GENERALIZACIONES DE MINIMOS CUADRADOS PARCIALES
CON APLICACIÓN EN CLASIFICACION SUPERVISADA

por:

José Carlos Vega Vilca

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

Doctor en Filosofía

en

Ciencias e Ingeniería de la Información y Computación

UNIVERSIDAD DE PUERTO RICO
Recinto Universitario de Mayagüez
2004

Aprobada por:

_____	_____
Edgar Acuña, Ph.D Presidente, Comité Graduado	Fecha
_____	_____
Raúl Macchiavelli, Ph.D Miembro, Comité Graduado	Fecha
_____	_____
Rodolfo Romañach, Ph.D Miembro, Comité Graduado	Fecha
_____	_____
Fernando Vega, Ph.D Miembro, Comité Graduado	Fecha
_____	_____
Andrés Calderón, Ph.D Representante de Estudios Graduados	Fecha
_____	_____
Jaime Seguel, Ph.D Director de Programa	Fecha
_____	_____
José A. Mari Mutt, Ph.D Director de Estudios Graduados	Fecha

Abstract

The development of technologies such as microarrays has generated a large amount of data. The main characteristic of this kind of data it is the large number of predictors (genes) and few observations (experiments). Thus, the data matrix **X** is of order $n \times p$, where n is much smaller than p. Before using any multivariate statistical technique, such as regression and classification, to analyze the information contained in this data, we need to apply either feature selection methods and/or dimensionality reduction using orthogonal variables, in order to eliminate multicollineality among the predictor variables that can lead to severe prediction errors, as well as to a decrease of the computational burden required to build and validate the classifier.

Principal component analysis (PCA) is a technique that has being used for some time to reduce the dimensionality. However, the first components that have the most variability of the data structure do not necessarily improve the prediction when it is used for regression and classification (Yeung and Ruzzo, 2001). Partial least squares (PLS), introduced by Wold (1975), was an important contribution to reduce dimensionality in a regression context using orthogonal components. The certainty that first PLS components improve the prediction has made PLS a widely technique used particularly in the area of chemistry, known as Chemometrics. Nguyen and Rocke (2002), working on supervised classification methods for microarray data, reduced the dimensionality by applying first feature selection using statistical techniques such as difference of means and analysis of variance, after which they applied PLS regression considering the vector of classes (a categorical variable) as a response vector (continuous variable). This procedure is not adequate since the predictions are not necessarily integers and they must be rounded up, losing accuracy. In spite of these shortcomings, regression PLS yields reasonable results.

In this thesis work we implement generalizations of regression PLS as a dimensionality reduction technique to be applied in supervised classification. We extend a technique introduced by Bastien et al. (2002), who combined PLS with ordinal logistic regression

for multiclass problems. However, since it is very uncommon to have ordered classes, in this work it has been combined PLS with nominal logistic regression. It was also considered the multivariate PLS along with logistic regression, as well as the construction of PLS components from linear discriminant analysis, and projection pursuit. The proposals presented in this thesis improve two recent results by Fort and Lambert (2004), and Ding and Gentleman (2004), combining logistic regression and PLS that are suitable only for datasets with two classes. A library of R functions was built to carry out the different proposals.

Resumen

El desarrollo de tecnologías tales como *microarrays* ha generado una gran cantidad de datos. La característica principal de este tipo de datos es que tiene un gran número de predictoras (genes) y pocas observaciones (experimentos). Así, la matriz de datos es de orden $n \times p$, donde n es mucho menor que p. Antes de usar alguna técnica estadística multivariada, tal como regresión y clasificación, para analizar la información contenida en esos datos, se necesita aplicar métodos de selección de variables y reducción de la dimensionalidad usando variables ortogonales para eliminar multicolinealidad entre las variables predictoras. Esta multicolinealidad podría causar severos errores de predicción. Por otro lado, la reducción de la dimensionalidad del conjunto de datos permite disminuir la carga computacional que se origina al construir y validar el clasificador.

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica que ha sido utilizada por mucho tiempo para de reducir la dimensionalidad. Sin embargo, las primeras componentes que se supone tienen la más alta variablidad no necesariamente mejoran la predicción cuando se usa en regresión o en clasificación (Yeung y Ruzzo, 2001). La técnica regresión por Mínimos Cuadrados Parciales (PLS, por sus siglas en inglés) introducida por Wold (1975), fue una contribución muy importante en la reducción de la dimensionalidad en regresión múltiple. La seguridad de que las primeras componentes ortogonales mejoran la predicción lo ha convertido en un método muy usado, sobre todo en el área de la química llamada *Chemometrics*. Nguyen y Rocke (2002) trabajaron con métodos de clasificación supervisada para datos de *microarray*, reduciendo la dimensionalidad y aplicando primero selección de variables usando técnicas estadísticas tales, como diferencia de medias y análisis de varianza. Posteriormente estos autores aplicaron regresión PLS considerando el vector de clases (una variable categórica) como un vector respuesta (variable continua). Este procedimiento no es adecuado porque las predicciones no necesariamente serán enteras y habría que redondear, perdiendo precisión, pero aún así sus esfuerzos por solucionar el problema de los datos son loables y han logrado éxito en sus resultados.

En este trabajo se implementan generalizaciones de regresión PLS como una técnica de reducción de la dimensionalidad para ser aplicado en clasificación supervisada. Ésta es una extensión de la técnica introducida por Bastien et al. (2002), quienes combinaron PLS con regresión logística ordinal para el problema de multiclases. Sin embargo, no es muy común tener ordenadas las clases; y por lo tanto, en este trabajo se combina PLS con regresión logística nominal. También se considera PLS multivariado a partir de regresión logística, así como la construcción de componentes PLS a partir del Análisis Discriminante Lineal y componentes PLS a partir de *Projection Pursuit*. Esta propuesta también mejora dos trabajos recientes de Fort y Lambert (2004), y Ding y Gentleman (2004), que combinan regresión logística y PLS que están disponibles sólo para dos clases. Se construyó una librería de funciones en R que llevan a cabo las diferentes propuestas.

Dedicatoria

A la memoria de mi padre; siempre estará en mi corazón

A mi querida madre, por sus sabios consejos y enseñanzas.

A Yrma Beatriz, mi querida esposa, por su cariño y comprensión.

A Claudio Andrés y Diana Cristina, mis hijos; ellos son mi mayor fuente de inspiración.

A mis queridos hermanos: Andrés, Bertha, Nelly, Jaime, Charo, Doris y Martín.

Agradecimientos

A Dios por ser mi guía, por hacer de mí lo que soy.

Al Dr. Edgar Acuña Fernández, presidente de mi Comité Graduado, por su orientación, apoyo constante y sus valiosas sugerencias en el desarrollo de la presente tesis.

A los miembros de mi Comité Graduado: Dr. Raúl Macchiavelli, Dr. Fernando Vega y Dr. Rodolfo Romañach; por sus recomendaciones y valiosas sugerencias para un mejor desarrollo y presentación de esta tesis.

A la Oficina de Investigación Naval (ONR) por apoyarme económicamente através del Grant N00014-03-1-0359

A todos los que fueron mis profesores del Departamento de Matemática e Ingeniería del Recinto Universitario de Mayagüez de la Universidad de Puerto Rico; en reconocimiento de sus labores como docentes y de sus virtudes como personas.

A todos los profesores del Departamento de Estadística de la Universidad Agraria La Molina, de Lima-Perú, por el apoyo moral y por los sentimientos de consideración hacia mi persona.

A todas las personas que de una u otra manera hicieron posible la culminación de la presente tesis.

Tabla de Contenido

Lista de Tablas xi

Lista de Algoritmos xii

Lista de Figuras xiii

1. Introducción 1

1.1 Justificación 1

1.2 Objetivos 3

1.3 Organización de la tesis 4

2. Revisión de literatura 6

2.1 Introducción 6

2.2 Regresión por Componentes Principales 6

2.3 Regresión por Mínimos Cuadrados Parciales (Regresión PLS) 10

2.3.1 Regresión PLS univariada (PLS1) 11

2.3.2 Propiedades observadas en PLS1 15

2.3.3 Regresión PLS, caso multivariado (PLS2) 16

2.3.4 Selección del número de componentes 19

2.4 Clasificación 21

2.4.1 Tasa de error de clasificación 23

2.5 Regresión Logística 24

2.5.1 Regresión Logística Ordinal 24

2.5.2 Regresión Logística Nominal 25

2.5.3 Predicción en regresión logística 25

2.6 Otras Técnicas relacionadas con el uso de PLS para clasificación supervisada 25

3. Regresión Logística PLS 28

3.1 Introducción 28

3.2 Regresión Logística Ordinal PLS (OLRPLS) 29
3.2.1 Descripción del algoritmo OLRPLS 30
3.3 Regresión Logística Nominal PLS (NLRPLS) 35
3.3.1 Descripción del algoritmo NLRPLS 36
3.4 Propiedades de los componentes PLS 38
3.4.1 Matriz de transformación a componentes PLS 41
3.5 Regresión Logística PLS Multivariado (MLRPLS) 43
3.5.1 Descripción del algoritmo MLRPLS 45
3.6 Selección del número de componentes PLS 46

4. Otros métodos de obtención de componentes PLS para clasificación 49
4.1 Introducción 49
4.2 Análisis Discriminante Lineal (LDA) 49
4.2.1 Componentes PLS a partir de LDA (LDAPLS) 51
4.2.2 Descripción del algoritmo LDAPLS 52
4.3 Regresión *Projection Pursuit* (PPR) 53
4.3.1 Componentes PLS a partir de PPR (PPRPLS) 55
4.3.2 Descripción del algoritmo PPRPLS 55

5. Metodología 58
5.1 Introducción 58
5.2 Manejo de las bases de datos 58
5.3 Cálculo de componentes PLS 61
5.4 Aplicación de clasificadores 62
5.5 Determinación de la tasa de error de clasificación 62
5.6 Determinación del número de componentes PLS 62
6.7 Implementación de programas 63

6. Aplicación y Resultados 64
6.1 Introducción 64

6.2 TE_{VC} usando componentes PLS a partir de OLR 67
6.3 TE_{VC} usando componentes PLS a partir de NLR 68
6.4 TE_{VC} usando componentes PLS a partir de RL, caso multivariado 69
6.5 TE_{VC} usando componentes PLS a partir de LDA 71
6.6 TE_{VC} usando componentes PLS a partir de PPR 72
6.7 Las mejores TE_{VC} usando componentes PLS 74
6.8 Gráfico de las dos y tres primeras componentes PLS: *microarrays* 75

7. Conclusiones y Recomendaciones 93
7.1 Conclusiones 93
7.1.1 Contribuciones 94
7.2 Trabajos futuros 95

8. Ética 96
8.1 Introducción 96
8.2 Ética de la Investigación 97
8.3 Ética de la tesis 100

Bibliografía 101

Lista de Tablas

Tabla 5.1 Descripción de la base de datos en estudio 57

Tabla 6.1 TE_{VC} usando todas las predictoras originales 64

Tabla 6.2 TE_{VC} usando componentes principales 64

Tabla 6.3 TE_{VC} usando componentes PLS a partir de OLR 65

Tabla 6.4 TE_{VC} usando componentes PLS a partir de NLR 66

Tabla 6.5 TE_{VC} usando componentes PLS a partir de LR, caso multivariado 68

Tabla 6.6 TE_{VC} usando componentes PLS a partir de LDA 69

Tabla 6.7 TE_{VC} usando componentes PLS a partir de PPR 70

Tabla 6.8 Las mejores TE_{VC} usando componentes PLS 72

Tabla 6.9 Comparación de tasas de error de clasificación 72

Lista de Algoritmos

Algoritmo 2.1 Componentes PLS univariado (PLS1) 11

Algoritmo 2.2 Componentes PLS multivariado (PLS2) 17

Algoritmo 3.1 Componentes PLS a partir de OLR (OLRPLS) 29

Algoritmo 3.2 Componentes PLS a partir de NLR (NLRPLS) 35

Algoritmo 3.3 Matriz de transformación a componentes PLS 42

Algoritmo 3.4 Componentes PLS, caso Multivariado (MLRPLS) 43

Algoritmo 4.1 Componentes PLS a partir de LDA (LDAPLS) 48

Algoritmo 4.2 Componentes PLS a partir de PPR (PPRPLS) 52

Lista de Figuras

Figura 6.1 Gráfico de dos y tres componentes: Datos Golub2. Algoritmo NLRPLS 74

Figura 6.2 Gráfico de dos y tres componentes: Datos Colon. Algoritmo NLRPLS 75

Figura 6.3 Gráfico de dos y tres componentes: Datos Golub3. Algoritmo NLRPLS 76

Figura 6.4 Gráfico de dos y tres componentes: Datos Breastcc. Algoritmo NLRPLS 77

Figura 6.5 Gráfico de dos y tres componentes: Datos Golub2. Algoritmo MLRPLS 78

Figura 6.6 Gráfico de dos y tres componentes: Datos Colon. Algoritmo MLRPLS 79

Figura 6.7 Gráfico de dos y tres componentes: Datos Golub3. Algoritmo MLRPLS 80

Figura 6.8 Gráfico de dos y tres componentes: Datos Breastcc. Algoritmo MLRPLS 81

Figura 6.9 Gráfico de dos y tres componentes: Datos Golub2. Algoritmo LDAPLS 82

Figura 6.10 Gráfico de dos y tres componentes: Datos Colon. Algoritmo LDAPLS 83

Figura 6.11 Gráfico de dos y tres componentes: Datos Golub3. Algoritmo LDAPLS 84

Figura 6.12 Gráfico de dos y tres componentes: Datos Breastcc. Algoritmo LDAPLS 85

Figura 6.13 Gráfico de dos y tres componentes: Datos Golub2. Algoritmo PPRPLS 86

Figura 6.14 Gráfico de dos y tres componentes: Datos Colon. Algoritmo PPRPLS 87

Figura 6.15 Gráfico de dos y tres componentes: Datos Golub3. Algoritmo PPRPLS 88

Figura 6.16 Gráfico de dos y tres componentes: Datos Breastcc. Algoritmo PPRPLS 89

Capítulo 1

Introducción

1.1 Justificación

Este trabajo de tesis se justifica por el desarrollo de tecnologías, tal como las investigaciones en *microarray*; esta tecnología consiste en el análisis del nivel de expresión de decenas de miles de genes o sus fragmentos en forma simultánea. El nivel de expresión de un gen indica la existencia de éste y cuantifica que tan activo es el gen dentro del organismo, de esta manera se puede estudiar como afecta cada gen las distintas características del organismo, o predecir los efectos de un conjunto de genes según su nivel de actividad. La tecnología *microarray* ha generado abundancia de datos y gran necesidad de metodologías para analizar y explotar la información contenida en esos datos, caracterizados por muchas mediciones de variables (genes) y pocas observaciones (experimentos). Es decir, se originan matrices de datos $\mathbf{X}(n \times p)$, donde n es mucho menor que p. En esta situación se hace necesaria la aplicación de técnicas de selección de variables y sobre todo de reducción de la dimensionalidad con variables ortogonales entre sí, antes de aplicar alguna técnica estadística de análisis multivariado, debido a dos razones: primero, para eliminar problemas de multicolinealidad de las variables predictoras que pueden causar severos errores de predicción y segundo, para disminuir la carga computacional que se origina al construir y validar el clasificador. Asimismo, en clasificación supervisada aplicada a matrices de datos usuales, caracterizadas por muchas predictoras, pero donde n es mucho mayor que p, se han invertido grandes esfuerzos en la construcción de diferentes tipos de funciones clasificadoras, las cuales gastan ingentes cantidades de tiempo en su validación; es decir en estimar su tasa de error de mala clasificación. Aquí, también se hace necesario la aplicación de técnicas de selección de

variables o reducción de la dimensionalidad, para disminuir el tiempo de estimación de la tasa de error de la función clasificadora y acelerar el proceso de predicción.

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica que ha sido utilizada por mucho tiempo con la finalidad de reducir la dimensionalidad. Sin embargo, las primeras componentes que se supone tienen la más alta variablidad no necesariamente mejoran la predicción cuando se usa en regresión o en clasificación. En clasificación no supervisada, trabajos como el de Yeung y Ruzzo (2001) demuestran que el uso de componentes principales en vez de las variables predictoras originales, no necesariamente mejora y en muchos casos degrada la calidad esperada de clasificación, ellos llegan al extremo de no recomendar su uso.

La técnica regresión por Mínimos Cuadrados Parciales (PLS, por sus siglas en inglés) introducida por Wold (1975), fue una contribución muy importante en la reducción de la dimensionalidad en regresión multiple. La seguridad de que las primeras componentes ortogonales mejoran la predicción lo ha convertido en un método muy usado, sobre todo en un área de la química llamada *Chemometrics*.

En clasificación supervisada, la abundancia de datos ha generado la necesidad de implementar metodologías de reducción de la dimensionalidad para factibilizar el análisis de la información contenida en esos datos. En ese sentido investigadores en el campo de la clasificación supervisada que usan datos de *microarrays*, como Nguyen y Rocke (2002a,b,c), generaron una metodología para solucionar el problema de pocas observaciones y muchas variables predictoras en sus datos. Trabajaron en primer lugar con la selección de variables, usando técnicas estadísticas como pruebas de diferencias de medias y análisis de variancia y después de ello aplicaron la reducción de la dimensionalidad, usando la técnica regresión PLS, considerando el vector de clases (categórico) como si fuera vector de respuestas en regresión (continua). El anterior procedimiento no es adecuado porque las predicciones no necesariamente serán enteras y

habría que redondear, perdiendo precisión, pero aún así sus esfuerzos en solucionar el problema de los datos son loables y han logrado éxito en sus resultados.

Por los motivos anteriores, en esta tesis se implementan generalizaciones de la regresión PLS como una técnica de reducción de la dimensionalidad para ser aplicada en problemas de clasificación supervisada. Se siguen los lineamientos trazados por Bastien, Esposito Vinzi y Tenenhaus (2002), quienes mostraron que el principio de regresión PLS, puede ser extendido a la regresión logística que usualmente trabaja con dos clases, pero puede ser generalizado a más de dos clases usando regresión logística ordinal, la cual es aplicada cuando hay un orden natural en las categorías de la variable respuesta. Sin embargo lo más común en clasificación supervisada es que las clases no tengan un ordenamiento natural entre sí. Por tal motivo, en esta tesis se implementa un algoritmo para construir componentes PLS a partir de la regresión logística nominal, componentes PLS a partir de la regresión logística como extensión de la regresión PLS multivariada, componentes PLS a partir de la función discriminante lineal así como de la regresión *projection pursuit*.

Recientemente ha habido un par de propuestas: de Fort y Lambert-Lacroix (2003), y el de Ding y Gentleman (2004), para aplicar componentes PLS a clasificación supervisada; que a diferencia de nuestra propuesta éstas sólo son aplicables cuando hay dos clases en el conjunto de datos.

1.2 Objetivos

Objetivo General

Implementar una técnica de reducción de la dimensionalidad que sigue las ideas fundamentales de la regresión PLS, a partir de Regresión Logística Nominal, regresión noparamétrica y función discriminante lineal para ser aplicada al problema de clasificación supervisada.

Objetivos específicos:

- Desarrollar el algorítmo de Regresión Logística Nominal PLS, aplicable cuando no hay un orden natural en las categorías de la variable respuesta, lo que constituye el caso más real cuando se trabaja en clasificación supervisada.

- Explorar variaciones de la regresión PLS con respuesta multivariada, para ser aplicada en clasificación supervisada.

- Desarrollar algoritmos para regresión no paramétrica PLS y aplicarlos a clasificación supervisada.

- Estudiar el efecto sobre la estimación de la tasa de error de clasificación de la regresión logística que usa como predictoras las componentes PLS, las cuales son obtenidas con la metodología propuesta.

- Estudiar y comparar las metodologías de generación de componentes PLS propuestas, usando como criterio de comparación la estimación de la tasa de error de mala clasificación y el número de componentes PLS usado para lograr la reducción de la dimensionalidad de la matriz de datos. Estas tasas de error de clasificación son obtenidas a partir de la aplicación de diferentes clasificadores sobre la matriz de componentes PLS.

- Construir una librería de programas en lenguaje R, en el ambiente Windows, basados en las metodologías propuestas, que puedan realizar todos los cálculos necesarios.

1.3 Organización de la tesis

Esta tesis está organizada en siete capítulos. En el segundo capítulo se revisan conceptos fundamentales, tales como: regresión por componentes principales, regresión por mínimos cuadrados parciales, clasificación y regresión logística.

El tercer capítulo está dedicado a la generación de componentes PLS a partir de la regresión logística ordinal y al desarrollo e implementación de la generación de componentes PLS a partir de la regresión logística nominal, metodología que constituye una de las contribuciones de esta tesis.

En el cuarto capítulo se proponen otras metodologías para la construcción de componentes PLS que serán usadas en clasificación supervisada; los componentes son obtenidos a partir del Análisis Discriminante Lineal y de la Regresión *Projection Pursuit.*

El quinto capítulo está referido a la metodología de la investigación donde se presentan las tareas fundamentales, que fueron realizadas para la elaboración de la presente tesis.

El sexto capítulo contiene la aplicación y resultados obtenidos en esta tesis; se muestra el trabajo experimental desde las metodologías planteadas en el tercer y cuarto capítulo, para probar la funcionalidad de los algoritmos propuestos.

El séptimo capítulo contiene las conclusiones y recomendaciones a las que se llegó con el desarrollo de la presente tesis.

El octavo capítulo contiene aspectos fundamentales de ética, que valen la pena ser reflexionados por toda persona dedicada a la investigación para que sus actos o los resultados de los mismos, sean éticamente correctos.

Capítulo 2

Revisión de literatura

2.1 Introducción

En la construcción de un modelo de regresión lineal múltiple basado en una matriz de datos **X**, de orden $n \times p$, se pueden presentar dos problemas: multicolinealidad y alta dimensionalidad de sus variables predictoras. En este capítulo se revisan dos metodologías relativamente similares y usadas en la solución de estos problemas: Regresión por Componentes Principales y Regresión por Mínimos Cuadrados Parciales. Ambos métodos transforman las variables predictoras en variables artificiales llamadas componentes o variables latentes, las cuales son ortogonales y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras. Luego usando solamente las variables latentes se construye el modelo de regresión estimado.

Uno de los objetivos del presente trabajo es mostrar que el principio de regresión por mínimos cuadrados parciales puede ser extendido a la regresión logística para ser aplicado al problema de clasificación supervisada. Por esta razón en este capítulo también incluimos una revisión de conceptos de clasificación y regresión logística. En la última sección de este capítulo se incluye una revisión de temas relacionados con el uso de PLS para clasificación.

2.2 Regresión por Componentes Principales

La regresión por componentes principales es un método que aplica mínimos cuadrados sobre un conjunto de variables artificiales llamadas componentes principales, obtenidas a

partir de la matriz de correlación. Sea $\mathbf{X}$ la matriz de predictoras estandarizada por columnas. La matriz de correlaciones está dada por $\mathbf{R}=(n\text{-}1)^{-1}\mathbf{X}'\mathbf{X}$; esta matriz es simétrica y semi definida positiva. Usando descomposición espectral de una matriz cuadrada y simétrica se tiene que

$$\mathbf{R} = \mathbf{\Gamma}\,\mathbf{\Lambda}\,\mathbf{\Gamma}' \tag{2.1}$$

donde $\mathbf{\Gamma}= (\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_p)$ es una matriz ortogonal de orden $p\times p$, cada $\boldsymbol{\gamma}_i$ es llamado autovector y tiene norma 1. La matriz $\mathbf{\Lambda} = diag\ (\lambda_1 \ldots \lambda_p)$ es diagonal de orden $p\times p$; los λ_i son llamados autovalores y $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Los autovectores forman una base en $\Re^p$, es decir cualquier vector en $\Re^p$ puede ser generado como una combinación lineal de estos autovectores. Por ortogonalidad de la matriz $\mathbf{\Gamma}$, la expresión (2.1) puede ser escrita como:

$$\mathbf{\Gamma}'\mathbf{R}\,\mathbf{\Gamma} = \mathbf{\Lambda} \tag{2.2}$$

$$\begin{pmatrix} \boldsymbol{\gamma}_1' \\ \vdots \\ \boldsymbol{\gamma}_p' \end{pmatrix} \mathbf{R} \begin{pmatrix} \boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_p \end{pmatrix} = \mathbf{\Lambda}$$

$$\begin{pmatrix} \boldsymbol{\gamma}_1'\mathbf{R}\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_1'\mathbf{R}\boldsymbol{\gamma}_p \\ \vdots & \ddots & \vdots \\ \boldsymbol{\gamma}_p'\mathbf{R}\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_p'\mathbf{R}\boldsymbol{\gamma}_p \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$$

De la relación anterior se puede verificar la siguiente equivalencia para $i, j = 1, \ldots, p$

$$\boldsymbol{\gamma}_i'\mathbf{R}\boldsymbol{\gamma}_j = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases} \tag{2.3}$$

La matriz de componentes principales $\mathbf{C}$ de orden $n\times p$, es obtenida transformando la matriz $\mathbf{X}$, de la siguiente manera:

$$\mathbf{C} = \mathbf{X}\,\boldsymbol{\Gamma} \tag{2.4}$$

$$= \mathbf{X}\,(\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_p)$$

$$\mathbf{C} = (\mathbf{X}\,\boldsymbol{\gamma}_1 \ \ldots \ \mathbf{X}\,\boldsymbol{\gamma}_p) \tag{2.5}$$

Cada $\mathbf{X}\boldsymbol{\gamma}_i$, para $i = 1,\ldots, p$ es llamada componente principal. De (2.3) se concluye que las componentes principales son ortogonales entre sí.

Fundamento de Componentes Principales

La idea es maximizar la varianza de la componente principal $\mathbf{X}\boldsymbol{\gamma}$ sujeto a que el autovector $\boldsymbol{\gamma}$, satisfaga $\boldsymbol{\gamma}'\boldsymbol{\gamma} = 1$

$$var\,(\mathbf{X}\,\boldsymbol{\gamma}) = \boldsymbol{\gamma}'\, var(\mathbf{X})\,\boldsymbol{\gamma}$$

$$= \boldsymbol{\gamma}'\,[(n\text{-}1)^{-1}\mathbf{X}'\mathbf{X}]\,\boldsymbol{\gamma}$$

$$var\,(\mathbf{X}\,\boldsymbol{\gamma}) = \boldsymbol{\gamma}'\mathbf{R}\,\boldsymbol{\gamma} \tag{2.6}$$

Sea ϕ una función que incluye la varianza de la componente principal a ser maximizada y una penalidad que contiene la restricción y al multiplicador de Lagrange, λ.

$$\phi = \boldsymbol{\gamma}'\,\mathbf{R}\,\boldsymbol{\gamma} \ - \ \lambda\,(\boldsymbol{\gamma}'\boldsymbol{\gamma} - 1) \tag{2.7}$$

La maximización de ϕ determina al vector $\boldsymbol{\gamma}$ que maximiza $\boldsymbol{\gamma}'\mathbf{R}\boldsymbol{\gamma}$. Derivando (2.7) con respecto a $\boldsymbol{\gamma}$, se tiene que

$$\frac{\partial \phi}{\partial \boldsymbol{\gamma}} = 2\mathbf{R}\,\boldsymbol{\gamma} - 2\lambda\,\boldsymbol{\gamma} = \mathbf{0}$$

$$\mathbf{R}\,\boldsymbol{\gamma} = \lambda\,\boldsymbol{\gamma} \tag{2.8}$$

De (2.8) se obtiene $\boldsymbol{\gamma}'\,\mathbf{R}\,\boldsymbol{\gamma} = \lambda$. La relación entre el autovector $\boldsymbol{\gamma}$ y el autovalor λ es determinada por los siguientes teoremas, que aparecen, por ejemplo, en Mardia et al. (1997).

Teorema 2.1 No existe vector normalizado **a**, que haga que la varianza de la transformación $\mathbf{Xa}$, sea más grande que λ_1, la varianza de la primera componente principal $\mathbf{X}\boldsymbol{\gamma}_1$

Prueba

Sea $\mathbf{a} = \boldsymbol{\Gamma}\mathbf{c} = c_1\boldsymbol{\gamma}_1 + \cdots + c_p\boldsymbol{\gamma}_p$, donde $\mathbf{c} = (c_1 \cdots c_p)'$ es un vector de constantes y $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_p)$ es la matriz de autovectores de la matriz de correlaciones **R**, los cuales forman una base en $\Re^p$. Ya que $\mathbf{a}'\mathbf{a} = 1$, por lo tanto $\mathbf{c}'\,\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\,\mathbf{c} = 1$, esto implica que $\mathbf{c}'\mathbf{c} = 1$, debido a la ortogonalidad de $\boldsymbol{\Gamma}$.

$$
\begin{aligned}
var(\mathbf{Xa}) &= \mathbf{a}'\, var(\mathbf{X})\, \mathbf{a} \\
&= \mathbf{a}'\, \mathbf{R}\, \mathbf{a} \\
&= \mathbf{c}'\, \boldsymbol{\Gamma}'\mathbf{R}\, \boldsymbol{\Gamma}\, \mathbf{c} \\
&= \mathbf{c}'\, \boldsymbol{\Lambda}\, \mathbf{c} \\
&= \sum_{i=1}^{p} \lambda_i\, c_i^2 \qquad (2.9)
\end{aligned}
$$

Puesto que λ_1 es el autovalor más grande, el máximo de la expresión (2.9) sujeto a $\mathbf{c}'\mathbf{c} = \sum c_i^2 = 1$ es λ_1 , es decir **c** = (1, 0, …, 0)'. Por lo tanto la varianza de la primera componente principal es maximizado a λ_1 cuando $\mathbf{a} = \boldsymbol{\gamma}_1$ ■

Un argumento similar al anterior muestra que la varianza de la última componente principal es λ_p cuando $\mathbf{a} = \boldsymbol{\gamma}_p$. El autovalor λ_p es el valor más pequeño de todas las varianzas de las demás componentes principales. Las componentes principales intermedias tienen propiedad de varianza maximal, dada por el siguiente teorema.

Teorema 2.2 Si $\boldsymbol{\alpha} = \mathbf{Xa}$ es una componente principal, la cual no está correlacionada con las primeras *k*-componentes principales, entonces la varianza de $\boldsymbol{\alpha}$ es maximizada cuando $\boldsymbol{\alpha}$ es la (*k*+1)-ésima componente principal.

Prueba

Los vectores $\mathbf{a} = c_1\boldsymbol{\gamma}_1 + \cdots + c_p\boldsymbol{\gamma}_p$ y $\mathbf{c} = (c_1 \cdots c_p)'$ son como en el teorema anterior.

$\boldsymbol{\alpha} = \mathbf{Xa}$ es no correlacionada con $\mathbf{X}\boldsymbol{\gamma}_i$, para i=1,…,k. Entonces $cor(\mathbf{Xa}, \mathbf{X}\boldsymbol{\gamma}_i) = 0$, implica que $cov(\mathbf{Xa}, \mathbf{X}\boldsymbol{\gamma}_i) = 0$, entonces $\mathbf{a}'var(\mathbf{X})\ \boldsymbol{\gamma}_i = 0$. Por lo tanto $\mathbf{a}'\mathbf{R}\boldsymbol{\gamma}_i = 0$ y por la expresión (2.3) se establece que $\boldsymbol{\gamma}_i \neq \mathbf{a}$, y en consecuencia $\mathbf{a}'\boldsymbol{\gamma}_i = 0$, $\forall\ i$ = 1,…,k. De esta última relación se obtiene que $c_i = 0$, $\forall\ i$ = 1,…,k. Por lo tanto $var(\boldsymbol{\alpha}) = var(\mathbf{Xa}) = \mathbf{a}'\mathbf{Ra}$, alcanza su valor máximo λ_{k+1} , cuando $\mathbf{a} = \boldsymbol{\gamma}_{k+1}$, es decir cuando $c_{k+1} = 1$. ■

2.3 Regresión por Mínimos Cuadrados Parciales (Regresión PLS)

La regresión por mínimos cuadrados parciales (regresión PLS, por sus siglas en inglés), fue introducida por Herman Wold (1975) para ser aplicada en ciencias económicas y sociales. Sin embargo gracias a las contribuciones de su hijo Svante Wold, ha ganado popularidad en el área de la química conocida como *Chemometrics*, en donde se analizan datos que se caracterizan por muchas variables predictoras, con problemas de multicolinealidad, y pocas unidades experimentales en estudio.

