



Standardization and Singular Value  
Decomposition in Canonical Correlation  
Analysis

**Melinda Borello**

Johanna Hardin, Advisor  
David Bachman, Reader

Submitted to Pitzer College in Partial Fulfillment  
of the Degree of Bachelor of Arts

April 24, 2013

Department of Mathematics



# Abstract

Canonical correlation analysis (CCA) is a type of multivariate analysis based on the correlation between linear combinations of variables in two data sets. Biological applications of CCA include large scale genomic studies that can have multiple phenotypic or genotypic data. In these cases, CCA can lead to results that lack interpretability since CCA considers all variables. These types of analyses usually have an enormous number of variables, where the number of genes exceeds tens of thousands. Sparse canonical correlation analysis (SCCA) aims to solve the problem of interpretability by providing sparse solutions. In this paper, I examine the relationship between running CCA and SCCA with raw unstandardized data and running both methods using data that have been standardized to have a mean of zero and a standard deviation of one. I also show how both CCA and SCCA relate to singular value decomposition (SVD) by looking at an algorithm for SVD in the context of CCA and SCCA.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>1 Canonical Correlation Analysis</b>	<b>1</b>
<b>2 Sparse Canonical Correlation Analysis</b>	<b>9</b>
<b>3 Singular Value Decomposition Algorithm</b>	<b>15</b>
<b>Bibliography</b>	<b>21</b>



# Acknowledgments

I'd like to thank Johanna Hardin for offering to be my thesis adviser at the last minute when I thought it was too late for me to write a thesis. Her guidance throughout the research and writing process has been invaluable. I would also like to thank Associate Professor of Mathematics, Stephen Garcia, for his help with SVD and other linear algebra questions that we came upon throughout the semester. Without some of his explanations, Chapter 3 wouldn't exist.

Of course, I would like to thank my mother, for all of my accomplishments are really hers. Thanks Mom, for everything.





## Chapter 1

# Canonical Correlation Analysis

Canonical correlation analysis (CCA) measures the relationship between two sets of variables. CCA accomplishes this by focusing on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. Canonical correlation can also be thought of as an extension of bivariate correlation allowing more than two continuous variables in each set. CCA seeks to answer how the best linear combination of one set relates to the best linear combination of the other set of variables.

The following presentation of canonical correlation analysis follows that which is offered by Johnson and Wichern (1992). Consider a group of  $p$  variables represented by the  $(p \times 1)$  random vector  $\mathbf{X}$ , and a second group of  $q$  variables represented by the  $(q \times 1)$  random vector  $\mathbf{Y}$ . Assume that  $p \leq q$ . The random vectors have  $\text{Cov}(\mathbf{X}) = \Sigma_{11}$ ,  $\text{Cov}(\mathbf{Y}) = \Sigma_{22}$ , and  $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \Sigma_{12}$ , with  $E(\mathbf{X}) = \mu_{\mathbf{x}}$  and  $E(\mathbf{Y}) = \mu_{\mathbf{y}}$ . For coefficient vectors

## 2 Canonical Correlation Analysis

---

$\mathbf{a}$  and  $\mathbf{b}$  we can form the linear combinations,  $G = \mathbf{a}'\mathbf{X}$  and  $H = \mathbf{b}'\mathbf{Y}$ . Then

$$\max_{a,b} \text{Corr}(G, H)$$

is attained by the linear combinations (first canonical variate pair)

$$G_1 = \mathbf{a}'_1\mathbf{X} \quad \text{and} \quad H_1 = \mathbf{b}'_1\mathbf{Y}.$$

The  $k$ th pair of canonical variates,  $k = 2, 3, \dots, p$ ,

$$G_k = \mathbf{a}'_k\mathbf{X} \quad \text{and} \quad H_k = \mathbf{b}'_k\mathbf{Y}$$

where

$$\mathbf{a}_k = \Sigma_{11}^{-1/2}\mathbf{u}_k \quad \text{and} \quad \mathbf{b}_k = \Sigma_{22}^{-1/2}\mathbf{v}_k$$

for  $k = 1, \dots, p$ , maximizes

$$\text{Corr}(G, H)$$

among those linear combinations uncorrelated with the preceding  $1, 2, \dots, k - 1$  canonical variables. That is, the second canonical variate pair are the linear combinations  $G_2$  and  $H_2$  which maximize all linear combinations which are uncorrelated with the first canonical variate pair. Then we have

$$\max_{a,b} \text{Corr}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}) = \text{Corr}(\mathbf{a}'_k\mathbf{X}, \mathbf{b}'_k\mathbf{Y}) = \text{Corr}(G_k, H_k) = \rho_k^*.$$

Here the  $(p \times 1)$  vectors,  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ , and the  $(q \times 1)$  vectors,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ , are the left and right singular vectors, respectively, of matrix  $K = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ .

