, the number of observation can not be smaller than the sample size (Bai & Shi, 2011). This is a problem for applicability of the methods, since in certain fields, like that of genetic studies, the number of available variables are high, while the cost and effort required for acquiring observation are limiting. Treating such high dimensional, low sample data plausibly is therefore an issue that needs solving.

One proposed general solution to this is the usage of the so called Random Subspace Method (RSM). RSM has been applied on decision tree models for machine learning (Ho, 1988), and other high dimensional data such as fMRI (Kuncheva et al., 2010). The simple nature and general procedure of the method makes it possible to combine RSM with multivariate statistical tests as well (Thulin, 2014). However, while it is theoretically possible to combine RSM with other methods, the consequence of RSM to the interpretation and validity of statistic is not uniform. This study will attempt to explain one particular scenario, RSM applied to canonical correlation analysis (CCA). As CCA is a general method of modeling linear relationships between variables (Sherry & Henson, 2005), finding out how RSM works with it will ideally provide insight into the method’s usability for multivariate or univariate linear statistical tests in a general sense.

The aim of this study is to investigate how the first canonical correlation coefficient obtained by RSM can explain data in the population. That is, is the coefficient obtained a valid estimator for population correlation in isolation? If not, then can it be used as a variable in a model to estimate population correlation? And third, can the canonical correlation coefficient be used as a test statistic to distinguish between population with or without actual correlation? All these questions are part of an overarching question how useful the information in canonical correlation RSM coefficient is when wanting to judge the nature of population from a sample.

This text will consist of firstly a theoretical discussion section, in which fundamental notions and intuition behind RSM and CCA will be presented. Next there will an exploratory investigation of effect of CC statistics, that is, theoretical zero canonical correlation of population matrix. The main study will then be analyzed by Monte Carlo method for population matrices with true correlation. Finally, result will be applied to test for true correlation of real biological datasets, to assess the method’s usability in practice. As this study is mainly explorative, greater emphasis will be assigned to visual plotted result rather than p-value and strict significance.

For the following text the abbreviation CCA will be used to denote the entire method, while CC will be used to refer to the correlation coefficient formed from CCA (in most cases the first canonical correlation coefficient).
2 Explanation of method

2.1 Canonical correlation

CCA can be said to be a general method of which most parametric tests can be considered a sub-variant to (Knapp, 1978). For two sets of variables CCA finds the linear combination sum (or variate), so that the sums of both sets are maximally linearly correlated. While linear correlation is a mutual relationship, and in strict sense there is no dependent or independent set of variables, one can still interpret the resulting canonical $R^2$ as the rate of variation of one set of the data explained by the other. The strength of CCA can be said to lie in its generality in finding linear correlation. Indeed, the method has been applied in modified form to primarily practical fields such as image retrieval, pattern recognition and computer vision (Chen, 2005).

CCA is calculated by using the sample covariance matrix for both sets of variables, as well as the covariance between the two sets. In doing this, the inverse of the covariance matrix is also used. Inverse of square matrices are defined if and only if the rank of the matrix is equal to the matrix order (Harville, 1997). Each row in the column needs to be linearly independent (in algebraic and not probabilistic sense of independence) from each other.

An intuitive explanation why this is problematic for high dimensional data can be given by the nature of matrix rank and covariance matrices. If one were to summarize an observation dataset in the form of a table, wherein the columns all denote values of a single variable, and the row represent one observation, then this table could be seen as a matrix. Firstly, the row rank of a matrix will always equal column rank (Emanuelsson, 2004). This means that no matrix can have a rank greater than $\min(\text{row}, \text{column})$. For the observation data table thus, given that the number of variables (column size) is greater than number of observation (row size), the maximum rank such a table can theoretically have is equal to the number of observations $n$. A covariance matrix can be considered as the mean corrected observation matrix multiplied by the transpose of itself, divided by $n$. Knowing that rank of the product of two matrices is equal to the smallest of the rank of the factor matrices, the resulting sample covariance matrix will thus always have a rank smaller or equal to $n$. Given that $n$ is smaller than the number of variables $p$, the rank of the covariance matrix is therefore always smaller than the order of the matrix. Thus, for high dimensional observations inverses to covariance matrices are not defined.

2.2 Random subspace method

To counter-act the issue of singular sample covariance matrix there are a number of suggested solutions. One way is to apply dimensional reduction prior to CCA (Sun et al., 2005). Dimension reduction, while attractive, often use similar methods of calculation such as CCA, and can themselves be plagued by high dimensional issues of covariance matrix (Lee, Lee & Park, 2012). Another method is the so called regularization of CCA. This involves adjusting the covariances by adding a constant diagonal matrix, so that the resulting new matrix has a reliable inverse (Nielsen et al., 1998). However, the choice of the regularizing constant can
have effect on the result of CCA, putting certain demand on finding the optimal value (Cano & Lee, 2014).

The easiest solution is however to simply limit the data, and exclude variables prior to analysis. However, this will naturally cause data loss. There are ways to see beforehand which variables are potentially interesting by individually analysing the correlation coefficient of the variables. However, such resulting statistics can often be weak (Thulin, 2014) or fail to capture relevant covariance effect within the set of "explaining" variables (Ho, 1988).

Another alternative, that will be the topic of this study, carries the simplicity of data elimination, while not removing the common correlation by using bigger sets of variables. Instead of taking individuals variables, a subset of variables, the size of which is denoted by \( k \), is used as the data for CCA. While one such subspace is insufficient in capturing all the information in the dataset, repeated randomized subspaces could be drawn, from each of which statistics can be formed (Ho, 1988). This approach called random subspace method (RSM) has been proven successful when applied to Hotelling's T-test comparison of multivariate mean (Thulin, 2014).