La idea motivadora de PLS fue heurística, por este motivo algunas de sus propiedades son todavía desconocidas a pesar de los progresos alcanzados por Helland (1988), Hoskuldson (1988), Stone y Brooks (1990) y otros. La metodología PLS generaliza y combina características del Análisis de Componentes Principales y Análisis de Regresión Múltiple. La demanda por esta metodología y la evidencia de que trabaja bien, van en aumento y así, la metodología PLS está siendo aplicada en muchas ramas de la ciencia. En PLS, a diferencia de Componentes Principales, los datos de entrada además de la matriz de predictoras **X**, deben contener una matriz de respuestas **Y**.

X : matriz de variables predictoras, de orden $n \times p$

Y : matriz de variables dependientes, de orden de $n \times q$

2.3.1 Regresión PLS univariada (PLS1)

Es el caso de aplicación de regresión PLS, cuando **Y** es un vector (q=1). Puede ser visto como una transformación de las variables predictoras **X**, considerando su relación con el vector de respuestas **Y** de orden $n\times1$, obteniéndose como resultado una matriz de componentes o variables latentes no correlacionadas, $\mathbf{T}=(\mathbf{T}_1,\ldots,\mathbf{T}_p)$ de orden $n\times p$. Se debe notar que esto contrasta con el Análisis de Componentes Principales, en el cual las componentes son obtenidas usando sólo la matriz de predictoras **X**. El número de variables latentes $\mathbf{T}_1$, ..., $\mathbf{T}_k$, donde k≤p, es determinado generalmente por el método de validación cruzada dejando una observación afuera, también llamado PRESS (*Prediction Sum of Squares*). La ecuación de regresión estimada tomará la siguiente forma:

$$\hat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \quad \cdots \quad + \beta_k T_k \tag{2.10}$$

El siguiente algoritmo para PLS1 es adaptado de Garthwaite (1994) y Trygg (2001). La entrada de datos corresponde a las matrices **X** e **Y** las cuales han sido centradas y normalizadas a la unidad, por columnas

1. Entrada : $\mathbf{X}(n\times p)$, $\mathbf{Y}(n\times1)$
2. Para i = 1 hasta p
3. $\mathbf{w} = cov(\mathbf{Y}, \mathbf{X})$: normalizar $\mathbf{w}$ ($\|\mathbf{w}\| = 1$)
4. $\mathbf{T} = \mathbf{X}\mathbf{w}$
5. $v = (\mathbf{T}'\mathbf{Y})/(\mathbf{T}'\mathbf{T})$
6. $\mathbf{b} = (\mathbf{T}'\mathbf{X})/(\mathbf{T}'\mathbf{T})$
7. $\mathbf{X} = \mathbf{X} - \mathbf{T}\,\mathbf{b}$
8. $\mathbf{Y} = \mathbf{Y} - \mathbf{T}\,v$
9. Fin i

Algoritmo 2.1 : Componentes PLS univariado (PLS1)

Descripción del algoritmo PLS1

Con base en el algoritmo anterior se presenta una descripción del proceso. La matriz de datos puede ser escrita como $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p)$, donde $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ son las columnas de la matriz **X**. A continuación se describen los principales pasos del algoritmo:

<u>*Paso 3*</u>. Se calcula el vector $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$, donde el elemento w_i corresponde a la covarianza de la variable respuesta **Y** con cada una de las variables predictoras ($\mathbf{X}_i$)

$$w_i = cov\,(\mathbf{Y}, \mathbf{X}_i) \qquad i = 1, ..., p$$
$$w_i = \frac{SP(\mathbf{Y}, \mathbf{X}_i)}{n-1} \cdot \frac{SC(\mathbf{X}_i)}{SC(\mathbf{X}_i)} = \hat{\beta}_i\, var(\mathbf{X}_i) \qquad (2.11)$$
$$w_i = coef\,(\mathbf{X}_i).var(\mathbf{X}_i), \text{ del modelo RLI: } \mathbf{Y} \sim \mathbf{X}_i$$

Donde *SP* y *SC* son suma de productos y suma de cuadrados respectivamente. Por lo tanto cada w_i es igual al coeficiente de Regresión Lineal simple (RLI) del modelo: $\mathbf{Y} \sim \mathbf{X}_i$, multiplicado por la varianza de la predictora $\mathbf{X}_i$. Finalmente $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$ es normalizado a la unidad.

<u>*Paso 4*</u>. Se calcula la componente PLS, $\mathbf{T} = \mathbf{Xw} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p) \cdot (w_1, w_2, \cdots, w_p)'$

Es decir $\mathbf{T} = \sum_{i=1}^{p} \mathbf{X}_i\, w_i$ (2.12)

<u>*Paso 5*</u>. Se calcula el coeficiente de regresión simple de **Y** sobre **T**.

$$v = \frac{SP(\mathbf{T}, \mathbf{Y})}{SC(\mathbf{T})} \quad \rightarrow \quad \hat{\mathbf{Y}} = v\mathbf{T} \qquad (2.13)$$

<u>*Paso 6*</u>. Se calcula el vector $\mathbf{b} = (b_1, b_2, \cdots, b_p)$; cada elemento de **b** corresponde al coeficiente de regresión simple de $\mathbf{X}_i$ sobre **T**

$$b_i = \frac{SP(\mathbf{T}, \mathbf{X}_i)}{SC(\mathbf{T})} \quad \rightarrow \quad \hat{\mathbf{X}}_i = b_i\, \mathbf{T}\ , \qquad i = 1, ..., p \tag{2.14}$$

Paso 7-8. Actualización de la matriz de predictoras y el vector respuesta

$$\mathbf{X} = \mathbf{X} - \hat{\mathbf{X}} \ = \ \mathbf{X} - \mathbf{T}\mathbf{b} \tag{2.15}$$

$$\mathbf{Y} \ = \ \mathbf{Y} - \hat{\mathbf{Y}} \ = \ \mathbf{Y} - v\mathbf{T}$$

donde $\hat{\mathbf{Y}}$ se obtiene de (2.13) y $\hat{\mathbf{X}} = (\hat{\mathbf{X}}_1 \cdots \hat{\mathbf{X}}_p)$ es obtenida de (2.14)

h-ésima componente PLS1 **:** $\mathbf{T}_h$

Aquí se supone que las componentes $\mathbf{T}_1$, $\mathbf{T}_2$, …, $\mathbf{T}_{h\text{-}1}$ fueron calculados en las h-1 iteraciones anteriores. Durante la h-ésima iteración este algoritmo calcula $\mathbf{w}(h)$, $\mathbf{T}_h$, $v(h)$, $\mathbf{b}(h)$, $\mathbf{X}(h)$, $\mathbf{Y}(h)$, usando el vector de respuestas y la matriz de predictoras de la iteración anterior: $\mathbf{Y}(h\text{-}1)$ y $\mathbf{X}(h\text{-}1)$. Cuando h =1, los datos necesarios para hacer estos cálculos son $\mathbf{X}(0)$ y $\mathbf{Y}(0)$, los cuales son la matriz de predictoras y el vector de respuestas estandarizadas por columnas, de datos iniciales.

En cada iteración del algoritmo PLS1 se calcula una variable latente. A continuación se presenta la h-ésima iteración del algoritmo PLS1

1. $\mathbf{w}(h) = \mathbf{X}'(h-1)\, \mathbf{Y}(h-1)$ $\quad \Rightarrow$ normalizar $\mathbf{w}(h)$
2. $\mathbf{T}_h = \mathbf{X}(h-1)\, \mathbf{w}(h)$ $\quad \Rightarrow$ h-ésima variable latente
3. $v(h) = \mathbf{T}_h'\, \mathbf{Y}(h-1) / \mathbf{T}_h'\, \mathbf{T}_h$ (2.16)
4. $\mathbf{b}(h) = \mathbf{T}_h'\, \mathbf{X}(h-1) / \mathbf{T}_h'\, \mathbf{T}_h$
5. $\mathbf{X}(h) = \mathbf{X}(h-1) - \mathbf{T}_h\, \mathbf{b}(h)$
6. $\mathbf{Y}(h) = \mathbf{Y}(h-1) - \mathbf{T}_h\, v(h)$

En el paso 2, se calcula la h-ésima variable latente $\mathbf{T}_h$ de dimensión $n\times 1$. Se debe observar que en los pasos 5 y 6, el algoritmo actualiza la matriz de predictoras y el vector de respuestas respectivamente, los cuales serán utilizados en la próxima iteración, h+1. Estas actualizaciones también son conocidas como la matriz y el vector de residuales de la iteración h.

Fundamento de PLS1

La idea es maximizar la covarianza al cuadrado entre la variable latente **T** = **Xw**, y la variable respuesta **Y**, sujeto a **w′w** = 1. La variable latente **T** está definida como una combinación lineal de las predictoras, tal que **w** ≠ **0**. Sea **A** el vector de covarianzas de **X** e **Y**, de orden $p \times 1$. El análisis de regresión establece la dependencia de **Y** sobre las predictoras **X**, por lo que **A** ≠ **0**

$$\begin{aligned} \left[cov(\mathbf{Xw}\,,\,\mathbf{Y})\right]^2 &= \left[\,\mathbf{w}'\,cov(\mathbf{X}\,,\,\mathbf{Y})\,\right]^2 \\ &= \left[\,\mathbf{w}'\,\mathbf{A}\,\right]^2 \\ &= \mathbf{w}'\,\mathbf{A}\,\mathbf{A}'\,\mathbf{w} \end{aligned} \qquad (2.17)$$

Sea ϕ una función que incluye la covarianza al cuadrado entre la variable latente **T** = **Xw** y la variable respuesta **Y** a ser maximizada y una penalidad que contiene la restricción y el multiplicador de Lagrange, λ.

$$\phi = \mathbf{w}'\,\mathbf{A}\mathbf{A}'\,\mathbf{w} - \lambda\,(\mathbf{w}'\mathbf{w} - 1)$$

La maximización de ϕ determina al vector **w** que maximiza **w′AA′w**, la covarianza al cuadrado entre la variable latente y el vector de respuestas.

$$\begin{aligned} \frac{\partial \phi}{\partial \mathbf{w}} &= 2\,\mathbf{A}\mathbf{A}'\mathbf{w} - 2\lambda\,\mathbf{w} = \mathbf{0} \\ \mathbf{A}\mathbf{A}'\,\mathbf{w} &= \lambda\,\mathbf{w} \end{aligned} \qquad (2.18)$$

y usando la restricción **w′w** = 1, en la expresión anterior, se tiene que

$$\mathbf{w}'\,\mathbf{A}\mathbf{A}'\,\mathbf{w} = \lambda \qquad (2.19)$$

Al multiplicar por la izquierda la expresión (2.18) por **A**’

$$\begin{aligned} \mathbf{A}'\,\mathbf{A}\mathbf{A}'\,\mathbf{w} &= \lambda\,\mathbf{A}'\mathbf{w} \\ (\mathbf{A}'\mathbf{A} - \lambda)\,\mathbf{A}'\,\mathbf{w} &= 0 \\ \mathbf{A}'\mathbf{A} - \lambda = 0 \quad \text{ó} \quad \mathbf{A}'\mathbf{w} &= 0 \end{aligned} \qquad (2.20)$$

Como $\mathbf{A'w}$ no puede ser cero, ya que se está buscando maximizar, entonces $\mathbf{A'A} - \lambda = 0$, de donde se obtiene la siguiente expresión

$$\lambda = \mathbf{A'A} = \| \mathbf{A} \|^2 \tag{2.21}$$

De la expresión anterior $\lambda^2 = (\mathbf{A'A})(\mathbf{A'A}) = \lambda \| \mathbf{A} \|^2$, entonces:

$$\mathbf{A'AA'A} = \lambda \| \mathbf{A} \|^2$$

$$\frac{\mathbf{A'}}{\| \mathbf{A} \|} \mathbf{AA'} \frac{\mathbf{A}}{\| \mathbf{A} \|} = \lambda \tag{2.22}$$

De (2.19) y (2.22), se puede reconocer que el vector $\mathbf{w}$ que maximiza $\mathbf{w'AA'w}$, la covarianza al cuadrado de la variable latente y el vector de respuestas, es el vector de covarianzas normalizado

$$\mathbf{w} = \frac{\mathbf{A}}{\| \mathbf{A} \|} = \frac{\mathbf{X'Y}}{\| \mathbf{X'Y} \|} \tag{2.23}$$

2.3.2 Propiedades observadas en PLS1

Asumiendo que:

- $\mathbf{U}$ es un vector columna de unos, de dimensión *n*.
- $\mathbf{X}(0)$ y $\mathbf{Y}(0)$, es la matriz de predictoras y el vector de respuestas, respectivamente, de datos iniciales centrados y normalizados a la unidad por columnas. Entonces se cumple: $\mathbf{X}'(0)\,\mathbf{U} = \mathbf{0}_{p\times 1}$, $\mathbf{Y}'(0)\,\mathbf{U} = 0$

Se cumplen las siguientes propiedades:

P1. El *h*-ésimo vector latente $\mathbf{T}_h$, siempre está centrado, es decir la suma de sus elementos es cero.

$$\mathbf{T}_h'\,\mathbf{U} = 0$$

P2. La matriz de predictoras siempre está centrada en cualquier iteración, es decir la suma de cada una de sus columnas es cero.

$$\mathbf{X}'(h)\,\mathbf{U} = \mathbf{0}_{p\times 1}$$

P3. El vector de respuestas siempre está centrado en cualquier iteración, es decir la suma de sus elementos es cero.

$$\mathbf{Y}'(h)\,\mathbf{U} = 0$$

P4. En la *h*-ésima iteración, se cumple que el vector latente $\mathbf{T}_h$ es ortogonal con cada una de las columnas de la matriz de predictoras

$$\mathbf{T}_h'\,\mathbf{X}(h) = \mathbf{0}_{1\times \mathrm{p}}$$

P5. En la *h*-ésima iteración, se cumple que el vector latente $\mathbf{T}_h$ es ortogonal con el vector de respuestas

$$\mathbf{T}_h'\,\mathbf{Y}(h) = 0$$

P6. Cada par de variables latentes son ortogonales, es decir el producto escalar de dos variables latentes cualesquiera es igual a cero. Sean dos variables latentes $\mathbf{T}_k$ y $\mathbf{T}_\ell$ donde $k \neq \ell$

$$\mathbf{T}_k'\,\mathbf{T}_\ell = 0$$

P7. La matriz $\mathbf{Z} = (\mathbf{z}_1 \ \ldots \ \mathbf{z}_p)$ de orden $p\times p$, que transforma variables predictoras en componentes PLS o variables latentes, puede ser hallada iterativamente.

$$\mathbf{z}_1 = \mathbf{w}(1)$$
$$\mathbf{z}_h = [\,\mathbf{I} - \textstyle\sum_{j=1}^{h-1}\mathbf{z}_j\,\mathbf{b}(j)\,]\,\mathbf{w}(h) \quad ; \quad h > 1$$

2.3.3 Regresión PLS, caso multivariado (PLS2)

Es una generalización de la regresión PLS univariado y se diferencia de ésta porque aquí se tiene una matriz de variables respuesta $\mathbf{Y}(n\times q)$, además de la matriz de predictoras $\mathbf{X}(n\times p)$, con $q < p$. El propósito del PLS multivariado es encontrar un conjunto de componentes $\mathbf{T}_1, \ldots, \mathbf{T}_k$, donde $k\leq p$, que rindan buenos modelos lineales para todas las variables respuesta $\mathbf{Y}$. El modelo estimado es de la siguiente forma:

$$\hat{Y}_j = \beta_{j0} + \beta_{j1}T_1 + \beta_{j2}T_2 + \quad \cdots \quad + \beta_{jk}T_k \qquad j = 1,...,q \qquad (2.24)$$

El siguiente algoritmo está basado en Hoskuldsson (1988) y Garthwaite (1994), y ha sido aumentado para un mejor entendimiento. Las **X** e **Y** son centradas y normalizadas a la unidad, por columnas

1. Input: **X**(n×p), **Y**(n×q)
2 Hacer k=0
3. Para i = 1 hasta ⌈p/q⌉ : ⌈ ⌉ es la función "ceiling", que redondea al entero superior
4. Para j = 1 hasta q
5. Sea **V** la j-ésima columna de **Y**
6. **w** = *cov* (**V**, **X**) : normalizar **w** (||**w**|| = 1)
7. **T** = **Xw**
8. **c** = *cov* (**T**, **Y**) : normalizar **c** (||**c**|| = 1)
9. $\mathbf{V}_{\text{nuevo}} = \mathbf{Yc}$
10. Si || **V** – $\mathbf{V}_{\text{nuevo}}$ || > ε ➔ Hacer **V** = $\mathbf{V}_{\text{nuevo}}$, Ir al paso 6
11. **V** = $\mathbf{V}_{\text{nuevo}}$
12. **b** = (**T**'**X**) / (**T**'**T**)
13. v = (**T**'**V**) / (**T**'**T**)
14. **X** = **X** – **T b**
15. **Y** = **Y** – v **T c**'
16. k=k+1
17. if (k = p) ➔ Terminar
18. Fin j
19. Fin i

Algoritmo 2.2 : Componentes PLS multivariado (PLS2)

Descripción del algoritmo PLS2

Con base en el algoritmo anterior se presenta una descripción del proceso. La matriz de datos puede ser escrita como $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p)$, donde $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ son las columnas de la matriz **X** , y la matriz de respuestas $\mathbf{Y} = (\mathbf{Y}_1 \cdots \mathbf{Y}_q)$. A continuación se describen los principales pasos del algoritmo:

Paso 6. Se calcula el vector $\mathbf{w} = (w_1, w_2, \quad \cdots \quad, w_p)'$; cada elemento de **w** corresponde a la covariancia de la j-ésima variable respuesta $\mathbf{Y}_j$, representada por el vector **V** , con cada una de las variables predictoras ($\mathbf{X}_i$)

$$w_i = cov\,(\mathbf{V}, \mathbf{X}_i) \qquad i = 1, ..., p$$
$$w_i = coef\,(\mathbf{X}_i).\text{var}(\mathbf{X}_i), \text{ del modelo RLS: } \mathbf{V} \sim \mathbf{X}_i \qquad (2.25)$$

Este resultado es obtenido usando el mismo argumento que quedó demostrado en la expresión (2.11). Finalmente $\mathbf{w} = (w_1, w_2, \quad \cdots \quad, w_p)'$ es normalizado a la unidad.

Paso 7. Se calcula la variable latente $\mathbf{T} = \mathbf{X}\mathbf{w} = (\mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_p) \cdot (w_1\ w_2 \cdots w_p)'$,

Es decir $\mathbf{T} = \sum_{i=1}^{p} \mathbf{X}_i\, w_i$ (2.26)

Paso 8. Se calcula el vector $\mathbf{c} = (c_1, c_2, \quad \cdots \quad, c_q)'$; cada elemento de **c** corresponde a la covarianza de la componente **T**, obtenida en el paso anterior, con cada una de las variables respuestas ($\mathbf{Y}_j$)

$$c_j = cov\,(\mathbf{T}\ , \mathbf{Y}_j) \qquad j = 1, ..., q$$
$$c_i = coef\,(\mathbf{Y}_i).\text{var}(\mathbf{Y}_i), \text{ del modelo RLS: } \mathbf{T} \sim \mathbf{Y}_i \qquad (2.27)$$

Este resultado es la aplicación de lo obtenido en la expresión (2.11). Finalmente $\mathbf{c} = (c_1, c_2, \quad \cdots \quad, c_q)'$ es normalizado a la unidad. De manera equivalente, **c** puede ser calculado como la normalización del vector **Y**'**T**

Paso 9. Se calcula un nuevo vector $\mathbf{V}_{nuevo} = \mathbf{Y}\mathbf{c} = (\mathbf{Y}_1 \cdots \mathbf{Y}_q) \cdot (c_1 \cdots c_q)'$, que reemplazará al vector **V** . Este nuevo vector también es expresado como sigue:

$$\mathbf{V}_{nuevo} = \sum_{j=1}^{q} \mathbf{Y}_j\, c_j \qquad (2.28)$$

Paso 12. Se calcula el vector $\mathbf{b} = (b_1, b_2, \cdots, b_p)$; cada elemento de **b** corresponde al coeficiente de regresión simple de $\mathbf{X}_i$ sobre la variable latente **T**

$$b_i = \frac{SP(\mathbf{T}, \mathbf{X}_i)}{SC(\mathbf{T})} \quad \rightarrow \quad \hat{\mathbf{X}}_i = b_i \mathbf{T} \quad , \quad i = 1, \ldots, p \tag{2.29}$$

Paso 13. Se calcula el coeficiente de regresión simple de **V** sobre **T,** donde SP y SC son la suma de productos y la suma de cuadrados, respectivamente.

$$v = \frac{SP(\mathbf{T}, \mathbf{V})}{SC(\mathbf{T})} \quad \rightarrow \quad \hat{\mathbf{V}} = v\mathbf{T} \tag{2.30}$$

Paso 14-15. Actualización de la matriz de predictoras y del vector respuesta

$$\mathbf{X} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{T}\mathbf{b} \tag{2.31}$$

$$\mathbf{Y} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \hat{\mathbf{V}}\mathbf{c}' = \mathbf{Y} - v\mathbf{T}\mathbf{c}'$$

donde $\hat{\mathbf{Y}}$ es obtenida de (2.30) y (2.27) y $\hat{\mathbf{X}} = (\hat{\mathbf{X}}_1 \cdots \hat{\mathbf{X}}_p)$ es obtenida de (2.29)

2.3.4 Selección del número de componentes

El número de componentes PLS necesario para estimar un buen modelo de regresión, a partir del algoritmo PLS1, se elige por el criterio de minimización de la suma de cuadrados de residuales. Los criterios más usados son:

- **Estimación del PRESS (*Prediction Sum of Squares*) :** Es un caso particular del método validación cruzada, consiste de los siguientes pasos:

1. Estimar el modelo de regresión, excluyendo la *i*-ésima observación, *i*=1, 2, ..., *n*
2. Calcular la predicción de la observación que no fue incluida: $\hat{y}_{(i)}$, *i*=1, 2, ..., *n*
3. Calcular el residual correspondiente: $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$, *i*=1, 2, ..., *n*
4. El PRESS promedio es calculado por: $\frac{1}{n}\sum_{i=1}^{n} e_{(i)}^2$

- **Estimación de la suma de cuadrados de residuales por Validación Cruzada (SCRvc) :** Es un método general de estimación, consiste de los siguientes pasos:

1. Permutar la muestra y dividirla en k partes; cada parte V_j , j = 1,…, k tiene aproximadamente n/k observaciones. Los valores más usados de k son 3, 10 ó n; cuando k=n, el cálculo se llama PRESS
2. Estimar el modelo de regresión, excluyendo una j-ésima parte (j = 1,…,k)
3. Con el modelo estimado calcular las predicciones de las observaciones, que no fueron incluidas para estimar el modelo: $\hat{y}_i^{(j)}$, j = 1, …, k, tal que $\mathbf{x}_i \in V_j$
4. Calcular la suma de cuadrados de residuales (SCR) correspondiente:

$$SCR_j = \sum_{\{i:\mathbf{x}_i \in V_j\}} (y_i^{(j)} - \hat{y}_i^{(j)})^2 \ , \ j = 1, \ldots, k$$

5. El SCRvc promedio es calculado por $\frac{1}{n}\sum_{j=1}^{k} SCR_j$

El número de componentes PLS que minimiza la suma de cuadrados de residuales se elige de la siguiente manera:

- Con base en la matriz de predictoras $\mathbf{X}(n \times p)$ y el vector de clases $\mathbf{Y}(n \times 1)$, se halla la matriz de componentes o variables latentes $\mathbf{T}(n \times p)$
- Estimar el promedio de la suma de cuadrados de residuales PRESS o SCR_{VC} del modelo de regresión $\mathbf{Y}$ sobre las primeras h-componentes $\mathbf{T}_1$, ..., $\mathbf{T}_h$. Entonces PRESS(h), h = 1,…, p.
- El número de componentes PLS (h*), que serán utilizados es obtenido por la siguiente regla:

$$h^* = \min\{\, h > 1 : \ \text{PRESS}(h+1) - \text{PRESS}(h) > 0 \,\} \qquad (2.32)$$

Duckworth (1998) menciona un método de selección basado en el cálculo del PRESS (SCRvc) usando las h-primeras componentes PLS; es decir se debe calcular el PRESS(h),

para $h = 1, \ldots, p$. Usando la expresión (2.32) se determina PRESS(h*), el cual es un valor mínimo y finalmente se establece el valor conocido como F ratio$_h$

$$F\,\text{ratio}_h = \frac{PRESS(h)}{PRESS(h^*)} \qquad h = 1, \ldots, p \tag{2.33}$$

Entonces el número de componentes PLS se obtiene bajo el supuesto de que la variable aleatoria X tiene distribución F con (a, a) grados de libertad, donde a es el tamaño de la muestra de entrenamiento. El número de componentes PLS está dado por la siguiente regla:

$$h^{**} = \min\{\, h\colon \Pr(X < F\,\text{ratio}_h) < 0.75\} \tag{2.34}$$

Esposito Vinzi y Tenenhaus (2001) menciona un método propuesto por Wold en el software SIMCA que consiste en retener la componente $\mathbf{T}_h$ si el PRESS en el paso h, es significativamente más pequeña que el RESS (*Residual Sum of Squares*) en el paso h-1. Se retiene la h-ésima componente PLS si el índice de Stone-Geisser (Q^2) es al menos 0.0975. Es decir, retener T_h si $Q^2 > 0.0975$

$$Q^2 = 1 - \frac{PRESS(h)}{RESS(h-1)} \tag{2.35}$$

2.4 Clasificación

Es un problema de análisis multivariado que consiste en asignar individuos u objetos en uno de G grupos o clases. Para esto se hace uso de una función llamada clasificador, la cual se construye con base a los datos observados que conforman la muestra en estudio. Hay dos tipos de problemas de clasificación

- ***Clasificación supervisada*:** En este caso se dispone de un conjunto de observaciones multivariadas, para las cuales se conocen a priori las clases a las que pertenecen, es decir la variable respuesta está definida.

- ***Clasificación no supervisada*:** En este caso se dispone de un conjunto de observaciones multivariadas, pero no se conocen las clases a las que pertenecen. Aquí, no existe variable respuesta.

En esta tesis se usará solamente clasificación supervisada y los clasificadores que se consideran son los siguientes: Análisis Discriminante Lineal (LDA, por sus siglas en inglés), el clasificador usando los k-vecinos más cercanos (KNN) y la regresión logística nominal.

El análisis discriminante lineal es un clasificador que se construye bajo el supuesto de que cada uno de los G grupos tiene distribución normal multivariada con matriz de covarianzas común y vector de medias es diferentes en cada grupo. Dado un objeto $\mathbf{x}_0$, el procedimiento de clasificación lo ubicará en el grupo con mayor probabilidad posterior de clasificación, lo que debido a proporcionalidad es equivalente a decir que el objeto será ubicado en el grupo donde la función discriminante lineal $\delta_g(\mathbf{x}_0) = c + \beta' \mathbf{x}_0$ sea mayor, para $g = 1, \ldots, G$. Mayores detalles serán dados en la sección 4.2

El clasificador por k-vecinos más cercanos no requiere un modelo para ser ajustado. Para un objeto x_0, el procedimiento de clasificación sería: primero, hallar los k objetos que están a una distancia más cercana a x_0, usualmente k es un número impar; segundo si la mayoría de estos k objetos pertenece a una determinada clase o grupo, entonces el objeto x_0 también pertenece a ella. En caso de empate se clasifica al azar. Hay dos problemas en el método KNN, la elección de la distancia o métrica y la elección de k. La métrica más comúnmente usada es la euclideana, y usualmente es aplicada sobre datos reescalados para eliminar posibles problemas si las variables predictoras fueron medidas en unidades muy distantes entre sí.

El modelo de regresión logística surge para modelar la probabilidad posterior de los G grupos a través de una función lineal en $\mathbf{x}_0$, mientras que al mismo tiempo se asegura que la suma de estas probabilidades posteriores es uno. El modelo y mayores detalles se presentan en la sección 2.5

2.4.1 Tasa de error de clasificación

La tasa de error de clasificación es la probabilidad de que el clasificador clasifique mal una observación de la población a la cual pertenece la muestra usada para construir el clasificador. Existen varios métodos de estimar la tasa de error de clasificación; dos de ellos se describen a continuación:

- ***Estimación de la tasa de error por resustitución*** **(TE_{RES}) :** El método consiste en hallar un clasificador usando todas las observaciones que conforman la muestra; luego se clasifican estas mismas observaciones y por comparación con su verdadera clase se obtiene una proporción de observaciones mal clasificadas. Comparado con otros métodos de estimación de errores, éste es un método que encuentra un estimador demasiado optimista y puede conducir a falsas conclusiones si el tamaño de muestra no es muy grande comparado con el número de variables envueltas en el clasificador.

- ***Estimación de la tasa de error por validación cruzada*** **(TE_{VC}) :** El método consiste en dividir la muestra en r partes (usualmente $r = 10$) para estimar el modelo de clasificación usando todas menos una de las partes; luego se clasifican las observaciones que se dejaron de lado; el promedio de las clasificaciones erradas dará el estimado de la tasa de error por validación cruzada. Comparado con otros métodos de estimación de errores, este es un método que encuentra un estimador con poco sesgo, pero con bastante variabilidad.

2.5 Regresión logística

En Regresión Logística (Dobson, 2002), cada fila de la matriz de predictoras corresponde a las observaciones del vector aleatorio p-dimensional $\mathbf{x} = (x_1\ x_2 \cdots x_p)'$, las entradas del vector de respuestas **Y**, corresponde a la observación de la variable y, la cual representa una categoría, codificada dentro del conjunto $\{1, 2, \cdots, G\}$, que se llamará grupo o clase, para efectos de clasificación supervisada. Si la variable respuesta es categórica con dos clases (G = 2), se tiene el modelo de regresión logística dicotómico, definido de la siguiente manera:

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) \quad = \quad c + \beta_1 x_1 + \beta_2 x_2 + \ \cdots \ + \beta_p x_p \tag{2.36}$$

Una forma equivalente de representar el modelo anterior, es el siguiente:

$$P(y=1) \quad = \quad \frac{\exp\left(c + \beta_1 x_1 + \beta_2 x_2 + \ \cdots \ + \beta_p x_p\right)}{1 + \exp\left(c + \beta_1 x_1 + \beta_2 x_2 + \ \cdots \ + \beta_p x_p\right)} \tag{2.37}$$

Si la variable respuesta es categórica, con más de dos clases, el modelo de regresión logística es generalizado a Regresión Logística Nominal o Regresión Logística Ordinal.

2.5.1 Regresión Logística Ordinal

Este modelo es usado cuando hay un obvio orden natural en las categorías de la variable respuesta. Hay varios modelos diferentes en regresión logística ordinal; aquí será usado el llamado modelo de chances proporcionales. La probabilidad de clasificar una observación en una de las G clases, según este modelo, es obtenido de:

$$P(y \le g) \quad = \quad \frac{\exp\left(c_g + \beta_1 x_1 + \beta_2 x_2 + \ \cdots \ + \beta_p x_p\right)}{1 + \exp\left(c_g + \beta_1 x_1 + \beta_2 x_2 + \ \cdots \ + \beta_p x_p\right)} \tag{2.38}$$

$$g = 1, 2, \cdots, G-1$$

Se debe notar que P($y \leq G$) = 1. Además el modelo de regresión logística ordinal también puede ser presentado de la siguiente forma:

$$\log\left(\frac{P(y \leq g)}{P(y > g)}\right) = c_g + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (2.39)$$
$$g = 1, 2, \cdots, G-1$$

2.5.2 Regresión Logística Nominal

Este modelo es usado cuando no hay un orden natural en las categorías de la variable respuesta. Aquí una categoría es elegida arbitrariamente como la categoría de referencia. Supongamos que ésta es la primera categoría, entonces la probabilidad de clasificar una observación en una de las G clases es obtenida del modelo:

$$\log\left(\frac{P(y = g)}{P(y = 1)}\right) = c_g + \beta_{1g} x_1 + \beta_{2g} x_2 + \cdots + \beta_{pg} x_p \qquad (2.40)$$
$$g = 2, 3, \cdots, G$$

2.5.3 Predicción en regresión logística

Una vez que se han estimado los parámetros de la regresión logística, ordinal o nominal, se puede hacer la predicción de una observación $\mathbf{x} = (x_1\ x_2 \cdots x_p)'$, lo cual consiste en la clasificación de dicha observación en una de las G clases. Para lograr este objetivo se estiman las probabilidades de pertenecer a cada una de las G clases y se aplica la siguiente regla:

$$\mathbf{x} \in \text{clase } g^* \Leftrightarrow g^* = \arg\max_{g} \ \mathrm{P}(y = g) \qquad (2.41)$$

2.6 Otras técnicas relacionadas con el uso de PLS para clasificación supervisada

En el caso de problemas de clasificación con dos clases y donde hay un mayor número de variables predictoras que de observaciones se han tratado de hacer modificaciones a la

regresión logística para que ésta pueda seguir siendo usada. Hay tres opciones que se pueden escoger:

1) Usar los mismos métodos que se proponen para resolver el problema que ocurre cuando el número de predictoras es menor que el número de observaciones pero no es posible encontrar una estimación finita de los parámetros. La ocurrencia de estas situaciones dependen de la configuración de los puntos muestrales (ver Albert y Anderson, 1984). El método más usado es el de la penalidad de Firth (1993), el cual se basa en una modificación de la función *score* que aparece en el procedimiento de mínimos cuadrado reponderados usado para obtener los estimadores máximos verosímiles en un modelo lineal generalizado. Heinze and Schemper (2002) mostraron que la penalidad de Firth da estimaciones finitas de una logística binaria.

