Notes on Sparse Multivariate Methods

Multivariate Data Analysis (MAE0330)

Heitor Baldo

Institute of Mathematics and Statistics
University of São Paulo

December 3, 2021

Summary

The Big-p Problem

The Big-p Problem ($n << p$)

When we have a data set with a very large number of variables (parameters) p in relation to the number of observations (individuals) n, that is, $n << p$, we commonly say that we have a big-p problem (sometimes big-p, little-n).

The techniques of Principal Component Analysis (PCA), Discriminant Analysis (DA) and Canonical Correlation Analysis (CCA) work well in the task of dimensionality reduction for the classical case ($n > p$), however, in the case where $n << p$, these techniques are not convenient.

An alternative to overcome this problem is the use of sparse methods, which are adaptations of these techniques for the case $n << p$ using penalties and regularizations.

The Big-p Problem ($n << p$)

Using penalization and regularization techniques we obtain the sparse versions of PCA, DA and CCA (which we will discuss in the next slides):

- Sparse Principal Component Analysis (*Sparse PCA* or sPCA);
- Sparse Discriminant Analysis (*Sparse DA* or sDA);
- Sparse Canonical Correlation Analysis (*Sparse CCA* or sCCA).

Sparsity: A vector x (or matrix X) is said to be sparse if many of its entries x_i (x_{ij}) are equal to zero.

Sparse Principal Component Analysis

A Review of Principal Components and Principal Coordinates

Let $X = X_{n\times p}$ be a data matrix. We have already seen that we can obtain the k-th *principal component* (PC), denoted by Z_k, by the spectral decomposition of the covariance matrix, i.e. $\Sigma = VDV' \Rightarrow Z_k = XV_k$, where V_k are the column vectors of V (eigenvectors).

Equivalently, Z_k can be obtained through the singular value decomposition (SVD) of X, i.e. $X = U\Lambda^{1/2}V'$:

$$Z_k = U_k\Lambda_{kk}^{1/2}, \tag{1}$$

where U_k are the column vectors of U and $\Lambda_{kk}^{1/2}$ are the singular values. In this case, Z_k are called *principal coordinates* (PCo). We can obtain equation (1) using *multidimensional scaling* from the matrix of Euclidean distances between the observations.

Equivalence between PC and PCo: The PCo analysis of the Euclidean distance matrix ($n \times n$ matrix) is equivalent to the PC analysis of the covariance matrix ($p \times p$ matrix) .

Sparse Principal Component Analysis

When we have a problem $n << p$, the disadvantage of performing "classical" PCA comes from the fact that the PCs are linear combinations of all p input variables, and since the number p is very large, the computational effort required to perform the computations is exaggeratedly large. An alternative to this problem is to use Sparse Principal Component Analysis.

For $n << p$, the interest is to make a selection of the most important variables, for the purpose of reducing the dimensionality (reduce p). Therefore, more than obtaining the reduction vectors through PCs, we want to obtain this reduction by means that penalize those variables that must be eliminated (brought to null), that is, obtain eigenvectors V that assign zero load to some variables. To do this, we use regression algorithms: penalized solutions and regularized solutions.

Sparse Principal Component Analysiss - LASSO

Penalized PC (LASSO)

We want to predict the principal components (Z_k) based on linear combinations of the data matrix X with vectors β (i.e., we want to find β's such that $X\beta \approx Z_k$).

Lagrangian Penalty:

$$\hat{\beta}_{lasso} = \text{arg min}_{\beta}\{||Z_k - X\beta||_2^2 + \lambda||\beta||_1\}, \tag{2}$$

where $||.||_2$ is the Euclidean norm, $||.||_1$ is the ℓ^1 norm and λ is the penalty parameter: if $\lambda \to 0 \Rightarrow$ Least Squares solution, if $\lambda \to \infty \Rightarrow \beta \to 0$. The PCs Z_k of equation (2) are known, obtained by multidimensional scaling in $\mathbb{R}^{n\times n}$ (i.e. Z_k is the k-th principal coordinate).

Limitation: The number of non-zero β is at most n.

