Minería de Datos

Preprocesamiento: Reducción de dimensionalidad

Dr. Edgar Acuña
Departamento de Ciencias Matemáticas
Universidad de Puerto Rico-Mayaguez

E-mail: edgar@math.uprm.edu, eacunaf@gmail.com

Website: math.uprm.edu/~edgar

Reducción de la Dimensionalidad

- ***Selección de variables*:** El principal objetivo de la selección de variables es reducir la dimensionalidad del espacio de variables, seleccionando variables relevantes y no redundantes. Una variable es redundante cuando da información contenida en alguna otra variable. Una variable es irrelevante si da muy poca información. Esto es, la selección de variables selecciona “q" variables del conjunto completo de “p" variables tal que $q \leq p$. Idealmente $q <<< p$.

- ***Extracción de Variables*:** Se construye un conjunto más pequeño de nuevas variables aplicando una transformación linear (o no linear) al conjunto original de variables. El método más conocido es el de Análisis de Componentes Principales (PCA). Otros: PLS, Curvas Principales.

Selección de Variables

Consideraremos solamente problemas de Clasificación Supervisada.

Objetivo: Seleccionar un subconjunto pequeño de variables tal que:

a) La precisión del clasificador en el conjunto de datos no varíe de forma significativa.

b) La distribución condicional resultante de una clase C, dado el vector de variables seleccionado G, esté tan cerca como sea posible a la distribución condicional original dadas todas las variables F.

Ventajas de la Selección de Variables

- El costo computacional de la clasificación será reducido ya que el número de variables será menor que antes.
- La complejidad del clasificador es reducida ya que se eliminan las variables redundantes y las irrelevantes.
- Ayuda a lidiar con el efecto de la "Maldición de la dimensionalidad", que ocurre en procesos que requieren tener un gran número de instancias en comparación con el número de variables.

Pasos en la selección de variables

1. **Proceso de Generación**: La búsqueda del subconjunto óptimo puede ser: completo, heurístico, aleatorio.
2. **Función de Evaluación**: Medidas de distancia, medidas de información, medidas de consistencia, medidas de dependencia, tasa de error de clasificación.
3. **Criterio de parada**: Un umbral dado, número predefinido de iteraciones, tamaño predefinido del mejor subconjunto de variables.
4. **Proceso de Validación** (Opcional) Verificar si el subconjunto es válido.

Guía para escoger un método de selección de variables

- Habilidad para manejar distintos tipos de variables (continua, binaria, nominal, ordinal)
- Habilidad para manejar múltiples clases.
- Habilidad para manejar grandes conjuntos de datos.
- Habilidad para manejar datos ruidosos.
- Baja complejidad en tiempo.

Categorización de los métodos de selección de variables (Dash and Liu, 1997)

Medidas de Evaluación	*Generación*		
	Heuristica	*Completa*	*Random*
Distancia	**Relief**	**Branch and Bound**	-
Información	**Trees**	**MDL**	-
Dependencia	**POEIACC**	-	-
Consistencia	**FINCO**	**Focus**	**LVF**
Tasa de Error de Clasificación	**SFS, SBS,SFFS**	**Beam Search**	**Genetic Algorith**

Los métodos de la ultima fila son también conocidos como los métodos “wrapper” (Kohavi). Los otros son llamados metodos de filtrado

Métodos de Filtrado

Estos no requieren un clasificador, en lugar de ello usan medidas que permiten seleccionar las variables que distinguen mejor a la clase.

- RELIEF
- Las Vegas Filter (LVF)
- FINCO
- Otros : Branch & Bound, Focus,

El método RELIEF

- Kira and Rendell (1992) para problemas de dos clases y generalizado a problemas multi-clases por Kononenko (1994) y Kononenko, et al. (1997).
- Genera subconjuntos de variables en forma heurística.
- Una variable tiene un peso de relevancia que es grande si puede distinguir claramente dos instancias que pertenecen a distintas clases pero no dos instancias que están en la misma clase.
- Usa una medida de distancia (Euclideana, Manhattan)

El método RELIEF (procedimiento)

- Se selecciona al azar un número dado ***Nsample*** de instancias del conjunto de entrenamiento D que contiene F variables.
- Los pesos de relevancia W_j de cada variable son inicializados en cero.
- Para cada instancia seleccionada **x**, se debe identificar dos instancias particulares:

  Nearhit: La instancia mas cercana a **x** que pertenece a la misma clase.