2) Reducir la dimensión del espacio de variables predictoras, usando métodos tales como el de Componentes Principales.

3) Maximizar el logaritmo de la función de verosimilitud bajo restricciones, introduciendo en dicha función un término de penalidad similar a lo que se hace en regresión "ridge" y es llamado regresión logística penalizada (Eilers y otros 2001). En este método se trabaja con todas las variables predictoras. A la función de verosimilitud se le resta una penalidad tipo "ridge" de tal manera que las estimaciones de los coeficientes no se vuelvan demasiado grandes. El problema aquí es que todas las variables predictoras intervienen en los cálculos y ello puede hacer lento el proceso de obtener las estimaciones, además de disminuir el rendimiento del clasificador.

Marx (1996) propuso una extensión del PLS para variables de respuesta categóricas en el contexto de regresión lineal generalizada (GLR, por sus siglas en inglés). Su método está basado en la sustitución de los dos ajustes por mínimos cuadrados del PLS por mínimos cuadrados reponderados iterativamente (IRLS, por sus siglas en inglés). Sin embargo Fort

y Lambert-Lacroix (2003) han observado que el algoritmo de Marx no necesariamente converge y lo muestran usando el conjunto de datos de Golub. Además la aplicación del método está sujeta a que las clases sigan una cierta distribución.

Fort and Lambert-Lacroix (2003) proponen un método llamado RIDGE-PLS que es la combinación de regresión logística penalizada ridge y PLS pero como las misma autoras lo indican sólo funciona en el caso de problemas de clasificación con dos clases.

En un trabajo reciente, Ding y Gentleman (2004) proponen una modificación del método de Marx, basándose en la penalidad de Firth para evitar soluciones infinitas en las estimaciones de los parámetros de la logística. Sin embargo su propuesta muestra varias inconsistencias.

Malthouse (1995) introdujo mínimos cuadrados parciales nolineales usando redes neurales del tipo FNN, pero esos modelos son aplicados exclusivamente a problemas de regresión y no de clasificación supervisada, que es nuestro interés.

Capítulo 3

Regresión Logística PLS

3.1 Introducción

En el capítulo 2 se introdujo la regresión por mínimos cuadrados parciales PLS1 y PLS2, correspondientes a regresión PLS univariada y regresión PLS multivariada, respectivamente; en ambos casos la metodología PLS soluciona el problema de regresión de pocas observaciones comparado con el número de variables predictoras y el problema de multicolinealidad. Es claro que para la aplicación de la metodología de regresión PLS, el vector o matriz de respuestas cdebe contener datos continuos. Cuando el vector de respuestas representa a una variable categórica codificada dentro del conjunto {1, 2,…, G}, el modelo de regresión que puede ser aplicado es la Regresión Logística (LR, por sus siglas en inglés), que al ser combinado con la metodología Regresión PLS se obtendrá la denominada Regresión Logística PLS (LRPLS), con el propósito de solucionar los mismos problemas existentes en el análisis de regresión PLS. En ese sentido Bastien, Esposito Vinzi y Tenenhaus (2002) usaron la Regresión Logística Ordinal (OLR) en más de dos clases, aplicable cuando hay un orden natural en las categorías de la variable respuesta y construyeron un algoritmo que calcula variables latentes para ser aplicadas en clasificación supervisada, dejando abierto el problema de la determinación del número óptimo de variables latentes necesarias.

Lo más común en clasificación supervisada es que las categorías de la variable respuesta no tengan un ordenamiento natural entre sí. Por este motivo un objetivo fundamental de este trabajo de tesis es desarrollar un algoritmo para regresión logística nominal PLS, (NLRPLS) aplicable cuando no hay un orden natural en las categorías de la variable

respuesta, lo que constituye el caso más real cuando se trabaja en clasificación supervisada.

Algunos investigadores en el campo de la clasificación supervisada que analizan bases de datos con muchas variables predictoras reducen la dimensionalidad de las mismas aplicando una metodología denominada Discriminante PLS, que consiste primero en aplicar regresión PLS usando el vector de respuestas de tipo categórico (grupos o clases) como si fuese de tipo cuantitativo y después en aplicar un clasificador sobre las variables latentes obtenidas en el paso anterior.

El primer paso considerado en Discriminante PLS, es inadecuado, ya que los datos del vector de respuestas son categóricos y la buena aplicación de regresión PLS, exige que los datos sean continuos. La LRPLS es una metodología adecuada para lograr reducir la dimensionalidad de predictoras en clasificación supervisada, ya que permite considerar variables de respuestas categóricas

3.2 Regresión Logística Ordinal PLS (OLRPLS)

Es un método introducido por Esposito-Vinzi y Tenenhaus (2001) y Bastien, Esposito-Vinzi y Tenenhaus (2002). La OLRPLS es la extensión de la regresión PLS, aplicable cuando la variable respuesta es categórica ordinal, y es usada en clasificación supervisada como una herramienta que soluciona problemas de fuerte multicolinealidad entre las variables predictoras y/o problemas de pequeño número de observaciones comparado con el número de variables.

La matriz de predictoras $\mathbf{X}(n \times p)$, es centrada y normalizada a la unidad por columnas (estandarizada); el vector de respuestas categóricas ordinal $\mathbf{Y}(n \times 1)$, no es alterado. El siguiente algoritmo formaliza la metodología propuesta por Bastien, Esposito-Vinzi y Tenenhaus (2002). Así como en regresión PLS de Wold (1975), no es de preocupación la validación del modelo de regresión logístico sino la obtención de la ponderación que relaciona al vector de clases y cada variable predictora. En esta sección se presenta una

modificación del algoritmo de estos autores que simplifica los cálculos en el proceso de actualización de la matriz de predictoras.

1. Entrada : $\mathbf{X}(n\times p)$, $\mathbf{Y}(n\times 1)$
2. Para $i = 1$ hasta p
3. Para $j = 1$ hasta p
4. Sea $\mathbf{X}_j$ la j-ésima columna de $\mathbf{X}$
5. Si $i = 1$ ➔ $w_j = coef(\mathbf{X}_j)$, modelo *OLR*: $\mathbf{Y}\sim \mathbf{X}_j$
6. Si $i > 1$ ➔ $w_j = coef(\mathbf{X}_j)$, modelo *OLR*: $\mathbf{Y}\sim\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{i-1}, \mathbf{X}_j$
7. End j
8. Normalizar $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$
9. $\mathbf{T}_i = \mathbf{X}\,\mathbf{w}$
10. Para $j = 1$ hasta p
11. $\mathbf{b}_j = [coef(\mathbf{T}_1) \ldots, coef(\mathbf{T}_i)]$, del modelo *RLI*: $\mathbf{X}_j \sim \mathbf{T}_1, \ldots, \mathbf{T}_i$
12. End j
13. $\mathbf{X} = \mathbf{X} - \mathbf{T}\,\mathbf{B}$
14. End i

Donde:
$\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_i]$ matriz de orden $n\times i$
$\mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_2\ \cdots\ \mathbf{b}_p]$ matriz de orden $i\times p$

Algoritmo 3.1 : Componentes PLS a partir de OLR (*OLRPLS*)

donde *OLR*: regresión logística ordinal y *RLI*: regresión lineal por mínimos cuadrados ordinarios.

3.2.1 Descripción del algoritmo OLRPLS

Con base en el algoritmo anterior se presenta una descripción del proceso. Se considera $\mathbf{X}(0)$, la matriz de predictoras de datos iniciales, estandarizadas por columnas; $\mathbf{X}(h\text{-}1)$, es la matriz de datos actualizada para calcular la h-ésima componente PLS. Básicamente el algoritmo OLRPLS, realiza los siguientes cálculos:

h-ésima componente PLS usando Regresión Logística Ordinal : $\mathbf{T}_h$

Supongamos que las componentes $\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_{h\text{-}1}$, fueron calculados en los h-1 pasos anteriores. Para calcular la componente $\mathbf{T}_h$, el algoritmo en estudio realiza lo siguiente :

1. Calcula el modelo estimado de OLR, de la variable categórica **Y** sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y la variable predictora $\mathbf{X}_j(h\text{-}1)$. El valor de w_j es el coeficiente de $\mathbf{X}_j$(h-1)

$$w_j = coef(\mathbf{X}_j), \text{ modelo } OLR\text{: } \mathbf{Y}\sim\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h\text{-}1}, \mathbf{X}_j(h\text{-}1) \quad , \quad j=1,\ldots,p \qquad (3.1)$$

El modelo de OLR estimado de **Y** sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y la variable $\mathbf{X}_j(h\text{-}1)$, consta de G-1 ecuaciones; la expresión (3.2), deja notar que los coeficientes de $\mathbf{X}_j(h\text{-}1)$ son los mismos; por tanto $w_j = b_h$, para $j=1,2,\ldots,p$

$$w_j: \begin{cases} \log\left(\dfrac{P(y\le 1)}{P(y>1)}\right) = c_{11} + b_1 T_1 + \quad \cdots \quad + b_{h-1} T_{h-1} + b_h X_j(h-1) \\ \log\left(\dfrac{P(y\le 2)}{P(y>2)}\right) = c_{21} + b_1 T_1 + \quad \cdots \quad + b_{h-1} T_{h-1} + b_h X_j(h-1) \\ \vdots \\ \log\left(\dfrac{P(y\le G-1)}{P(y>G-1)}\right) = c_{G-1,1} + b_1 T_1 + \quad \cdots \quad + b_{h-1} T_{h-1} + b_h X_j(h-1) \end{cases} \qquad (3.2)$$

Se obtiene $\mathbf{w}(h) = (w_1, w_2, \quad \cdots \quad, w_p)'$, que debe ser normalizado a la unidad.

2. Calcula la h-ésima componente PLS, usando los pesos $\mathbf{w}(h)$ obtenidos en el paso anterior.

$$\mathbf{T}_h = \mathbf{X}(h-1)\,\mathbf{w}(h) \qquad (3.3)$$

3. Actualiza la matriz de predictoras $\mathbf{X}(h)$, necesaria para hallar $\mathbf{T}_{h+1}$ mediante el análisis de regresión lineal múltiple (*RLI*) de cada variable predictora de $\mathbf{X}(h\text{-}1)$, sobre las componentes $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_h$,

$$\mathbf{b}_j = [coef(\mathbf{T}_1) \ldots, coef(\mathbf{T}_h)], \text{ del modelo } RLI\text{: } \mathbf{X}_j\text{(h-1)} \sim \mathbf{T}_1, \ldots, \mathbf{T}_h \qquad (3.4)$$

Cada vector de coeficientes $\mathbf{b}_j$ se obtiene desde cada uno de los siguientes modelos estimados de regresión lineal múltiple, sin término constante debido a que las variables que hacen de predictoras y respuesta están centradas, según las propiedades que serán vistas en la sección 3.4.

$$\begin{aligned}
\mathbf{b}_1 \quad &\Rightarrow \quad \hat{\mathbf{X}}_1(h-1) \quad = \quad b_{11}\,\mathbf{T}_1 + b_{21}\,\mathbf{T}_2 + \,\cdots\, + b_{h1}\,\mathbf{T}_h \\
\mathbf{b}_2 \quad &\Rightarrow \quad \hat{\mathbf{X}}_2(h-1) \quad = \quad b_{12}\,\mathbf{T}_1 + b_{22}\,\mathbf{T}_2 + \,\cdots\, + b_{h2}\,\mathbf{T}_h \\
&\vdots \\
\mathbf{b}_p \quad &\Rightarrow \quad \hat{\mathbf{X}}_p(h-1) \quad = \quad b_{1p}\,\mathbf{T}_1 + b_{2p}\,\mathbf{T}_2 + \,\cdots\, + b_{hp}\,\mathbf{T}_h
\end{aligned} \tag{3.5}$$

Los vectores originados están formados por los coeficientes de regresión de las variables latentes consideradas.

$$\mathbf{b}_1 = \begin{pmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{h1} \end{pmatrix} \;,\; \mathbf{b}_2 = \begin{pmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{h2} \end{pmatrix} \;,\; \ldots \;,\; \mathbf{b}_p = \begin{pmatrix} b_{1p} \\ b_{2p} \\ \vdots \\ b_{hp} \end{pmatrix} \tag{3.6}$$

Se obtiene la matriz $\mathbf{B} = \left(\mathbf{b}_1\,, \mathbf{b}_2\,, \cdots, \mathbf{b}_p\right)$ de orden $h \times p$ para actualizar la matriz de predictoras o matriz de residuales $\mathbf{X}(h)$ que será utilizada para calcular la próxima variable latente $\mathbf{T}_{h+1}$

$$\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - [\ \mathbf{T}_1\ ,\ \ldots\ ,\ \mathbf{T}_h\ ]\ \mathbf{B} \tag{3.7}$$

Usando la propiedad de ortogonalidad de los vectores latentes y una redefinición de la matriz **B**, para que pueda ser vista como un arreglo de vectores fila, como en la expresión (3.8) a continuación:

$$\mathbf{B} = \begin{pmatrix} \mathbf{a}_1(h) \\ \mathbf{a}_2(h) \\ \vdots \\ \mathbf{a}_h(h) \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{h1} & b_{h2} & \cdots & b_{hp} \end{pmatrix} \tag{3.8}$$

Se puede verificar que estos vectores fila $\mathbf{a}_i(h)$, $i = 1, \ldots, h$ pueden ser calculados de la siguiente manera:

$$\mathbf{a}_i(h) = \frac{\mathbf{T}_i' \mathbf{X}(h-1)}{\mathbf{T}_t' \mathbf{T}_i} \tag{3.9}$$

Reemplazando la expresión (3.8) en (3.7), se obtiene una forma equivalente de actualización de la matriz $\mathbf{X}(h)$, que aparece en el paso 13 del algoritmo 3.1

$$\mathbf{X}(h) = \mathbf{X}(h-1) - \mathbf{T}_1 \mathbf{a}_1(h) - \cdots - \mathbf{T}_{h-1} \mathbf{a}_{h-1}(h) - \mathbf{T}_h \mathbf{a}_h(h), \; h \geq 1 \tag{3.10}$$

El siguiente teorema demuestra que la actualización de la matriz de predictoras $\mathbf{X}(h)$ de la expresión (3.7), puede ser simplificado a $\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\mathbf{b}$, donde $\mathbf{b} = \mathbf{a}_h(h)$, es un vector fila de dimensión p.

Teorema 3.1 Dada la actualización de la matriz de predictoras $\mathbf{X}(h)$, como en la expresión (3.10), se cumple que: $\mathbf{a}_1(h) = \mathbf{a}_2(h) = \ldots = \mathbf{a}_{h\text{-}1}(h) = \mathbf{0}$, por lo tanto la actualización de la matriz de predictoras queda simplificada a la siguiente expresión:

$$\mathbf{X}(h) = \mathbf{X}(h-1) - \mathbf{T}_h \mathbf{a}_h(h) \;\; , h \geq 1 \tag{3.11}$$

Prueba

Usando inducción matemática sobre el número de iteraciones h

Para $h = 2$

De (3.10), $\mathbf{X}(2) = \mathbf{X}(1) - \mathbf{T}_1 \mathbf{a}_1(2) - \mathbf{T}_2 \mathbf{a}_2(2)$

De (3.9), $\mathbf{a}_1(2) = \mathbf{T}_1' \mathbf{X}(1) / \mathbf{T}_1' \mathbf{T}_1$

De (3.10), $\mathbf{X}(1) = \mathbf{X}(0) - \mathbf{T}_1 \mathbf{a}_1(1)$

De (3.9), $\mathbf{a}_1(1) = \mathbf{T}_1' \mathbf{X}(0) / \mathbf{T}_1' \mathbf{T}_1$

Se debe probar que $\mathbf{a}_1(2) = \mathbf{0}_{1\times p}$

$$\mathbf{a}_1(2) = \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\left(\mathbf{X}(0) - \mathbf{T}_1\,\mathbf{a}_1(1)\right)\right] \quad = \quad \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\left(\mathbf{X}(0) - \mathbf{T}_1 \;\; \mathbf{T}_1'\mathbf{X}(0)/\mathbf{T}_1'\mathbf{T}_1\right)\right]$$

$$\mathbf{a}_1(2) = \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\,\mathbf{X}(0) - \mathbf{T}_1'\,\mathbf{T}_1 \;\; \mathbf{T}_1'\mathbf{X}(0)/\mathbf{T}_1'\mathbf{T}_1\right] \quad = \quad \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\,\mathbf{X}(0) - \mathbf{T}_1'\mathbf{X}(0)\right] \quad = \quad \mathbf{0}_{1\times p}$$

Por lo tanto:

$$\mathbf{X}(2) \quad = \quad \mathbf{X}(1) - \mathbf{T}_2\,\mathbf{a}_2(2)$$

<u>Para $h = k$</u>

De (3.10), $\mathbf{X}(k) = \mathbf{X}(k-1) - \mathbf{T}_1\,\mathbf{a}_1(k) - \;\cdots\; - \mathbf{T}_{k-1}\,\mathbf{a}_{k-1}(k) - \mathbf{T}_k\,\mathbf{a}_k(k)$

Se cumple: $\mathbf{a}_1(k) = \mathbf{a}_2(k) = \ldots = \mathbf{a}_{k\text{-}1}(k) = \mathbf{0}$

Por lo tanto: $\mathbf{X}(k) = \mathbf{X}(k-1) - \mathbf{T}_k\,\mathbf{a}_k(k)$

<u>Para $h = k + 1$</u>

De (3.10), $\mathbf{X}(k+1) = \mathbf{X}(k) - \mathbf{T}_1\,\mathbf{a}_1(k+1) - \;\cdots\; - \mathbf{T}_k\,\mathbf{a}_k(k+1) - \mathbf{T}_{k+1}\,\mathbf{a}_{k+1}(k+1)$

Probar que: $\mathbf{a}_1(k+1) = \mathbf{a}_2(k+1) = \ldots = \mathbf{a}_k(k+1) = \mathbf{0}$

$$\mathbf{a}_1(k+1) = \frac{\mathbf{T}_1'\,\mathbf{X}(k)}{\mathbf{T}_1'\mathbf{T}_1} \quad = \quad \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\left(\mathbf{X}(k-1) - \mathbf{T}_k\,\mathbf{a}_k(k)\right)\right]$$

$$\mathbf{a}_1(k+1) = \quad \frac{1}{\mathbf{T}_1'\mathbf{T}_1}\left[\mathbf{T}_1'\,\mathbf{X}(k-1) - \mathbf{T}_1'\mathbf{T}_k\,\mathbf{a}_k(k)\right]$$

Por propiedad P3, de la sección 3.4, ortogonalidad de componentes PLS: $\mathbf{T}_1'\mathbf{T}_k = 0$

$$\mathbf{a}_1(k+1) = \quad \frac{\mathbf{T}_1'\,\mathbf{X}(k-1)}{\mathbf{T}_1'\mathbf{T}_1} = \mathbf{a}_1(k) = \mathbf{0}$$

Así sucesivamente, se cumple:

$$\mathbf{a}_k(k+1) = \frac{\mathbf{T}_k'\,\mathbf{X}(k)}{\mathbf{T}_k'\,\mathbf{T}_k} \quad = \quad \frac{1}{\mathbf{T}_k'\,\mathbf{T}_k}\left[\mathbf{T}_k'\,\left(\mathbf{X}(k-1) - \mathbf{T}_k\,\mathbf{a}_k(k)\right)\right]$$

$$\mathbf{a}_k(k+1) = \frac{1}{\mathbf{T}_k'\,\mathbf{T}_k}\left[\mathbf{T}_k'\,\mathbf{X}(k-1) - \mathbf{T}_k'\,\mathbf{T}_k\,\mathbf{a}_k(k)\right], \text{ pero de (3.9): } \mathbf{a}_k(k) = \frac{\mathbf{T}_k'\,\mathbf{X}(k-1)}{\mathbf{T}_k'\,\mathbf{T}_k}$$

$$\mathbf{a}_k(k+1) = \frac{1}{\mathbf{T}_k'\,\mathbf{T}_k}\left[\mathbf{T}_k'\,\mathbf{X}(k-1) - \mathbf{T}_k'\,\mathbf{X}(k-1)\right] = \mathbf{0} \qquad \blacksquare$$

El teorema anterior implica lo siguiente:

- La matriz **B**, de la expresión (3.8), queda simplificada de la siguiente manera:

$$\mathbf{B} = \begin{pmatrix} \mathbf{a}_1(h) \\ \mathbf{a}_2(h) \\ \vdots \\ \mathbf{a}_h(h) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_{h1} & b_{h2} & \cdots & b_{hp} \end{pmatrix} \tag{3.12}$$

- La actualización de la matriz de residuales del algoritmo 3.1, expresada en los pasos del 10 al 13, es simplificada por la siguiente expresión:

$$\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\,\mathbf{b} \tag{3.13}$$

Donde $\mathbf{b} = (b_{h1}\ \ b_{h2}\ \ \cdots\ \ b_{hp}) = \mathbf{T}_h'\,\mathbf{X}(h-1)/\mathbf{T}_h'\,\mathbf{T}_h = \mathbf{a}_h(h)$

3.3 Regresión Logística Nominal PLS (NLRPLS)

La metodología presentada en esta sección constituye el tema fundamental del presente trabajo. Se basa en la extensión de la OLRPLS y tiene el mismo objetivo de reducir la dimensionalidad de la matriz de datos. Es aplicable cuando no hay un orden natural en las categorías de la variable respuesta, lo cual es el caso más real cuando se trabaja en clasificación supervisada. En la aplicación de la regresión logística nominal se eligió la primera categoría como referencia, por tanto el modelo queda expresado como en (2.36).

La matriz de predictoras $\mathbf{X}(n\times p)$, es centrada y normalizada a la unidad por columnas y el vector de respuestas categóricas nominal $\mathbf{Y}(n\times 1)$, no es alterado. El siguiente algoritmo

calcula las componentes PLS, usando Regresión Logística Nominal (NLR), el cual considera la simplificación expresada en (3.13) para actualizar la matriz de predictoras.

1. Input : **X**(nxp) , **Y**(nx1)
2. Para i = 1 hasta p
3. Para j = 1 hasta p
4. Sea $\mathbf{X}_j$ la j-ésima columna de **X**
5. Si i = 1 ➔ modelo *NLR*: $\mathbf{Y} \sim \mathbf{X}_j$
   Sea *g** grupo con predicción máxima
   Si $g^* \neq 1 \Rightarrow w_j = coef(\mathbf{X}_j)$ desde $\log[P(y=g^*)/P(y=1)]$
   Si $g^* = 1 \Rightarrow w_j =$ promedio$[coef(\mathbf{X}_j)]$ desde la *G*-1 ecuaciones
6. Si i > 1 ➔ modelo *NLR*: $\mathbf{Y} \sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{i-1}, \mathbf{X}_j$
   Sea *g** grupo con predicción máxima
   Si $g^* \neq 1 \Rightarrow w_j = coef(\mathbf{X}_j)$ desde $\log[P(y=g^*)/P(y=1)]$
   Si $g^* = 1 \Rightarrow w_j =$ promedio$[coef(\mathbf{X}_j)]$ desde la *G*-1 ecuaciones
7. Fin *j*
8. Normalizar $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$
11. $\mathbf{T} = \mathbf{Xw}$
12. $\mathbf{b} = \mathbf{T'X/T'T}$
13. $\mathbf{X} = \mathbf{X} - \mathbf{T}\,\mathbf{b}$
14. Fin *i*

Algoritmo 3.2 : Componentes PLS a partir de NLR (*NLRPLS*)

3.3.1 Descripción del algoritmo NLRPLS

Con base en el algoritmo anterior se presenta una descripción del proceso. Se considera **X**(0), la matriz de predictoras de datos iniciales, estandarizadas por columnas; **X**(*h*-1), es la matriz de datos actualizada para calcular la *h*-ésima componente PLS. Básicamente el algoritmo realiza los siguientes cálculos:

h-ésima componente PLS usando Regresión Logística Nominal : $\mathbf{T}_h$

Supongamos que las componentes $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h-1}$, fueron calculados en los *h*-1 pasos anteriores. Para calcular la componente $\mathbf{T}_h$, el algoritmo en estudio realiza lo siguiente:

1. Calcula el modelo de RLN de la variable categórica $\mathbf{Y}$ sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y la variable predictora $\mathbf{X}_j(h\text{-}1)$. Sean $modelo_j$ y $predicción_j$ el modelo de NLR y el vector de predicción por resustitución, respectivamente

$$modelo_j = \mathbf{Y} \sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h-1}, \mathbf{X}_j(h\text{-}1) \quad ; \quad j = 1,\ldots, p \tag{3.14}$$
$$predicción_j = (\ 1\ 1\ 1\ \ldots\ \ldots\ 2\ 2\ 2\ \ldots\ \ldots\ G\ G\ G\ )$$

Al comparar el vector $predicción_j$ versus $\mathbf{Y}$, se puede definir $n_j(g)$ como el número de observaciones bien clasificadas dentro del grupo g , para g=1,2,…,G

En OLR, desde el modelo $\mathbf{Y}\sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h-1}, \mathbf{X}_j(h\text{-}1)$ se obtienen G-1 coeficientes para la variable $\mathbf{X}_j(h\text{-}1)$, los cuales son idénticos entre sí, por lo que tomar el coeficiente de esta variable es directo; este modelo está representado por la expresión (3.2). Sin embargo, en NLR, desde el modelo $\mathbf{Y}\sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h-1}, \mathbf{X}_j(h\text{-}1)$ se obtienen G-1 coeficientes diferentes $b_{1,h}$, $b_{2,h}$, …, $b_{G\text{-}1,h}$, para la variable $\mathbf{X}_j(h\text{-}1)$, esto es representado según modelo de la expresión (3.15).

$$w_j : \begin{cases} \log\left(\dfrac{P(y=2)}{P(y=1)}\right) = c_{11} + b_{11}T_1 + \cdots + b_{1,h-1}T_{h-1} + b_{1,h}X_j(h-1) \\ \log\left(\dfrac{P(y=3)}{P(y=1)}\right) = c_{21} + b_{21}T_1 + \cdots + b_{2,h-1}T_{h-1} + b_{2,h}X_j(h-1) \\ \vdots \\ \log\left(\dfrac{P(y=G)}{P(y=1)}\right) = c_{G-1,1} + b_{G-1,1}T_1 + \cdots + b_{G-1,h-1}T_{h-1} + b_{G-1,h}X_j(h-1) \end{cases} \tag{3.15}$$

2. El coeficiente w_j que será elegido de los $b_{1,h}$, $b_{2,h}$, …, $b_{G\text{-}1,h}$ posibles, depende del poder de predicción de la variable $\mathbf{X}_j(h\text{-}1)$. Sea g^* el grupo donde la predicción es máxima, es decir donde $n_j(g^*)$ es máximo, entonces el coeficiente w_j para j=1,…,p, es elegido según la siguiente regla:

$$w_j = \begin{cases} b_{g^*-1,h} & si \quad g^* = 2, 3, \cdots, G \\ \dfrac{b_{1,h} + \cdots + b_{G-1,h}}{G-1} & si \quad g^* = 1 \end{cases} \tag{3.16}$$

Según el modelo de NLR, dado por la expresión (3.15) el grupo 1 es considerado como grupo referencia; de ahí que cuando se produce predicción máxima en este grupo, el peso es obtenido como promedio de los G-1 coeficientes de $\mathbf{X}_j(h\text{-}1)$. Finalmente se obtiene $\mathbf{w}(h) = (w_1, w_2, \quad \cdots \quad, w_p)'$, normalizado a la unidad.