Sparse Principal Component Analysis - LASSO

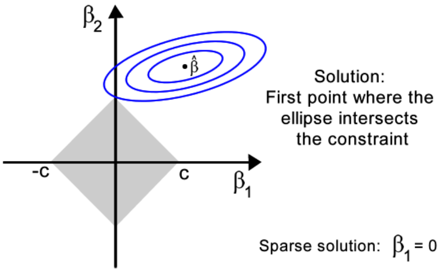
Penalty in the restriction form:

Similarly, we can formalize the model by explaining the restriction in the vector β. For the two-dimensional case, we have:

$$\hat{\beta}_{2\times 1} = \arg\min_{\beta} \sum_{i=1}^{n} (Z_{ik} - X_i'\beta)^2,$$

$$|\beta_1| + |\beta_2| \leq c.$$

Pictorially:



Penalized PC (lasso)

Sparse Principal Component Analysis - Ridge Regression

Regularized PC (Ridge Regression)

Replacing the ℓ^1 norm with the Euclidean norm in the LASSO model, we obtain a *regularized* estimate for β known as Ridge Regression:

Lagrangian regularization:

$$\hat{\beta}_{ridge} = \arg\min_{\beta}\{||Z_k - X\beta||_2^2 + \lambda||\beta||_2^2\}, \tag{3}$$

where λ is the regularization parameter: if $\lambda \to 0 \Rightarrow$ Least Squares solution, if $\lambda \to \infty$ $\Rightarrow \beta \to 0$. The PCs Z_k of equation (3) are known, obtained by multidimensional scaling (in $\mathbb{R}^{n\times n}$).

Sparse Principal Component Analysis - Ridge Regression

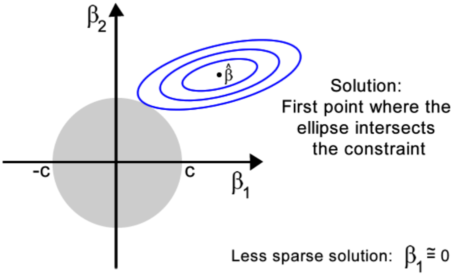
Regularization in the restriction form:

Analogously, we can formalize the model by explaining the restriction in the vector β. For the two-dimensional case, we have:

$$\hat{\beta}_{2\times 1} = \arg\min_\beta \sum_{i=1}^{n} (Z_{ik} - X_i'\beta)^2,$$

$$\beta_1^2 + \beta_2^2 \leq c.$$

Pictorially:



Regularized PC (Ridge Regression)

Sparse Principal Component Analysis - Elastic Net

Penalized and Regularized PC (Elastic Net; Zou et al. [8])

The following model, known as Elastic Net, is a generalization of the LASSO model, and was introduced by Zou and Hastie [8]:

$$\hat{\beta}_{en} = \text{arg min}_{\beta}\{||Z_k - X\beta||_2^2 + \lambda_1||\beta||_2^2 + \lambda_2||\beta||_1\}, \quad (4)$$

where λ_1 is the regularization parameter and λ_2 is the penalty parameter. We can fix λ_1 and λ_2 or obtain them by cross-validation.

Advantage: All variables can be selected (there is no limitation on the number of non-zero charges).

Sparse Principal Component Analysis

From the estimates of the β vectors obtained by one of the previous models ($\hat{\beta}_{lasso}, \hat{\beta}_{ridge}, \hat{\beta}_{en}$), we obtain the approximation for the principal components Z_k:

$$\hat{Z}_k = X\hat{V}_k, \tag{5}$$

where

$$\hat{V}_k = \frac{\hat{\beta}}{||\hat{\beta}||_2}. \tag{6}$$

For more details on sparse PCA, see Zou et al. [9] and Hastie et al. [4].

Sparse Principal Component Analysis - Example (Breast.TCGA)

Implementation in R - ElasticNet

Sparse PCA is implemented in R in the elasticnet [10] package. In the following example, we use data from R's Bioconductor (`Breast.TCGA`).

Data: `Breast.TCGA`: Three databases (X1=mRNA, X2=miRNA and X3=Protein) evaluated on 150 individuals:

X1=mRNA: (`breast.TCGA$data.train$mRNA`) $n = 150$, $p = 200$. ($n << p$)

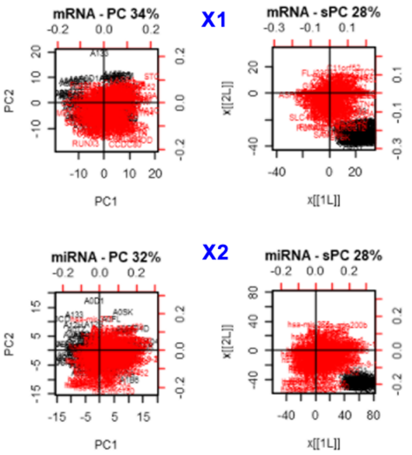
X2=miRNA: (`breast.TCGA$data.train$mRNA`) $n = 150$, $p = 184$. ($n << p$)