The singular values  $\rho_1^* \geq \rho_2^* \geq \dots \geq \rho_p^*$ , of matrix  $K$  are the canonical correlations.

In some instances, one may want to work with standardized variables, allowing ease of comparison of the variables to each other. There may be other motivations for standardization, such as standardization can simplify computations. Consider if we standardized the original variables as follows:

$$\mathbf{Z}_X = V_{11}^{-1/2}(\mathbf{X} - \mu_x) \quad \text{and} \quad \mathbf{Z}_Y = V_{22}^{-1/2}(\mathbf{Y} - \mu_y)$$

where  $V_{11}^{-1/2}$  is the  $(p \times p)$  diagonal matrix with one over the standard deviation on its diagonal, i.e.,

$$V_{11}^{-1/2} = \begin{bmatrix} 1/\sqrt{\sigma_{x11}} & 0 & \dots & 0 \\ 0 & 1/\sqrt{\sigma_{x22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sqrt{\sigma_{xpp}} \end{bmatrix}.$$

The matrix  $V_{22}^{-1/2}$  is similarly defined as a  $(q \times q)$  matrix with one over the standard deviation of  $\mathbf{Y}$ ,  $\frac{1}{\sqrt{\sigma_{y_{ii}}}}$ , on its diagonal.

Here,  $\rho_{11} = \text{Cov}(\mathbf{Z}_X)$ ,  $\rho_{22} = \text{Cov}(\mathbf{Z}_Y)$  and  $\rho_{12} = \text{Cov}(\mathbf{Z}_X, \mathbf{Z}_Y)$ , with  $E(\mathbf{Z}_X) = E(\mathbf{Z}_Y) = 0$ . Let the  $(p \times 1)$  vectors,  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$  and the  $(q \times 1)$  vectors,  $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p$  be the left and right singular vectors of  $L$ , respectively, where  $L = \rho_{11}^{-1/2} \rho_{12} \rho_{22}^{-1/2}$ . Coefficient vectors,  $\alpha_k$  and  $\beta_k$ , form the  $k$ th pair of canonical variates

$$M_k = \alpha_k' \mathbf{Z}_X, \quad N_k = \beta_k' \mathbf{Z}_Y.$$

## 4 Canonical Correlation Analysis

---

Then we have

$$\max_{\alpha, \beta} \text{Corr}(\alpha' \mathbf{Z}_X, \beta' \mathbf{Z}_Y) = \text{Corr}(\alpha'_k \mathbf{Z}_X, \beta'_k \mathbf{Z}_Y) = \text{Corr}(M_k, N_k) = \rho_k^*.$$

Here  $\rho_1^* \geq \rho_2^* \geq \dots \geq \rho_p^*$  are the singular values of matrix  $L$  and  $\alpha'_k = \mathbf{e}'_k \boldsymbol{\rho}_{11}^{-1/2}$  and  $\beta'_k = \mathbf{f}'_k \boldsymbol{\rho}_{22}^{-1/2}$ .

**Theorem 1.1** *Given coefficient vectors  $\boldsymbol{\alpha} = \boldsymbol{\rho}_{11}^{-1/2} \mathbf{e}$  and  $\boldsymbol{\beta} = \boldsymbol{\rho}_{22}^{-1/2} \mathbf{f}$  from scaled data,  $\mathbf{Z}_X$  and  $\mathbf{Z}_Y$ , and coefficient vectors  $\mathbf{a} = \boldsymbol{\Sigma}_{11}^{-1/2} \mathbf{u}$  and  $\mathbf{b} = \boldsymbol{\Sigma}_{22}^{-1/2} \mathbf{v}$  from raw data,  $\mathbf{X}$  and  $\mathbf{Y}$ , if  $V_{11}^{1/2}$  and  $V_{22}^{1/2}$  are diagonal matrices with  $i^{\text{th}}$  diagonal element  $\sqrt{\sigma_{x_{ii}}}$  and  $\sqrt{\sigma_{y_{ii}}}$ , respectively, then  $\alpha_k = V_{11}^{1/2} \mathbf{a}_k$  and  $\beta_k = V_{22}^{1/2} \mathbf{b}_k$ , i.e., coefficients from scaled data are equal to scaled coefficients from raw data in classical canonical correlation analysis.*