  Nearmiss: La instancia mas cercana a **x** que pertenece a una clase diferente.

El método RELIEF (cont)

- Luego los pesos W_j's (i=1,..F) son actualizados usando la relación

$$W_j = W_j - \text{diff}(x_j, \text{Nearhit}_j)^2/NS + \text{diff}(x_j, \text{Nearmiss}_j)^2/NS$$

Si la variable X_k es nominal o binaria, entonces:

- $\text{diff}(x_{ik},x_{jk}) = 1$ para $x_{ik} \neq x_{jk}$
  $= 0$ en caso contrario.

Si la variable X_k es continua u ordinal entonces:

- $\text{diff}(x_{ik},x_{jk}) = (x_{ik}-x_{jk})/c_k$, donde $c_k = \text{range}(X_k)$

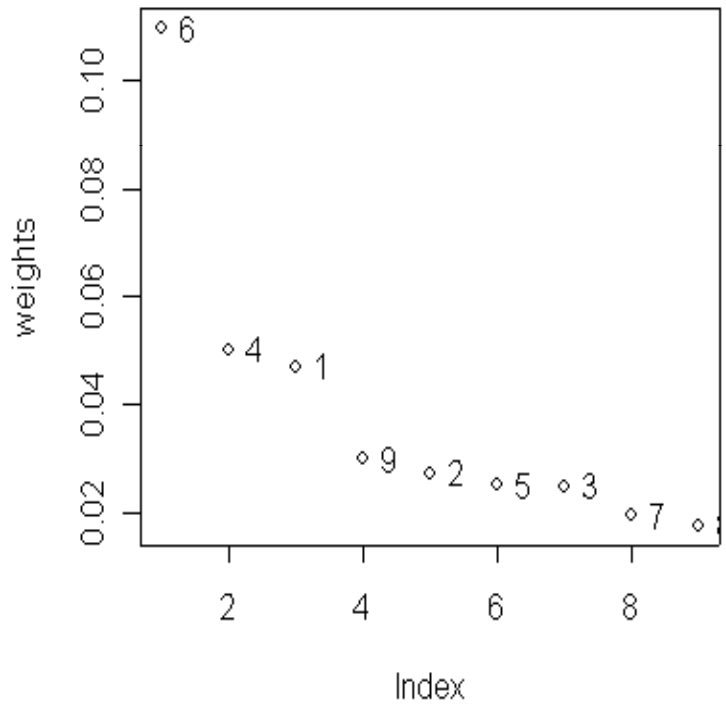
Decisión: Si $W_j \geq \tau$ (un valor preestablecido) entonces se selecciona la variable f_j

Conjunto de datos: Breast-Wisconsin

- 699 instancias, 9 variables y dos clases (benigno o maligno). 16 instancias han sido eliminadas por tener valores faltantes.
- 1. Clump Thickness 2. Uniformity of Cell Size, 3. Uniformity of Cell Shape, 4. Adhesion Marginal Adhesion, 5. Single Epithelial Cell Size, 6. Bare Nuclei, 7. Bland Chromatin 8. Normal. nucleoli 9. Mitoses.
- Cada variable tiene valores en un rango de 0 a 10.