The basic idea behind random subspace method works, when a function of several repeated random subspace statistics can be a valid estimator for the statistic of the dataset as a whole. That is, the method is appropriate when an individual random subspace may not capture the entire information of the data, but when repeated sampling converges into the "true" population value. In the Hotelling's multivariate T-test using RSM the "combining" function was a weighted average of all subspace T-squared statistics (Thulin, 2014). For CCA, the statistic of interest is the variate correlation. However, averaging the variate correlations will likely not converge into the true value. One subspace correlation will inevitably be lower than the sample correlation of the entire dataset. Thus, the greatest correlation found among the subspaces should pose a better estimate for the total correlation, and for the coming analysis this will the be "summarizing" function of choice. This is however not an entirely unproblematic choice because of several reasons presented in the following section.

2.3 Overfitting, population CCA, sample CCA

In the above section we have concluded that we want to use RSMCCA to estimate the true population correlation between two sets of variables. However, is this a reasonable notion to investigate? One needs not look deeper into empirical data to conclude this is problematic. First we can consider the distinction between population correlation and sample correlation. The sample CC is not an unbiased estimator of population CC. This can easily be illustrated by the fact that we can imagine a population CC between two sets of variables being zero, but the expected value of a sample CC can never be exactly 0, because CC can never obtain negative values, while it can obtain values above 0 by chance. Thus, at the very least, for CCA on data that has zero inter-set correlation in the population, the sample CC will inevitably be greater than population CC.

Another issue of sample CC overestimation is that for any added variable to the model, the resulting larger model will inevitably have equal or higher CC than that of the smaller model. Therefore, for the case of high dimensional data sample where \( p \) is large, estimating population CC is particularly a problem. In fact it
is not only that the resulting coefficient needs to be adjusted, but also for the following reason, the meaning of the interpretation itself must be reconsidered. We have already noticed that the rank of continuous stochastic observation table (treated as a matrix) is equal to that of $\min(p, n)$. Thus for high dimensional low sample data the rank will equal $n$. We also know that for $p$ algebraically linearly independent vectors all in vector space $\mathbb{R}^n$ a linear sum of those vectors can equal all vectors in $\mathbb{R}^n$ (Emanuelsson, 2004). A CCA can be said to be complete when the linear combination of one of the variable set is completely equal (for every observation) to a linear combination from the other set. During such cases the CCA can be considered trivial, and from a statistical point of view redundant. This situation hold when for two observation matrices $a$, $b$ there are vectors $c$ and $d$ such that the following holds:

$$
\begin{pmatrix}
    a_{1,1} & a_{1,2} & \cdots & a_{1,p_1} \\
    a_{2,1} & a_{2,2} & \cdots & a_{2,p_1} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n,1} & a_{n,2} & \cdots & a_{n,p_1}
\end{pmatrix}
\begin{pmatrix}
    c_1 \\
    c_2 \\
    \vdots \\
    c_{p_1}
\end{pmatrix}
= 
\begin{pmatrix}
    b_{1,1} & b_{1,2} & \cdots & b_{1,p_2} \\
    b_{2,1} & b_{2,2} & \cdots & b_{2,p_2} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{n,1} & b_{n,2} & \cdots & b_{n,p_2}
\end{pmatrix}
\begin{pmatrix}
    d_1 \\
    d_2 \\
    \vdots \\
    d_{p_2}
\end{pmatrix}
$$

This equation is equivalent to:

$$
\begin{pmatrix}
    a_{1,1} & a_{1,2} & \cdots & a_{1,p_1} & b_{1,1} & \cdots & b_{1,p_2} \\
    a_{2,1} & a_{2,2} & \cdots & a_{2,p_1} & b_{2,1} & \cdots & b_{2,p_2} \\
    \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
    a_{n,1} & a_{n,2} & \cdots & a_{n,p_1} & b_{n,1} & \cdots & b_{n,p_2}
\end{pmatrix}
\begin{pmatrix}
    c_1 \\
    c_2 \\
    \vdots \\
    c_{p_1}
\end{pmatrix}
= 
\begin{pmatrix}
    0 \\
    0 \\
    \vdots \\
    0
\end{pmatrix}
$$

We have already seen that for high dimensional datasets the matrix to the left hand has $n$ independent vectors, and the null vector to the right hand side is a member of vector space $\mathbb{R}^n$, therefore there will always exist a vector $c$ and $d$ so that the equation is fulfilled. From this one can see that for CCA in which $n$ is equal or smaller than $p$, the resulting maximal CC of the sample will always be 1, and therefore completely linear.

Recall that maximal RSM CC will always underestimate the sample CC, in light of above result this information becomes trivial when we have now found out that the sample CC for high dimensional data will always be 1. So RSM CC underestimates sample CC, while sample CC in turn overestimates the population CC (or in fact will equal 1). Because of this the relationship between RSM CC and population CC becomes obscure.

Another problem of this result is that the size of each individual subspace becomes hugely relevant for the outcome of RSM CC. The size of each subspace (denoted $k$) will be chosen by the researcher using CCA, and a $k$ close to $n$ will inevitably overfit the variation, leading to CC close to 1. However, a low $k$ will fail to capture all the information together. The choice of $k$ is therefore not only influential to the outcome of RSM CC, but it often alone vary the expected CC from a value close to 0 to something extremely near 1.

The idea of overfitted sample CCA, as well as the high impact on outcome by the choice of $k$, provide difficulties in the performance and interpretation of RSM
CCA. Due to the resulting CC not having an easily explained relationship to population CC it is hard to treat RSM CCA as a statistical test, or method of general estimation of some population parameter. Rather, it could be more fruitful to think of the process as that of continual and more researcher driven data exploration. In the following section I will argue that the complexity and abundance of parameters involved in RSM CCA is not always a limitation. In arguing for this I will look closer into the distribution of RSM CC during cases of null population CC, that is when the covariance matrix for the data generation equals an identity matrix. If it is illustratable that for a given choice of parameter the expected distribution of RSM CC for zero correlative data is stable, then it could be possible to apply Monte Carlo testing of zero correlation by RSMCC.