**Proof:** We will show that  $\mathbf{e}'_k \boldsymbol{\rho}_{11}^{-1/2} = \mathbf{u}'_k \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2}$  or equivalently, that  $\boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2} = \boldsymbol{\rho}_{11}^{-1/2}$  and that  $\mathbf{e}_k = \mathbf{u}_k$ . By showing these two equalities hold, we will prove the theorem since without loss of generality,

$$\begin{aligned} \alpha_k &= \mathbf{e}'_k \boldsymbol{\rho}_{11}^{-1/2} = \mathbf{u}'_k \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2} \\ &= \mathbf{a}'_k V_{11}^{1/2} \\ \alpha'_k &= (V_{11}^{1/2} \mathbf{a}_k)' \\ \alpha_k &= V_{11}^{1/2} \mathbf{a}_k. \end{aligned}$$

Consider  $\boldsymbol{\rho}_{11}$ :

$$\begin{aligned}\boldsymbol{\rho}_{11} &= \text{Cov}(\mathbf{Z}_X) \\ &= \text{Cov}(V_{11}^{-1/2}(\mathbf{X} - \mu_x)) \\ &= V_{11}^{-1/2} \text{Cov}(\mathbf{X} - \mu_x) V_{11}^{-1/2'}\end{aligned}\tag{1.1}$$

$$\begin{aligned}&= V_{11}^{-1/2} \text{Cov}(\mathbf{X}) V_{11}^{-1/2} \\ &= V_{11}^{-1/2} \boldsymbol{\Sigma}_{11} V_{11}^{-1/2} \\ &= V_{11}^{-1/2} \boldsymbol{\Sigma}_{11}^{1/2} \boldsymbol{\Sigma}_{11}^{1/2} V_{11}^{-1/2}\end{aligned}\tag{1.2}$$

Since  $\boldsymbol{\rho}_{11}$  is positive semi-definite, taking the square root of  $\boldsymbol{\rho}_{11}$  results in

$$\begin{aligned}\boldsymbol{\rho}_{11}^{1/2} &= V_{11}^{-1/2} \boldsymbol{\Sigma}_{11}^{1/2} \\ \boldsymbol{\rho}_{11}^{-1/2} &= \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2}\end{aligned}$$

Line (1.2) follows from (1.1) because  $V_{11}^{-1/2}$  is diagonal and thus symmetric, so it is equal to its transpose. A similar argument for  $\boldsymbol{\rho}_{22}$  also shows that  $\boldsymbol{\rho}_{22}^{-1/2} = \boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2}$ .

Next, does  $\mathbf{u}_k = \mathbf{e}_k$ ? In other words, does the left singular vector of  $K$  equal the left singular vector of  $L$ ? We just proved that  $\boldsymbol{\rho}_{11}^{-1/2} = \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2}$  and  $\boldsymbol{\rho}_{22}^{-1/2} = \boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2}$ . Recall that  $\boldsymbol{\rho}_{12} = \text{Cov}(\mathbf{Z}_X, \mathbf{Z}_Y)$ . Thus we can

## 6 Canonical Correlation Analysis

---

write  $L$  as:

$$\begin{aligned}
 L &= \boldsymbol{\rho}_{11}^{-1/2} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1/2} \\
 &= \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2} \text{Cov}(\mathbf{Z}_X, \mathbf{Z}_Y) \boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2} \\
 &= \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2} \text{Cov}(V_{11}^{-1/2}(\mathbf{X} - \mu_x), V_{22}^{-1/2}(\mathbf{Y} - \mu_y)) \boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2} \\
 &= \boldsymbol{\Sigma}_{11}^{-1/2} V_{11}^{1/2} V_{11}^{-1/2} \text{Cov}(\mathbf{X} - \mu_x, \mathbf{Y} - \mu_y) V_{22}^{-1/2} \boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2} \quad (1.3)
 \end{aligned}$$