Ejemplo de Relief: Breastw

```
> relief(breastw,600,0)
      Variables que aparecen en por lo menos la mitad de las
      repeticiones ordenadas por su peso de relevancia
      promedio:
     feature frequency     weight
[1,]       6        10 0.10913169
[2,]       4        10 0.05246502
[3,]       1        10 0.04682305
[4,]       9        10 0.03171399
[5,]       2        10 0.02869547
[6,]       3        10 0.02566461
[7,]       5        10 0.02512963
[8,]       7        10 0.02096502
[9,]       8        10 0.01708025
Variables Seleccionadas
[1] 6 4 1 9 2 3 5 7 8
> relief(breastw,600,0.04)
      Variables que aparecen en por lo menos la mitad de las
      repeticiones ordenadas por su peso de relevancia
      promedio:
feature frequency     weight
[1,]       6        10 0.10844239
[2,]       4        10 0.05293210
[3,]       1        10 0.04853909
Variables Seleccionadas
[1] 6 4 1
```

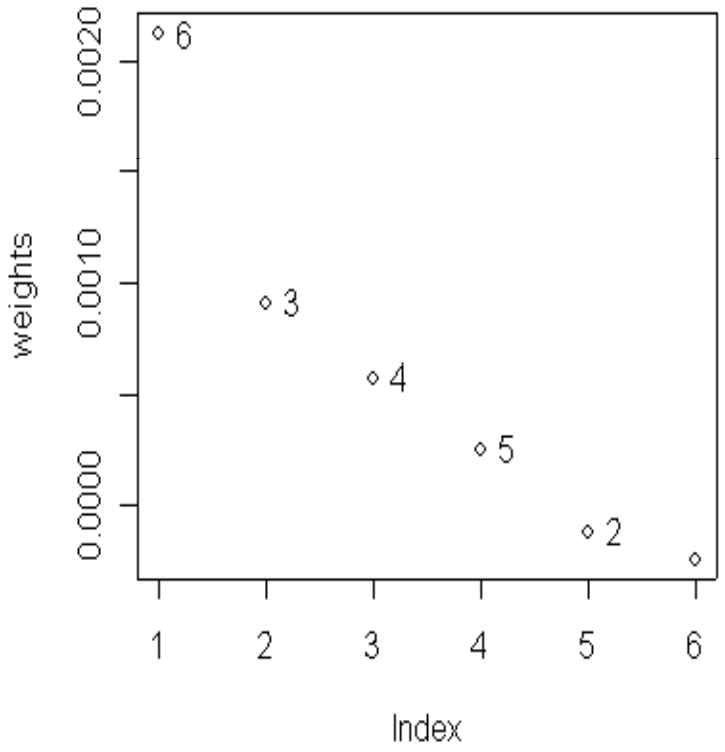


Ejemplo de Relief: Bupa

```
> relief(bupa,345,0.0003)
```

Variables que aparecen en por lo menos la mitad de las repeticiones ordenadas por si peso de relevancia promedio:

```
     feature frequency      weight
[1,]       6         6 0.0021190217
[2,]       3         8 0.0009031895
[3,]       4         8 0.0005711548
Variables Seleccionadas
[1] 6 3 4
>
```



Método Relief: problema multi-clase

Primero se debe encontrar un ***Nearmiss*** por cada clase distinta desde **x**, y luego se promedia su contribución usando los pesos basados en anteriores. Los pesos son actualizados usando:

$$W_j = W_j - diff\,(x_j, \text{Nearhit})^2 + \sum_{C \neq class(x_j)} \frac{P(C)}{1 - P(class(x_j))}\, diff\,(x_j, \text{Nearmiss}\,(C))^2$$

Conjunto de datos Vehicle

- 846 instancias, 18 variables continuas y 4 Clases (double decker bus, Chevrolet van, Saab 9000 y Opel Manta 400).
- [,1] Compactness [,2] Circularity [,3] Distance Circularity [,4] Radius ratio [,5] p.axis aspect ratio [,6] max.length aspect ratio[,7] scatter ratio [,8] elongatedness [,9] pr.axis rectangularity [,10] max.length rectangularity [,11] scaled variance along major axis[,12] scaled variance along minor axis[,13] scaled radius of gyration[,14] skewness about major axis[,15] skewness about minor axis[,16] kurtosis about minor axis[,17] kurtosis about major axis[,18] hollows ratio.