3 Pre-analysis

3.1 Potentially relevant parameters

Before going into the main analysis of this study a closer investigation regarding parameters involved in RSMCCA is necessary. We have already touched on how RSMCC, by its nature, has different result depending on how many variables are contained within a single analysis. For the question of RSMCCA it is further necessary to highlight other choice of parameters. To summarize, this section will be dedicated to getting a glimpse of how the choice of \( k \) (number of independent variables that is sampled for each CCA), \( t \) (number of random sample trials), \( p \) (number of total variables), and \( n \) (number of sample observations), influence the outcome of RSMCCA.

Two aspects which are highly relevant, but will not be discussed in detail here is that of multicollinearity within set, and that of number/shape of true correlation. Multicollinearity is potentially influential to result, as it is the way in which the information of large data CCA can get "tangled". While no proper investigation will be done regarding multicollinearity, a hypothesis can be that large degree of collinearity within set of variables will yield lower CC coefficients, since the collinearity can be considered "overlap" of variance information.

The parameter \( p \) will be considered, however how the \( p \) is divided into both variable sets will not be varied. For all RSM CC in this text the "dependent" variable set will consist of five variables, and the independent variable set will be of size \( p - 5 \). This means that random subspace draws will only be made from the independent variable set. For practical purposes it is time consuming to test bilateral RSMCCA, as the number of possible combination of subspaces gets inflated. In many case of real life CCA analysis both the dependent and independent variable sets will contain a high number of variables, and thus it could be necessary to random subspace both of these simultaneously. While this study does not investigate the effect of such bilateral RSM, there should in theory not be a large difference in conducting such from a unilateral RSMCCA.

3.2 RSMCC for \( n \) and \( k \)

Figure 1 shows the beanplot distribution of the largest CC for data generated from a 100x100 identity matrix and analyzed with RSMCCA. Each "bean" represent the
distribution of highest CC for a given value of $n$. The size of $k$ is 5, and the number of repetitions, $t$ is 1000. The mean highest CC decreases as the number of samples increases. When $n$ is 100 most of the first CC appears centred around the value of 0.5, and for $n = 20$ the mean value is higher than 0.9.

![Figure 1: Distribution of highest CC for different $n$, $k = 5$](image)

Figure 2 shows the beanplot distribution of the largest first CC for the same population matrix as above. Instead of varying $n$ however for this different $k$ is chosen. Again it is visible that the expected first highest CC increases as the number of $k$ nears $n$. These two results combined make a strong case that RSMCCA under null distribution is heavily influenced by the ratio of $k$ compared to $n$. So having too few $n$ is a problem for RSMCCA, as it should be with regular CC.

### 3.3 RSMCC and $p$

The follow up question is whether the number of variables influence expected CC. Figure 3 shows the bean plot of highest CC generated from same distribution as above, however with varying number of variables. To compensate for greater number of possible combination the number of trials the number of trials the number of RSM trials equals $p * 20$. While certain deviations are visible from the data with smaller number of variables, mean max CC for each value of $p$ appears relatively similar from $p 10000$ to 40000.

### 3.4 RSMCC and $t$

Figure 4 shows the trajectory of highest CC coefficients over repeated subspace trials. Each line colour represent a value of $k$, and the value of $p$ is here 100. We can see that CC increases in rapid pace for the first few trials, and then for the first 10000 trials there are frequent smaller jumps upwards. After 20000 trials the value of the highest CC appears to be more or less stable, and when a superior CC is found the difference appears small. It is not given whether the size of $k$ influence the speed in which CC stabilizes, although logically the greater number of possible
Figure 2: Distribution of highest RSM CC under zero correlation for different $k, n = 50$

Figure 3: Change of maximal CC as variables increases
combinations of variables available given larger \( k \) would intuitively make sense as to provide greater possibility of CC increase even in later trials.

![Figure 4: Change of maximal CC as following trial number, \( p = 100 \)](image)

Figure 4 shows corresponding result for data with 1000 variables, again under zero population correlation. Compared to the smaller dataset we see higher degree of CC increases in later trials. Again, this is reasonable as there are simply more potential combination of variables in the bigger dataset. In other regards however the trajectory seems quite similar to our smaller data. The effect of size of \( k \) seems identical. When the later increases are included the larger data has slightly higher average CC, although the difference is not extremely large (consistent with previous result). We see more or less that at 20000 trials we can get reasonable approximation to when CC becomes stable.

### 3.5 Summary

The main effect of overfitting on CC appears to be defined by the size of \( k \) compared to \( n \). For larger \( k \) the CC is close to 1, and even when \( k \) is 2 and \( n \) is 50 the expected CC under non-correlation appear over 0.6. While these numbers are all quite near 1 it is not necessarily a problem, as you could replace CC coefficient with CC-squared, and get more even distribution of the statistic for different values of \( k \). Another result is that for any given combination where \( k \) and \( n \) is defined, the highest first CC under null hypothesis is quite robustly centred around a value. This is visible from the beanplots presented above, but also by the lineplots showing the change of the highest CC as a function of trial numbers.

Furthermore, it is noteworthy that the total size of datasets do not appear to influence the overfitting to excessive degree. For high-dimensional data this tolerance for higher number of parameters could provide useful, as there will be lesser
need to limit the data prior to analysis. However, it should be mentioned that the number of trials required before CC stabilizes will increase for larger datasets, putting certain computational demand.

All the results above are for data in which all variables have zero true correlation to one another, so given presence of true inter-set or intra-set correlation the trajectories of CC may look different. For sufficiently large size of $p$ the number of possible combinations of subspaces are so large the true choices maximal correlation are likely not found unless the number of trials are much larger. However, under such circumstances it is more effective to systematically test all combinations, and indeed such an approach would defeat the entire purpose of RSM. The following parts will investigate, that even when RSM only tests a fraction of all possible variable combinations, is the maximal CC found still able to distinguish between data in which inter-set correlation is present and when it is not present?