$$= \boldsymbol{\Sigma}_{11}^{-1/2} \text{Cov}(\mathbf{X}, \mathbf{Y}) V_{22}^{-1/2} (\boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2})' \quad (1.4)$$

$$= \boldsymbol{\Sigma}_{11}^{-1/2} \text{Cov}(\mathbf{X}, \mathbf{Y}) V_{22}^{-1/2} V_{22}^{1/2(')} \boldsymbol{\Sigma}_{22}^{-1/2(')} \quad (1.5)$$

$$= \boldsymbol{\Sigma}_{11}^{-1/2} \text{Cov}(\mathbf{X}, \mathbf{Y}) V_{22}^{-1/2} V_{22}^{1/2} \boldsymbol{\Sigma}_{22}^{-1/2}$$

$$= \boldsymbol{\Sigma}_{11}^{-1/2} \text{Cov}(\mathbf{X}, \mathbf{Y}) \boldsymbol{\Sigma}_{22}^{-1/2}$$

$$= \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2}$$

$$= K$$

Moving from line (1.3) to (1.4), note that as proven above  $\boldsymbol{\Sigma}_{22}^{-1/2} V_{22}^{1/2} = \boldsymbol{\rho}_{22}^{-1/2}$  which is positive definite and symmetric. Line (1.5) follows from (1.4) since  $\boldsymbol{\Sigma}_{22}^{-1/2}$  and  $V_{22}^{1/2}$  are both symmetric. Since  $L = K$  it is clear that  $L$  and  $K$  have the same singular vectors since they have the same singular value decomposition. Therefore,  $V_{11}^{1/2} \mathbf{a}_k$  in fact is the coefficient vector  $\boldsymbol{\alpha}_k$  for the  $k^{\text{th}}$  canonical variate  $M_k$  constructed from the standardized variable  $\mathbf{Z}_X$ , and  $V_{22}^{1/2} \mathbf{b}_k$  is the coefficient vector  $\boldsymbol{\beta}_k$  for the  $k^{\text{th}}$  canonical variate  $N_k$  constructed from the standardized variable  $\mathbf{Z}_Y$ .

**Claim 1.2** *If (i)  $\mathbf{a}_k$  and  $\mathbf{b}_k$  maximize  $\text{Corr}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y})$  over all  $\mathbf{a}$  and  $\mathbf{b}$  subject to  $\mathbf{a}$  and  $\mathbf{b}$  being uncorrelated to  $\mathbf{a}_i$  and  $\mathbf{b}_i$  where  $i = 1, 2, \dots, k-1$  (ii)  $\text{Corr}(G_k, H_k) =$*

$\rho_k^*$ , then, (i)  $\max_{\alpha, \beta} \text{Corr}(\alpha' \mathbf{Z}_X, \beta' \mathbf{Z}_Y) = \text{Corr}(\alpha'_k \mathbf{Z}_X, \beta'_k \mathbf{Z}_Y)$  subject to  $\alpha$  and  $\beta$  being uncorrelated to  $\alpha_i$  and  $\beta_i$  where  $i = 1, 2, \dots, k-1$  (ii)  $\text{Corr}(M_k, N_k) = \rho_k^*$  as well, i.e., the canonical correlations are unchanged by the standardization.

We will now show that raw and standardized canonical variates produce the same maximum correlation.

From the raw data,  $G_k$  and  $H_k$  are the  $k^{\text{th}}$  pair of canonical variates that maximize  $\text{Corr}(G, H)$ . If this correlation is equal to  $\rho_k^*$  then,

$$\begin{aligned}
 \rho_k^* &= \text{Corr}(G_k, H_k) \\
 &= \text{Corr}(\mathbf{a}'_k \mathbf{X}, \mathbf{b}'_k \mathbf{Y}) \\
 &= \text{Corr}(\mathbf{a}'_k (\mathbf{X} - \mu_x), \mathbf{b}'_k (\mathbf{Y} - \mu_y)) \\
 &= \text{Corr}(\mathbf{a}'_k V_{11}^{1/2} V_{11}^{-1/2} (\mathbf{X} - \mu_x), \mathbf{b}'_k V_{22}^{1/2} V_{22}^{-1/2} (\mathbf{Y} - \mu_y)) \\
 &= \text{Corr}(\mathbf{a}'_k V_{11}^{1/2} \mathbf{Z}_X, \mathbf{b}'_k V_{22}^{1/2} \mathbf{Z}_Y) \\
 &= \text{Corr}(\alpha'_k \mathbf{Z}_X, \beta'_k \mathbf{Z}_Y) \\
 &= \text{Corr}(M_k, N_k)
 \end{aligned}$$

Thus canonical correlations are unchanged by the standardization.