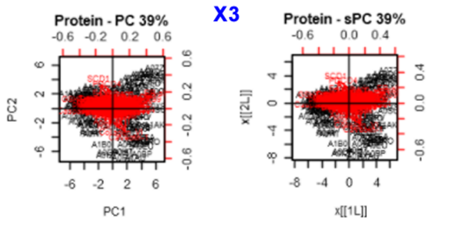
X3=Protein: (`breast.TCGA$data.train$mRNA`) $n = 150$, $p = 142$. ($n > p$)

Subtypes of cancers: (breast.TCGA$data.train$subtype) basal: 45; Her2: 30; LumA: 75.

Sparse Principal Component Analysis - Biplots

Principal Coordinates (PCo) | Sparse PC





The biplots for the X3 (Protein) data are identical, because in this case we have n > p, that is, it is not a big-p problem.

PCo: prcomp() (Stats)

```
x1.pc <- prcomp(x1)

biplot(x1.pc$x[,1:2],
x1.pc$rotation[,1:2],
var.axex=TRUE, main="mRNA -
PC 34%")
```

Sparse PCs: SPCA() (ElasticNet)

```
x1.spca <- spca(x1, K = 2,
type = "predictor",
sparse = "penalty",
para = rep(1e-05, 2))
```

Sparse Discriminant Analysis

Discriminant Analysis

In discriminant analysis (Fisher linear), as we have grouped observations, we consider the decomposition of the covariance matrix into two components: covariance due to the between groups effect and covariance due to the within groups effect, $\Sigma_{p\times p} = \Sigma_{B_{p\times p}} + \Sigma_{W_{p\times p}}$. From this, we are interested in solving the following optimization problem:

$$\max_l \frac{l'\Sigma_B l}{l'\Sigma_W l}. \tag{7}$$

In other words, we want to find vectors l such that maximize the ratio (7). This problem is equivalent to finding the eigenvalues and eigenvectors of $\Sigma_W^{-1}\Sigma_B$, which is equivalent to finding solutions of the determinant equation:

$$|\Sigma_W^{-1}\Sigma_B - \lambda I_p| = 0. \tag{8}$$

We assume homoscedasticity in the groups.

Sparse Discriminant Analysis

However, in the case where $n << p$ (big-p), the inverse of the covariance matrix within groups, Σ_W, does not exist (it is singular), since the rank of this matrix is in maximum n. An alternative to correct the problem of the incomplete rank of Σ_w is to use *Sparse Discriminant Analysis* (sDA). In the following, we present the sDA models proposed by Witten et al. [7] and Clemmensen et al. [1].

Regularization through Ω matrix

We can find a positive-definite diagonal matrix Ω such that

$$|(\Sigma_W + \Omega) - dI_p| = 0; \qquad d > 0. \tag{9}$$

If all the eigenvalues of a matrix are positive, then it is invertible (non-singular). Algorithms for obtaining the matrix Ω are discussed in Hastie et al. [3].

Sparse Discriminant Analysis

Hence our optimization problem, $\max_{\beta_k} \frac{\beta_k' \Sigma_B \beta_k}{\beta_k' \Sigma_W \beta_k}$, becomes:

$$\max_{\beta_k} \frac{\beta_k' \Sigma_B \beta_k}{\beta_k' (\Sigma_W + \Omega) \beta_k}. \tag{10}$$

Equivalently, we can find a positive-definite matrix Ω such that the discriminant vectors of the optimization problem

$$\max_{\beta_k} \{\beta_k' \Sigma_B \beta_k\}, \tag{11}$$

where $\beta_k'(\Sigma_W + \Omega)\beta_k = 1$ and $\beta_k'(\Sigma_W + \Omega)\beta_l = 0$, $\forall l < k$, can be calculated, even when $n << p$.

Sparse Discriminant Analysis

Furthermore, we want the load vectors (discriminant vectors) β_k to be *sparse*. A way to obtain these vectors is by applying the ℓ^1 (LASSO) penalty to the previous optimization problem, resulting in the following problem:

$$\max_{\beta_k}\{\beta_k'\Sigma_B\beta_k - \gamma||\beta_k||_1\}, \tag{12}$$

where $\beta_k'(\Sigma_W + \Omega)\beta_k = 1$ and $\beta_k'(\Sigma_W + \Omega)\beta_l = 0$, $\forall l < k$, can be calculated, even when $n << p$. This method was proposed by Witten and Tibshirani [7].