3. Calcular la h-ésima componente $\mathbf{T}_h$ de regresión logística nominal PLS, usando los pesos $\mathbf{w}(h) = (w_1, w_2, \ldots, w_p)'$, obtenidos en el paso anterior.

$$\mathbf{T}_h = \mathbf{X}(h\text{-}1)\, \mathbf{w}(h)$$

4. Actualizar la matriz de predictoras $\mathbf{X}(h)$, necesaria para hallar $\mathbf{T}_{h+1}$. En la sección 3.2, se demostró que este paso de actualización de matriz de predictoras $\mathbf{X}(h)$, es simplificado a $\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\, \mathbf{b}$, donde $\mathbf{b} = \mathbf{T}_h'\, \mathbf{X}(h-1) / \mathbf{T}_h'\, \mathbf{T}_h$

3.4 Propiedades de las componentes PLS

Estas propiedades son aplicables a las componentes PLS generados a partir de los modelos OLR, NLR, Regresión Logística con respuesta multivariada que será vista en la sección 3.5; así como desde los modelos de Análisis Discriminante Lineal y Regresión *Projection Pursuit*, que serán vistos en el próximo capítulo. Para enunciar las siguientes propiedades es necesario tener presente lo siguiente:

- $\mathbf{U}$ es un vector columna de unos, de dimensión n.
- $\mathbf{X}(0)$ de orden $n \times p$, es la matriz de predictoras, de datos iniciales, centrada y normalizada a la unidad por columnas; también se le denomina matriz estandarizada por columnas. Se cumple $\mathbf{X}'(0)\, \mathbf{U} \quad = \quad \mathbf{0}_{p\times 1}$

- La actualización de la matriz de residuales debe realizarse usando la versión simplificada, dada por la expresión (3.13)

P1. El h-ésimo vector latente $\mathbf{T}_h$, está centrado, es decir, la suma de sus elementos es cero. También, la matriz de predictoras siempre está centrada en cualquier iteración. Usando inducción matemática sobre el número de latentes

Para $i = 1$

$$\begin{aligned}\mathbf{T}_1' \mathbf{U} &= [\mathbf{X}(0)\ \mathbf{w}(1)]' \mathbf{U} \\ &= \mathbf{w}'(1)\mathbf{X}'(0)\ \mathbf{U} \\ &= \mathbf{w}'(1)\,[\mathbf{0}] = 0\end{aligned}$$

$$\begin{aligned}\mathbf{X}'(1)\ \mathbf{U} &= [\mathbf{X}(0) - \mathbf{T}_1\ \mathbf{b}(1)]'\ \mathbf{U} \\ &= \mathbf{X}'(0)\ \mathbf{U} - \mathbf{b}'(1)\ \mathbf{T}_1'\ \mathbf{U} \\ &= \mathbf{0} - \mathbf{b}'(1)\ 0 = \mathbf{0}\end{aligned}$$

Se cumple para $i = 1$. Asumiendo que la propiedad se cumple para $i = h$, se debe probar que se cumple para $i = h + 1$

Para $i = h$

$$\begin{aligned}&\mathbf{T}_h' \mathbf{U} = 0 \\ &\mathbf{X}'(h)\mathbf{U} = \mathbf{0}\end{aligned}$$

Para $i = h$+1

$$\begin{aligned}\mathbf{T}_{h+1}'\ \mathbf{U} &= [\mathbf{X}(h)\ \mathbf{w}(h+1)]'\ \mathbf{U} \\ &= \mathbf{w}'(h+1)\ \mathbf{X}'(h)\ \mathbf{U} \\ &= \mathbf{w}'(h+1)\ \mathbf{0}\ =\ 0\end{aligned}$$

$$\begin{aligned}\mathbf{X}'(h+1)\mathbf{U} &= [\mathbf{X}(h) - \mathbf{T}_{h+1}\ \mathbf{b}(h+1)]'\ \mathbf{U} \\ &= \mathbf{X}'(h)\ \mathbf{U} - \mathbf{b}'(h+1)\ \mathbf{T}_{h+1}'\ \mathbf{U} \\ &= \mathbf{0} - \mathbf{b}'(h+1)\cdot 0\ =\ \mathbf{0}\end{aligned}$$ ■

P2. En la h-ésima iteración, se cumple que el vector latente $\mathbf{T}_h$ es ortogonal con cada una de las columnas de la matriz de predictoras: $\mathbf{T}_h' \mathbf{X}(h) = \mathbf{0}_{1\times p}$

$$
\begin{aligned}
\mathbf{T}_h' \, \mathbf{X}(h) &= \mathbf{T}_h' \, [\mathbf{X}(h-1) - \mathbf{T}_h \, \mathbf{b}(h)] \\
&= \mathbf{T}_h' \, \mathbf{X}(h-1) - \mathbf{T}_h' \, \mathbf{T}_h \, \mathbf{b}(h) \\
&= \mathbf{T}_h' \, \mathbf{X}(h-1) - \mathbf{T}_h' \, \mathbf{T}_h \left(\frac{\mathbf{T}_h' \, \mathbf{X}(h-1)}{\mathbf{T}_h' \, \mathbf{T}_h} \right) \\
&= \mathbf{0} \qquad \blacksquare
\end{aligned}
$$

P3. Cada par de variables latentes son ortogonales, es decir el producto escalar de dos variables latentes cualesquiera es igual a cero.

$$
\mathbf{T}_1' \, \mathbf{T}_2 = \mathbf{T}_1' \, [\mathbf{X}(1) \, \mathbf{w}(2)] = \mathbf{T}_1' \, \mathbf{X}(1) \, \mathbf{w}(2) = \mathbf{0} \, \mathbf{w}(2) = 0
$$

$$
\begin{aligned}
\mathbf{T}_1' \, \mathbf{T}_3 &= \mathbf{T}_1' \, \mathbf{X}(2) \, \mathbf{w}(3) \\
&= \mathbf{T}_1' \, [\mathbf{X}(1) - \mathbf{T}_2 \, \mathbf{b}(2)] \, \mathbf{w}(3) \\
&= \mathbf{T}_1' \, \mathbf{X}(1) \, \mathbf{w}(3) - \mathbf{T}_1' \mathbf{T}_2 \, \mathbf{b}(2) \, \mathbf{w}(3) \\
&= \mathbf{0} \, \mathbf{w}(3) - 0 \, \mathbf{b}(2) \, \mathbf{w}(3) \\
&= 0 \\
&\vdots
\end{aligned}
$$

Esta propiedad es generalizada para dos variables latentes $\mathbf{T}_i$ y $\mathbf{T}_j$, tal que se cumple lo siguiente:

- $i < j$, siendo $j\text{-}i = m \Rightarrow i = j\text{-}m$
- $\mathbf{T}_i' \, \mathbf{T}_k = 0$, $\forall \; i < k < j$, siendo si $k\text{-}i < m$

$$
\begin{aligned}
\mathbf{T}_i'\,\mathbf{T}_j &= \mathbf{T}_i'\,\mathbf{X}(j-1)\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,[\mathbf{X}(j-2) - \mathbf{T}_{j-1}\,\mathbf{b}(j-1)]\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,\mathbf{X}(j-2)\,\mathbf{w}(j) - \mathbf{T}_i'\,\mathbf{T}_{j-1}\,\mathbf{b}(j-1)\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,\mathbf{X}(j-2)\,\mathbf{w}(j) - 0\,\mathbf{b}(j-1)\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,[\mathbf{X}(j-3) - \mathbf{T}_{j-2}\,\mathbf{b}(j-2)]\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,\mathbf{X}(j-3)\,\mathbf{w}(j) - \mathbf{T}_i'\,\mathbf{T}_{j-2}\,\mathbf{b}(j-2)\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\,\mathbf{X}(j-3)\,\mathbf{w}(j) - 0\,\mathbf{b}(j-2)\,\mathbf{w}(j) \\
&\;\;\vdots \\
&= \mathbf{T}_i'\,[\mathbf{X}(j-m) - \mathbf{T}_{j-m+1}\,\mathbf{b}(j-m+1)]\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\mathbf{X}(j-m)\mathbf{w}(j) - \mathbf{T}_i'\,\mathbf{T}_{j-m+1}\mathbf{b}(j-m+1)\mathbf{w}(j) \\
&= \mathbf{T}_i'\,\mathbf{X}(j-m)\mathbf{w}(j) - 0\,\mathbf{b}(j-m+1)\,\mathbf{w}(j) \\
&= \mathbf{T}_i'\mathbf{X}(j-m)\mathbf{w}(j) \;=\; \mathbf{0}\,\mathbf{w}(j) \\
&= 0 \qquad\qquad \blacksquare
\end{aligned}
$$

3.4.1 Matriz de transformación a componentes PLS

En análisis de componentes principales, la matriz que transforma variables predictoras en componentes principales, es la matriz ortogonal $\boldsymbol{\Gamma}$, dada en (2.4). En análisis PLS, la matriz que transforma variables predictoras en componentes PLS o variables latentes, puede ser hallada iterativamente. Sea $\mathbf{Z} = (\mathbf{z}_1 \ \ldots \ \mathbf{z}_p)$ de orden $p \times p$, la matriz que transforma variables predictoras en variables latentes

$$
\begin{aligned}
\mathbf{T} &= \mathbf{X}(0)\mathbf{Z} \qquad (3.16) \\
&= \mathbf{X}(0)\,(\mathbf{z}_1 \ldots \mathbf{z}_p) \\
\mathbf{T} &= [\mathbf{X}(0)\mathbf{z}_1 \ \ldots \ \mathbf{X}(0)\mathbf{z}_p] \qquad (3.17)
\end{aligned}
$$

En las expresiones (3.16) y (3.17), $\mathbf{X}(0)$ es la matriz de predictoras de datos iniciales, $\mathbf{T}(n \times p)$ es la matriz de componentes PLS, $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_p)$, siendo $\mathbf{T}_h = \mathbf{X}(0)\mathbf{z}_h$ la h-ésima componente PLS, para valores de $h = 1, \ldots, p$; esto es equivalente a la expresión (3.3), por lo tanto se debe cumplir que: $\mathbf{T}_h = \mathbf{X}(h\text{-}1)\mathbf{w}(h) = \mathbf{X}(0)\mathbf{z}_h$

Teorema 3.2 : La *h*-ésima componente PLS está dada por la siguiente expresión:

$$\mathbf{T}_h = \mathbf{X}(0)\mathbf{z}_h \tag{3.18}$$

Donde: $\mathbf{z}_h = \begin{cases} \mathbf{w}(1) \quad \text{para } h = 1 \\ [\mathbf{I} - \mathbf{z}_1\mathbf{b}(1) - \mathbf{z}_2\mathbf{b}(2) - \cdots - \mathbf{z}_{h-1}\mathbf{b}(h-1)]\,\mathbf{w}(h) \quad \text{para } h > 1 \end{cases}$

Prueba

Usando inducción matemática sobre el número de iteraciones *h*

Para *h* = 1

$$\mathbf{T}_1 = \mathbf{X}(0)\,\mathbf{w}(1) = \mathbf{X}(0)\,\mathbf{z}_1 \quad \Rightarrow \quad \mathbf{z}_1 = \mathbf{w}(1)$$

Para *h* = *k*

Se cumple: $\mathbf{T}_k = \mathbf{X}(0)\mathbf{z}_k$

Donde: $\mathbf{z}_k = [\mathbf{I} - \mathbf{z}_1\mathbf{b}(1) - \mathbf{z}_2\mathbf{b}(2) - \cdots - \mathbf{z}_{k-1}\mathbf{b}(k-1)]\,\mathbf{w}(k)$

Para *h* = *k* + 1

Se debe demostrar que $\mathbf{T}_{k+1} = \mathbf{X}(0)\mathbf{z}_{k+1}$

Donde: $\mathbf{z}_{k+1} = [\mathbf{I} - \mathbf{z}_1\mathbf{b}(1) - \mathbf{z}_2\mathbf{b}(2) - \cdots - \mathbf{z}_k\mathbf{b}(k)]\,\mathbf{w}(k+1)$

$$\begin{aligned}
\mathbf{T}_{k+1} &= \mathbf{X}(k)\,\mathbf{w}(k+1) \\
&= [\mathbf{X}(k-1) - \mathbf{T}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
&= [\mathbf{X}(k-1) - \mathbf{X}(0)\mathbf{z}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
&= [\mathbf{X}(k-2) - \mathbf{T}_{k-1}\,\mathbf{b}(k-1) - \mathbf{X}(0)\mathbf{z}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
&= [\mathbf{X}(k-2) - \mathbf{X}(0)\mathbf{z}_{k-1}\,\mathbf{b}(k-1) - \mathbf{X}(0)\mathbf{z}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
&\vdots \\
&= [\mathbf{X}(0) - \mathbf{X}(0)\mathbf{z}_1\,\mathbf{b}(1) - \cdots - \mathbf{X}(0)\mathbf{z}_{k-1}\,\mathbf{b}(k-1) - \mathbf{X}(0)\mathbf{z}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
&= \mathbf{X}(0)\,[\mathbf{I} - \mathbf{z}_1\,\mathbf{b}(1) - \cdots - \mathbf{z}_{k-1}\,\mathbf{b}(k-1) - \mathbf{z}_k\,\mathbf{b}(k)]\,\mathbf{w}(k+1) \\
\mathbf{T}_{k+1} &= \mathbf{X}(0)\mathbf{z}_{k+1} \qquad \blacksquare
\end{aligned}$$

El siguiente algoritmo calcula la matriz $\mathbf{Z} = (\mathbf{z}_1 \ \ldots \ \mathbf{z}_p)$ de orden $p \times p$, que transforma variables predictoras a componentes PLS. El algoritmo trabaja iterativamente y en cada iteración calcula una columna de **Z**. La matriz **I** es la identidad de orden $p \times p$

1. Input : $\mathbf{X}$(nxp) , $\mathbf{Y}$(nx1)
2. Para $h = 1$ hasta p
3. Calcular ponderaciones $\mathbf{w}(h)$, normalizado
4. Calcular $\mathbf{T}_h = \mathbf{X}(h\text{-}1)\mathbf{w}(h)$
5. Si $h = 1$ ➔ $\mathbf{z}_1 = \mathbf{w}(1)$
6. Si $h > 1$ ➔ $\mathbf{z}_h = [\mathbf{I} - \mathbf{z}_1\mathbf{b}(1) - \cdots - \mathbf{z}_{h-1}\mathbf{b}(h-1)]\,\mathbf{w}(h)$
7. $\mathbf{b}(h) = \mathbf{T}_h' \mathbf{X}(h-1) / \mathbf{T}_h' \mathbf{T}_h$
8. $\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\,\mathbf{b}(h)$
9. Fin h

Algorítmo 3.3 : Matriz de transformación a componentes PLS

El algoritmo 3.3, puede ser implementado para obtener la matriz de transformación a componentes PLS desde los modelos de OLR, NLR y Regresión Logística Multivariada, que será vista en la próxima sección; así como desde los modelos del Análisis Discriminante Lineal y Regresión *Projection Pursuit*, que serán vistos en el próximo capítulo.

En cada uno de los modelos mencionados anteriormente, se tiene bien definido el cálculo de $\mathbf{w}(h)$ y $\mathbf{T}_h$, que son el vector de ponderaciones y la componente PLS respectivamente, expresado en los pasos 3 y 4 del algoritmo 3.3. La actualización de la matriz de predictoras también está bien definida y está expresada en los pasos 7 y 8 del mismo algoritmo.

3.5 Regresión Logística PLS Multivariada (MLRPLS)

En esta sección se propone una extensión de la regresión PLS multivariada (PLS2), vista en la sección 2.3.3. El vector de respuestas categóricas $\mathbf{Y}(n\times1)$ que contiene G grupos o clases es presentado como una matriz de orden $n\times(G\text{-}1)$. La g-ésima columna está formada por "unos", si la observación pertenece a la g-ésima clase y "ceros", en caso contrario, para g =1, 2, …, G-1. El algoritmo 3.3 describe el cálculo de componentes PLS a partir de regresión logística como una extensión de la regresión PLS multivariada.

$$\mathbf{Y} = \begin{pmatrix} 2 \\ G-1 \\ 3 \\ \vdots \\ G \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} \equiv \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times(G-1)} \tag{3.19}$$

La matriz de predictoras **X** es centrada y normalizada a la unidad por columnas; el vector de respuestas categóricas **Y** es presentado como en la expresión (3.19), donde cada columna es dicotómica. Una vez más se hace notar que cuando la respuesta es dicotómica los modelos de regresión logística ordinal y nominal coinciden y es indiferente aplicar cualquiera de los dos modelos. El algoritmo se muestra a continuación:

1. Input $\mathbf{X}(n\times p)$, $\mathbf{Y}(n\times G)$
2. conteo = 0
3. Para k = 1 hasta $\lceil p/G\text{-}1 \rceil$; $\lceil\ \rceil$ es la función "ceiling", que redondea al entero superior
4. Para j = 1 hasta G-1
5. Sea **V** la j-ésima columna de **Y**
6. Para i =1 hasta p
7. Sea $\mathbf{X}_i$ la i-ésima columna de **X**
8. Si $k*j$ =1 ➔ $w_i = coef(\mathbf{X}_i)$, modelo RL: $\mathbf{V}\sim\mathbf{X}_i$
9. Si $k*j$ >1 ➔ $w_i = coef(\mathbf{X}_i)$, modelo RL: $\mathbf{V}\sim\mathbf{T}_1$, $\mathbf{T}_2$, … , $\mathbf{T}_{conteo}$, $\mathbf{X}_i$
10. Fin i
11. Normalizar $\mathbf{w} = (w_1\ w_2 \cdots w_p)'$
12. $\mathbf{T} = \mathbf{Xw}$
13. Si $k*j$ =1, modelo RL: $\mathbf{Y}\sim\mathbf{T}_1$ ➔ $\mathbf{V}_{\text{nuevo}} = \hat{\mathbf{Y}}[grupo\ 1]$
14. Si $k*j$ >1, modelo RL: $\mathbf{Y}\sim \mathbf{T}_1$, $\mathbf{T}_2$, … , $\mathbf{T}_{conteo}$, $\mathbf{T}$ ➔ $\mathbf{V}_{\text{nuevo}} = \hat{\mathbf{Y}}[grupo\ j]$
15. Si #diferencias($\mathbf{V}$,$\mathbf{V}_{nuevo}$) > 0 ➔ $\mathbf{V} = \mathbf{V}_{nuevo}$. Ir al paso 6
16. Si #diferencias($\mathbf{V}$,$\mathbf{V}_{nuevo}$) = 0 ➔ **T** es definitivo
17. $\mathbf{b} = (\mathbf{T}'\mathbf{X}) / (\mathbf{T}'\mathbf{T})$
18. $\mathbf{X} = \mathbf{X} - \mathbf{T}\,\mathbf{b}$
19. conteo = conteo +1
20. Si (conteo = p) ➔ Terminar
21. Fin j
22. Fin k

Algoritmo 3.4 : Componentes PLS, caso Multivariado (*MLRPLS*)

3.5.1 Descripción del algoritmo MLRPLS

Con base en el algoritmo anterior se presenta una descripción del proceso. Se considera **X**(0), la matriz de predictoras de datos iniciales, estandarizadas por columnas; **X**(*h*-1), es la matriz de datos actualizada para calcular la *h*-ésima componente PLS. Básicamente el algoritmo realiza los siguientes cálculos:

h-ésima componente de Regresión Logística PLS Multivariado : $\mathbf{T}_h$

Aquí la variable *conteo* indica el número de componentes PLS ya calculados en las iteraciones anteriores. Sea *conteo* = *h*-1, entonces las componentes $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$, fueron calculados en los *h*-1 pasos anteriores. Para calcular la componente $\mathbf{T}_h$, el algoritmo en estudio realiza lo siguiente:

1. Calcula los coeficientes de regresión logística de la variable dicotómica **V** sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y la variable predictora $\mathbf{X}_j(h\text{-}1)$. El valor de w_j es el coeficiente de $\mathbf{X}_j(h\text{-}1)$

$$w_j = coef(\mathbf{X}_j), \text{ modelo LR: } \mathbf{V}\sim\mathbf{T}_1 , \mathbf{T}_2 , \ldots, \mathbf{T}_{h\text{-}1} , \mathbf{X}_j(h\text{-}1) \quad ; \; j = 1,..,p \tag{3.20}$$

El valor de w_j es obtenido como coeficiente del modelo de regresión logística dicotómica estimado. Una vez más se puede verificar la coincidencia de los modelos de regresión logística ordinal y nominal con dos clases.

$$\begin{aligned}
w_1 &: \log\left(\frac{P(v=1)}{1-P(v=1)}\right) = c_1 + b_{11}T_1 + \quad \cdots \quad + b_{1,h-1}T_{h-1} + b_{1,h}X_1(h-1) \\
w_2 &: \log\left(\frac{P(v=1)}{1-P(v=1)}\right) = c_2 + b_{21}T_1 + \quad \cdots \quad + b_{2,h-1}T_{h-1} + b_{2,h}X_2(h-1) \\
&\vdots \\
w_p &: \log\left(\frac{P(v=1)}{1-P(v=1)}\right) = c_p + b_{p,1}T_1 + \quad \cdots \quad + b_{p,h-1}T_{h-1} + b_{p,h}X_p(h-1)
\end{aligned} \tag{3.21}$$

De donde $w_1 = b_{1,h}$, $w_2 = b_{2,h}$, …, $w_p = b_{p,h}$ y por lo que el vector de ponderaciones queda definido como $\mathbf{w}(h) = (w_1, w_2, \cdots, w_p)'$, que debe ser normalizado a la unidad.

2. Calcular la h-ésima componente PLS, $\mathbf{T}_h = \mathbf{X}(h-1)\,\mathbf{w}(h)$. Esta componente será provisional si el número de diferencias entre los vectores dicotómicos $\mathbf{V}$ y $\mathbf{V}_{nuevo}$ es mayor que cero, pero definitivo en caso contrario. El vector $\mathbf{V}_{nuevo}$ es calculado en el próximo paso.

3. Actualización del vector dicotómico $\mathbf{V}$. Se estima el modelo de regresión logística de $\mathbf{Y}$ de orden $n\times(G\text{-}1)$, sobre las variables $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$, $\mathbf{T}_h$ y se hace la predicción de $\mathbf{Y}$, el cual es representado por $\hat{\mathbf{Y}}$ de orden $n\times(G\text{-}1)$. La actualización del vector $\mathbf{V}$, denominado $\mathbf{V}_{\text{nuevo}}$, consiste en elegir la j-ésima columna de $\hat{\mathbf{Y}}$, que representa el j-ésimo grupo en estudio.

$$\text{Y}\sim \mathbf{T}_1 , \mathbf{T}_2 , \ldots , \mathbf{T}_{h\text{-}1} , \mathbf{T}_h \;\rightarrow\; \mathbf{V}_{\text{nuevo}} = \hat{\mathbf{Y}}[grupo\ j] \;\; ; \;\; j = 1,\ldots, G\text{-}1 \qquad (3.22)$$

4. Evaluar la convergencia de $\mathbf{V}$ y $\mathbf{V}_{nuevo}$. Dado que ambos son vectores dicotómicos que representan al mismo grupo, se puede establecer que ambos vectores convergen si el número de puntos discordantes, o también llamado suma de diferencias es cero, en caso contrario se redefine $\mathbf{V} = \mathbf{V}_{nuevo}$ y se debe volver al punto 1.

5. Actualización de la matriz de predictoras. Se ha demostrado que este paso de actualización de matriz de predictoras, $\mathbf{X}(h)$ es simplificado a $\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\,\mathbf{b}$, donde $\mathbf{b} = \mathbf{T}_h'\,\mathbf{X}(h-1)/\mathbf{T}_h'\,\mathbf{T}_h$.

3.6 Selección del número de componentes PLS

En esta sección se proponen algunos métodos que son propuestos por similitud a los utilizados en Regresión PLS y desarrollados en la sección 2.3.4 con el fin de determinar

el número de componentes PLS a partir de regresión logística nominal, así como las componentes PLS generados a partir de la función discriminante lineal y de la regresión *projection pursuit*; que serán vistos en el próximo capítulo.

a) Para estimar la Tasa de Error por Validación Cruzada (TE_{VC}). En la regla dada por la expresión (2.32) se puede modificar PRESS por la estimación de la tasa de error por Validación Cruzada (TEvc). Con base en la matriz de predictoras $\mathbf{X}(n\times p)$ y el vector de clases $\mathbf{Y}(n\times 1)$, se halla la matriz de componentes PLS; sobre ésta matriz y el vector de clases se estima la tasa de error por validación cruzada que consiste de los siguientes pasos:

6. Permutar la muestra y dividirla en k partes; cada parte V_j , j = 1, …, k tiene aproximadamente n/k observaciones. Los valores más usados de k son 3, 10 ó n.
7. Estimar el modelo de clasificación, excluyendo una j-ésima parte (j = 1,…,k)
8. Con el modelo de clasificación estimado se calculan las predicciones de las observaciones, que no fueron incluidas para estimar el modelo: $\hat{y}_i^{(j)}$, j = 1, …, k.
9. Calcular el número de malas clasificaciones correspondiente:

$$Error_j = \#diferencias(y_i^{(j)}, \hat{y}_i^{(j)}) \text{ , } j = 1, \ldots, k$$

10. El TEvc promedio es calculado por $\frac{1}{n}\sum_{j=1}^{k} Error_j$

El número de componentes PLS que minimiza la tasa error por validación cruzada se elige de la siguiente manera: Una vez determinadas las k componentes PLS por uno de los métodos establecidos se estima la tasa de error por validación cruzada para un clasificador, con las h-primeras componentes PLS, $TE_{VC}(h)$, para h = 1, 2, …, k, obteniéndose los valores $TE_{VC}(1)$, $TE_{VC}(2)$, …, $TE_{VC}(k)$. La metodología que se presenta considera los siguientes puntos, para lograr la reducción de la dimensionalidad

1. Estimar la tasa de error por validación cruzada $TE_{VC}(h)$, h = 1,…, p usando un clasificador sobre las h-primeras componentes, $\mathbf{T}_1$, ..., $\mathbf{T}_h$

2. Una vez que se han estimado las tasas de error por validación cruzada $TE_{VC}(h)$, para $h = 1, \ldots, p$; el número de componentes PLS que serán utilizados es obtenido por la siguiente regla:

$$h^* = \min\{\, h > 1 :\ TE_{VC}(h+1) - TE_{VC}(h) > 0 \,\} \tag{3.23}$$

b) Para estimar el valor de F ratio$_h$. Por similitud a la regla usada en regresión PLS se obtiene una regla reemplazando PRESS por TEvc en la expresión (2.33), por lo tanto:

$$F\,\text{ratio}_h = \frac{TEvc(h)}{TEvc(h^*)} \qquad h = 1, \ldots, p \tag{3.24}$$

donde TEvc(h*) es un valor mínimo obtenido desde la expresión (3.23). Entonces el número de componentes PLS se obtiene bajo el supuesto de que la variable aleatoria X tiene distribución F con (a, a) grados de libertad, donde a es el tamaño de la muestra de entrenamiento. El número de componentes PLS está dado por la siguiente regla:

$$h^{**} = \min\{\, h: \Pr(X < F\,\text{ratio}_h) < 0.75\} \tag{3.25}$$

c) Para Estimar del índice de Stone-Geisser (Q^2). Por similitud a la regla usada en regresión PLS se obtiene una regla reemplazando PRESS por TEvc y RESS (*Residual Sum of Squares*) por TE_{RES} (Tasa de error por resustitución), por lo tanto:

$$Q^2(h) = 1 - \frac{TE_{VC}(h)}{TE_{RES}(h-1)} \tag{3.26}$$

El número de componentes PLS está dado por la siguiente regla:

$$h^{**} = \min\{\, h: Q^2(h) > 0.975 \,\} \tag{3.27}$$

Capítulo 4

Otros métodos de obtención de componentes PLS para clasificación

4.1 Introducción

En el capítulo 3 se presentó una metodología para construir componentes PLS a partir de la Regresión Logística y ser usado en clasificación supervisada, lo cual constituye una extensión de la regresión PLS de Wold (1975). En este capítulo se proponen otros métodos para la obtención de componentes PLS, los cuales siguen siendo ortogonales y cumplen cada una de las propiedades presentadas en la sección 3.4. Las componentes PLS serán obtenidos a partir del Análisis Discriminante Lineal y desde la Regresión *Projection Pursuit*.

Los métodos para la construcción de componentes PLS se sintetizan en la búsqueda del vector de ponderaciones **w**, donde sus elementos resaltan la importancia de cada variable predictora en un modelo donde la variable respuesta es el vector o la matriz de clases. En Análisis Discriminante Lineal las ponderaciones se obtienen a partir de la función discriminante que es una cantidad directamente proporcional a la probabilidad posterior. En Regresión *Projection Pursuit*, las ponderaciones se obtienen a partir de los vectores de proyección en cada función *ridge*.

4.2 Análisis Discriminante Lineal (LDA)

El fundamento del Análisis Discriminante Lineal (LDA, por sus siglas en inglés) está basado en la Teoría de Decisión que necesita conocer la probabilidad posterior $P(y=g/\mathbf{x}=\mathbf{x}_0)$ es decir la probabilidad de clasificar algún vector de observaciones $\mathbf{x}_0$, en

una clase $g \in \{1, 2, \ldots, G\}$, con probabilidades a priori $P(y=g) = \Pi_g$ siendo $\sum_{g=1}^{G} \Pi_g = 1$.

Se supone que en cada clase g, la densidad $f_g(\mathbf{x})$ es normal multivariada con vector de medias $\boldsymbol{\mu}_g$ y matriz de covarianzas Σ, común para todas las clases.

$$f_g(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{1/2}} \; \exp\left\{ -\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_g)' \, \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_g) \right\} \tag{4.1}$$

La aplicación del Teorema de Bayes es necesaria para calcular la probabilidad posterior de clasificación

$$P(y = g / \mathbf{x} = \mathbf{x}_0) \quad = \quad \frac{f_g(\mathbf{x}_0) \Pi_g}{\sum_{g=1}^{G} f_g(\mathbf{x}_0) \, \Pi_g} \qquad g = 1, \ldots, G \tag{4.2}$$

La clasificación del vector observación $\mathbf{x}_0$ está dado por la siguiente regla:

$$\mathbf{x}_0 \in g^* \;\Leftrightarrow\; \arg\max_{g \in \{1, \cdots G\}} P(y = g^* / \mathbf{x} = \mathbf{x}_0) \tag{4.3}$$

En la expresión (4.2) se puede observar que el denominador es constante y por lo tanto la probabilidad posterior es una cantidad directamente proporcional a $f_g(\mathbf{x}_0) \Pi_g$

$$\begin{aligned}
P(y = g / \mathbf{x} = \mathbf{x}_0) \;&\propto\; f_g(\mathbf{x}_0) \Pi_g \\
&\propto\; \exp\left\{ -\tfrac{1}{2} (\mathbf{x}_0 - \boldsymbol{\mu}_g)' \, \Sigma^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_g) \right\} \Pi_g \\
&\propto\; \exp\left\{ \mathbf{x}_0' \, \Sigma^{-1} \boldsymbol{\mu}_g - \tfrac{1}{2} \, \boldsymbol{\mu}_g' \, \Sigma^{-1} \boldsymbol{\mu}_g \right\} \Pi_g \\
&\propto\; \exp\left\{ \mathbf{x}_0' \, \Sigma^{-1} \boldsymbol{\mu}_g - \tfrac{1}{2} \, \boldsymbol{\mu}_g' \, \Sigma^{-1} \boldsymbol{\mu}_g + \log \Pi_g \right\} \\
&\propto\; \exp\left\{ \mathbf{x}_0' \, \boldsymbol{\beta}_g + c_g \right\} = \exp\left\{ \delta_g(\mathbf{x}_0) \right\} \\
&\propto\; \delta_g(\mathbf{x}_0)
\end{aligned} \tag{4.4}$$

Por el resultado obtenido en (4.4), una regla de clasificación equivalente a la presentada en (4.3), es la siguiente:

$$\mathbf{x}_0 \in g^* \;\Leftrightarrow\; \arg\max_{g \in \{1, \cdots G\}} \left\{ \delta_{g^*}(\mathbf{x}_0) \right\} \tag{4.5}$$

En general, la expresión $\delta_g(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_g + c_g$, con $\boldsymbol{\beta}_g = \Sigma^{-1}\boldsymbol{\mu}_g$ y $c_g = -\frac{1}{2}\boldsymbol{\mu}_g\Sigma^{-1}\boldsymbol{\mu}_g + \log\Pi_g$, es llamada *función discriminante lineal*. Donde $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$ es un vector aleatorio, $\boldsymbol{\beta}_g = (\beta_{1g}, \beta_{2g}, \cdots, \beta_{pg})'$ es un vector de coeficientes y c_g es un término constante. Por tanto la función discriminante lineal para cada grupo *g*, queda expresada de la siguiente forma:

$$\delta_g(\mathbf{x}) = c_g + \beta_{1g}\, x_1 + \beta_{2g}\, x_2 + \cdots + \beta_{pg}\, x_p \quad ; \quad g = 1, 2, \ldots, G \qquad (4.6)$$

4.2.1 Componentes PLS a partir de LDA (LDAPLS)

El siguiente algoritmo calcula componentes PLS usando LDA, donde cada elemento del vector de ponderaciones $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$ es obtenido de los coeficientes de la función discriminante dada en (4.6). La matriz de predictoras $\mathbf{X}(n\times p)$, es centrada y normalizada a la unidad por columnas, el vector de respuestas categóricas nominal $\mathbf{Y}(n\times 1)$, no es alterado.