4 Main study

4.1 Hypothesis and research question

Initially we have defined the research question, "what is the relationship between RSMCC and population correlation?". During theoretical discussion we have found out that whatever the relationship may be, it is not likely that of an unbiased estimator. The follow up question is therefore if there exist any form of relationship at all between RSMCC and population correlation. To test this we will want to investigate whether highest RSMCC can be used as a test statistic to distinguish between samples derived from population with true correlation and samples from non-correlation population matrix.
1. For any given dataset in which the true inter-set canonical correlation is above 0, the RSMCCA highest CC will not be different from that under zero correlation.

2. The rate of discovery of non-zero canonical correlation by RSMCCA is equal or lower than that by fixed lambda regularized canonical correlation.

4.2 Experiment design

4.2.1 Material

The general procedure of data generation is presented in algorithm 1 in the form of pseudo-code. 1500 data frames are generated from different population correlation matrices. Each data frame generated this way contains 50 samples ($n = 50$), and the number of variables in the correlation matrix will equal 1000 ($p = 1000$). Variables 1-5 are the dependent variable set, and all remaining 995 variables are in the independent variable set. Out of all independent variables a maximum of 5 holds a population correlation to the dependent variables that is non-zero, these variables will henceforth be referred to as "target" variables.

**Result:** RSM statistic data over $d$ simulated data

```plaintext
for i to 1500 do
  Randomly pick target matrix with true correlation (0.15 to 0.95),
  dimension = 10x10;
  Randomly pick distraction matrix, dimension = 990x990;
  Combine target matrix with distraction matrix to create population
  correlation matrix, dimension = 1000x1000;
  Generate 50 observations from population correlation matrix (= Maindata);
  for K = 2 to 5 do
    for t to d do
      Y = Maindata[variable1 to variable5];
      X = Maindata[K random samples from (variable6 to variable1000)];
      Perform canonical correlation with variables Y and X (= RSMdata);
    end
    Extract summarizing statistic from RSMdata (Enddata);
    Return Enddata;
  end
end
```

**Algorithm 1:** RSM data generation

The target correlation matrices is generated by varying the values of maximally five inter-set correlations. Each dependent variable can correlate with one target variable. A total of 20000 matrices is produced in this manner, and for each of these the true canonical correlation coefficient is calculated by generating 10000 samples and performing regular CC (by cancor function in R). After this, $d$ number of random draws is taken from a uniform distribution with the lowest bound 0.15 and highest bound 0.95. The population matrix with true CC closest to the value of the draw is chosen as the matrix to be included in the analysis. By this, the matrices included in the analysis will be approximately uniformly represented CC
between 0.15 and 0.95.

Together with a choice of target-dependent variable matrix a separate matrix for the non-target independent variables are also generated. The distraction variables are all non-correlated with dependent or target variables, but have smaller correlation noise within group. The 990x990 distraction correlation matrix is created using the random.correlation function from the clusterGeneration package in R.

Once all correlation matrices are ready the sample data will be generated as multivariate normal distribution, with rvnorm function on R. Each dataset created this way is analysed by RSMCCA with different values for \( k \), ranging from 2 to 5. Summarizing statistics is recorded after having repeated 20000 RSM trials on one data. Among these statistics are: largest CC, mean CC, squared-mean CC, standard deviation of CC, median CC and mean of upper quarter CC, skewness of CC. Along with these a regularized canonical correlation analysis is also performed on the same data (modifying lambda = 0.5 on all), and the first CC is recorded for each dataset.

### 4.2.2 Significance check

Significance check from the RSMCCA, and RCC statistics created above is done by Monte Carlo method. Prior to data experiment several thousand of same statistics as the RSMCCA and RCC data is generated for \( n = 50, p = 1000, t = 2000 \), from an identity matrix of dimension 1000x1000. The result from the Monte Carlo simulation will be used as estimate for distribution of statistics under zero correlation (this distribution will henceforth be referred to as null-distribution). If say, one data set created with a given correlation matrix returns maximum CC of 0.8 (for some \( k \)), given that this value is higher or equal to the 95th percentile of the null-distribution of the same statistics, this data set will be considered to be correctly classified. In other words, for each dataset of RSMCC one would perform a significance check of alpha = 0.05, using an approximated critical limit for when data follows that of zero correlation population matrix. The same procedure is repeated for RCC statistics.

After detection/non-detection has been judged for all data sets, these will then be analysed in two forms, each analysis respectively corresponding to one of our previously defined null hypotheses. The first is: Is detection rate in RSMCCA significantly higher than chance? And the second: Is the detection rate of RSMCCA significantly higher than RCC? Both RSMCCA detection and RCC detection will be modelled in a logistic regression model with true population CCA as explaining parameter. This is to judge how successful detection probability is affected by the magnitude of correlation that is actually present in the population matrix used for data generation. If the modeled mean of detection is significantly above that (through confidence interval) of chance (here 0.05), then the null hypothesis can be rejected.

The latter half of the result will be concerned with fitting a suitable model to estimate true population CC from the RSM statistics. Here beta regression, using the betareg package on R, will be the main tool of modelling. The fit model is then applied to novel simulated data to analyse prediction accuracy.
4.3 Result

4.3.1 Critical limits for alpha = 0.05

Beanplot on figure 6 shows the null distribution of maximum CC for \( n = 50, \ p = 1000, \ t = 20000, \) and with \( k \) ranging from 2 to 5. Variation of maximal CC appears largest at smaller \( k \), and as \( k \) increase the average largest CC increases. The upper 5-th percent, or the 95 percentile, will serve as the critical limit for significance tests with alpha = 0.05 (single tailed) for coming analysis. The critical limits are for \( k = 2, 0.746743, \) for \( k = 3, 0.7743859, \) for \( k = 4, 0.7984569, \) for \( k = 5, 0.8135282. \) For any maximal CC above these values the null hypothesis of two sets not being correlated is rejected.