We know that  $G_k$  and  $H_k$  maximize the correlation for the raw data, but do  $M_k$  and  $N_k$  maximize the correlation for the standardized data? Suppose not. Suppose that there was a larger correlation,  $\rho_k^{**}$ , between the two standardized canonical variates,  $M_k$  and  $N_k$ . Then  $\rho_k^{**}$  must also be the maximum correlation for the raw canonical variates,  $G_k$  and  $H_k$ , since as shown above, the correlation for  $M_k$  and  $N_k$  is equal to the correlation for  $G_k$  and  $H_k$ .

Consider matrices  $K$  and  $L$ . As shown above,  $K = L$ , and thus they have the same singular vectors. It follows that they would also have the same singular values,  $\rho_k^*$ . Furthermore, these singular values are the correlation for the  $k$ th canonical variate pair (detailed in Chapter 3). Hence, the pairs  $G_k$  and  $H_k$ , and  $M_k$  and  $N_k$  have the same (maximum) correlation value.

Ultimately, CCA provides the same results whether using scaled data or using raw data and scaling coefficients by  $V_{11}^{-1/2}$  and  $V_{22}^{-1/2}$  (respectively) at the end. Both raw and standardized variables maximize the correlation of linear combinations of the variables from both data sets.



## Chapter 2

# Sparse Canonical Correlation Analysis

As presented above, canonical correlation analysis uses all variables from both sets,  $\mathbf{X}$  and  $\mathbf{Y}$ , to create canonical vectors. Yet when trying to apply CCA to real data, CCA often fails to produce interpretable results. Data used in microarray data analysis and genome-wide linkage analysis usually have an enormous number of variables, where the number of genes exceeds tens of thousands. Thus results from CCA lack biological interpretability.

One way to combat this issue is by using sparse canonical correlation analysis (SCCA). SCCA helps solve problems of interpretability by providing sparse sets of the associated variables, i.e., canonical vectors contain sparse loadings. Following a method to select appropriate sparseness parameters, the sparse solution contains variables that are deemed more im-

portant than others. Thus the solution reduces the dimensionality which improves interpretability.

The iterative algorithm below is that presented by Parkhomenko et al. (2009). This algorithm uses soft-thresholding as its penalty function. Note that there are a number of different penalty functions that one can use for SCCA which have been outlined and compared by Chalise and Friedley (2012). Following classical CCA, consider two sets of variables  $\mathbf{X}$  and  $\mathbf{Y}$ , with  $p$  variables in  $\mathbf{X}$  and  $q$  variables in  $\mathbf{Y}$ . As before, let  $K = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$ , where  $\text{Cov}(\mathbf{X}) = \Sigma_{11}$ ,  $\text{Cov}(\mathbf{Y}) = \Sigma_{22}$ , and  $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \Sigma_{12}$ . The first sparse canonical vectors are identified using the following algorithm:

1. Select sparseness parameters,  $\lambda_u$  and  $\lambda_v$
2. Select initial values  $\mathbf{u}^0$  and  $\mathbf{v}^0$  and set  $i = 0$
3. Update  $\mathbf{u}$ :

(a)  $\mathbf{u}^{i+1} \leftarrow K\mathbf{v}^i$

(b) Normalize:  $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

- (c) Apply soft-thresholding to obtain sparse solution:

$$\mathbf{u}_j^{i+1} \leftarrow (|\mathbf{u}_j^{i+1}| - \frac{1}{2}\lambda_u)_+ \text{Sign}(\mathbf{u}_j^{i+1}) \text{ for } j = 1, \dots, p$$