Sparse Discriminant Analysis

Another sparse discriminant analysis (sDA) method, proposed by Clemmensen et al. [1], is defined sequentially as follows. The k-th pair (θ_k, β_k) is the solution to the problem:

$$\min_{\beta_k, \theta_k} \left\{ ||G\theta_k - X\beta_k||_2^2 + \gamma\beta_k'\Omega\beta_k + \lambda||\beta_k||_1 \right\}, \tag{13}$$

where $\frac{1}{n}\theta_k' G' G\theta_k = 1$ and $\theta_k' G' G\theta_l = 0$, $\forall l < k$, where $\theta_{k_{N\times 1}}$ are the group weight vectors, $G_{n\times N}$ is a group incidence matrix (composed by 0's and 1's) and γ and λ are the non-negative regularization and penalization parameters. The ℓ^1 penalty on β_k results in sparsity when λ is large.

The β_k vector that resolves (13) is called the *k-th discriminant vector* of the sDA.

Sparse Discriminant Analysis

To solve (13), we use a simple iterative algorithm to obtain a local optimum for (13). The algorithm involves keeping θ_k fixed and optimizing with respect to β_k, and keeping β_k fixed and optimizing with respect to θ_k. For fixed θ_k, we obtain:

$$\min_{\beta_k}\left\{||G\theta_k - X\beta_k||_2^2 + \gamma\beta_k'\Omega\beta_k + \lambda||\beta_k||_1\right\}. \qquad (14)$$

Note that for $\Omega = I$, (14) is exactly an ElasticNet problem.

Sparse Discriminant Analysis - Example (Breast.TCGA)

Implementation in R - sparseLDA

sDA is implemented in R in the sparseLDA package [12]. In the following example, we use the same data (`Breast.TCGA`) that was used in the previous example.

Remember that the data from this set is classified into three groups of subtype of cancers:

G1: basal: 45; **G2: Her2:** 30; **G3: LumA:** 75.

In other words, we have $N = 3$ groups, $G = G1 \cup G2 \cup G3$, with a total of $\#(G) = 45 + 30 + 75 = 150$ individuals.

Sparse Discriminant Analysis - Scores e Loads

X1 (mRNA):

Scores (discriminant functions):

$X\hat{\beta}_1$ $X\hat{\beta}_2$

	LD1	LD2
A0FJ	1.822563	-0.90566072
A0G0	1.817337	-0.32230828
A0DA	2.772156	-2.23228938
A0B3	2.229491	-0.74279549
A0I2	3.422815	-2.11782785
A0RT	2.787478	-1.96265989
A131	1.487769	2.00316138
A124	1.494891	-0.88192122
A1B6	2.224953	0.05620507
A1AZ	3.487032	0.09911661
A0YM	3.206956	-1.17263109
A04P	1.871571	-0.46004985
A04T	3.113374	-0.41541073
A0AT	2.106453	0.81464297
...		

Variable loads:

$\hat{\beta}_1$	$\hat{\beta}_2$
0.00000000	-0.6549641
-1.14725442	0.0000000
-1.48943562	0.0000000
0.06294696	0.0000000
0.00000000	-0.7050653
0.00000000	2.0153849
...	

Discriminant variables: sda() (sparseLDA)

```
sda.x1 <- sda(x1t, yt,
lambda = 1e-6, stop = -3,
maxIte = 25, trace = TRUE)
```

Group weight matrix:

$\hat{\theta}_1$	$\hat{\theta}_2$
1.3460511	-0.7163423
0.3229238	2.0027743
-0.9289263	-0.3495289

Incidence matrix

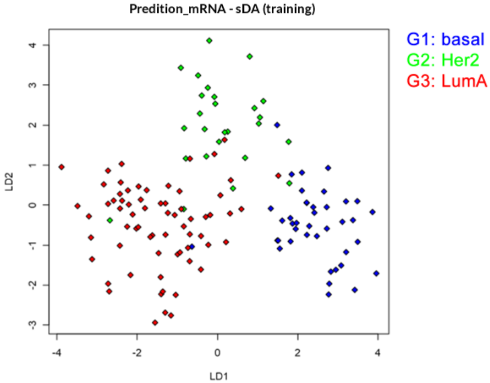
$G_{nxN} = G_{150x3}$

1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
...		