1. Input : **X**(nxp) , **Y**(nx1)
2. Para $i = 1$ hasta p
3. Para $j = 1$ hasta p
4. Sea $\mathbf{X}_j$ la j-ésima columna de **X**
5. Si i = 1 ➔ modelo LDA: $\mathbf{Y}\sim\mathbf{X}_j \Rightarrow \delta_1(\mathbf{x}), \delta_2(\mathbf{x}), \ldots, \delta_G(\mathbf{x})$
   g* = clase con el menor número de errores
   $w_j = coef(\mathbf{X}_j)$, en $\delta_{g*}(\mathbf{x})$
6. Si i > 1 ➔ modelo LDA: $\mathbf{Y}\sim\mathbf{T}_1, \ldots, \mathbf{T}_{h\text{-}1}, \mathbf{X}_j \Rightarrow \delta_1(\mathbf{x}), \delta_2(\mathbf{x}), \ldots, \delta_G(\mathbf{x})$
   g* = clase con el menor número de errores
   $w_j = coef(\mathbf{X}_j)$, en $\delta_{g*}(\mathbf{x})$
7. Fin j
8. Normalizar $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$
11. $\mathbf{T} = \mathbf{Xw}$
12. $\mathbf{b} = \mathbf{T'X/T'T}$
13. $\mathbf{X} = \mathbf{X} - \mathbf{T}\,\mathbf{b}$
14. Fin i

Algoritmo 4.1 : Componentes PLS a partir de LDA (LDAPLS)

4.2.2 Descripción del algoritmo LDAPLS

En base al algoritmo anterior se presenta una descripción del proceso. Se considera **X**(0), la matriz de predictoras de datos iniciales, estandarizadas por columnas; **X**(*h*-1), es la matriz de datos actualizada para calcular la *h*-ésima componente PLS. Básicamente el algoritmo realiza los siguientes cálculos:

h-ésima componente PLS usando LDA : $\mathbf{T}_h$

Supongamos que las componentes $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$, fueron calculados en los *h*-1 pasos anteriores. Para calcular la componente $\mathbf{T}_h$, el algoritmo en estudio realiza lo siguiente :

1. Calcular el modelo de LDA de la variable categórica **Y** sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y cada variable predictora $\mathbf{X}_j(h\text{-}1)$. Sean $modelo_j$ y $predicción_j$ el modelo de LDA y el vector de predicción por resustitución, respectivamente

$$modelo_j = \mathbf{Y} \sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h-1}, \mathbf{X}_j(h\text{-}1) \quad ; \quad j = 1,\ldots, p \qquad (4.7)$$
$$predicción_j = (\,1\ 1\ 1\ \ldots\ \ldots\ 2\ 2\ 2\ \ldots\ \ldots\ G\ G\ G\,)$$

   Al comparar el vector $predicción_j$ versus **Y**, se puede definir $n_j(g)$ como el número de observaciones bien clasificadas dentro del grupo *g* , para *g*=1,2,…,*G*

   En el modelo LDA por cada variable predictora $\mathbf{X}_j(h\text{-}1)$ se obtienen *G* coeficientes diferentes $b_{1,h}$, $b_{2,h}$, … , $b_{G,h}$, según modelo de la expresión (4.8).

$$w_j : \begin{cases} \delta_1(\mathbf{x}) = c_{11} + b_{11}T_1 + \cdots + b_{1,h-1}T_{h-1} + \quad b_{1,h}X_j(h-1) \\ \delta_2(\mathbf{x}) = c_{21} + b_{21}T_1 + \cdots + b_{2,h-1}T_{h-1} + \quad b_{2,h}X_j(h-1) \\ \quad \vdots \\ \delta_G(\mathbf{x}) = c_{G,1} + b_{G,1}T_1 + \cdots + b_{G,h-1}T_{h-1} + \ b_{G,h}X_j(h-1) \end{cases} \qquad (4.8)$$

3. El valor elegido como w_j es uno de los coeficientes de $\mathbf{X}_j(h\text{-}1)$: $b_{1,h}$, $b_{2,h}$, … , $b_{G,h}$, y depende del poder de predicción de la variable $\mathbf{X}_j(h\text{-}1)$ dentro del grupo g. Sea g^* el grupo donde la predicción es máxima, es decir donde $n_j(g^*)$ es máximo, entonces el coeficiente w_j es elegido desde la función $\delta_{g^*}(\mathbf{x})$, es decir $w_j = b_{g^*,h}$

3. Calcular la h-ésima componente PLS, $\mathbf{T}_h$, usando los pesos $\mathbf{w}(h) = (\, w_1, w_2, \ldots, w_p \,)'$, obtenidos en el paso anterior.

$$\mathbf{T}_h = \mathbf{X}(h\text{-}1)\, \mathbf{w}(h)$$

4. Actualizar la matriz de predictoras $\mathbf{X}(h)$, necesaria para hallar $\mathbf{T}_{h+1}$, de la misma forma que en los métodos anteriores

$$\mathbf{b}(h) = \mathbf{T}_h' \mathbf{X}(h-1) / \mathbf{T}_h' \mathbf{T}_h$$
$$\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\, \mathbf{b}(h)$$

4.3 Regresión Projection Pursuit (PPR)

En Regresión *Projection Pursuit*, (PPR, por sus siglas en inglés) de Friedman y Stuetzle (1981), la matriz de predictoras $\mathbf{X} = [\mathbf{x}]$, es de orden $n \times p$ y el vector de respuestas categórica $\mathbf{Y}$, de dimensión n, que contiene G grupos o clases, es presentado como una matriz de clases de orden $n \times G$, donde cada columna está formado por "unos", si la observación pertenece a la g-ésima clase y "ceros" en caso contrario. Cada fila de la matriz de predictoras corresponde a las observaciones del vector aleatorio p-dimensional $\mathbf{x} = (x_1\ x_2 \cdots x_p)'$; cada fila de la matriz de clases corresponde a un vector que representa una clase $\mathbf{y} = (y_1\ y_2 \cdots y_G)'$. El modelo PPR es el siguiente:

$$y_g = \bar{y}_g + \sum_{m=1}^{M} \beta_m^g\ \phi_m(\boldsymbol{\alpha}_m' \mathbf{x}) \quad ; \quad g = 1, 2, \ldots, G \tag{4.9}$$

donde:

$\bar{y}_g = \frac{1}{n}\sum_{i=1}^{n} y_{g,i}$: promedio de la *g*-ésima columna de **Y**

M : número de términos

ϕ_m : función predictora, *smooth* o función *ridge*

$\beta_m^g \in \boldsymbol{\beta}^g = (\beta_1^g, \beta_2^g, \cdots, \beta_M^g)$: coeficientes de cada función *ridge*

$\boldsymbol{\alpha}_m = (\alpha_1, \alpha_2, \cdots, \alpha_p)'$: vector de proyecciones, normalizado

La parte *projection* del término *Projection Pursuit*, indica que el vector de observaciones **x**, es proyectado sobre los vectores $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_M$ para conseguir las longitudes $\boldsymbol{\alpha}'\mathbf{x}$ de las proyecciones, y la parte *pursuit* indica que los vectores de proyecciones $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_M$ fueron hallados usando técnicas de optimización.

Más formalmente en regresión *projection pursuit*, se debe satisfacer el modelo de esperanza condicional

$$E[y_g \mid x_1, \cdots, x_p] = \mu_{y_g} + \sum_{m=1}^{M} \beta_m^g \, \phi_m(\boldsymbol{\alpha}_m'\mathbf{x}) \qquad (4.10)$$

donde $\mu_{y_g} = E(y_g)$. Las funciones *ridge* ϕ_m han sido estandarizadas, tienen media cero y varianza uno. Esto es:

$$E[\phi_m(\boldsymbol{\alpha}_m'\mathbf{x})] = 0 \quad , \quad E[\phi_m^2(\boldsymbol{\alpha}_m'\mathbf{x})] = 1 \quad ; \quad m = 1, 2, \ldots, M \qquad (4.11)$$

Los parámetros del modelo: β_m^g , ϕ_m , $\boldsymbol{\alpha}_m$ para m = 1, 2, …, M , dado en la expresión (4.10) minimizan el error cuadrático medio (ECM), sobre todo posible valor de β_m^g , ϕ_m , $\boldsymbol{\alpha}_m$.

$$ECM = E\left[\; y_g - \mu_g - \sum_{m=1}^{M} \beta_m^g \, \phi_m(\boldsymbol{\alpha}'\mathbf{x}) \; \right] \qquad (4.12)$$

Mas detalles acerca de la generación de ϕ pueden ser encontrados en el manual de S-plus 6 para Windows.

4.3.1 Componentes PLS a partir de PPR (PPRPLS)

El siguiente algoritmo calcula componentes PLS usando PPR, donde cada elemento del vector de ponderaciones $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$ es obtenido desde las componentes de los vectores de proyecciones $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_M$ del modelo PPR dado en la expresión (4.9). La matriz de predictoras $\mathbf{X}(n \times p)$, es centrada y normalizada a la unidad por columnas; el vector de respuestas categóricas nominal **Y** es presentado como matriz de clases y representado por la expresión (3.16). En la aplicación del modelo PPR se hace necesario definir el número de términos M.

1. Input : **X**(nxp) , **Y**(nx*G*)
2. Para i = 1 hasta p
3. Para j = 1 hasta p
4. Sea $\mathbf{X}_j$ la j-ésima columna de **X**
5. Si i = 1 ➔ modelo PPR: $\mathbf{Y} \sim \mathbf{X}_j$
   Si $M = 1 \Rightarrow w_j = coef(\mathbf{X}_j)$
   Si $M > 1 \Rightarrow w_j =$ promedio$[coef(\mathbf{X}_j)]$
6. Si i > 1 ➔ modelo PPR: $\mathbf{Y} \sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{i-1}, \mathbf{X}_j$
   Si $M = 1 \Rightarrow w_j = coef(\mathbf{X}_j)$
   Si $M > 1 \Rightarrow w_j =$ promedio$[coef(\mathbf{X}_j)]$
7. Fin *j*
8. Normalizar $\mathbf{w} = (w_1, w_2, \cdots, w_p)'$
11. **T** = **Xw**
12. **b** = **T'X/T'T**
13. **X** = **X** – **T b**
14. Fin *i*

Algoritmo 4.2 : Componentes PLS a partir de PPR (PPRPLS)

4.3.2 Descripción del algoritmo PPRPLS

Con base en el algoritmo anterior se presenta una descripción del proceso. Se considera **X**(0), la matriz de predictoras de datos iniciales, estandarizadas por columnas; **X**(*h*-1), es la matriz de datos actualizada para calcular la *h*-ésima componente PLS. Básicamente el algoritmo realiza los siguientes cálculos:

h-ésima componente PLS usando PPR : $\mathbf{T}_h$

Supongamos que los componentes $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$, fueron calculados en los h-1 pasos anteriores. Para calcular la componente $\mathbf{T}_h$, el algoritmo en estudio realiza lo siguiente :

1. Calcula el modelo de PPR de la matriz categórica $\mathbf{Y}$ sobre $\mathbf{T}_1$, $\mathbf{T}_2$, ..., $\mathbf{T}_{h\text{-}1}$ y la variable predictora $\mathbf{X}_j(h\text{-}1)$. Sea el modelo PPR en estudio

$$\mathbf{Y} \sim \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{h\text{-}1}, \mathbf{X}_j(h\text{-}1) \qquad ; \quad j = 1,\ldots, p \tag{4.13}$$

Cuando el número de términos es $M = 1$, el modelo PPR de la expresión (4.13) es de la siguiente forma:

$$\begin{aligned}
y_1 &= \bar{y}_1 + \beta_1^1 \, \phi_1(\alpha_1 T_1 + \cdots + \alpha_{h-1} T_{h-1} + \alpha_h X_j) \\
y_2 &= \bar{y}_2 + \beta_1^2 \, \phi_1(\alpha_1 T_1 + \cdots + \alpha_{h-1} T_{h-1} + \alpha_h X_j) \\
&\vdots \\
y_G &= \bar{y}_G + \beta_1^G \, \phi_1(\alpha_1 T_1 + \cdots + \alpha_{h-1} T_{h-1} + \alpha_h X_j)
\end{aligned} \tag{4.14}$$

Cuando el número de términos es $M > 1$, el modelo PPR de la expresión (4.13) es de la siguiente forma:

$$\begin{aligned}
y_1 &= \bar{y}_1 + \beta_1^1 \, \phi_1(\alpha_{11} T_1 + \cdots + \alpha_{1,h} X_j) + \cdots + \beta_M^1 \, \phi_M(\alpha_{M,1} T_1 + \cdots + \alpha_{M,h} X_j) \\
y_2 &= \bar{y}_2 + \beta_1^2 \, \phi_1(\alpha_{11} T_1 + \cdots + \alpha_{1,h} X_j) + \cdots + \beta_M^2 \, \phi_M(\alpha_{M,1} T_1 + \cdots + \alpha_{M,h} X_j) \\
&\vdots \\
y_G &= \bar{y}_G + \beta_1^G \, \phi_1(\alpha_{11} T_1 + \cdots + \alpha_{1,h} X_j) + \cdots + \beta_M^G \, \phi_M(\alpha_{M,1} T_1 + \cdots + \alpha_{M,h} X_j)
\end{aligned} \tag{4.15}$$

2. Calcula la ponderación w_j para $j = 1, \ldots, p$ como coeficiente de $\mathbf{X}_j(h\text{-}1)$ en el modelo PPR con M términos. En el modelo PPR con $M = 1$, dado por la expresión (4.14), el valor elegido como w_j es un valor único, el coeficiente de $\mathbf{X}_j(h\text{-}1)$, es decir es $w_j = \alpha_h$.

En el modelo PPR con $M > 1$, dado por la expresión (4.15), el valor elegido como w_j es un promedio de coeficientes de $\mathbf{X}_j(h\text{-}1)$, estos coeficientes son obtenidos desde cada uno de las M funciones *ridge*.

$$w_j = \frac{\alpha_{1,h} + \cdots + \alpha_{M,h}}{M} \tag{4.16}$$

3. Calcula la h-ésima componente PLS, $\mathbf{T}_h$, usando los pesos $\mathbf{w}(h) = (w_1, w_2, \ldots, w_p)'$, obtenidos en el paso anterior:

$$\mathbf{T}_h = \mathbf{X}(h\text{-}1)\, \mathbf{w}(h)$$

4. Actualiza la matriz de predictoras $\mathbf{X}(h)$, necesaria para hallar $\mathbf{T}_{h+1}$, de la misma forma que en los métodos anteriores,

$$\mathbf{b}(h) = \mathbf{T}_h' \mathbf{X}(h-1) / \mathbf{T}_h' \mathbf{T}_h$$
$$\mathbf{X}(h) = \mathbf{X}(h\text{-}1) - \mathbf{T}_h\, \mathbf{b}(h)$$

Capítulo 5

Metodología

5.1 Introducción

En este capítulo se presenta una descripción de las tareas fundamentales, tales como el manejo inicial de las bases de datos, necesarias para la aplicación de las metodologías propuestas; implementación de algoritmos conducentes a determinar las componentes PLS, lo que constituye el principal aporte de este trabajo, algoritmos que fueron necesarios para reducir la dimensionalidad de las matrices de datos; implementación de clasificadores, necesarios para medir el poder de clasificación cuando se trabaja con componentes PLS; determinación de la tasa de error de clasificación, necesaria para determinar qué algoritmo de componentes PLS transformó mejor a los datos iniciales; determinación del número de componentes, necesario para disminuir la dimensionalidad de la matriz de predictoras; y finalmente la tarea de implementación de programas, necesaria para probar los planteamientos teóricos de esta tesis.

5.2 Manejo de las bases de datos

Para la aplicación de las metodologías presentadas en esta tesis se utilizaron diez bases de datos reales que ya han sido analizadas por varios investigadores en el marco de la clasificación supervisada y en el uso de diferentes tipos de clasificadores. Estas bases de datos pueden clasificarse en dos grandes grupos, de acuerdo a la relación entre el número de predictoras (p) y número de observaciones (n).

1) Cuando $p < n$. Referido a bases de datos usuales, donde el número de observaciones es mayor que el número de predictoras. Estas bases de datos son: Sonar, Ionósfera, Heartc, Vehicle, Segment y Landsat.

2) Cuando $p >> n$. Referido a bases de datos obtenidas de experimentos en *microarrays*; estos datos se caracterizan por una gran cantidad de predictoras y un número muy pequeño de observaciones. Estas bases de datos son: Golub2, Colon, Breastcancer, Golub3.

El primer grupo de datos están a disposición en "*The Repository of Maching Learning Databases*" el cual es mantenido por el Departamento de Ciencias de Computadoras de la Universidad de California en Irvine, Blake y Merz (1998). El segundo grupo de datos está disponible en varios lugares en la Internet, así por ejemplo los datos Golub 2 y 3, están disponibles en la página del *Center for Genome Research* del Instituto Whitehead asociado al MIT, (http://www.broad.mit.edu/cancer/datasets.html)； estos datos han sido analizados usando varias técnicas estadísticas. Los datos Colon están disponibles en la página del *Gene Expresión Project* de la Universidad de Princeton, (http://microarray.princeton.edu/oncology) y fueron discutidos en Alon et al. (2000). Los datos Breastcancer están disponibles en la página del *Benedum Oncology Informatics Center* de la Universidad de Pittsburgh (http://www.upci.upmc.edu/facilities/cis) y fueron considerados por Hedenfalk et al. (2001)

Una descripción general de las bases de datos se presenta a continuación y en la Tabla 5.1

Sonar.- datos donde se clasifican las señales de sonar provenientes de un cilindro metálico y de una roca aproximadamente cilíndrica. Tiene 60 atributos numéricos

Ionosfera.- se toman observaciones de electrones libres en la ionosfera que muestran una "buena" o "mala" evidencia de algún tipo de estructura en la ionosfera. Hay 32 atributos numéricos.

Heartc.- datos de presencia o ausencia de enfermedad del corazón determinada por los resultados de varias pruebas médicas aplicadas a pacientes. Hay dos clases, siete atributos numéricos y seis atributos categóricos.

Vehicle.- datos de cuatro modelos de vehículos: un Double decker, Chevrolet furgoneta, Saab 9000 y un Opel Manta 400; de acuerdo a dieciocho atributos numéricos.

Segmentation.- datos de siete tipos de segmentación de imágenes al aire libre: ladrillo, cielo, follaje, cemento, ventana, camino y hierba. Hay dieciséis atributos numéricos.

Landsat.- datos de imágenes de satélite. Hay seis clases: suelos rojos, cosecha de algodón, suelo gris, suelo gris húmedo, suelo con vegetación de rastrojo y suelo gris muy húmedo. Hay treinta y seis atributos numéricos.

Golub2.- datos de experimentos en *microarrays*, obtenidos de dos tipos de cáncer *acute lymphoblastic leukemia* (ALL) y *acute myeloid leukemia* (AML) (Golub et al. 1999)

Golub3.- las mismas predictoras de Golub2; las respuestas son las mismas para el tipo de cáncer AML, pero dentro del tipo de cáncer ALL se distinguen dos subtipos de cáncer: T-*cell* ALL y B-*cell* ALL (Golub et al. 1999)

Tabla 5.1 Descripción de las bases de datos en estudio

Nombre	**Objetos**	**Predictoras**	**Clases**	**Descripción**
Sonar	208	60	2	Señales de sonar
Ionosfera	351	32	2	Estructuras en la ionosfera
Heartc	297	13	2	Enfermedad del corazón
Vehicle	846	18	4	Modelos de autos
Segment	2310	16	7	Segmentación de Imágenes
Landsat	4435	36	6	Imágenes de satélite
Golub2	72	3571	2	Microarrays
Colon	62	2000	2	Microarrays
Breastcancer	22	3226	3	Microarrays
Golub3	72	3571	3	Microarrays

Colon.- datos de experimentos en *microarrays*, obtenidos de dos tipos de tejidos de colon: tumor y normal (Alon et al. 2000)

Breastcancer.- datos de experimentos en *microarrays*, obtenidos de dos clases de cáncer de mamas de tipo hereditario: BRCA1 y BRCA2; y de una clase de cáncer esporádico (Hedenfalk et al. 2001)

5.3 Cálculo de componentes PLS

La aplicación de la metodología propuesta en esta tesis, consiste en transformar cada matriz de datos analizada en una matriz de componentes PLS, los cuales tienen la propiedad de ser no correlacionados, esta característica es importante para el mejoramiento de la predicción, eliminación de multicolinealidad y la reducción de la dimensionalidad de la matriz de predictoras. En esta tesis se extiende la metodología propuesta por Esposito-Vinzi y Tenenhaus (2002), consistente en la construcción de componentes PLS a partir de regresión logística ordinal (OLR). También se presentan otras metodologías que constituyen el aporte de esta tesis, tales como: componentes PLS a partir de regresión logística nominal (NLR) y componentes PLS multivariado a partir de regresión logística (LR) y que fueron presentados en el capítulo 3. En el capítulo 4, se presentan otras metodologías para la construcción de componentes PLS, que también constituyen el aporte de la tesis, tales como: componentes PLS a partir del análisis discriminante lineal (LDA) y componentes PLS a partir de regresión *projection pursuit* (PPR). En resumen, se presentan cinco formas de transformar matrices de datos en matrices de componentes PLS, estas son:

1) Componentes PLS a partir de OLR
2) Componentes PLS a partir de NLR
3) Componentes PLS multivariado a partir de LR
4) Componentes PLS a partir de LDA
5) Componentes PLS a partir de PPR

Cada uno de estos métodos anteriores calcula uno a uno cada componente, a través de algoritmos iterativos. También se presenta un algoritmo que consiste en hallar una matriz de transformación a componentes PLS, con lo que la matriz de componentes PLS queda definida por la multiplicación de la matriz de datos iniciales y la matriz de transformación.

5.4 Aplicación de clasificadores

Los clasificadores que se utilizaron fueron el discriminante lineal (LDA), los *k*-vecinos mas cercanos (KNN) , para valores de $k = 1, 3, 5$ y la regresión logística nominal (NLR).

Estos clasificadores fueron aplicados a las bases de datos originales y a las bases de datos transformados en componentes PLS, teniendo presente cada uno de los métodos de transformación.

5.5 Determinación de la tasa de error de clasificación

Para determinar la tasa de error de clasificación se usó el método de validación cruzada, dividiendo a la matriz de datos en 10 partes. La aplicación del método de validación cruzada implica permutar los objetos de la matriz de datos, lo que origina un resultado diferente en cada aplicación; por tal motivo el proceso consistió en repetir 20 veces la validación cruzada, a fin de obtener resultados más confiables. Finalmente se promediaron los errores estimados que corresponden a la tasa de error de clasificación por validación cruzada 10 y además se calculó la desviación estandar como medida de variabilidad. Se empleó la metodología descrita en la sección 3.6

5.6 Determinación del número de componentes PLS

Con la seguridad de que la metodología PLS reduce la dimensionalidad de la matriz de predictoras de una base de datos, se generan sólo un número k de componentes PLS, el cual es mucho menor que el número total de predictoras. Una vez generado las k componentes PLS por uno de los métodos establecidos, se estiman las tasas de errores por validación cruzada a partir de un clasificador (LDA, KNN, LR) y las h-primeras componentes PLS, $TE_{VC}(h)$ para $h = 1, 2, \ldots, k$, obteniéndose los valores $TE_{VC}(1)$, $TE_{VC}(2)$, …, $TE_{VC}(k)$. La metodología que se presenta considera los siguientes puntos, para lograr la reducción de la dimensionalidad

- Con base en la matriz de predictoras $\mathbf{X}(n \times p)$ y el vector de clases $\mathbf{Y}(n \times 1)$, se halla la matriz de componentes o variables latentes $\mathbf{T}(n \times p)$
- Se estima la tasa de error por validación cruzada $TE_{VC}(h)$, $h = 1,\ldots, p$ usando un clasificador: LDA, KNN o NLR sobre las h-primeras componentes, $\mathbf{T}_1, ..., \mathbf{T}_h$

- Una vez que se han estimado las tasas de error por validación cruzada $TE_{VC}(h)$, para $h = 1, 2, \ldots, p$; el número de componentes PLS que serán utilizados es obtenido por la regla dada en la expresión (3.23)

Respecto a la regla dada por la expresión (3.25), acerca del cálculo del índice $F\ ratio_h$, se puede mencionar que no pudo ser implementada debido a que las tasas de error en los datos de *microarrays* alcanzan un valor mínimo del 0.00% y al ser reemplazada en la expresión (3.24) se obtienen un valor indeterminado.

Respecto a la regla dada por la expresión (3.27), acerca del cálculo del índice de *Stone-Geisser*, se puede mencionar que fue implementada pero se obtuvieron malos resultados debido a que la expresión (3.26) valores muy grandes y negativos.

5.7 Implementación de programas

Para poder aplicar las metodologías propuestas en este trabajo fue necesario la implementación de diversas funciones que puedan ser integradas en una librería y que permitan llevar a cabo los cálculos necesarios para realizar las tareas computacionales requeridas para probar los planteamientos teóricos de esta tesis. Las funciones implementadas corresponden a la puesta en marcha de los diferentes algoritmos propuestos en esta tesis para la generación de las componentes PLS. De la misma forma se implementaron funciones para calcular la matriz de transformación y para la determinación de la tasa de error por validación cruzada usando cada uno de los clasificadores en estudio.

La programación se llevó a cabo usando el lenguaje R, en el ambiente Windows, en computadoras con doble procesador Pentium Xeon corriendo a 3.06 GHz y con 3 GB de memoria RAM. Se hicieron uso de funciones propias del lenguaje R, tales como: lda, knn, ppr, multinom y lrm; obtenidas de las librerias MASS, class, base, nnet y Design, respectivamente.

Capítulo 6

Aplicación y Resultados

6.1 Introducción

Con la finalidad de probar la funcionalidad de los algoritmos propuestos se procedió a realizar el trabajo experimental siguiendo las metodologías planteadas en los capítulos 3 y 4. La aplicación práctica tiene dos fases:

1. Generación de las componentes PLS, por cada una de las cinco técnicas expuestas en los capítulos 3 y 4.
2. Aplicación de los clasificadores a las componentes PLS generados, como si estos fueran las predictoras. Los clasificadores usados son: LDA, KNN y NLR.

Se calcula la tasa de error de clasificación por validación cruzada (TE_{VC}) para cada clasificador en estudio y además, se determina el número de componentes PLS que son necesarios para concretizar la reducción de la dimensionalidad de la matriz de predictoras. A continuación se presenta un ejemplo de la salida del programa computacional que calcula 10 componentes PLS a partir de NLR, de los datos Ionosfera y luego se aplica el clasificador KNN, considerando un vecino más cercano (K=1)

```
> datos: Ionosfera
> componentes PLS desde NLR
> clasificador KNN (K=1)

$Tasa de errores por Validación Cruzada
 [1] 26.42450 16.05413 14.37322 10.69801 10.85470 10.51282 10.51282
 [8] 10.32764 10.88319 11.60969

$Desviación estandar
 [1] 1.1009916 0.8435064 0.6501656 0.4283489 0.3799920 0.4612440
 [7] 0.4518872 0.3682883 0.4195316 0.4015823
```

En esta salida se observa que el clasificador KNN aplicado a las cuatro primeras componentes PLS rinde una tasa de error por validación cruzada de 10.70%. Este valor es obtenido como promedio de 20 repeticiones del método de validación cruzada con una desviación estandar de 0.43%. Según lo explicado en la sección 5.6 la dimensionalidad de los datos ionosfera queda reducida a cuatro componentes PLS que tienen la propiedad de ser ortogonales. A continuación se muestra la matriz de correlaciones de las cuatro componentes PLS de los datos ionosfera.

```
> datos: ionosfera
> matriz de correlaciones: 4 primeros componentes PLS
               V1           V2           V3            V4
V1  1.000000e+00 4.769495e-17 9.667338e-18 -3.772006e-18
V2  4.769495e-17 1.000000e+00 1.440581e-16  1.767344e-16
V3  9.667338e-18 1.440581e-16 1.000000e+00  9.823815e-17
V4 -3.772006e-18 1.767344e-16 9.823815e-17  1.000000e+00
```

Para tener una idea general de la eficiencia de cada unos de los clasificadores y del comportamiento de los datos en estudio se procedió a realizar la clasificación usando todas las predictoras, en su estado original. Se determinó la tasa error de clasificación por validación cruzada de cada uno de los clasificadores. Estos resultados se presentan en la Tabla 6.1, en la que resalta la ausencia de resultados para datos de *microarrays* cuando se usa el clasificador NLR, esto debido a la limitación computacional de trabajar con muchas predictoras de este clasificador.

Para tener un punto de comparación de la metodología propuesta en cuanto a la reducción de la dimensionalidad, en la Tabla 6.2 se presentan las tasas de error por validación cruzada de los clasificadores sobre las componentes principales de los datos en estudio, obtenidas usando la función *prcomp* de R. Así en los datos: Sonar, Ionosfera y Heartc se usaron 10, 10 y 7 componentes, los cuales explican el 73.9%, 77.2% y 74.9% de la variabilidad total, respectivamente. En los datos: Vehicle, Segment y Landsat se usaron 8 componentes, los cuales explican el 97.5%, 96.6% y 97.1% de la variabilidad total, respectivamente. Finalmente, en los datos: Golub2, Colon, Golub3 y Breastcc se usaron 10 componentes, los cuales explican el 48.2%, 81.4%, 48.2% y 71.8% de la variabilidad total, respectivamente.

Tabla 6.1 TE_{VC} usando todas las predictoras originales*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	25.50 (1.69)	17.69 (0.79)	18.97 (0.95)	19.33 (1.11)	25.55 (1.75)
Ionosfera	14.33 (0.62)	12.98 (0.54)	15.36 (0.63)	15.44 (0.54)	16.24 (1.04)
Heartc	16.68 (0.54)	41.98 (1.06)	37.29 (1.22)	34.04 (1.04)	16.80 (0.51)
Golub2	2.22 (1.05)	1.67 (0.73)	2.01 (0.95)	2.08 (0.96)	--
Colon	21.77 (1.99)	20.48 (1.74)	16.85 (2.59)	17.10 (1.69)	--
Golub3	5.63 (1.23)	3.13 (1.09)	4.72 (1.05)	3.33 (0.95)	--
Breastcc	47.95 (7.15)	45.45 (2.95)	48.64 (2.14)	46.36 (5.23)	--
Vehicle	21.99 (0.55)	35.09 (0.75)	34.69 (0.68)	35.14 (0.91)	20.00 (0.59)
Segment	8.53 (0.10)	3.53 (0.18)	4.71 (0.25)	5.71 (0.20)	4.89 (0.18)
Landsat	15.66 (0.11)	9.62 (0.18)	9.36 (0.19)	9.52 (0.23)	18.87 (0.47)

* *valor entre paréntesis: desviación estándar*

Tabla 6.2 TE_{VC} usando componentes principales*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar [10]	21.13 (1.37)	12.07 (0.88)	14.88 (1.48)	15.91 (1.26)	21.25 (0.93)
Ionosfera [10]	15.41 (0.37)	9.53 (0.57)	11.57 (0.53)	13.09 (0.41)	16.47 (0.64)
Heartc [7]	15.45 (0.27)	22.26 (1.16)	19.34 (1.05)	17.14 (0.79)	16.14 (0.52)
Golub2 [10]	5.14 (0.65)	5.76 (0.82)	5.28 (1.53)	7.64 (1.46)	10.21 (1.70)
Colon [10]	15.48 (1.52)	33.79 (2.31)	23.79 (2.61)	23.71 (3.14)	17.58 (2.09)
Golub3 [10]	5.21 (0.76)	7.01 (1.39)	8.06 (1.72)	9.72 (2.02)	10.28 (1.52)
Breastcc [10]	38.18 (6.66)	37.05 (6.13)	42.27 (5.92)	49.32 (5.95)	33.41 (8.11)
Vehicle [8]	35.24 (0.64)	32.67 (0.66)	31.60 (0.88)	29.32 (0.86)	33.71 (0.59)
Segment [8]	15.92 (0.15)	3.45 (0.11)	4.39 (0.21)	5.14 (0.14)	10.88 (0.16)
Landsat [8]	17.13 (0.08)	10.40 (0.12)	9.56 (0.16)	9.47 (0.15)	15.57 (0.09)

* *valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes principales

6.2 TE_{VC} usando componentes PLS a partir de OLR

La generación de componentes PLS a partir de la regresión logística ordinal, es la aplicación del método propuesto por Esposito-Vinzi y Tenehaus (2001), ellos aplicaron su metodología a una matriz de datos de orden 34×4, las 4 variables predictoras son cuantitativas y las clases son tres tipos de calidad de vino: malo, regular y bueno. Una vez construidos las componentes PLS, aplicaron el clasificador logístico ordinal.

Con la finalidad de observar los resultados del algoritmo que genera componentes PLS a partir de OLR se asumió que las clases de los datos en estudio son categóricas ordinales, Las tasas de errores son mostradas en la Tabla 6.3, se puede observar que en general los resultados dejan ver el éxito de la reducción de la dimensionalidad usando componentes PLS. Las componentes PLS a partir de los modelos OLR y NLR coinciden para conjuntos de datos con dos clases

Tabla 6.3 TE_{VC} usando componentes PLS a partir de OLR*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	14.47 (1.06) [4]	17.04 (0.94) [4]	13.12 (0.83) [4]	12.43 (1.32) [4]	13.00 (0.79) [4]
Ionosfera	11.65 (0.52) (5)	10.84 (0.65) [4]	9.36 (0.36) [5]	10.89 (0.44) [2]	10.84 (0.52) [6]
Heartc	15.28 (0.17) [2]	23.70 (0.92) [3]	17.39 (1.06) [5]	16.97 (0.53) [3]	15.69 (0.28) [2]
Golub2	0.00 (0.00) [2]	0.00 (0.00) [2]	0.00 (0.00) [2]	0.28 (0.57) [3]	0.69 (0.71) [4]
Colon	2.58 (0.81) [5]	5.48 (1.32) [5]	7.10 (0.81) [5]	8.14 (0.64) [5]	1.93 (1.70) [4]
Golub3	1.46 (0.31) [2]	0.14 (0.43) [3]	0.42 (0.65) [3]	1.60 (0.82) [3]	3.33 (0.95) [4]
Brestcc	0.00 (0.00) [2]	0.00 (0.00) [2]	1.82 (2.28) [2]	9.77 (2.22) [3]	1.14 (2.02) [2]
Vehicle	22.85 (0.47) [13]	29.94 (0.61) [7]	30.43 (0.74) [5]	28.27 (0.58) [12]	24.07 (0.44) [10]
Segment	8.46 (0.09) [9]	5.51 (0.18) [7]	6.80 (0.16) [9]	6.82 (0.19) [6]	4.66 (0.18) [8]
Landsat	16.34 (0.07) [9]	10.54 (0.18) [12]	10.06 (0.18) [9]	10.08 (0.15) [9]	14.54 (0.14) [9]

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

6.3 TE_{VC} usando componentes PLS a partir de NLR

La generación de componentes PLS a partir de la regresión logística nominal, es la aplicación de uno de los métodos propuestos en esta tesis en la sección 3.3, siguiendo el algoritmo 3.2. Los clasificadores en estudio fueron aplicados a los datos en estudio y las tasas de errores son mostradas en la Tabla 6.4, de la cual se puede hacer los siguientes comentarios:

1) En comparación con los resultados obtenidos en las Tablas 6.1 y 6.2 se observa que en cada conjunto de datos, las componentes PLS a partir de NLR logran reducir la tasa de error usando sólo algunas componentes PLS con los tres clasificadores. La comparación que sobresale es la tasa de error de Breastcc, usando el clasificador LDA; con sólo cinco componentes PLS es 0.00%, con todas las predictoras es alrededor de 47% y con 10 componentes principales es alrededor de 38%.