![Highest CC under zero correlation](image)

Figure 6: Distribution of highest RSM CC under null hypothesis

Corresponding result for RCC (\( \lambda_1 = 0.5, \ \lambda_2 = 0.5 \)) is given by figure 7 assessed from 4000 simulations. As RCC does not hold \( k \) as parameter there is only one critical limit at alpha = 0.05, 0.8744795.

4.3.2 Maximal CC by true population correlation

Figure 8 shows the result from RSMCCA simulation using population correlation matrix with non-zero inter-set correlation, with 1500 simulations per \( k \). The charts show the resulting highest CC found given a true population correlation, with each picture representing one value of \( k \). The line on each chart represent a fitted local polynomial regression. As can be seen for low population correlation (0.25 or below) the average highest CC is similar to that of expected highest CC under null distribution. However, on all choices of \( k \) for these simulation there is an increase
of CC appearing to start from a threshold value of around 0.5 true population CC. By a true population CC of 0.9 or higher the mean highest RSM CC appear to converge to a similar level for all choices of $k$.

Figure 9 shows the same data as above, only for this it is not the RSMCC coefficient that is plotted, but rather RSMCC coefficient percentile as compared to null distribution. A value on 0.5 on Y axis indicate the RSMCC for corresponding data is equal to that of median highest CC under zero correlation. Again, for smaller true correlation the mean percentile appear close to 0.5 for all choices of $k$. But as true correlation move towards 0.5 and beyond, the percentile value rapidly increase, and by population correlation of 0.85 almost all RSMCC is at 100th percentile of the null distribution. The shape of the curves for all four choices of $k$ look similar, with $k = 2$ appearing to have somewhat higher mean percentile as compared to the others for smaller population correlation, and $k = 5$ being slightly lower than the remaining lines for all but the highest population correlations. It is worth noting that the RSMCC statistics under true correlation appear to diverge from that of null distribution at a relatively early stage. That is, the expected RSMCC for $k = 2$ under null hypothesis for our data is roughly 0.7. However, the resulting RSMCC for our data shows that RSMCC is distinct under the presence of true correlation, even when the true correlation is lower than 0.7.

One can further simplify above plot, by only showing whether the RSMCCA managed to detect correlation or not (that is highest RSMCC is above that of alpha threshold). This value would be equivalent to the statistical power of the hypothesis test "Is this RSM data different from null distribution?". Figure 10 shows a local polynomial smoothing line fitted to the binary detection data. Again, for all choices of $k$ the line shows very similar trajectory. The detection rate increases starting correlation 0.38 and continues to be near 1 by population correlation 0.75.
Figure 8: Highest RSM by true population correlation and $k$
Figure 9: Percentile of highest CC value compared to null distribution
for all choices of $k$.

Now that detection rate is understood one can create logistical regression to model detection probability. Such a model could corroborate or reject our first null hypothesis for the research, "RSM detects correlation at chance level". Now we have already seen that for this question to be answered we first need to consider the true population CC, as it is obvious it majorly influence to detection rate. Figure 11 shows the logit model created by having detection as dependent variable, true population correlation and $k$ (dummy-coded) as independent variables. True population correlation has a three starred significant positive effect on odds (and also worth mentioning is that the intercept appears significant different from 0), however the effect of categorical variables for $k$ are not significant. Figure 12 shows the resulting model plotted with 99-percent confidence interval for detection probability mean. For all $k$ the interval and line appears highly overlapping. And for value of true population correlation above 0.41 we see that all lower limits to the confidence intervals are higher than that of the detection by chance rate of 0.05. Thus, we can conclude that RSMCC, for our data, can detect correlation when the true population correlation is 0.41 or above. Hower it might be necessary to interpret this logit model with caution, as the confidence intervals for smaller value of true correlation are in fact lower than chance, or even extremely close to 0, which we know should not be the case.

Figure 13 shows the distribution of RCC score, from the same data as above, as a function of true population correlation. The fitted local regression slope appears close to parallel. Figure 14 shows the same data fitted to a logistic regression model based on succesful detection of deviation from null distribution (alpha = 0.05). The ribbon around the fitted line represent a 99-percent confidence interval of true probability of detection. The effect of true population correlation on

![Figure 10: Detection rate by true population CC](image-url)
Call:
glm(formula = as.factor(RSNdat) ~ True, family = "binomial",
      data = totalSimT)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.3305 -0.3965 -0.1045  0.3489  3.3963

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.7220   0.1953  -39.53  <2e-16 ***
True         12.6017   0.3129   40.27  <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8210.2 on 5999 degrees of freedom
  Residual deviance: 3617.3 on 5998 degrees of freedom
  AIC: 3621.3

Number of Fisher Scoring iterations: 6

Figure 11: Logistic model summary, correlation detection modelled by true population CC

Figure 12: Logit model detection probability plotted with confidence band
the model is significantly above zero \((p = 0.021)\), with an estimated log-odds effect of 0.5287. This effect is rather small, since even at true population = 1, the estimated mean probability of detection is smaller than 0.075. However, as can be seen through the confidence interval in the figure, the rate of detection does nonetheless significantly deviate from chance level \((0.05)\) at a true population CC of approximately 0.51 and above. Note that this result is for RCC under fixed value of lambda, and as such should be seen only as an indicator for comparison rather than accurate representation of RCC applicability.

Figure 13: RCC score \((\lambda = 0.5)\) as function of true population correlation

Figure 15 shows both RCC logistic regression line and RSM regression line plotted together. Note that for RSM regression line all \(k\) are pooled together, resulting in narrower confidence interval for probability mean. We can visually see that the overlap of confidence interval for mean probability given a value of true CC cease to overlap at early stage. After which there is dramatic difference in detection probability observable between both logit models.