relief(vehicle,400,0.012)
Variables que aparecen en por lo menos la mitad de las repeticiones ordenadas por si peso de relevancia promedio:

	feature	frequency	weight
[1,]	16	10	0.03375733
[2,]	18	10	0.03087840
[3,]	15	10	0.01991083
[4,]	17	10	0.01586413
[5,]	10	10	0.01521946
[6,]	12	9	0.01433016
[7,]	9	9	0.01372653
[8,]	3	10	0.01369564
[9,]	1	9	0.01337022
[10,]	7	8	0.01278588
[11,]	8	8	0.01267531
[12,]	2	5	0.01201989

Variables Seleccionadas
[1] 16 18 15 17 10 12 9 3 1 7 8 2

El método Relief (Cont)

Ventajas:

Trabaja bien con variables ruidosas y correlacionadas.
La complejidad en tiempo es linear en el número de variables y en número de muestras tomadas (Nsample).
Trabaja para cualquier tipo de variable.

Desventajas:

Elimina las variables irrelevantes pero no elimina las variables redundantes.

Elección del umbral.
Elección de Nsample.

Método Las Vegas Filter (LVF)

Liu y Setiono (1997)

- El subconjunto de variables es elegido al azar.
- La función de evaluación usada es una medida de inconsistencia.
- Dos instancias son inconsistentes si tienen los mismos valores en las variables pero pertenecen a clases diferentes.
- Las variables continuas del conjunto de datos tienen que ser previamente discretizadas.

Medida de Inconsistencia

La inconsistencia de un conjunto de datos con solo variables no-continuas está dado por:

$$\frac{\sum_{i=1}^{K} (|D_i| - h_i)}{N}$$

K: Número de combinaciones distintas de las N instancias.

$|D_i|$: Cardinalidad de la i-ésima combinación.

h_i: Frecuencia de la clase modal en la i-ésima combinación.

Inconsistencia: Ejemplo

```
> m1
    col1 col2 col3 col4 class
[1,]  1    2  2    1    1
[2,]  4    3  2    2    2
[3,]  4    3  2    2    1
[4,]  1    3  8    1    1
[5,]  9    3  8    2    2
[6,]  9    3  8    1    2
[7,]  9    3  1    2    1
> inconsist(m1)
[1] 0.1428571
```

Algoritmo LVF

Input : D = conjunto de datos , p = número de variables,
S = conjunto de todas las variables,
MaxTries = número máximo de ensayos,
Umbral = τ .
C_{best} = p , S_{best} = S
Para desde i= 1 hasta MaxTries
S_i = Subconjunto de S elegido al azar.
C = card(S_i)
Si (C < C_{best})
{ Si Inconsistencia (S_i, D) <τ
S_{best}= S_i , C_{best} = C }
si (C = C_{best} y Inconsistencia (S_i, D)$\leq \tau$)
S_{best}= S_i.
Output : S_{best}

Ejemplos

> dbupa=disc.ew(bupa,1:6)
>inconsist(dbupa)
[1] 0.01159420
> lvf(dbupa,.1,1000)
La inconsistencia del mejor subconjunto es:
0.05217391
El mejor subconjunto de variables es:
[1] 1 2 3 6
> > lvf(breastw,.01,2000)

La inconsistencia del mejor subconjunto es:

0.005856515

El mejor subconjunto de variables es:

[1] 1 6 8

Desventajas de LVF

- Selección del umbral. Un umbral con un valor pequeño implicará la selección de un número grande de variables.
- Un número grande de iteraciones disminuye la variabilidad de conjunto seleccionado pero hace lento el cómputo.

Método FINCO

FINCO (Acuna, 2002) combina la selección secuencial hacia adelante con una medida de inconsistencia como función de evaluación.

PROCEDIMIENTO

- Se inicializa el mejor subconjunto de variables T como un conjunto vacío.
- En el primer paso, se selecciona la variable que produce el menor nivel de inconsistencia.
- Luego, se selecciona la variable que junto con la primera variable seleccionada produzca el menor nivel de inconsistencia.
- El proceso continúa hasta que cada variable que aun no fue seleccionada junto con las variables que ya estén en T produzcan un nivel de inconsistencia menor que un umbral τ.