(d) Normalize:  $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

4. Update  $\mathbf{v}$ :

(a)  $\mathbf{v}^{i+1} \leftarrow K'\mathbf{u}^i$

(b) Normalize:  $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

(c) Apply soft-thresholding to obtain sparse solution:

$$\mathbf{v}_j^{i+1} \leftarrow (|\mathbf{v}_j^{i+1}| - \frac{1}{2}\lambda_v)_+ \text{Sign}(\mathbf{v}_j^{i+1}) \text{ for } j = 1, \dots, q$$

(d) Normalize:  $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

5.  $i \leftarrow i + 1$

6. Repeat steps 3-5 until convergence

where  $(x)_+$  is equal to  $x$  if  $x \geq 0$  and 0 if  $x < 0$ , and

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Parkhomenko et al. (2009) replace  $\Sigma_{11}$  and  $\Sigma_{22}$  by  $\text{diag}(\Sigma_{11})$  and  $\text{diag}(\Sigma_{22})$  thus  $K$  becomes an approximation of the sample correlation matrix of  $\mathbf{X}$  and  $\mathbf{Y}$  without any of the information of how the variables in  $\mathbf{X}$  are correlated or how the variables in  $\mathbf{Y}$  are correlated. Using  $\text{diag}(\Sigma_{11})$  avoids computational problems. The computation of  $K$  requires  $(\mathbf{X}'\mathbf{X})^{-1}$  and  $(\mathbf{Y}'\mathbf{Y})^{-1}$  which may not exist in cases where the number of variables is greater than the number of observations (when  $p$  or  $q$  are larger than  $n$ ). This is common in the biological applications discussed above, where one can have tens of thousands of genes, but only a few hundred observations. The authors also select initial values so that  $\mathbf{u}^0$  is the row means of  $K$  and  $\mathbf{v}^0$  is the column means of  $K$ , both standardized to have unit length.

Although the authors assume that  $\mathbf{X}$  and  $\mathbf{Y}$  have been standardized to  $\mathbf{Z}_\mathbf{X}$  and  $\mathbf{Z}_\mathbf{Y}$ , SCCA applied to raw data produces the same results as ap-

plied to standardized data. As in classical CCA, using scaled and centered data or using raw data does not affect  $K$  nor does it affect the left and right singular vectors of  $K$ . Since the initial values,  $\mathbf{u}^0$  and  $\mathbf{v}^0$ , come from  $K$  and recall that  $K = L$ , there is no difference in the algorithm for raw variables or standardized variables.

In this context, we are interested in the effect of scaling and centering the variables on the soft-thresholding step. The sparseness parameters  $\lambda_u$  and  $\lambda_v$  are chosen using  $k$ -fold cross-validation as outlined below:

1. Choose  $\lambda_u$  and  $\lambda_v$
2. Remove  $\frac{1}{k}$  of the data (testing sample)
3. Find canonical coefficients from  $\frac{k-1}{k}$  of the data (training sample)
4. Find canonical correlation using testing sample and coefficients from training sample
5. Repeat steps 2-4  $k$  times. Average across the  $k$  correlations, keeping this value
6. Repeat steps 1-5 for new  $\lambda_u$  and  $\lambda_v$

From this process, we find the optimal combination of  $\lambda_u$  and  $\lambda_v$  out of all specific pairs of sparseness parameters that correspond to the highest average test sample correlation. This process is not affected by scaling variables as it looks at the correlation between canonical vectors in testing and training samples. We have already proven that the correlation will be the same between raw canonical vectors and standardized canonical vectors.

---

Thus, the choices for sparseness parameters are not affected by the scaling of variables. (For more on the selection of sparseness parameters consult Parkhomenko et al. (2009)).

The relationship between the coefficients from raw data and coefficients from standardized data is the same in the sparse setting as it was in CCA. The difference is that  $\mathbf{u}$  and  $\mathbf{v}$  will have sparse loadings, so our coefficient vectors will also be sparse. As before, if  $\alpha_k$  and  $\beta_k$  are coefficients for standardized data  $\mathbf{Z}_X$  and  $\mathbf{Z}_Y$ , and  $\mathbf{a}_k$  and  $\mathbf{b}_k$  are coefficients for raw data,  $\mathbf{X}$  and  $\mathbf{Y}$ , then  $\alpha_k = V_{11}^{1/2} \mathbf{a}_k$  and  $\beta_k = V_{22}^{1/2} \mathbf{b}_k$ , where  $\text{Corr}(\alpha_k' \mathbf{Z}_X, \beta_k' \mathbf{Z}_Y) = \text{Corr}(\mathbf{a}_k' \mathbf{X}, \mathbf{b}_k' \mathbf{Y})$ . That is, Theorem 1.1 holds for SCCA.