Tabla 6.4 TE_{VC} usando componentes PLS a partir de NLR*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	14.42 (0.94) [4]	16.44 (0.72) [4]	12.64 (1.08) [4]	12.72 (1.40) [4]	12.86 (0.87) [4]
Ionosfera	11.84 (0.34) [4]	10.70 (0.43) [4]	9.57 (0.39) [5]	10.84 (0.44) [2]	10.88 (0.52) [6]
Heartc	15.25 (0.25) [2]	23.38 (0.70) [3]	20.29 (0.93) [3]	16.85 (0.48) [3]	15.59 (0.34) [3]
Golub2	0.00 (0.00) [2]	0.00 (0.00) [2]	0.00 (0.00) [3]	0.35 (0.62) [2]	1.39 (0.00) [4]
Colon	1.61 (0.00) [6]	7.98 (1.11) [5]	5.40 (1.08) [5]	8.39 (0.66) [6]	0.40 (0.89) [6]
Golub3	0.00 (0.00) [6]	0.21 (0.51) [5]	1.53 (0.62) [6]	0.00 (0.00) [6]	5.14 (1.43) [4]
Brestcc	0.00 (0.00) [5]	5.68 (3.26) [3]	17.95 (5.80) [3]	12.95 (4.24) [3]	3.86 (4.24) [3]
Vehicle	25.34 (0.41) [9]	30.24 (0.70) [6]	31.52 (0.97) [6]	27.68 (0.80) [9]	21.85 (0.53) [11]
Segment	8.61 (0.09) [12]	2.02 (0.15) [10]	3.06 (0.14) [12]	3.70 (0.13) [8]	6.98 (0.15) [6]
Landsat	16.30 (0.08) [6]	10.39 (0.20) [8]	9.87 (0.19) [12]	9.78 (0.18) [10]	14.59 (0.15) [10]

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

2) En comparación con los resultados obtenidos en la Tabla 6.2 se observa que en general el número de componentes principales utilizado es mayor que el número de componentes PLS.

3) Los resultados para conjuntos de dos clases coinciden con los obtenidos a partir del modelo OLR, presentados en la tabla 6.3. Para conjuntos con más de dos clases, el modelo OLR no es recomendado cuando el vector de clases es nominal, debido a que para cada asignación de clases se obtendán resultados diferentes. Por ejemplo, el conjunto Breastcc tiene tres clases asignadas a 1, 2, 3 y aparentemente brinda mejores resultados desde el modelo OLR; pero al permutar las clases 1 y 2, los resultados obtenidos son: 7.27 (2.72) [3], 5.45 (2.80) [2], 13.86 (3.12) [2], 10.91 (2.28) [2] y 4.55 (1.47) [4], para los clasificadores LDA, KNN(1), KNN(3), KNN(5) y NLR, respectivamente, según el orden de la tabla 6.3.

6.4 TE_{VC} usando componentes PLS a partir de LR, caso multivariado

La generación de componentes PLS desde la regresión logística, caso multivariado, es también la aplicación de uno de los métodos propuestos en esta tesis en la sección 3.5, siguiendo el algoritmo 3.3. Según este algoritmo cada componente PLS se obtiene por la convergencia a cero diferencias entre el vector de clases y el vector de clases estimado, como lo especificado en el paso 15 del algoritmo 3.3; en caso contrario el algoritmo sigue iterando hasta alcanzar la convergencia deseada. Las componentes PLS de los datos Heartc, Golub2, Colon, Golub3 y Breastcc se obtienen por convergencia a cero diferencias en menos de 10 iteraciones por cada componente. A continuación se ilustra la convergencia para los datos Heartc, que obtiene cero diferencias en 6 iteraciones para la primera componente, en 4 iteraciones para la segunda componente y así sucesivamente.

```
> datos: Heartc
> número de componentes PLS a partir de LR, caso multivariado: 10
> máximo número de iteraciones: 10
 componente PLS: 1  2  3  4  5  6  7  8  9 10
 convergencia:   0  0  0  0  0  0  0  0  0  0
 max. Iteración: 6  4  6  4  6  3  2  3  2  3
```

Las componentes PLS de los datos Sonar, Ionosfera, Vehicle, Segment y Landsat no logran convergencia a cero diferencias, por lo que en cada caso se determinó un número de iteraciones que logra una convergencia mínima para cada conjunto de datos. Así el máximo de iteraciones fueron 20, 20, 40, 20 y 40 para estos datos, respectivamente; incrementar este número de iteraciones empeora los valores de convergencia alcanzados. A continuación se ilustra la convergencia de los datos Landsat.

```
> datos: Landsat
> número de componentes PLS a partir de LR, caso multivariado: 13
> máximo número de iteraciones: 40
 componente PLS:  1  2  3  4  5  6  7  8  9 10 11 12 13
 convergencia:   36  0  0  0 80  0  0 11  0  0  0  0  9
 max. Iteración: 40  5 10  7 40  3  6 40  7  8  7  5 40
```

Las tasas de errores son mostradas en la Tabla 6.5, de la cual se pueden hacer los siguientes comentarios:

Tabla 6.5 TE_{VC} usando componentes PLS a partir de LR, caso multivariado*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	18.99 (0.67) [8]	15.46 (0.95) [6]	22.93 (1.31) [2]	20.63 (1.29) [2]	18.94 (0.55) [8]
Ionosfera	13.79 (0.31) [3]	11.11 (0.68) [3]	12.05 (0.71) [3]	11.67 (0.63) [3]	14.37 (0.38) [3]
Heartc	15.57 (0.34) [3]	24.76 (0.97) [2]	20.10 (0.56) [2]	18.65 (1.07) [3]	15.69 (0.40) [2]
Golub2	0.00 (0.00) [2]	1.39 (0.00) [2]	1.39 (0.00) [2]	1.39 (0.00) [2]	0.07 (0.31) [3]
Colon	1.61 (0.00) [6]	9.74 (1.37) [6]	11.29 (1.65) [4]	13.23 (1.70) [3]	1.53 (1.61) [6]
Golub3	0.00 (0.00) [3]	1.39 (0.00) [3]	1.39 (0.00) [4]	1.39 (0.00) [4]	2.78 (0.00) [2]
Brestcc	0.00 (0.00) [2]	0.45 (0.45) [2]	0.00 (0.00) [2]	1.36 (2.60) [2]	0.91 (2.38) [2]
Vehicle	24.07 (0.43) [12]	30.01 (0.63) [7]	29.92 (0.65) [7]	30.11 (0.74) [7]	26.74 (0.34) [8]
Segment	8.37 (0.09) [7]	3.06 (0.19) [7]	5.06 (0.23) [4]	5.61 (0.24) [4]	6.23 (0.12) [6]
Landsat	16.59 (0.09) [5]	10.44 (0.14) [11]	9.84 (0.16) [11]	9.94 (0.16) [11]	14.52 (0.14) [10]

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

1) En comparación con los resultados obtenidos en la Tabla 6.1 y en la Tabla 6.2 se observa que en cada conjunto de datos, las componentes PLS a partir de LR, caso multivariado, logran reducir la tasa de error usando sólo algunas componentes PLS con los tres clasificadores.

2) En comparación con los resultados obtenidos en la Tabla 6.4 se observa que con componentes PLS a partir de LR, caso multivariado, las tasas de error son en general equivalentes, aunque ligeramente mayor para los datos Sonar y Colon, pero bastante menor para los datos Breastcc con el clasificador KNN

6.5 TE_{VC} usando componentes PLS a partir de LDA

La generación de componentes PLS desde el Análisis Discriminante Lineal (LDA) es la aplicación de otro de los métodos propuesto en esta tesis, en la sección 4.2, siguiendo el algoritmo 4.1. Los clasificadores en estudio fueron aplicados a los datos en estudio y se calcularon las tasas de errores, las cuales son mostradas en la Tabla 6.6, de la cual se puede hacer los siguientes comentarios:

1) Los resultados obtenidos en general son superiores a los obtenidos en las Tablas 6.1 y Tabla 6.2, referido a tasas de error usando todas las predictoras y componentes principales, respectivamente.

2) En comparación con los resultados obtenidos en la Tabla 6.4 se observa que con componentes PLS a partir de LDA, las tasas de error son en general equivalentes, aunque mayores para el conjunto Breastcc con el clasificador KNN .

3) En comparación con los resultados obtenidos en la Tabla 6.5 se observa que con componentes PLS a partir de LDA, las tasas de error son en general equivalentes, aunque menores para el conjunto Sonar y mucho menores para el conjunto Colon con

clasificador KNN y mucho mayores para el conjunto Breastcc con los clasificadores KNN y NLR.

Tabla 6.6 TE_{VC} usando componentes PLS a partir de LDA*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	15.14 (0.99) [4]	11.99 (0.85) [7]	10.34 (0.67) [6]	12.40 (1.04) [4]	14.30 (1.01) [4]
Ionosfera	12.14 (0.41) [4]	10.23 (0.42) [6]	8.38 (0.53) [6]	8.92 (0.59) [6]	11.38 (0.46) [4]
Heartc	14.81 (0.00) [2]	20.76 (0.70) [2]	18.90 (0.73) [3]	17.44 (0.78) [3]	15.22 (0.14) [2]
Golub2	0.00 (0.00) [2]	0.00 (0.00) [2]	0.00 (0.00) [2]	0.07 (0.31) [3]	0.00 (0.00) [6]
Colon	0.00 (0.00) [8]	4.92 (0.82) [5]	6.13 (1.44) [5]	6.37 (0.64) [6]	0.40 (1.03) [6]
Golub3	0.00 (0.00) [4]	1.67 (0.97) [2]	2.15 (1.15) [2]	7.15 (1.51) [2]	2.36 (1.75) [4]
Brestcc	0.00 (0.00) [3]	10.00 (3.16) [2]	21.59 (2.50) [2]	23.64 (5.63) [2]	6.14 (3.98) [3]
Vehicle	27.21 (0.59) [8]	29.18 (0.59) [8]	30.93 (0.77) [5]	28.39 (0.68) [8]	19.77 (0.42) [13]
Segment	24.68 (0.16) [2]	4.44 (0.19) [10]	5.90 (0.25) [9]	6.50 (0.21) [6]	4.52 (0.12) [8]
Landsat	16.41 (0.07) [6]	9.80 (0.24) [10]	9.71 (0.19) [11]	9.61 (0.20) [12]	15.06 (0.09) [6]

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

6.6 TE_{VC} usando componentes PLS a partir de PPR

La generación de componentes PLS a partir de la regresión *Projection Pursuit* (PPR) es la aplicación del último método propuesto en esta tesis en la sección 4.3 y siguiendo el algoritmo 4.2, con un término $M = 1$. Los clasificadores en estudio fueron aplicados a los datos en estudio y las tasas de errores son mostradas en la Tabla 6.7, de la cual se puede hacer los siguientes comentarios:

1) Los resultados obtenidos en general son superiores a los obtenidos en las Tablas 6.1 y Tabla 6.2, referido a tasas de error usando todas las predictoras y componentes principales, respectivamente.

2) Los resultados obtenidos son ligeramente mejores que los obtenidos en la Tabla 6.4, bastante mejores que los resultados de la Tabla 6.5, con la excepción de ser muy malos para Breastccc. Respecto a los resultados de la Tabla 6.6, las componentes PLS a partir de PPR son bastante similares.

Tabla 6.7 TE_{VC} usando componentes PLS a partir de PPR*

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	NLR
Sonar	12.88 (0.79) [5]	9.88 (0.82) [6]	10.14 (0.93) [6]	11.25 (1.15) [6]	10.93 (0.70) [6]
Ionosfera	12.56 (0.24) [2]	10.61 (0.54) [5]	10.75 (0.60) [8]	10.97 (0.38) [4]	11.76 (0.25) [5]
Heartc	15.12 (0.29) [3]	20.92 (0.97) [2]	17.78 (1.07) [2]	16.90 (0.84) [3]	15.34 (0.34) [3]
Golub2	0.00 (0.00) [2]	0.00 (0.00) [2]	0.00 (0.00) [3]	0.28 (0.57) [2]	0.00 (0.00) [5]
Colon	2.34 (0.98) [6]	5.00 (0.72) [8]	6.77 (1.12) [6]	9.92 (0.79) [3]	1.53 (1.85) [6]
Golub3	1.39 (0.78) [5]	1.25 (0.62) [3]	2.50 (0.73) [3]	2.29 (0.82) [3]	4.24 (1.23) [3]
Brestcc	0.00 (0.00) [3]	11.36 (3.46) [2]	12.73 (4.07) [2]	14.54 (5.82) [2]	1.13 (2.02) [3]
Vehicle	26.23 (0.43) [9]	28.29 (0.65) [10]	26.65 (0.52) [9]	26.51 (0.80) [11]	25.76 (0.41) [9]
Segment	8.39 (0.13) [11]	3.13 (0.17) [8]	3.99 (0.22) [9]	5.15 (0.22) [8]	6.47 (0.11) [8]
Landsat	16.96 (0.07) [6]	11.87 (0.19) [9]	10.85 (0.21) [9]	10.68 (0.17) [7]	15.48 (0.10) [6]

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

3) Las componentes PLS a partir de PPR con dos términos, $M = 2$, no son considerados debido a problemas que pueden surgir cuando se genera la primera componente PLS. Como los coeficientes buscados corresponden al vector de proyecciones, los cuales son vectores normalizados, muchas veces el coeficiente dentro de la primera y segunda función *ridge* son 1 y -1 respectivamente y al aplicar la expresión (4.16), se obtiene un peso no deseado, igual a cero.

4) Las componentes PLS a partir de PPR con tres términos, $M = 3$ fueron considerados pero no se presentan debido a que en general las tasas de error obtenidas son mucho mayores que las obtenidas en la Tabla 6.7 con $M = 1$, con excepción de las tasas de error de los datos Breastcc que son mucho menores (casi cero) y equivalentes a lo presentado en la Tabla 6.5

6.7 Las mejores TE_{VC} usando componentes PLS

La obtención de las mejores tasas de error por validación cruzada para cada conjunto de datos depende de la metodología con que se generó las componentes PLS y del clasificador utilizado. A continuación se presenta un resumen de los resultados expuestos en las tablas 6.3, 6.4, 6.5 y 6.6; donde se muestran las tasas de error más bajas, alcanzadas por los diferentes conjuntos de datos

Tabla 6.8 Las mejores TE_{VC} usando componentes PLS*

	TE_{VC}	ALGORITMO PLS	CLASIFICADOR
Sonar	9.88 (0.82) [6]	PPRPLS	KNN (K=1)
Ionosfera	8.38 (0.53) [6]	LDAPLS	KNN (K=3)
Heartc	14.81 (0.00) [2]	LDAPLS	LDA
Golub2	0.00 (0.00) [2]	NLRPLS, MLRPLS, LDAPLS, PPRPLS	LDA, KNN (K=1,3)
Colon	0.00 (0.00) [8]	LDAPLS	LDA
Golub3	0.00 (0.00) [3]	MLRPLS	LDA
Breastcc	0.00 (0.00) [2]	MLRPLS	LDA, KNN (K=3)
Vehicle	21.85 (0.55) [11]	NLRPLS	NLR
Segment	2.02 (0.15) [10]	NLRPLS	KNN (K=1)
Landsat	9.80 (0.24) [10]	LDAPLS	KNN (K=1)

** valor entre paréntesis: desviación estándar*
valor entre corchetes: número de componentes PLS

A continuación se presenta un comparativo de tasas de error de clasificación de datos de *microarrays* obtenidas por otros autores.

Tabla 6.9 Comparación de tasas de error de clasificación

	Nguye-Rocke	Ding-Gentleman	Fort-Lambert
Golub2	1 error = 1.39% PLS – Regresión Logística		6 errores = 8.33% PLS – logística penalizada
Colon	4 errores = 6.45% PLS – Regresión Logística	6 errores = 9.68% IRWPLSF	5 errores = 8.06% PLS – logística penalizada
Golub3	0 errores = 0.00% PLS – Reg. Logística Nominal		
Breastcc	0 errores = 0.00% PLS – Reg. Logística Nominal		

En general las metodologías de Nguyen-Rocke, Ding-Gentleman y Fort-Lambert trabajan con una previa selección de variables predictoras; es decir, estas metodologías no consideran todas las variables predictoras para el cálculo de cada una de las componentes PLS. Además, sólo la metodología de Nguyen-Rocke ha podido trabajar en clasificación supervisada con más de dos clases

6.8 Gráfico de las dos y tres primeras componentes PLS: *microarrays*

En esta sección se presenta el gráfico de las dos y tres primeras componentes PLS de cada uno de los cuatro conjuntos de datos de *microarrays*. Estas componentes fueron generadas con cada una de las cuatro metodologías presentadas como aporte de esta tesis.

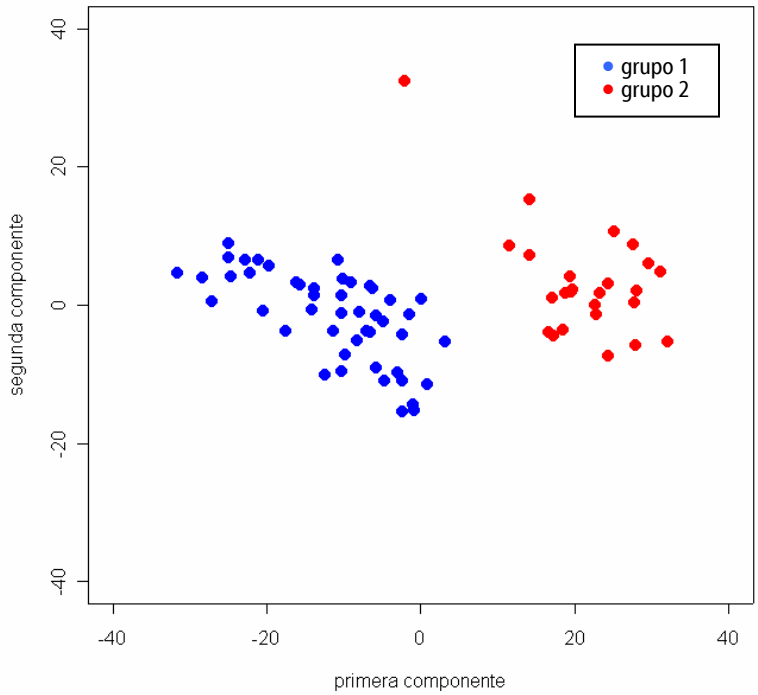
El gráfico de los datos Golub2, en las figuras 6.1, 6.5, 6.9 y 6.13; dejan ver la separabilidad casi perfecta de los grupos. Se podría señalar que la mejor separabilidad se logró usando tres componentes a partir del algoritmo LDAPLS.

El gráfico de los datos Colon, en los figuras 6.2, 6.6, 6.10 y 6.14; dejan ver que las cuatro metodologías no logran una buena separabilidad de grupos con dos o tres componentes. Según la tabla 6.8, la mejor separabilidad se logra con 8 componentes PLS

El gráfico de los datos Golub3, en los figuras 6.3, 6.7, 6.11 y 6.15; dejan ver la separabilidad de grupos. Se podría señalar que la mejor separabilidad se logró usando tres componentes a partir del algritmo PPRPLS.

El gráfico de los datos Btreastcc, en los figuras 6.4, 6.8, 6.12 y 6.16; dejan ver la separabilidad de grupos. Se podría señalar que la mejor separabilidad se logró usando dos componentes a partir del algritmo MLRPLS.

Figura 6.1 Gráfico de dos y tres componentes: Datos Golub2
Algoritmo NLRPLS



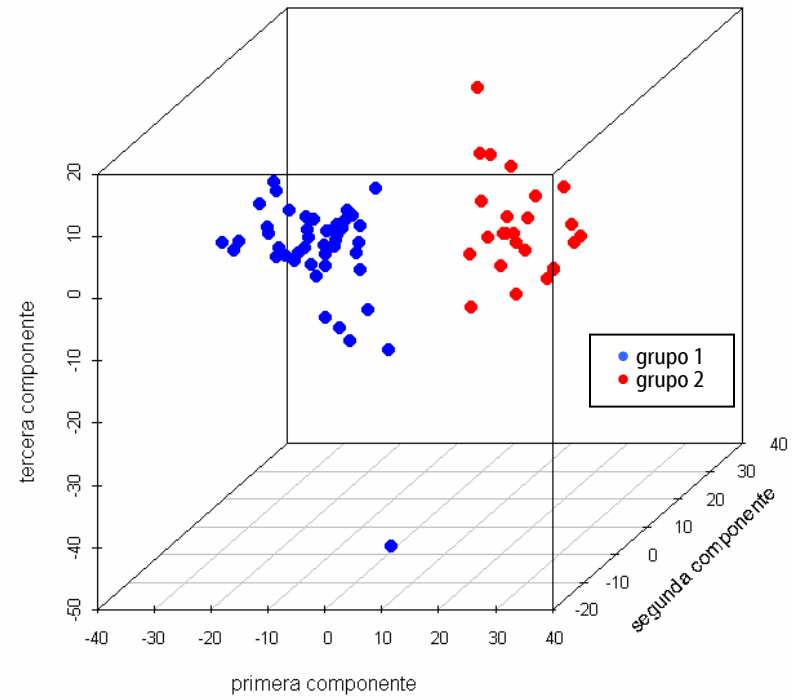
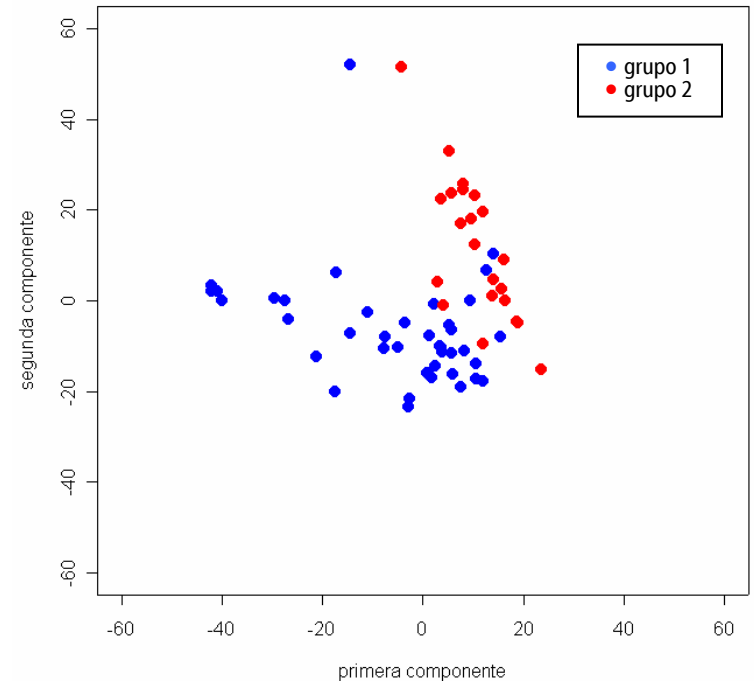


Figura 6.2 Gráfico de dos y tres componentes: Datos Colon Algoritmo NLRPLS



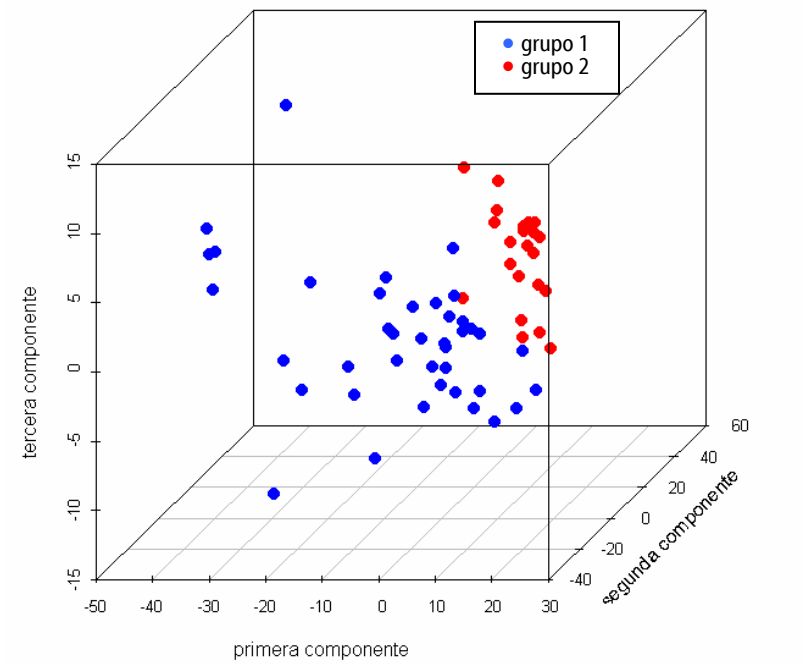
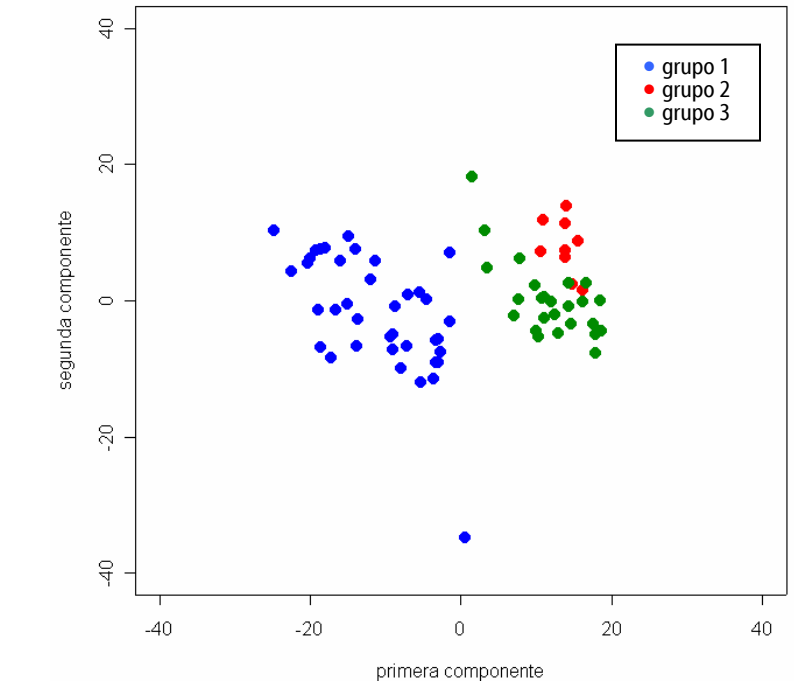


Figura 6.3 Gráfico de dos y tres componentes: Datos Golub3
Algoritmo NLRPLS



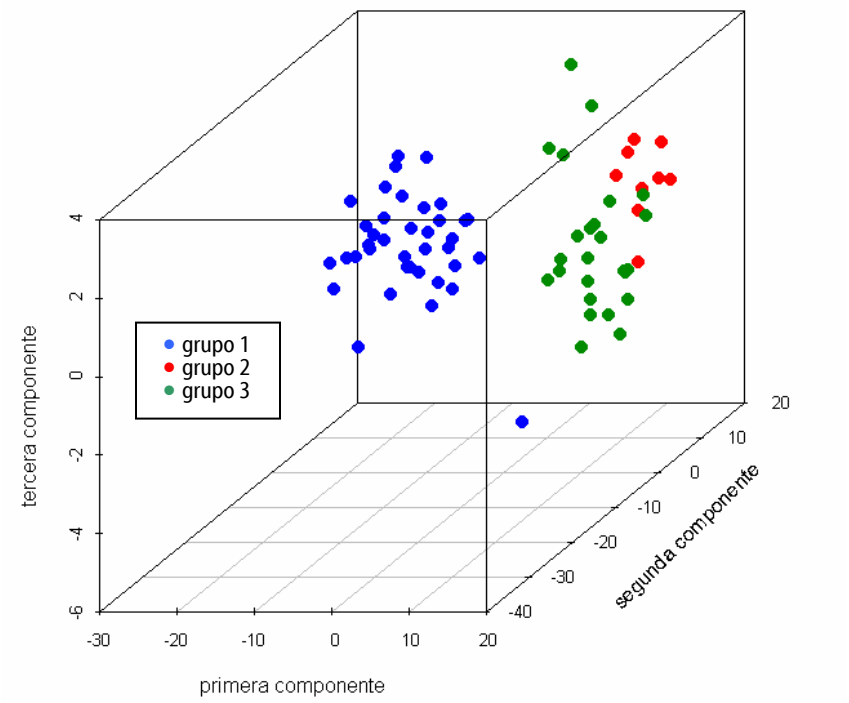
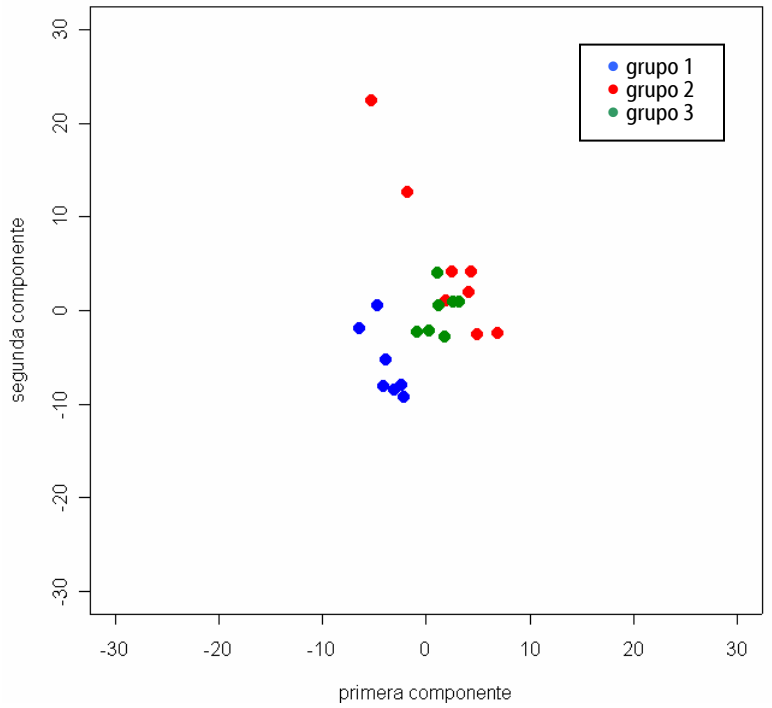