Sparse Discriminant Analysis - Classification

X1 (mRNA):

Representation of predicted groups for training data in sparse discriminant variables (LD1, LD2):



Classification accuracy (training data):

```
class.vector: Basal = 1; Her2 = 2, LumA = 3

class.vector Basal Her2 LumA
           1    38    1    1
           2     1   21    4
           3     1    4   62

yt
Basal  Her2  LumA
   40    26    67

Accuracy:
0.909774436090226
```

Sparse Canonical Correlation Analysis

Classical Canonical Correlation Analysis

Consider the data matrix $X_{n\times(p+q)} = (X_{1_{n\times p}} \quad X_{2_{n\times q}})$. Let $X^1_{p\times 1}$ and $X^2_{q\times 1}$ be the original variables such that:

$$\begin{pmatrix} X^1 \\ X^2 \end{pmatrix} \sim^{iid} \left(\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Canonical correlation analysis aims to solve the following optimization problem: find vectors a, b such that maximize the correlation coefficient $Corr(a'X^1, b'X^2)$, that is,

$$\max_{a,b} \left\{ \frac{Cov(a'X^1, b'X^2)}{\sqrt{Cov(a'X^1)}\sqrt{Cov(b'X^2)}} \right\} = \max_{a,b} \left\{ \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}} \right\}. \qquad (15)$$

Sparse Canonical Correlation Analysis

However, when we have $n << p$ and $n << q$, occurs the impasse that the matrices Σ_{11} and Σ_{22} are singular (non-invertible). Furthermore, classical CCA results in vectors U, V that are not sparse, and these vectors are not unique if p or q exceeds n. An alternative to overcoming this problem is to use *Sparse Canonical Correlation Analysis* (sCCA).

For sCCA, Witten et al. [6] proposed a penalized solution for the singular value decomposition (SVD) of matrices, called Penalized Matrix Decomposition (PMD).

This method does not involve the inverses of the covariance matrices, but the cross-product matrix $X_1'X_2$. Applying PMD to this cross-product matrix, we obtain a penalized method for CCA.

To this aim, we will work with centered and scaled columns X_1 and X_2. Also, we will use sample correlation, which, for centered $x, y \in \mathbb{R}^m$, is given by:

$$cor(x, y) = \frac{x'y}{\sqrt{x'x}\sqrt{y'y}}. \tag{16}$$

Sparse Canonical Correlation Analysis - PMD

Penalized Matrix Decomposition (PMD)

Consider the SVD decomposition, $X = UDV'$, $U'U = I_n$, $V'V = I_p$. Let U_k and V_k be the column vectors of U and V, respectively, and d_k be the diagonal elements of D. In [6], the following generalization of the approximation of X through least squares (first proposed by Eckart et al. [2]) was proposed:

$$\min_{U_k, V_k, d_k}\{||X - d_k U_k V_k'||_2^2\}, \tag{17}$$

with restrictions $||U_k||_2^2 \leq 1$, $||U_k||_1 \leq c_1$; $||V_k||_2^2 \leq 1$, $||V_k||_1 \leq c_2$.

Sparse Canonical Correlation Analysis - PMD

In [6], as a corollary of theorem 2.1, it was verified that the previous problem is equivalent to the following maximization problem:

$$\max_{U_k, V_k}\{U_k' X V_k\}, \tag{18}$$

with restrictions $||U_k||_2^2 \leq 1$, $||U_k||_1 \leq c_1$; $||V_k||_2^2 \leq 1$, $||V_k||_1 \leq c_2$.

One solution is to fix U and get V; fix V and get U:

- Fixed V_k: $\max_{U_k}\{U_k' X V_k\}$; $||U_k||_2^2 \leq 1$, $||U_k||_1 \leq c_1$, $1 \leq c_1 \leq \sqrt{n}$;

- Fixed U_k: $\max_{V_k}\{U_k' X V_k\}$; $||V_k||_2^2 \leq 1$, $||V_k||_1 \leq c_2$, $1 \leq c_2 \leq \sqrt{p}$.

This algorithm is spelled PMD(L_1, L_1).

Sparse Canonical Correlation Analysis - Penalized sCCA via PMD

Sparse canonical correlation analysis uses the PMD(L_1, L_1) algorithm (sCCA Penalized via PMD), considering the SVD decomposition of the matrix $X_1'X_2$ (sample covariance matrix), as follows (for the norm ℓ^1):

$$\max_{a_k,b_k}\{(X_1a_k)'X_2b_k\} = \max_{a_k,b_k}\{a_k'X_1'X_2b_k\}, \tag{19}$$

with restrictions $a_k'X_1'X_1a_k \leq 1$, $||a_k||_1 \leq c_1$ e $b_k'X_2'X_2b_k \leq 1$, $||b_k||_1 \leq c_2$.