4.3.3 Exploring other statistics

From above analysis it is clear that at the very least maximal CC among the RSM can be used to distinguish between non-zero or zero population canonical correlation (at least for population CC higher than 0.4). As such we can partly answer our main research question about relationship between RSMCC and population CC. The connection is sufficiently strong for RSMCC to be used as a test statistic for detecting population CC presence. It could be useful to investigate whether other statistics gathered by RSM analysis could, not only explain presence, but also the magnitude of true population correlation. This section will attempt to find variables of interest, that can be used for modeling true population CC through ap-
Figure 14: Logit model of detection probability on RCC as a function of true population CC

Figure 15: Comparison of logit model of RSM and RCC
appropriate technique. As this section is mainly explorative, there will be less focus on p-values, and greater emphasis will be put on visual plotted data.

Figure 16 shows the average CC from all trials of RSM simulation for same data, by the size of population correlation. A certain slope can be observed on the fitted regression lines, however the residual variation is quite large compared to explained variation. This result is further expanded by figure 17, showing the smallest CC found among all RSM trials for all data sets. There is little to no effect on the slope by increased true population CC. The median of all CC shows only slight effect by true population correlation (see figure 18).

The above result is potentially explained by the theory that, for the majority of RSM trials, the target variables are not chosen, and as such the result is the same as a would be expected under zero correlation. So the summary of CC over all trials should be weighed up by a fraction of all trials, trials which contain target independent variables. This idea can be corroborated by figure 19 showing the skewness of RSM CC distribution for each dataset as a function of true population CC. The figure shows a clear positive effect of higher population CC on skewness, indicating that the distribution of RSM under high correlating target variables are skewed positively, in other words having a longer upper tail. From this it is reasonable to assume that higher RSM CC correlate with greater RSM distribution skewness, and vice versa. Indeed, the Pearson correlation coefficient between skewness and highest CC is 0.854.
Figure 17: Smallest RSM score over all trials

Figure 18: Median RSM score over all trials
4.3.4 Estimating true population

As we now know there is substantial correlation between RSM statistic of skewness and maximal CC to real population CC, a follow up question is whether this connection is strong enough to yield practical predictive models. Before going into model building it is worth highlighting that this section only concerns our specific set of datasets, that is the dataset which has: $n = 50$, $p = 1000$ (dependent = 5, independent target variable = 5, independent distraction variable = 990), no multicollinearity between distraction variable and target variable, no multicollinearity within dependent variables, no multicollinearity within independent target variables, and normal distribution of all variables. Sufficient to say, the predictive model created here will not be directly applicable to other type of datasets. However, there might still be value in researching whether the predictive models. This is because, if RSM statistics are succesfully able to esimate population CC, then it should indicate that the information of population CC is so to speak hidden but obtainable within the RSM statistics.

Figure 20 shows the distribution of true population CC as a function of highest RSM CC for all $k$ in the analysis. While there seems to be a strong hint at correlation on all levels of $k$, at least starting from some threshold RSM CC not too low, the relationship does not live up to that of unbiased estimator as the CC diverge far from diagonal line. Figure 21 shows the difference between RSM CC and population CC. It is apparent RSM CC is biased to overestimate population CC, although the degree of overestimation seems to diminish as a function of higher population CC. Thus, there is a clear correlative structure present here, although
not necessarily clear cut and one-to-one.

Figure 20: True CC plotted as function of highest RSM CC

The question is whether RSM CC and skewness can be used not as direct estimators, but as variables in a predictive model that best explain the true population CC. Since population CC is a continuous value strictly limited between 0 and 1, it is not suitably to analyse this in the form of linear OLS regression. Instead, for this section a so called beta regression is used. Beta regression assumes that the dependent variable follows a beta distribution of some parameter and fits the independent variable so to best model the likely beta distribution parameter for each observation (Ferrari & Cribari-Neto, 2004). The independent variables are RSM maximum CC, skewness, interaction between maximum CC and skewness, and dummy variables for levels of $k$. Figure 22 shows model parameters. All variables, including the interaction term, significantly explain variation, and total model fit to the data can be interpreted by the pseudo $R^2$ score 0.7583. Figure 23 shows the true population CC plotted as a function of model prediction value. Under best fitting possible model all simulated datasets should be aligned along the diagonal slope. We see however that the residuals showcase some form of autocorrelating systematicity. In particular it is apparent very few datasets have a model fit of 0.3 or lower, even though much of the data indeed does have 0.3 population CC or smaller. Much of the data with true population CC of 0.6 lower is concentrated in a block, rather than being aligned diagonally. This indicates that the model does a poor job of fitting data for low (below 0.6) population CC. This is consistent with earlier finding regarding skewness and RSM CC, as both only increase as a function of true CC given that true CC exceeds a certain threshold limit.
Figure 21: Bias of highest RSM compared to true CC

```
betareg(formula = True ~ Max * Skewness + as.factor(K), data = totalSimT)

Standardized weighted residuals 2:

    Min 1Q Median 3Q Max
-3.4067 -0.8409  0.0356  0.6298  3.5113

Coefficients (mean model with logit link):

  Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -7.66668   0.18163  -42.21  <2e-16 ***
Max           9.66979   0.23186   41.82  <2e-16 ***
Skewness      7.32583   0.48211   15.20  <2e-16 ***
  as.factor(K)3 -0.18381   0.02170  -8.47  <2e-16 ***
  as.factor(K)4 -0.32638   0.02405 -13.57  <2e-16 ***
  as.factor(K)5 -0.43138   0.02675 -16.13  <2e-16 ***
Max:Skewness  -6.75273   0.52772  -12.80  <2e-16 ***

Phi coefficients (precision model with identity link):

   Estimate Std. Error z value Pr(>|z|)
(phi)     13.7240   0.24444   56.15  <2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  4565 on 8 Df
Pseudo R-squared: 0.7583
Number of iterations: 14 (BFGS) + 2 (Fisher scoring)
```

Figure 22: Beta regression coefficients
Under the assumptions that, at least for the explaining variables in the current model, there is no systematic relationship enough to fit a model for smaller population CC data, it could be necessary to rectify this prior to model creation. If lower population CC data is unmodelable (given circumstances), then there is a risk that the large noise at lower CC levels will influence the outcome of fitted variable coefficients. Figure 24 shows the resulting model created using the same variable as our earlier model, however for this all observations with true population CC below 0.5 is excluded beforehand. All variables are significant still, and the pseudo $R^2$ is 0.8234. When compared to the initial "whole" model fitted to same data subsection (figure 25), it is visible that for the new model the observation are centred more evenly around the regression line.