## Chapter 3

# Singular Value Decomposition Algorithm

One may notice that the sparse algorithm is nothing more than the standard algorithm for singular value decomposition (SVD) with an added soft-thresholding step to obtain a sparse vector. If we consider the standard algorithm, we have:

1. Select initial values  $\mathbf{u}^0$  and  $\mathbf{v}^0$  and set  $i = 0$

2. Update  $\mathbf{u}$ :

(a)  $\mathbf{u}^{i+1} \leftarrow K\mathbf{v}^i$

(b) Normalize:  $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

3. Update  $\mathbf{v}$ :

(a)  $\mathbf{v}^{i+1} \leftarrow K' \mathbf{u}^i$

## 16 Singular Value Decomposition Algorithm

---

(b) Normalize:  $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

4.  $i \leftarrow i + 1$

5. Repeat steps 2-4 until convergence.

Why does this provide the singular vectors of  $K$ ?

Recall that the SVD of  $K$  is  $K = UDV'$  where we choose  $U$  and  $V$  to be orthogonal matrices, i.e.,  $U' = U^{-1}$  and  $V' = V^{-1}$ . Thus, the column vectors of  $U$  and  $V$ ,  $\mathbf{u}_i$  and  $\mathbf{v}_i$ , are unit vectors and are the left and right singular vectors of  $K$ . Matrix  $D$  is a diagonal matrix with the singular values of  $K$  on its diagonal, i.e., the square roots of the eigenvalues of  $K'K$  or  $KK'$ . Alternating multiplying  $\mathbf{v}$  by  $K$  and  $\mathbf{u}$  by  $K'$  produces the singular values of  $K$  if the following is noted:

$$K\mathbf{v}_i = UDV'\mathbf{v}_i = d_i\mathbf{u}_i.$$

Observe that  $V'\mathbf{v}_i$  is a column vector of zeros except for a 1 in the  $i^{\text{th}}$  position, since this is equivalent to taking the inner product of  $\mathbf{v}_i$  with the columns of  $V$ . This then results in  $DV'\mathbf{v}_i$  which is a diagonal matrix with all zeros except for  $d_i$  on the diagonal in the  $i^{\text{th}}$  column. Thus  $UDV'\mathbf{v}_i$  is the  $i^{\text{th}}$  column of  $U$  multiplied by the singular value,  $d_i$ . Next we normalize to get a unit vector, giving us,

$$\frac{d_i\mathbf{u}_i}{\|d_i\mathbf{u}_i\|} = \frac{d_i\mathbf{u}_i}{|d_i|\|\mathbf{u}_i\|} = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} = \mathbf{u}_i$$



Similarly,

$$K' \mathbf{u}_i = VDU' \mathbf{u}_i = d_i \mathbf{v}_i$$

which we normalize to get  $\mathbf{v}_i$ .

The following process will produce the largest eigenvalue of  $D$  which is the largest singular value of  $K$ . In the context of CCA and SCCA, this is the largest correlation,  $\rho_1^*$ , between the canonical variates. By switching off multiplying  $\mathbf{u}$  and  $\mathbf{v}$  by  $K'$  and  $K$  respectively, each starting vector  $\mathbf{u}^0$  and  $\mathbf{v}^0$  will be scaled towards the direction of the largest singular value of  $K$  until  $\mathbf{u}$  and  $\mathbf{v}$  become the singular vectors of  $K$ .

Notice that if we construct a symmetric matrix  $A$ , such that

$$A = \begin{bmatrix} 0 & K' \\ K & 0 \end{bmatrix}$$

the problem is then reduced to an eigenvalue problem. We wish to solve

$$Ax = dx$$

where  $d$  is the dominant singular value of  $K$  (and the square root of the dominant eigenvalue of  $A$ ) and  $x$  is the vector  $(\mathbf{v} \ \mathbf{u})'$ . Note that applying

$A$  to  $x$  leads to

$$A \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{K}'\mathbf{u} \\ \mathbf{K}\mathbf{v} \end{bmatrix} \\ = \begin{bmatrix} d\mathbf{v} \\ d\mathbf{u} \end{bmatrix}$$

if  $\mathbf{u}$  and  $\mathbf{v}$  are the singular vectors of  $K$ . Thus putting  $K$  and  $K'$  into a block matrix results in a symmetric matrix  $A$  with dominant eigenvalue  $d$  and eigenvector  $(\mathbf{v} \ \mathbf{u})'$ . Now instead of finding singular vectors and singular values, we just need to find the eigenvectors and eigenvalues of  $A$ .