Assuming that for high-dimensional data the diagonal covariance matrix can be adopted (CCA-P Diagonal), the previous restrictions become:

$a_k'X_1'X_1a_k = a_k'a_k \leq 1$, pois $X_1'X_1 = I_p$, e $b_k'X_2'X_2b_k = b_k'b_k \leq 1$, pois $X_2'X_2 = I_q$.

Another approach to sCCA can be found in Suo et al. [5].

Sparse Canonical Correlation Analysis - Example (Breast.TCGA)

Implementation in R - PMA

sCCA is implemented in R in the PMA (*Penalized Multivariate Analysis*) package [11]. In the following example, we use the same data (`Breast.TCGA`) that was used in the previous two examples. However, we now want to analyze the pairwise correlation of the three multivariate databases:

Integration X1_X2: $\max_{a,b}\{cor(X_1 a, X_2 b)\}$

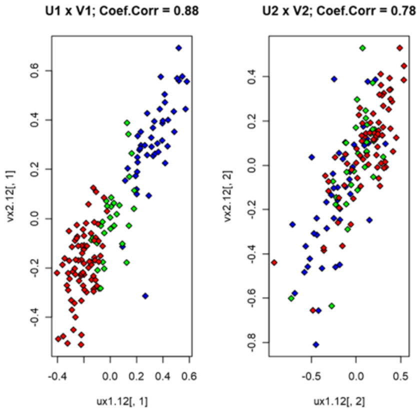
Integration X1_X3: $\max_{a,b}\{cor(X_1 a, X_3 b)\}$

Integration X2_X3: $\max_{a,b}\{cor(X_2 a, X_3 b)\}$

Sparse Canonical Correlation Analysis - sCCA on X1_X2

Integration X1_X2

Observations represented on canonical axes U1 x V1 e U2 x V2:



Sparse canonical vectors:

u and v maximize u'X1'X2v

v = (v1, v2):

0.0000000	0
0.0000000	0
0.0000000	0
0.0000000	0
0.0000000	0
0.1562996	0

...

u = (u1, u2):

0	0
0	0
0	0
0	0
0	0
0	0

...

sCCA via PMD: CCA() (PMA)

```
scca.12 <- CCA(x1,x2,typex=
"standard",typez="standard",
K=2)
```

Canonical variables: U = (U1 U2) = (X1*u1 X1*u2)

V = (V1 V2) = (X2*v1 X2*v2)

Canonical correlation coefficients:

Cor(X1*u1, X2*v1),	Cor(X1*u2, X2*v2):
0.88443973794229	0.779709063287576

References I

[1] Clemmensen, L., Hastie, T., Witten, D., & Ersboll, B. (2011). "Sparse Discriminant Analysis." `https://web.stanford.edu/~hastie/Papers/sda_resubm_daniela-final.pdf`

[2] ECKART, C. AND YOUNG, G. (1936). "The approximation of one matrix by another of low rank." *Psychometrika* 1, 211.

[3] Hastie, T., Buja, A., & Tibshirani, R. (1995). "Penalized discriminant analysis." *The Annals of Statistics* 23 (1), 73–102.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

[5] Suo, X., Minden, V., & Nelson, B. (2017). "Sparse canonical correlation analysis." https://arxiv.org/pdf/1705.10865.pdf

[6] Witten, D. M., Tibshirani, R., & Hastie, T. (2009). "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." *Biostatistics* (Oxford, England), 10(3), 515–534.

References II

[7] Witten, D., and Tibshirani, R. (2011). "Penalized classification using fisher's linear discriminant." *Journal of the Royal Statistical Society,* Series B.

[8] Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society*, Series B, 67, 301–320.

[9] Zou, H., Hastie, T., & Tibshirani, R. (2006). "Sparse Principal Component Analysis." `https://web.stanford.edu/~hastie/Papers/spc_jcgs.pdf`

[10] Package 'elasticnet'. `https://cran.r-project.org/web/packages/elasticnet/elasticnet.pdf`

[11] Package 'PMA'. `https://cran.r-project.org/web/packages/PMA/PMA.pdf`

[12] Package 'sparseLDA'. `https://cran.r-project.org/web/packages/sparseLDA/sparseLDA.pdf`