To investigate whether the above model is robust, new RSM simulations are created for which true population CC is 0.5 or above (the range of model). New data is generated the same way as the previous data for this analysis. Resulting fit prediction by model and error is shown in figure 26. The linear correlation coefficient between prediction and actual data was 0.9214226. 95 percent of all true population CC for the new data was within +- 0.112 of the predicted value from the model. Figure 27 plots the prediction error with a line of smoothed mean. The errors appear more or less even around the prediction line (mean error = 0.00736429), indicating the prediction is not excessively biased. Figure 28 shows the squared error as a function of predicted CC. Some heteroscedasticity appears to be observable, with fluctuation being largest at between 0.7 and 0.8 CC.
Figure 24: Fitted beta regression model on data for only true CC > 0.5

Figure 25: Fitted beta regression model created by unlimited data
Figure 26: Prediction using beta regression model on novel simulation data with 0.95 coverage interval

Figure 27: Prediction error from beta regression model
5 Discussion

5.1 Limitation in generalizability of result

We have seen that RSM can be used to detect presence of true correlation over a threshold CC value, and indeed the same variable used for detection can be used to create somewhat accurate point estimator of true CC. However, in interpreting this result it is necessary to be extremely careful. Recall that our simulation put several artificial limitation on how the generated data could look. Prime example of these are, small number of target variables, lack of multicollinearity between target variable and distraction variable, all normal distributed variables, only one-to-one correlation between inter-set variables. For real life application, each one of these assumptions will pose concern. While it is relatively easy to control assumptions such as multivariate normality of data, it is much harder to estimate any form of shape of correlation matrix a priori. The fact that a model fitted on one dataset is able to predict another dataset, with exact same parameter of generation, is also a more or less trivial result. In the following discussion section I will argue that, while the concerns mentioned above are valid, there is still valuable result to take from our RSM analysis.
5.2 Can RSM CC be used to detect real correlation in real chaotic data?

There are two main ways in which the result obtained for correlation detection would be deemed practically useless. First, that for real "ugly" data we will not see an effect or only see very little effect, of true population CC on highest observed RSM CC. Indeed, for our data there were 1 to 5 target variables who had higher correlation to the dependent variables, but if the target variables were more numerous, and each individual correlation smaller, then the same population CC might yield lower maximal RSM CC. While this is a concern we also found out that RSM analysis generates not one but several statistics. If having more spread out targets would generate smaller peaks, it is reasonable that they would instead influence RSM trials more evenly, and perhaps showing its effect in other statistics like skewness or mean RSM CC.

The second issue is whether the distribution of maximal RSM under zero population CC is really the same for more or less ugly data. If, say the addition of collinearity, changed the shape of the RSM distribution, then the confidence interval would be inapplicable from one data to another. To answer this it is not necessary to conclude that all null distributions be equal. Instead, we only need to know, under which condition of the data is the expected 95th percentile of the maximal CC the highest, given that the true population CC is zero. Considering that collinearity is essentially just data redundancy, it is hard to see how average null distribution would increase by increasing collinearity. It is possible that the distribution of RSM CC changes if the variables in the data follows a different distribution than that of normal distribution (as used here in our analysis), but given that one can control beforehand that the actual data to use is approximately normal this needs not to be a crucial concern.

5.3 What does a significant correlation tell us?

A second concern is, that even if the type 1 and type 2 errors isn't present, the resulting conclusion given by rejecting null hypothesis tell us nothing. Indeed, a statistic being able to distinguish between zero, and non-zero population CC is rather trivial. If one were to assemble 1000 variables and claim that you found that at least some of them have non-zero correlation between each other, then such a result would likely not excite the reader.

To address this, it is necessary to consider what "finding true CC" means to begin with. In our initial theoretical discussion we have already addressed the relationship between population CC, sample CC and RSM CC. When we say we want to find population CC from a multi-dimensional data, what we really are asking is to find the all variables for which correlation to dependent variables is non-zero, and collinearity with other independent variable is not 1. Because if we have found all variables that have true correlation, and is non-redundant, then that subset of variables will capture all linear relationships that exist in relationship to the dependent variables. We can consider such an ideally parsimonious and trimmed model as a goal model.

Now why is this relevant for RSM CC? Simply put, if maximal RSM CC under high population CC significantly deviates from that of maximal RSM CC under
no population CC, then this can only reasonably explained by the fact that within those maximal CC random subspaces, is likely to contain at least one target variable. That is, higher RSM CC is caused by variables that should be included in the goal model. Thus it should theoretically be possible to estimate at least some variables for the complete model through scavenging among the random subspaces which yield significantly high CC. In short, while the presence of non-zero correlation is itself rather trivial result, this result could potentially still be expanded to find which variables one would want to include when searching for complete CC model.

The statistic of skewness proved to be a stable predictor of deviation from null distribution in the case of our data generation. This is possibly explained by the high peaks caused by target variable including RSM trials. However, it is less obvious how the null distribution of skewness is affected by collinearity or change in correlation structure in-between groups, so any result found here regarding skewness is best regarded as anecdotal. With that said the numerous statistics obtainable by RSM, and the fact that you can create CI for all of them given certain assumptions, create valuable insight into the nature of correlation present in the data.