Figura 6.4 Gráfico de dos y tres componentes: Datos Breastcc
Algoritmo NLRPLS



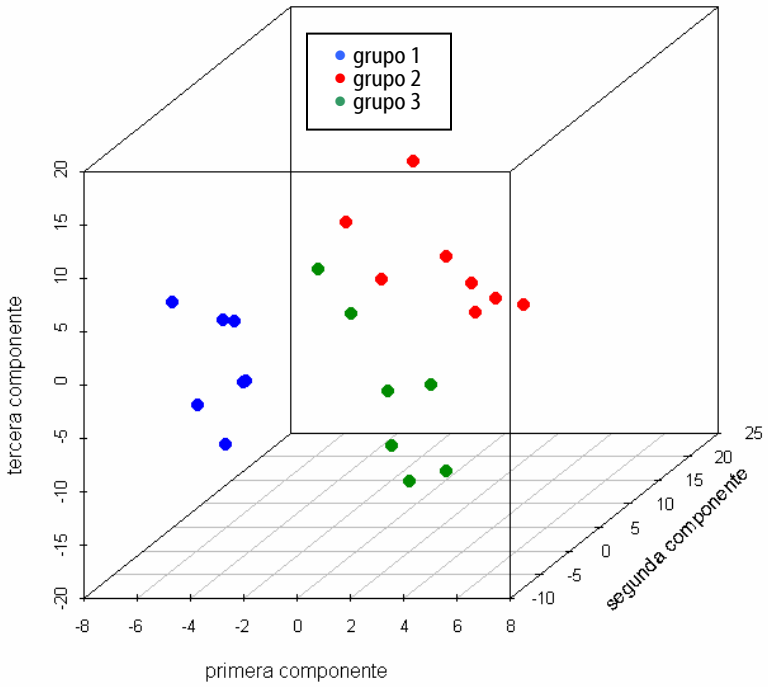
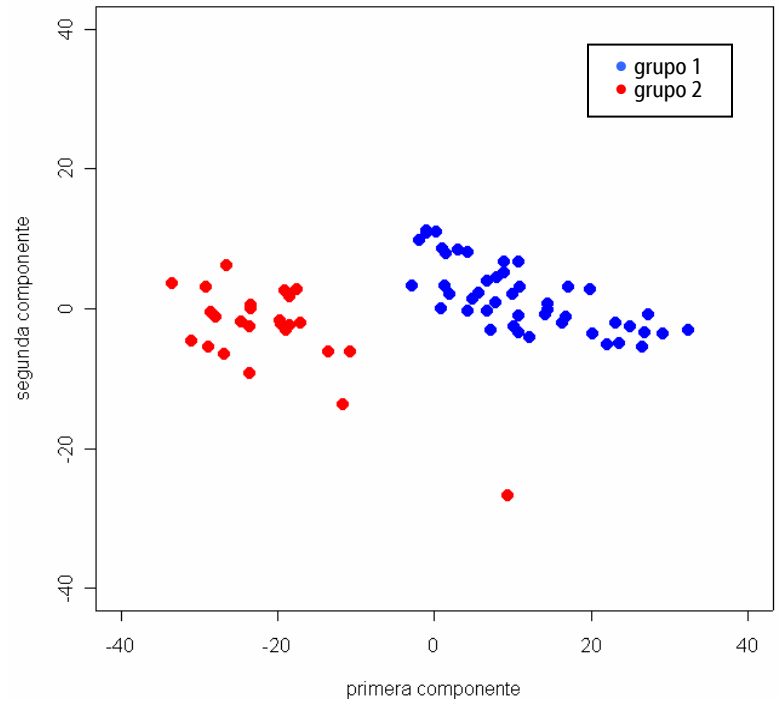


Figura 6.5 Gráfico de dos y tres componentes: Datos Golub2
Algoritmo MLRPLS



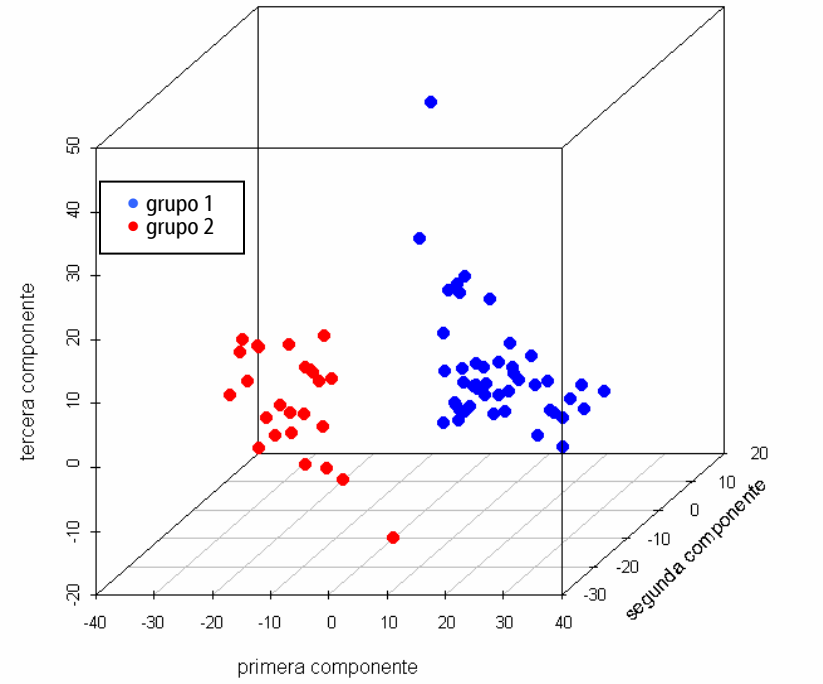
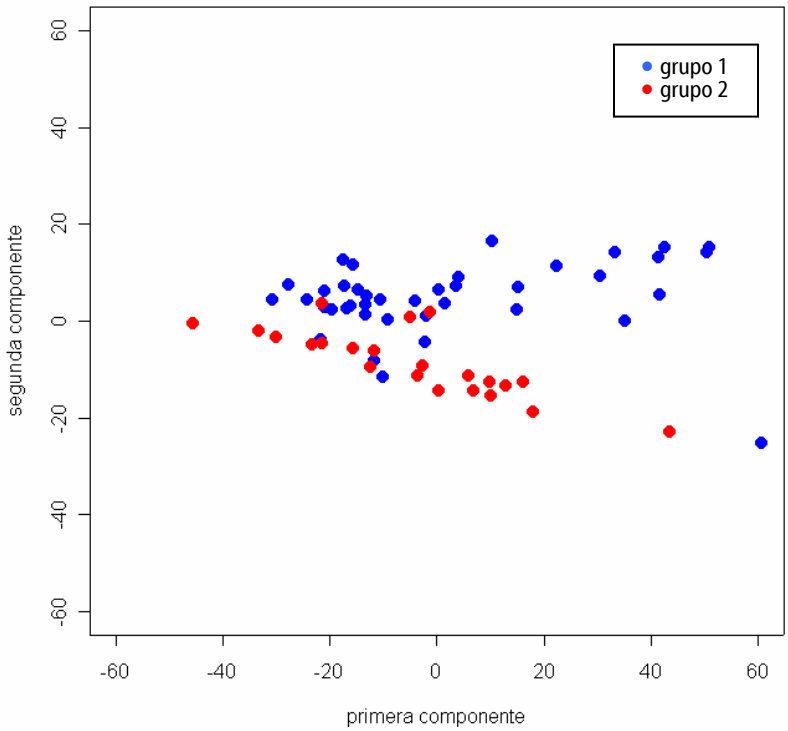


Figura 6.6 Gráfico de dos y tres componentes: Datos Colon Algoritmo MLRPLS



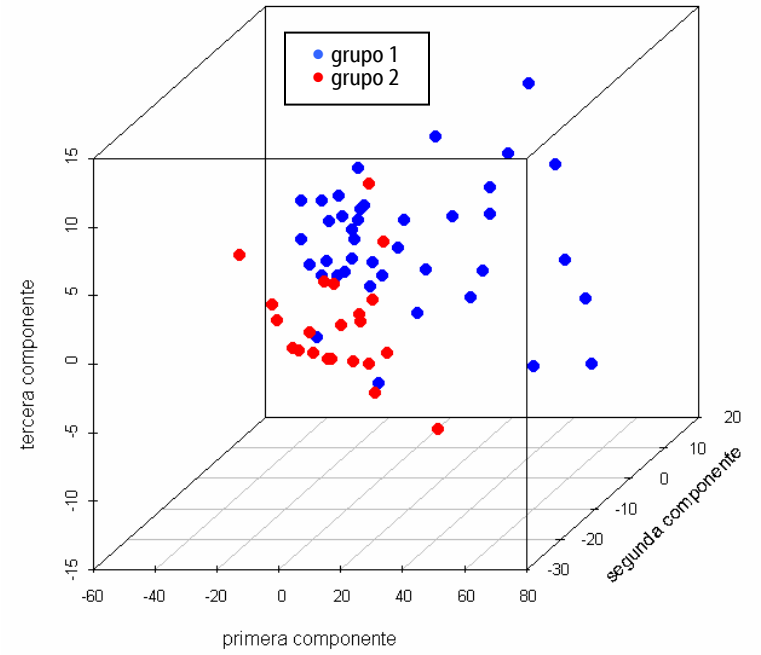
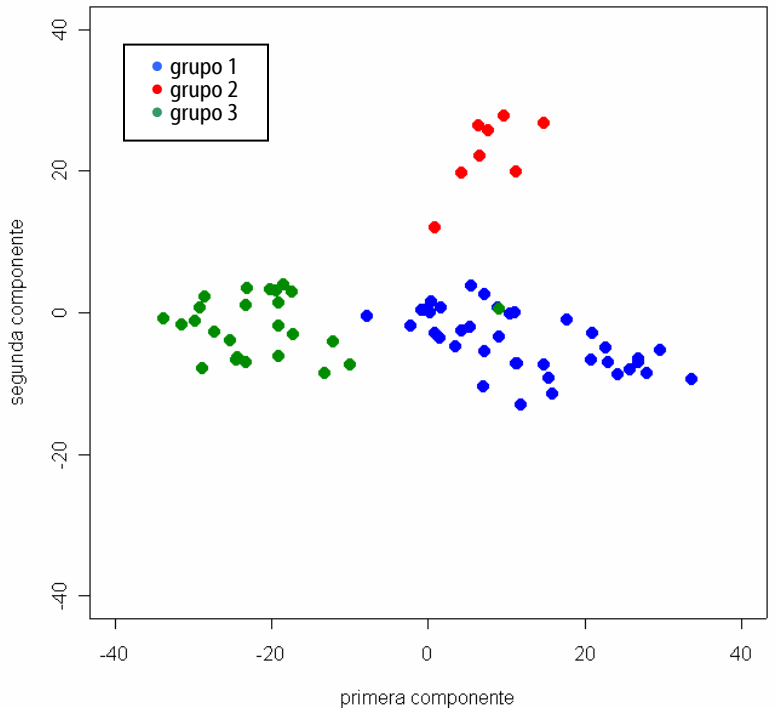


Figura 6.7 Gráfico de dos y tres componentes: Datos Golub3
Algoritmo MLRPLS



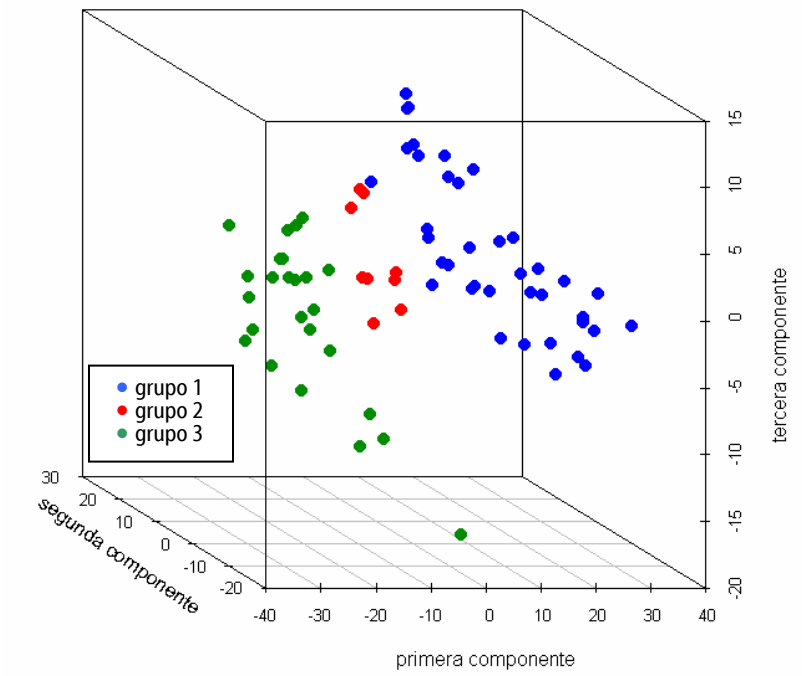
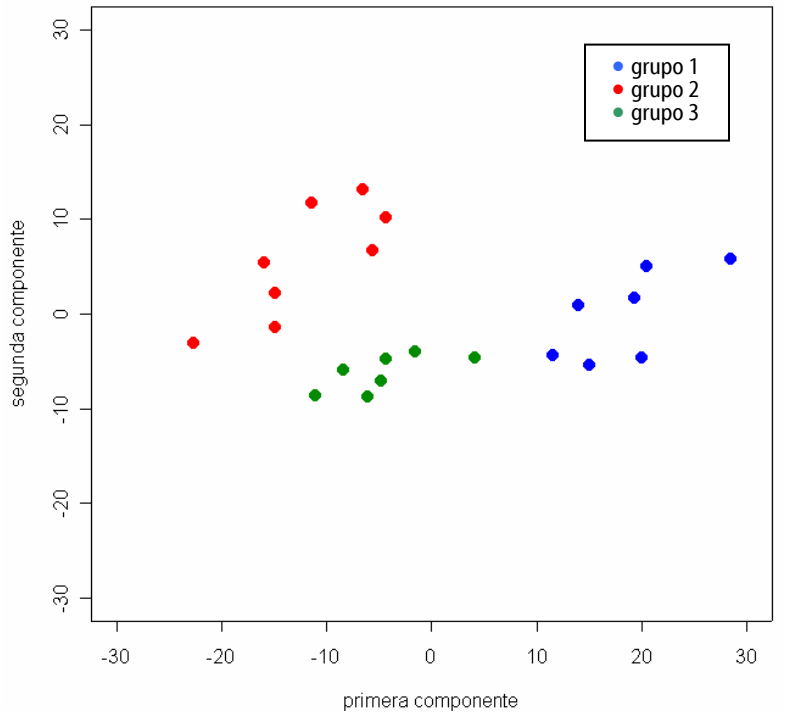


Figura 6.8 Gráfico de dos y tres componentes: Datos Breastcc Algoritmo MLRPLS



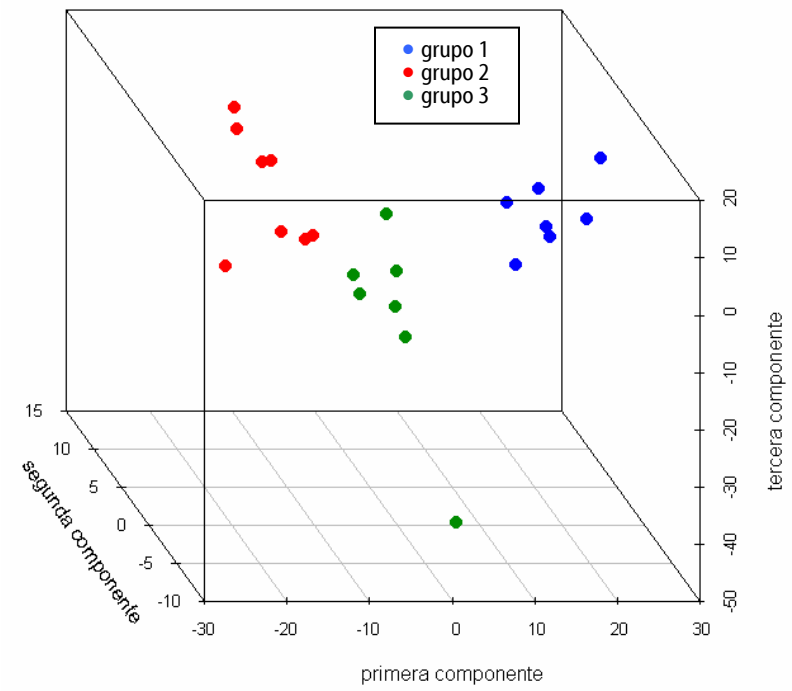
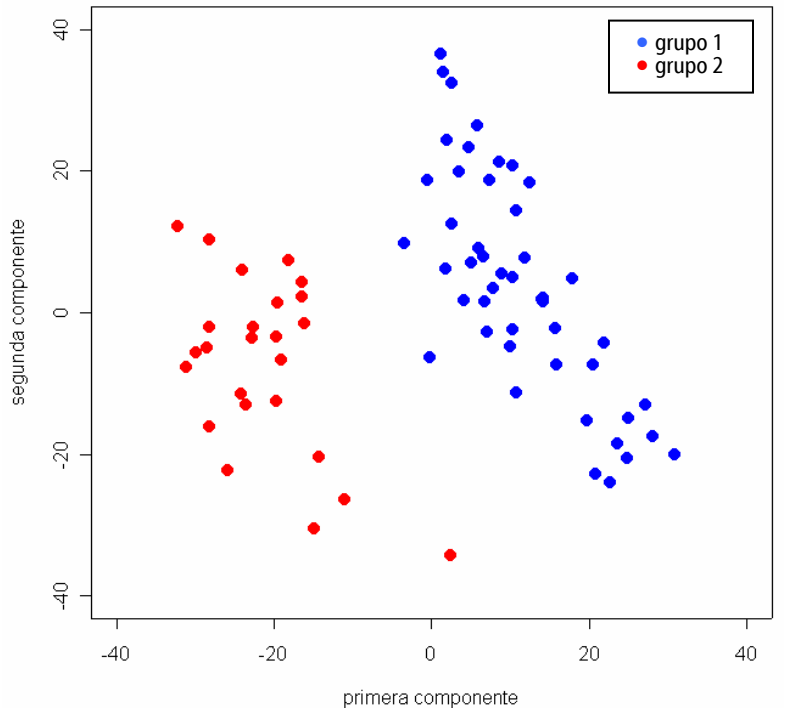


Figura 6.9 Gráfico de dos y tres componentes: Datos Golub2
Algoritmo LDAPLS



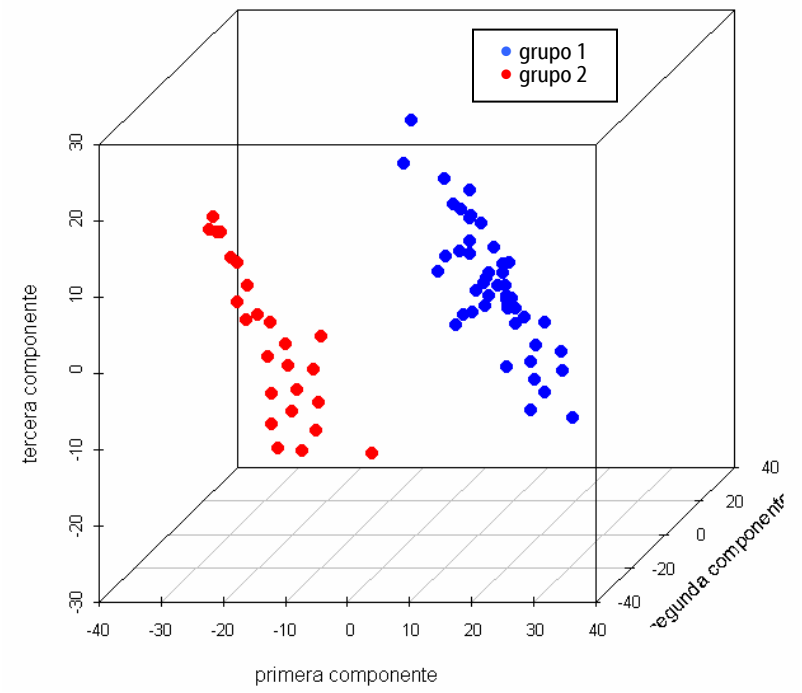
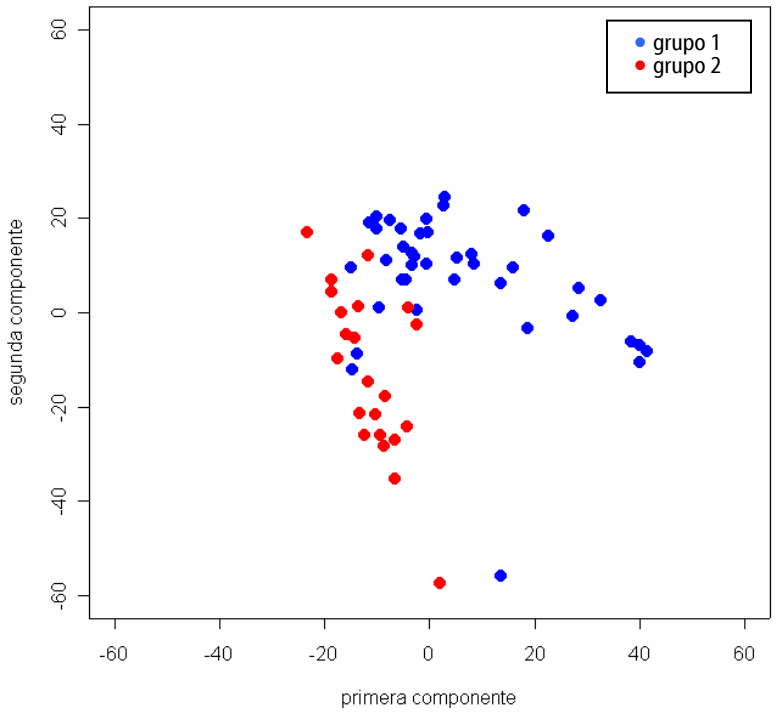


Figura 6.10 Gráfico de dos y tres componentes: Datos Colon Algoritmo LDAPLS



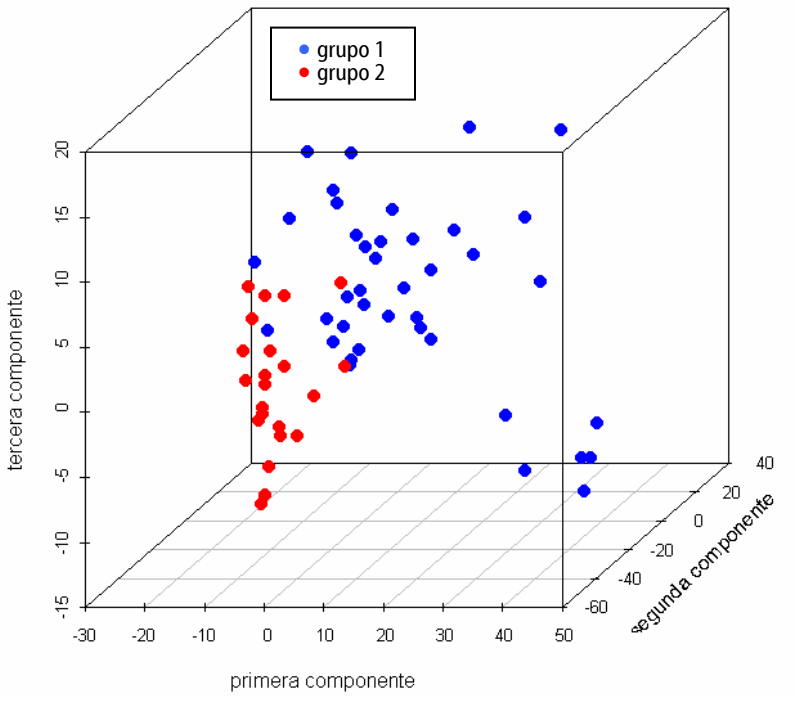
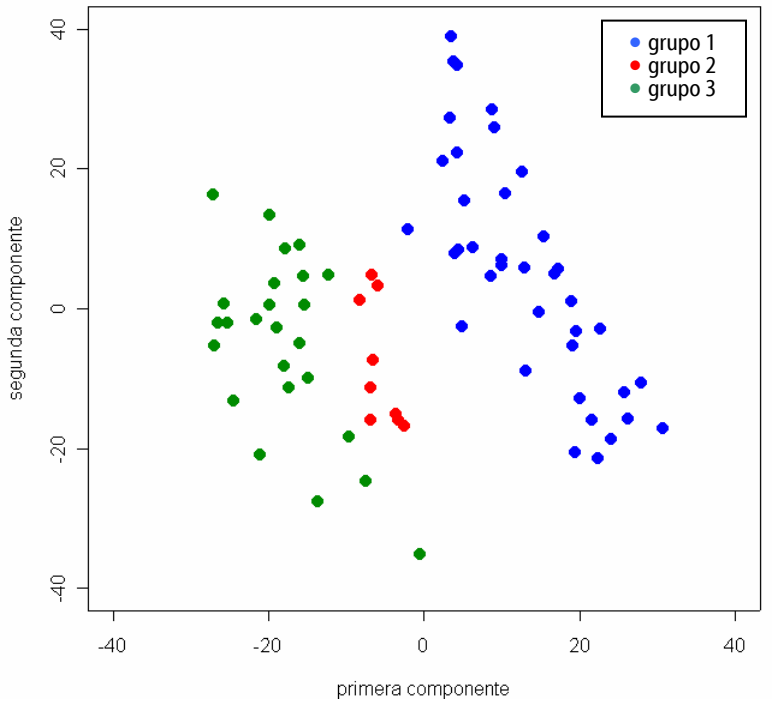


Figura 6.11 Gráfico de dos y tres componentes: Datos Golub3
Algoritmo LDAPLS



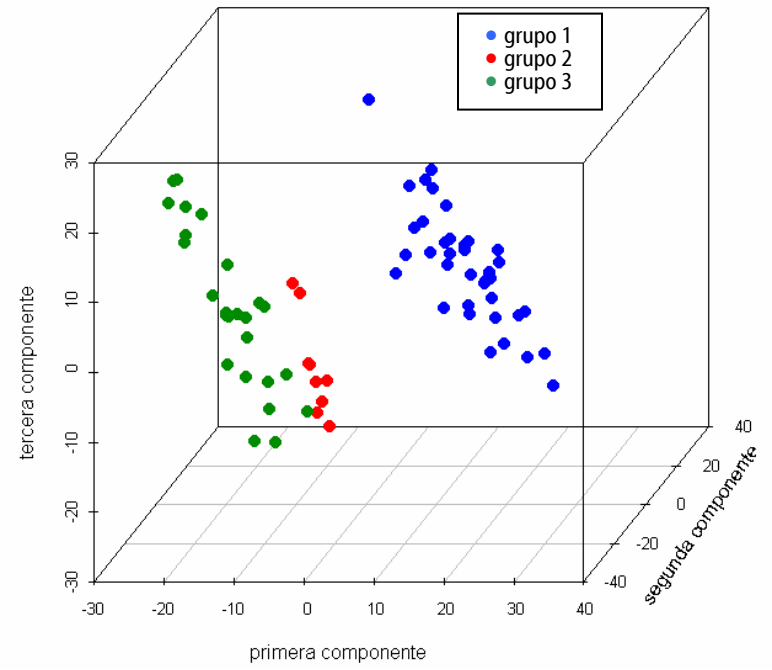


Figura 6.12 Gráfico de dos y tres componentes: Datos Breastcc Algoritmo LDAPLS

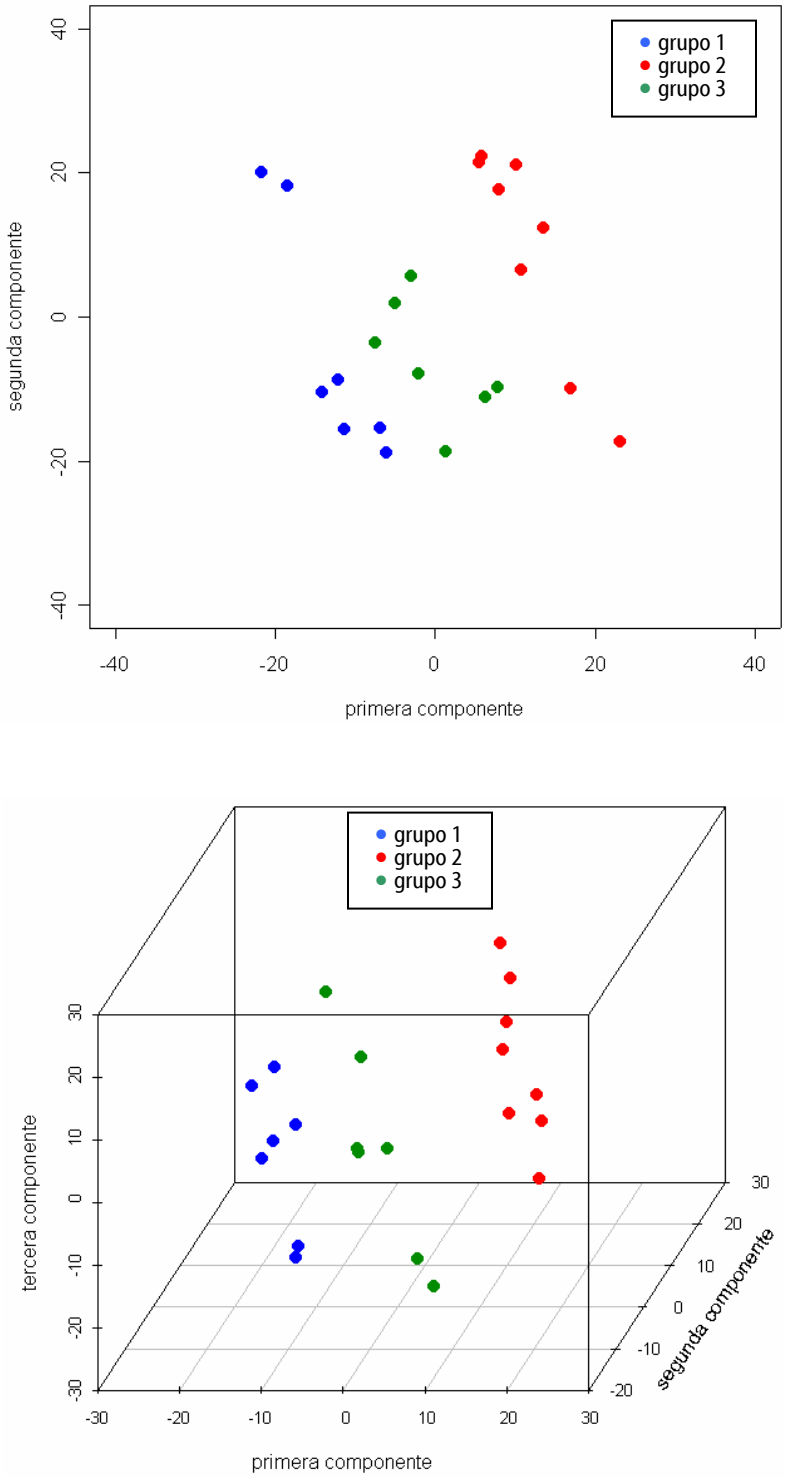
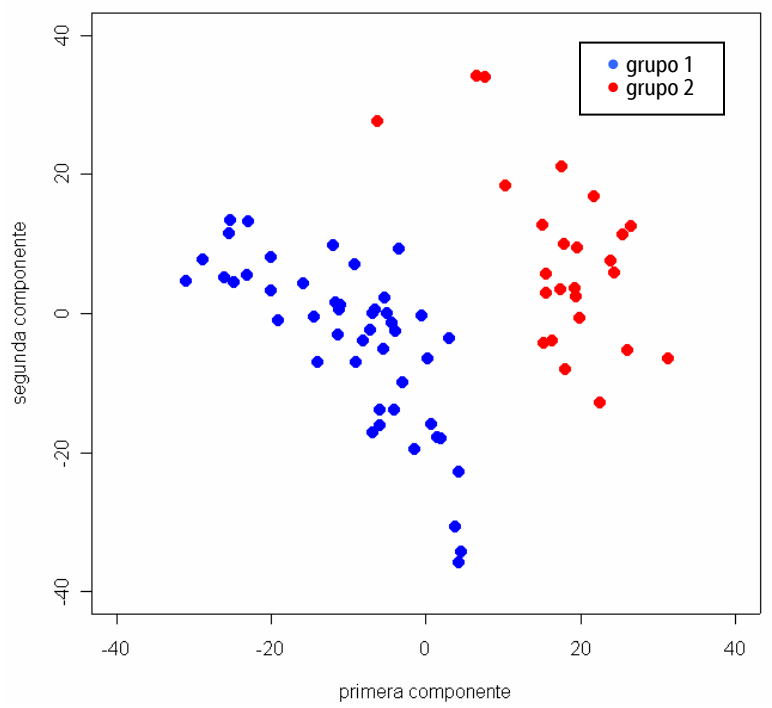