The above algorithm for SVD can be used for CCA and SCCA. It produces the components needed for the canonical variates, (singular vectors) and it produces the canonical correlations, (singular values). How does SVD produce canonical correlations? Recall that in CCA we seek coefficient vectors,  $\mathbf{a}$  and  $\mathbf{b}$ , such that

$$\text{Corr}(G, H) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}}$$

is maximized. Notice that we can reduce this expression as follows:

$$\max_{a,b} \text{Corr}(G, H) = \max_{a,b} \frac{\mathbf{a}' \Sigma_{12} \mathbf{b}}{\sqrt{\mathbf{a}' \Sigma_{11} \mathbf{a}} \sqrt{\mathbf{b}' \Sigma_{22} \mathbf{b}}} \quad (3.1)$$

$$= \max_{u,v} \frac{\mathbf{u}' \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2} \mathbf{v}}{\sqrt{\mathbf{u}' \Sigma_{11}^{-1/2} \Sigma_{11} \Sigma_{11}^{-1/2} \mathbf{u}} \sqrt{\mathbf{v}' \Sigma_{22}^{-1/2} \Sigma_{22} \Sigma_{22}^{-1/2} \mathbf{v}}} \quad (3.2)$$

$$= \max_{u,v} \frac{\mathbf{u}' K \mathbf{v}}{\sqrt{\mathbf{u}' I_{p \times p} \mathbf{u}} \sqrt{\mathbf{v}' I_{q \times q} \mathbf{v}}} \\ = \max_{u,v} \frac{\mathbf{u}' K \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (3.3)$$

Line 3.2 follows from 3.1 by a change of variables. Line 3.3 is equivalent to finding  $\mathbf{u}^*$  and  $\mathbf{v}^*$  that are unit vectors and maximizing over those vectors.

Under this condition, 3.3 boils down to

$$\max_{u^*, v^*} \mathbf{u}^{*'} K \mathbf{v}^*.$$

Substituting  $K$  by its singular value decomposition gives

$$\max_{u^*, v^*} \mathbf{u}^{*'} U D V' \mathbf{v}^*. \quad (3.4)$$

When  $\mathbf{u}^*$  and  $\mathbf{v}^*$  are the  $i^{\text{th}}$  singular vectors of  $K$ , from above calculations, 3.4 becomes,

$$\begin{aligned}\max_{\mathbf{u}^*, \mathbf{v}^*} \mathbf{u}^{*'} U D V' \mathbf{v}^* &= \mathbf{u}_1^{*'} U D V' \mathbf{v}_1^* \\ &= \mathbf{u}_1^{*'} d_1 \mathbf{u}_1^* \\ &= d_1 \|\mathbf{u}_1^*\| \\ &= d_1\end{aligned}$$

The first singular vector of  $K$  will produce the largest singular value, which we saw in the above algorithm. The iterative process to find  $\mathbf{u}_1$  and  $\mathbf{v}_1$  will always be pulled towards the direction of the largest singular value of  $K$ . Thus, SVD produces the canonical correlations.

To summarize, SVD is a helpful tool that can be used to accomplish CCA and SCCA. SVD provides an algorithm that helps find the linear combinations of variables in each data set that produces the maximum correlation. In both SCCA and CCA, standardizing the variables so as to have mean zero and standard deviation of one does not change the canonical correlation. The coefficients from the standardized data are equal to the coefficients from the raw data scaled by their standard deviations. Therefore, it is at the discretion of the researcher whether to use raw data or to use standardized data in CCA or SCCA.

# Bibliography

Chalise, P. and Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Comput. Statist. Data Anal.*, 56(2):245–254.

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Prentice Hall Inc., Englewood Cliffs, NJ, third edition.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, 8:Art. 1, 36.