5.4 Effect of choice $k$

For the main part of the analysis the value of $k$ varied between 2 to 5. In most data cases, the power of the test were roughly equal between the sizes of $k$. The size of each subspace for our data were sufficient even with 2. This is perhaps not a huge surprise considering the number of potential target variables in the data was 5 at its highest, and often the lion’s share of the correlation was carried by one to three independent variables. A question that needs answering is what the optimal choice of $k$ is. Is it so that the power of deviation test from null distribution is highest at smaller $k$? Or is perhaps the maximum power obtained at some other data set specific choice of $k$? One hypothesis is that the power of the test increases as size of $k$ approaches value close to the number of variables included in complete model, and from that point on the power diminishes as $k$ increases further. To test this it would be valuable to investigate RSMCCA on noisier, and overall more complex datasets where number of relevant target variables are more numerous than that used in this study.

In either case it is important that for RSM to be useful, that the parameters choices stay have easily explained and systematic relationship to end result. To control for this it is necessary to investigate if RSMCCA indeed can work during varied circumstances and provide similar result. The following section of this text will cover a short attempt at testing RSMCCA for real, noisy, data.

6 RSM CC on real data

6.1 Introduction

To partly answer some of the concerns mentioned in the discussion, the same Monte Carlo method used in the main analysis will here be applied to real biological data.
6.2 Material

The data set for analysis is p53 mutant data, which is a collection of 4826 2D electrostatic and surface features from p53 cell obtained over 16772 observations (Danziger et al., 2006). From 4826 features 1000 are randomly selected to be used for this analysis. These variables are then controlled so there are no missing value, and that they are all linearly independent from each other. Figure 29 shows correlation matrix for the first 200 variables. It is visible that many variables have absolute correlation that is near 1, and as such the true correlation between two groups using all these variables have a high probability of also being near one. To be able to draw informative conclusions from RSMCCA it is necessary that the true correlation for the population varies, and as such not all 1000 variables will be used in one analysis. Instead 5 variables are selected randomly as the dependent variable set, and another 95 variables are picked as independent variable set. This procedure is repeated several thousand time, and for each selection the true CC is recorded by cancort function in R using all observations. This value will serve as true population CC estimate. 63 variable combinations are chosen from the sampled variable combination, so that there is approximately even spread of true CC ranging from 0.55 to 0.99. The variables included show clear divergence from multivariate normal distribution (figure 30).

![Correlation matrix for first 200 p53 variables.](image)

Figure 29: Correlation matrix for first 200 p53 variables.

50 random observations are sampled from each variable combination picked above, and after they have been combined with 900 distraction variables these are each analyzed with RSM (n = 50, p = 1000, t = 20000, k ranging from 2 to 5). Summarizing statistics are collected after all trials are completed. As all RSMCCA
parameters are equal for this analysis as the main study the same critical limits are used for the significance tests.

6.3 Result

Figure 31 shows maximal RSM CC as plotted after true CC for the variable combinations. While certain spread can be observed in value, all variables for this analysis returned maximal CC above 0.85, with the majority of which being above 0.95. Comparing the values to corresponding critical limit for k, all variable combinations were significantly different from zero correlation (alpha = 0.05). As true canonical correlation was detected for all the RSMs no logistic regression model can be formed from this data.

To test whether this high CC generated from p53 data is actually responding to correlation in the sample, or if it is mere some form of overfitting result, control RSM CCA is conducted. For each of the same combinations of variables used in independent set above, 5 random normal distributed data is generated as dependent variable set. As the true population correlation between sets equals zero, this data should follow the same distribution of RSM as that under our initial Monte Carlo simulation. The result of RSM trials generated from such zero correlating data with p53 dependent variable is shown in 32, where the Y axis represent the value of highest RSM percentile as compared to original Monte Carlo distribution. If this RSM is identically distributed as the previous data simulation under zero correlation then all the RSM percentiles should have a uniform distribution from 0 to 1, with a mean of 0.5, which reasonably appears to be the case.
Figure 31: Highest RSM CC coefficient for p53 data after true CC

Figure 32: p53 RSM percentile compared to Monte Carlo distribution under zero correlation
6.4 Discussion

For our p53 data, which was highly unnormal and showed great degree of multicolinearity, RSM CC was successfully able to detect true correlation. However the extremely high CC coefficients, which appears to not be hugely influenced by either true CC level or value of k, is problematic. Given that we want RSM CC to be useful in predicting the real nature of correlation present, having statistic that goes close to 1, regardless of the true value of the data, gives only so much information. What’s more is that this result suggests that RSMCC fluctuates greatly as shape of data for analysis changes.

Switching dependent variable set to normal distributed simulated variables the coefficient stabilized away from 1. This indicates it at least is not an issue related solely the shape of one set of variables. One potential explanation is that this is not necessarily an issue related to RSM, but rather canonical correlation in general, only enchanced by the nature of high dimensional data. Canonical correlation is sensitive to data diverging from normality, in particular by outliers (Muirhead & Watermanux, 1979). This is the same issue that plagues the metric of linear correlation in general. Having 100 variables with irregular distribution and potential outliers increases the likeliness that one of them will "override" any other true correlation that would ideally have served as better estimator of true population CC. This is even more of the case for data with low samples.

While there is no argument for that above explanation is true, at the very least one can form a conclusion that the need to analyze the data prior to performing analysis grows even stronger when one goes from regular canonical correlation analysis to RSMCCA. Certainly we can conclude that RSMCCA is not necessarily easily interpretable, but it could under right circumstances provide useful explorative insight to the large dataset that the researcher has at hand. Indeed even if one were to not assign greater importance to the summarizing statistic as any test variable, you can still consider RSM as simply an express process of cutting-up and detail examining subsets from a large dimensional data, and in that regard the method is far from inane.

7 References


Data set retrieved from https://archive.ics.uci.edu/ml/datasets/p53+Mutants