Figura 6.13 Gráfico de dos y tres componentes: Datos Golub2
Algoritmo PPRPLS



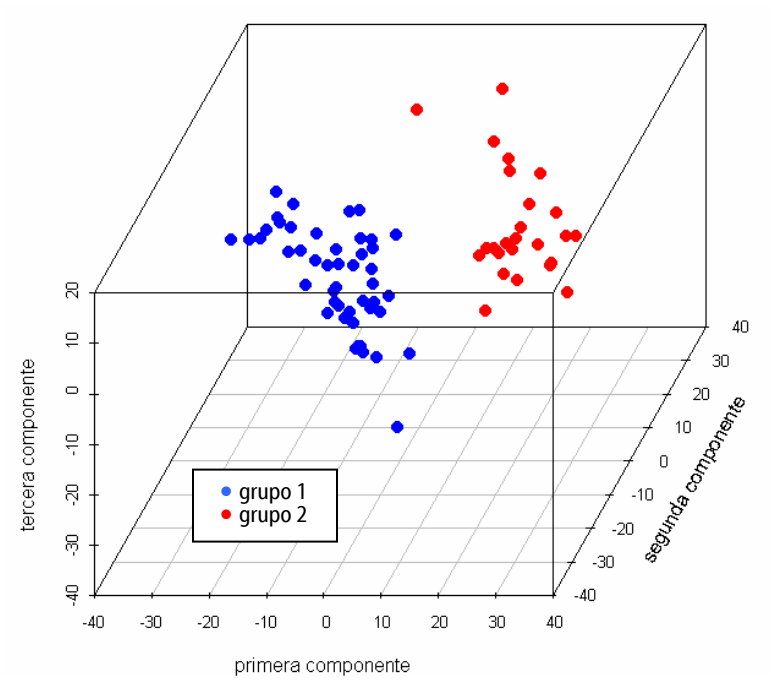
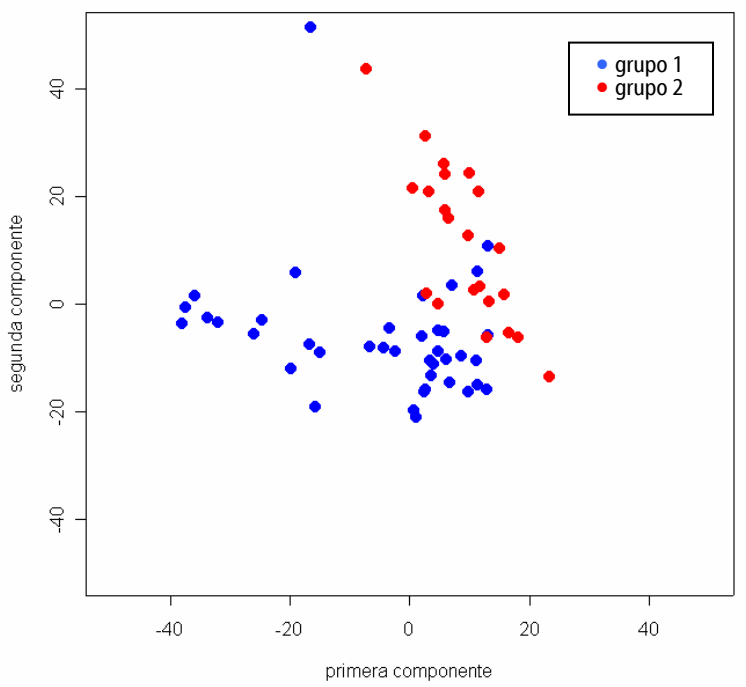


Figura 6.14 Gráfico de dos y tres componentes: Datos Colon Algoritmo PPRPLS



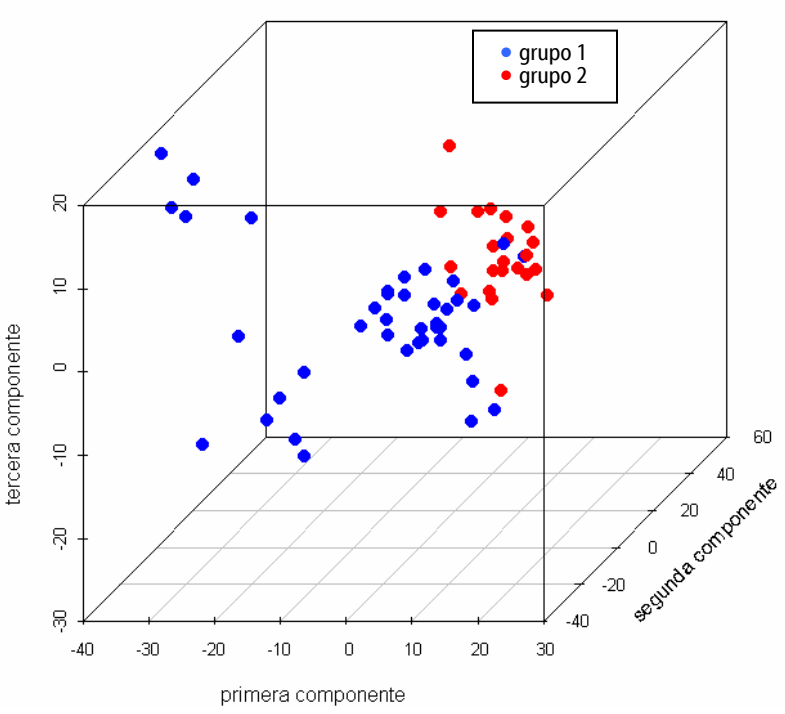


Figura 6.15 Gráfico de dos y tres componentes: Datos Golub3
Algoritmo PPRPLS

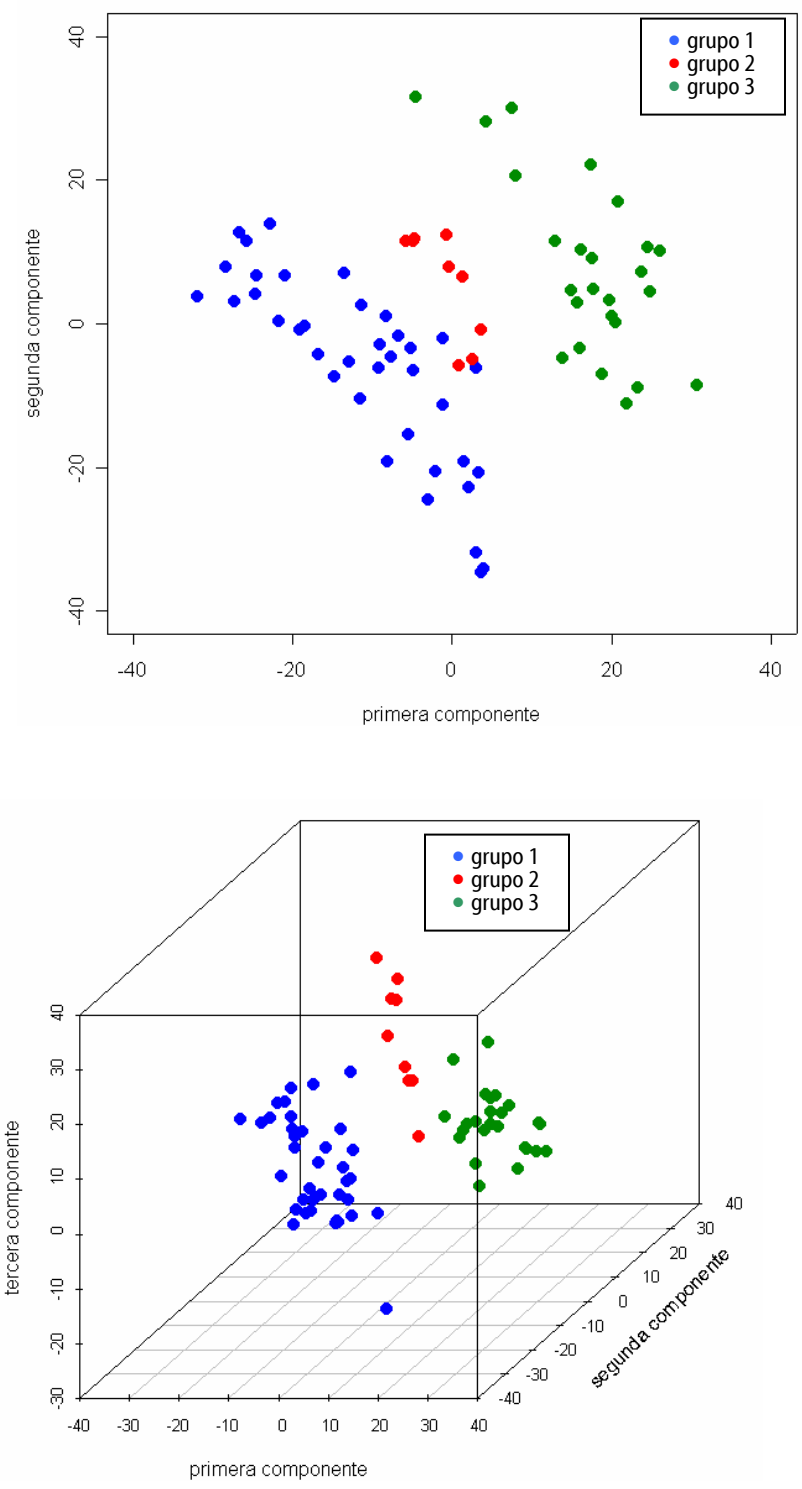
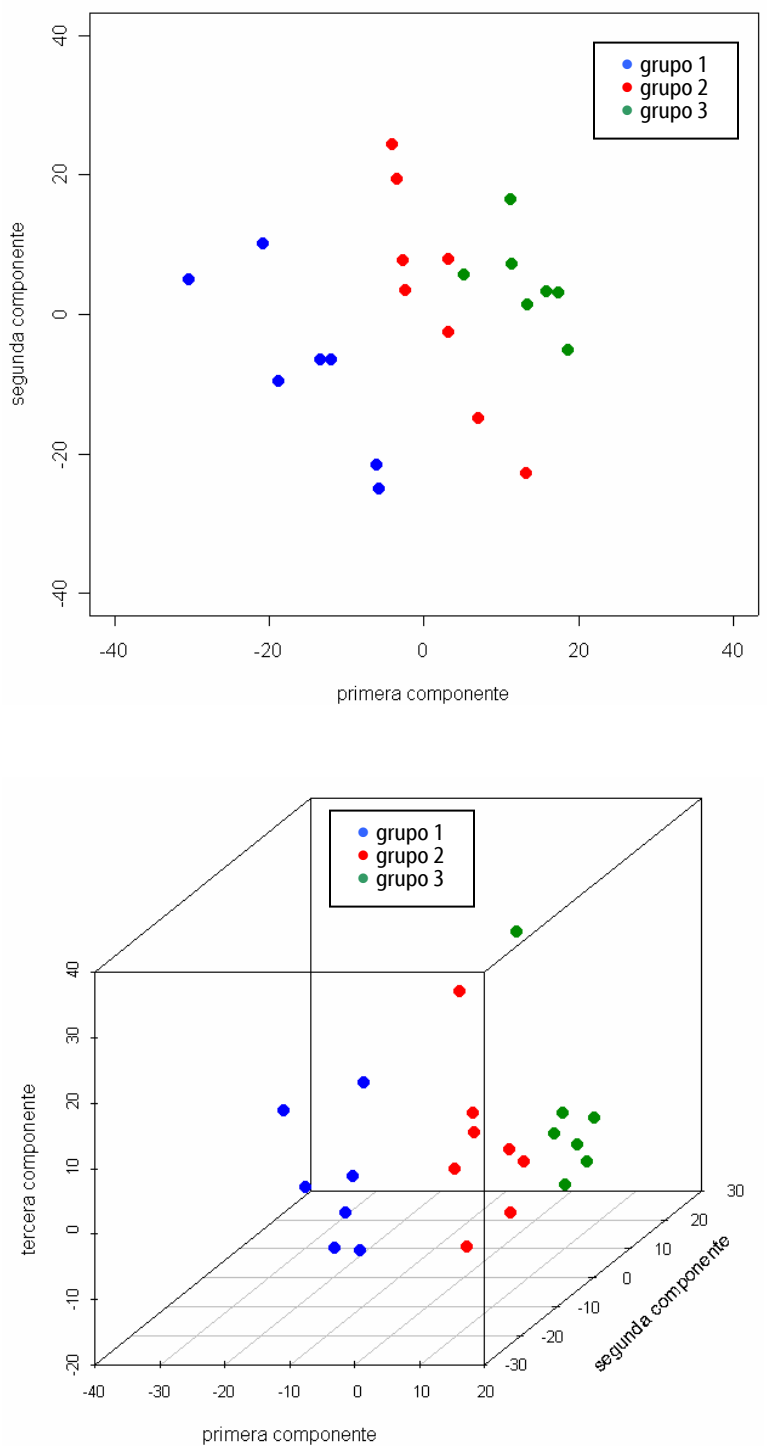


Figura 6.16 Gráfico de dos y tres componentes: Datos Breastcc Algoritmo PPRPLS



Capítulo 7

Conclusiones y Recomendaciones

7.1 Conclusiones

1. Las componentes PLS generados a partir de las metodologías presentadas en esta tesis son ortogonales entre sí. Esta característica es fundamental para reducir la dimensionalidad del espacio de predictoras y ser aplicados en clasificación supervisada.

2. Las componentes PLS generados a partir de las metodologías presentadas, son combinaciones lineales de las variables predictoras, así como lo son las componentes principales y obtienen sus ponderaciones a partir de las variables predictoras y del vector de clases; mientras que las componentes principales lo hacen sólo a partir de las variables predictoras.

3. La clasificación supervisada a partir de componentes PLS supera a la clasificación a partir de componentes principales. Las tasas de error de clasificación estimadas, así como el número de componentes son menores usando componentes PLS. Por tanto se recomienda el uso de componentes PLS en clasificación supervisada.

4. Las componentes PLS obtenidos a partir de la regresión logística ordinal y regresión logística nominal son los mismos siempre que el conjunto de datos en estudio considere sólo dos clases.

5. Las tasas de error de clasificación por validación cruzada para cada conjunto de datos dependen de la metodología con que se generó las componentes PLS y del

clasificador utilizado. No se pudo identificar una metodología ideal, que genere componentes PLS y haga que los clasificadores en estudio produzcan las más bajas tasa de error en todos los datos.

6. Por los resultados obtenidos se puede afirmar que esta propuesta mejora a dos recientes trabajos, de Fort y Lambert (2004) y Ding y Gentleman (2004) que combinan regresión logística y PLS que son disponibles sólo para dos clases

7.1.1 Contribuciones

Las contribuciones en esta tesis, son las siguientes:

1. Se da a conocer en forma explícita el fundamento de la teoría de regresión PLS

2. Se resaltan la propiedades más importantes en regresión PLS

3. Se simplifica el algoritmo que genera componentes PLS a partir de la regresión logística ordinal, de Esposito-Vinzi, en lo relacionado a la actualización de la matriz de residuales. Se formula y demuestra el teorema 3.1, que simplifica la actualización de la matriz de residuales y por consiguiente el proceso computacional del algoritmo 3.1

4. Se formula un algoritmo que genera componentes PLS a partir de la regresión logística nominal, aplicable cuando no hay un orden natural en las categorías de la variable respuesta, lo cual es lo más real en clasificación supervisada. Esta metodología es una extensión del algoritmo de Esposito-Vinzi y usa el teorema 3.1 para simplificar la actualización de la matriz de residuales.

5. Se proponen y demuestran propiedades de las componentes PLS que conducen a probar matemáticamente la ortogonalidad de los mismos.

6. Se formula un algoritmo que genera la matriz de transformación de variables predictoras en componentes PLS. La formulación de este algoritmo implica el planteamiento y demostración del teorema 3.2. La aplicación de esta matriz de

transformación es la predicción de categorías o la clasificación de nuevas observaciones.

7. Se formula un algoritmo que genera componentes PLS a partir de la regresión logística, caso multivariado, que considera al vector de clases como una matriz. Esta metodología surge como extensión de la teoría de regresión PLS multivariada (PLS2) y del algoritmo NLRPLS.

8. Se formulan otros dos algoritmos de obtención de componentes PLS para clasificación supervisada. El algoritmo que genera componentes PLS a partir del Análisis Discriminante Lineal y el algoritmo que genera componentes PLS a partir de la Regresión *Projection Pursuit*.

7.2 Trabajos Futuros

1. Profundizar en el estudio de las componentes PLS, que han sido obtenidos por los diferentes métodos en este trabajo; tales como, determinación de sus propiedades estadísticas, búsqueda de otras aplicaciones, etc.

2. Estudiar, a nivel de análisis exploratorio de datos, las características que hacen que un conjunto de datos sea más adecuado para generar un determinado tipo de componentes PLS

3. Estudiar el comportamiento de las componentes PLS, que han sido obtenidos por los diferentes métodos en este trabajo, con respecto a otros clasificadores.

4. Buscar otras fuentes de obtención de componentes PLS.

5. Estudiar acerca de nuevos métodos de selección de componentes PLS

Capítulo 8

Ética

8.1 Introducción

Han pasado los tiempos en que la propuesta de una nueva tecnología era asociada como sinónimo de aceptación y progreso. Hoy, las nuevas tecnologías son asociadas a factores negativos como daño, inseguridad, beneficios para unos y perjuicios para otros; todo esto debido a sucesos polémicos como la guerra biológica, la clonación, etc., que ha sensibilizado a los profesionales y a la sociedad civil sobre un concepto clave en la ética de las Ciencias e Ingeniería: Responsabilidad Profesional; es decir la responsabilidad moral como conocimiento individual adquirido, vinculado a la conciencia y a la subordinación de valores admitidos por la persona.

Un científico que tiene la responsabilidad moral por un asunto, debe utilizar su juicio y preparación para alcanzar o mantener los objetivos planteados. La meta de un profesional responsable es la creación de productos tecnológicos útiles y seguros, que no comprometan la salud pública, la seguridad ni el bienestar de la sociedad.

En base al trabajo de Buendía y Berrocal (2001), se desarrolla este capítulo que contiene aspectos fundamentales de ética, que valen la pena ser reflexionados, por toda persona dedicada a la investigación para que sus actos o los resultados de los mismos, sean éticamente correctos. El resultado de esta reflexión ha permitido enmarcar el desarrollo de las metodologías presentadas en este trabajo.

8.2 Ética de la investigación

En general la ética es considerada como sinónimo de Filosofía moral y por lo tanto una parte de la Filosofía encargada del estudio de conductas morales. Desde el punto de vista del conocimiento vulgar, no academicista ni científico, la ética está vinculada a cada uno de los actos que se realizan cada día, en diferentes ámbitos de la vida y por lo tanto la ética es una actividad que nos concierne a todos, en la medida que todo el mundo se enfrenta con situaciones que implican la toma de decisiones.

En este sentido, se puede reflexionar sobre aquellas normas que como investigadores deberíamos respetar para que nuestros actos o los resultados de los mismos, sean éticamente adecuados.

1) La investigación debe ser un acto ético

La investigación no es sólo un acto técnico; es ante todo el ejercicio de un acto responsable y desde esta perspectiva la ética de la investigación hay que planteársela como un subconjunto dentro de la moral general aunque aplicada a problemas mucho más restringidos que la moral general, puesto que nos estaríamos refiriendo a un aspecto de la ética profesional.

Pero la ética en una profesión es la obligación de una conducta correcta. Las múltiples situaciones a las que hay que dar respuesta desde cada profesión, muestran que la ética profesional es una parte de cada acto profesional individual que incluye un conflicto entre el efecto intencionado y el efecto conseguido. Así pues, desde el punto de vista de la investigación, un acto ético es el que se ejerce responsablemente, evitando generar perjuicios, que a veces se realiza inconscientemente, por estar vinculado el daño a los métodos que el investigador utiliza para la consecución de sus fines.

2) El investigador debe ceñirse al desarrollo del trabajo

Los investigadores deben basar sus conclusiones en pruebas válidas y fiables, siendo los resultados de dichas pruebas los únicos indicadores para la toma de decisiones. La

negación de esta propuesta hace que las actuaciones más censurables estén vinculadas al desarrollo del proceso de investigación. En la investigación experimental muchas veces existe manipulación de la variable independiente y contextos artificiales o selección de las condiciones en las que va a tener lugar la experiencia.

Cuando la investigación es considerada un proceso encaminado a la comprensión de la realidad, no ausente de valores y generadora de conocimiento, hace que la investigación que se realiza, esté en función de la interpretación que el investigador haga del tema, la cual estará siempre vinculada al contexto y a los valores del investigador, que impregnan todo el proceso.

3) Evitar problemas éticos

Se puede analizar los problemas éticos respecto a los participantes como unidades experimentales en la investigación, respecto al desarrollo del trabajo y respecto al propio investigador.

- Respecto a los participantes, se considera que su protección como sujetos de investigación exige respetar su autonomía, por lo que se les debe informar acerca de los fines que se persiguen con el desarrollo del proyecto, sin ningún tipo de coacción económica o de poder. Junto al valor de autonomía está el de la privacidad de los participantes que exige anonimato y confidencialidad de parte del investigador.

- Respecto al desarrollo del trabajo, los usos incorrectos en la investigación pueden aparecer tanto en la planificación como en el proceso o en la utilización de resultados. En la planificación de la investigación las intenciones del investigador pueden ser: provecho político, provecho personal, publicidad, relaciones públicas, prestigio, justificación de resultados, etc. En el proceso de la investigación las intenciones del investigador pueden ser: prorrogar decisiones críticas, trabajar con muestras intencionales con fines políticos o personales, sabotear la investigación

porque no responde a lo esperado, etc. En los resultados obtenidos las intenciones del investigador pueden ser: aceptar hipótesis que son falsas, modificar conclusiones, simplificar, exagerar u ocular resultados, presentar informes intencionados, etc.

- Respecto al propio investigador, que puede considerar que sus investigaciones van a ser muy importantes una vez realizadas, por lo tanto, cobra especial protagonismo la intencionalidad que se tiene en el trabajo. Las amplias expectativas del investigador generan a veces fraudes en las informaciones que se difunden en base a datos falsos. Los errores de una mala utilización de los resultados de la investigación generan daños a los participantes de la investigación, daños a los investigadores así como a la profesión de la investigación y daños a la sociedad en general.

4) Evitar daños a otros investigadores

El problema ético más conocido y el que más juicios ha levantado por el perjuicio que ocasiona a los propios colegas de profesión es el plagio. Existen tres tipos de plagio:

- Copiar literalmente un trabajo de investigación de otros colegas y presentarlo como propio.
- Utilizar trozos de textos o citas de otros autores sin citarlo
- Usar la propiedad intelectual de un autor, sin su permiso expreso.

Estas situaciones, han sido frecuentemente denunciadas y atentan gravemente contra la ética de la investigación. Hoy, con la posibilidad de acceder tan fácilmente a la información, el plagio podría parecer que se acrecienta pero justamente esta mayor accesibilidad a las investigaciones permiten un mayor control, junto con el desprestigio social que llevan aparejadas este tipo de conductas.

Quizás el acto de plagio más inmoral es el que se comete por abuso de autoridad. Esto referido a las publicaciones que los investigadores realizan como propias, sin citar a los

colaboradores, siendo en la mayoría obra de todos, o los plagios de trabajos de alumnos o compañeros utilizando el estatus o poder.

5) Evitar daños sociales

El problema ético generado por la manipulación de datos conduce a que se dañen los resultados y la veracidad de las conclusiones obtenidas, repercutiendo esto en el ámbito científico y social. La utilización de datos falsos puede deberse a dos razones:

- Para confirmar hipótesis, los investigadores pueden cambiar los datos obtenidos para poder confirmar hipótesis que son falsas. Esta conducta a veces es inducida por presiones externas que por haber financiado la investigación desean confirmar hipótesis beneficiosas para sus propósitos.
- Para conseguir mayor reputación, el investigador puede ofrecer resultados y conclusiones sobre datos inventados.

8.3 Ética de la tesis

En la presente tesis, se plantea la generación de una metodología estadístico computacional de propósito general que trabaja con datos obtenidos desde diferentes estudios, obtenidos por investigadores en Biología y Ciencias Sociales, con el objetivo de lograr un clasificador eficiente de los mismos. Concientes de la realidad en el campo del desarrollo tecnológico y los fundamentos básicos de ética, se puede afirmar que esta metodología propuesta está enmarcada dentro del principio ético de responsabilidad profesional, que es puesto a disposición de la comunidad científica para su mejor aplicación y desarrollo.

Bibliografía

1. Albert, A. y Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71:1-10

2. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96, 6745–6750.

3. Antoniadis, A., Lambert-Lacroix, S. y Leblanc, F. (2003). Effective Dimension Reduction Methods for Tumor Classification using gene Expression Data. Bioinformatics, 19(5): 563-570.

4. Bastien, P., Esposito Vinzi, V. y Tenenhaus, M. (2002). Régression Linéaire Généralisée PLS. HEC Research Papers Series, No. 766/2002, HEC School of Business and Management, Jouy-en-Josas, France

5. Blake, C. y Merz, C. (1998). UCI Repository of Machine Learning Databases. Department of Computer Science and Information, University of California, Irvine

6. Brereton R. (2003). *Chemometrics.* Data Analysis for the Laboratory Chemical Plant. John Wiley & Sons.

7. Buendía, L. y Berrocal, E. (2001). La Ética de la Investigación Educativa. http://www.uhu.es/agora/digital/numeros/01/01-articulos/miscelanea/herrera1.PDF

8. Datta, B. N. (1995). *Numerical Linear Algebra and Applications*. Brooks/Cole Publishing Company. An International Thomson Publishing Company

9. Ding, B. y Gentleman, R. (2004). Classification Using Generalized Partial Least Squares. Bioconductor Project Working Papers. http://www.bepress.com/bioconductor/paper5.

10. Dobson, A. (2002). *An Introduction to Generalized Linear Model.* Second Edition, Chapman & Hall/CRC

11. Duckworth, J. (1998) Spectroscopic Quantitative Analysis, in *Applied Spectroscopy*: A compact reference for practitioners. Jerry Workman Jr. and Art Sringsteen, Eds, Academic Press

12. Duda, R.O., Hart, P.E. y Stork, D.G. (2001). *Pattern Classification*. Second Edition, John Wiley, New York

13. Efron B. y Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York

14. Eilers PHC, Boer JM, van Ommen GJB, van Houwelingen JC (2001). Classification of microarray data with penalized logistic regression. *Proc. Int. Symp. Biomedical Optics* 20-26 January, 2001, San Jose, United States.

15. Esposito Vinzi, V. y Tenenhaus M. (2001). PLS Logistic Regression. In PLS and Related Methods, Proceedings of the PLS'01 International Symposium, Esposito Vinci V., Lauro C., Morineau A. & Tenenhaus M. (Eds.). CISIA-CERESTA Editeur, Paris, p. 117-130

16. Firth, D. (1993). Bias reduction of maximum likelihood estimates. (Corr: 95V82 p667). Biometrika, 80:27–38.

17. Fort, G. y Lambert-Lacroix S. (2003). Classification using Partial Least Squares with penalized logistic regression. Technical Report 0331, IAP Statistics Network, Interuniversity Attraction Pole.

18. Frank, I.E. y Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35, 109-148

19. Friedman, J. y Stuetzle, W. (1981). Projection Pursuit Regresión. JASA, 76, 817-823

20. Garthwaite, P.H. (1994). An Interpretation of Partial Least Square. Journal of the American Statistical Association, Vol. 89, No.425, pp. 122-127

21. Ghosh, D. (2002). Singular value decomposition regression modelling for classification of tumors from microarray experiments. Proceedings of the Pacific Symposium on Biocomputing 98, 11462–11467.

22. Golub, G. y Van Loan, C. (1990). *Matrix Computations*. Baltimore: Johns Hopkins University Press

23. Golub, T.,Slonim, P.,Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. y Lander, E. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286, 531-537.

24. Harrel, F., O'Connell, M., Pikounis, W., Pinheiro, J., Ripley, B., Slack, J., Therneau, T. y Venables, W. (2001). *S-Plus 6 for Windows.* Guide to Statistics, Volume 1.

25. Hastie, T. y Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London

26. Hastie, T., Tibshirani, R. y Friedman, J. (2001). *The Elements Statistical Learning*. Data Mining, Inference and Prediction, Springer Series in Statistics.

27. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O.,Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O., Borg, A., Trent, J., (2001). Gene expression profiles in hereditary breast cancer. N Engl J Med 344, 539–548.

28. Heinze, G. y Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine, 21:2409–2419.

29. Helland, I. (1988). On the Structure of Partial Least Squares Regression. Communications in Statistics, Simulation and Computation, 17(2), 581-607

30. Helland, I. (1990). Partial Least Squares Regression and Statistical Models. Scand. J. Statist., 17:97-114

31. Hervé A. (2003). Partial Least Square (PLS) Regression. in Lewis-Beck, M., Bryman, A., Futing, T. (eds.), Encyclopedia of Social Sciences Research Methods, Thousand Oaks

32. Hoskuldsson, A. (1988). PLS Regression Methods. Journal of Chemometrics, 2, 211-228

33. Hosmer, D. y Lemeshow, S. (1989). *Applied Logistics Regression.* John Wiley, New York

34. Huang, X., Pan, W. (2003). Linear regression and two-class classification with gene expression data. Bioinformatics 19, 2072–2078

35. Malthouse, E.C. (1995). Nonlinear Partial Least Square. Thesis, Doctoral dissertation, Department of Statistics, Northwestern University.

36. Mardia, K.V., Kent, J.T. y Bibby, J.M. (1997). *Multivariate Analysis*, Academic Press, London

37. Martens, H., Naes, T. (1989). *Multivariate Calibration*. Wiley, New York

38. Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics, 38:374–381.

39. McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition, Chapman and Hall, London

40. Naes, T., Martens, H. (1985). Comparison of prediction methods for multicollinear data. Communications in Statistics, Part B – Simulation and Computation 14, 545–576

41. Nguyen, D.V. y Rocke, D.M. (2002a). Classification of acute leukemia based on DNA microarray gene expressions using Partial Least Square. In Lin, S.M. and Johnson, K.F. (eds.), Methods of Microarray Data Analysis, Kluwer, Dordrecht, pp. 109-124

42. Nguyen, D.V. y Rocke, D.M. (2002b). Tumor classification by Partial Least Square using microarray gene expression data. Bioinformatics, 18, 39-50

43. Nguyen, D.V. y Rocke, D.M. (2002c). Multi-class cancer classification via Partial Least Square with gene expression profiles. Bioinformatics, 18, 1216-1226

44. Nguyen, D.V. y Rocke, D.M. (2002d). Partial Least Square proportional hazard regression for application to DNA microarray survival data. Bioinformatics 18, 1625-1632

45. Sharaf, M.A., Illman, D.L., Kowalski, B.R. (1986). *Chemometrics*. John Wiley, New York

46. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, Ser. B, 36, 111-147

47. Stone, M. y Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). Journal of the Royal Statistical Society, Ser. B, 52, 237-269

48. Tobias, R. (1995). An Introduction to Partial Least Squares Regression. In Proceedings of the Twentieth Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., 1250-1257

49. Trygg J. (2001). Parsimonious Multivariate Models. Thesis. Research Group for Chemometrics Department of Chemistry, Umea University.

50. Wall, M.E., Rechtsteiner, A. y Rocha, L.M. (2003). Singular value decomposition and principal component analysis. In A Practical Approach to Microarray Data Analysis, Berrar, D.P., Dubitzky, W., Granzow, M. (eds.), pp. 91-109

51. Webb, A. (2002). *Statistical Pattern Recognition*. Second Edition, John Wiley, New York

52. Wilson, M.D., Ustin, S.L. y Rocke, D.M. (2004) Classification of Contamination in Salt Marsh Plants Using Hyperspectral Reflectance. IEEE Transactions on Geosciences and Remote Sensing, vol. 42, No. 5, May 2004

53. Wold, H. (1975). Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Square Approach. In *Perspectives in probability and Statistics*, Papers in Honour of M. S. Bartlett, ed. J. Gani, London: Academic Press.

54. Wold, H. (1984). PLS Regression. In *Encyclopedia of Statistical Sciences*, Vol. 6, eds. N. L. Johnson and S. Kotz, New York: John Wiley, pp. 581-591

55. Wold, S., Martens, H., y Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. Lecture Notes in Mathematics, Springer Verlag, Heidelberg, pp. 286-293

56. Yeung, K.Y. y Ruzzo, W.L. (2001). An empirical study of Component Principal Analysis for clustering gene expression data. Bioinformatics, Vol. 17 no. 9, pp. 763-774.