Algoritmo FINCO

Input : D = Conjunto de datos, p = Número de variables en D,
S = conjunto de todas las variables, Umbral = τ .

Initialization:

Set k=0 and $T_k = \phi$

Inclusión: Para k=1 hasta p

Seleccionar variable x^+ tal que:

$$x^+ = \arg\min_{x \in S - T_k} Incons\ (T_k + x)$$

donde S- T_k es el subconjunto de variables que aun no fueron seleccionadas.

si Incons(T_k+x^+)<Incons(T_k) and Incons(T_k+x^+)>τ, entonces

$T_{k+1} = T_k + x^+$ y k:=k+1

sino parar

Output: T_k: subconjunto de variables seleccionadas

Ejemplos

```
> finco(dbupa,.05)
Variables seleccionadas y sus tasas de inconsistencia
$varselec
[1] 2 1 6 3
$inconsis
[1] 0.37681159 0.26376812 0.13333333 0.05217391
> finco(breastw,.01)
Variables seleccionadas y sus tasas de inconsistencia
$varselec
[1] 2 6

$inconsis
[1] 0.07027818 0.02635432
```

finco(breastw,.001)
Variables seleccionadas y sus tasas de inconsistencia
$varselec
[1] 2 6 1

$inconsis
[1] 0.070278184 0.026354319 0.005856515

- El umbral se elige como un valor un poco mas grande que la inconsistencia del conjunto de datos.

Métodos Wrapper

Wrappers usan la tasa de error de mala clasificación como la función de evaluación para el subconjunto de variables.

- Selección secuencial hacia adelante (SFS)
- Selección secuencial hacia atrás (SBS)
- Selección secuencial flotante hacia adelante (SFFS)
- Otros: SFBS, Take l-remove r, GSFS, GA, SA.

Selección secuencial hacia adelante (SFS)

- Se inicializa el mejor subconjunto de variables T como el conjunto vacío.
- La primera variable que ingresa a T es aquella con la tasa de reconocimiento más alta con un clasificador dado.
- La segunda variable que ingresa a T será aquella que junto con la variable seleccionada en el paso previo produzca la tasa de reconocimiento más alta.
- El proceso continúa y en cada paso ingresa solamente una variable a T hasta que la tasa de reconocimiento no crezca cuando el clasificador se construya usando las variables que ya estén en T mas cada una de las demás variables.

Ejemplos: Bupa y Breastw

```
 sfs(bupa,"knn") #knn classifier
El mejor subconjunto de variables es:
[1] 5 3 1
> sfs(bupa,"lda") #Linear discriminant classifier
El mejor subconjunto de variables es:
[1] 5 4 3 6
> sfs(bupa,"rpart") #decision tree classifier
El mejor subconjunto de variables es:
[1] 5 3 2
> sfs(breastw,"knn")
El mejor subconjunto de variables es:
[1] 6 1 3 7
> sfs(breastw,"lda")
El mejor subconjunto de variables es:
[1] 6 2 1 4
> sfs(breastw,"rpart")
El mejor subconjunto de variables es:
[1] 6 3 5
```

Selección secuencial hacia atrás (SBS)

- Inicialmente el mejor subconjunto de variables T incluyen todas las variables del conjunto de datos.
- En el primer paso, se realiza la clasificación sin considerar cada una de las variables, y se elimina la variable donde la tasa de reconocimiento es la más baja.
- El proceso continúa eliminando una variable en cada paso hasta que las tasas de reconocimiento con las variables usadas comienzan a decrecer.

  No eficiente para los clasificadores no paramétricos porque tiene un alto tiempo de computación

Selección secuencial flotante hacia adelante(SFFS)

Pudil, et al (1994). Trata de resolver el problema de anidamiento que aparece en SFS y SBS.

- Inicialmente el mejor subconjunto de variables T es el conjunto vacio.
- En cada paso se incluye una nueva variable en T usando SFS, pero es seguida por una verificación de una posible exclusión de variables que ya estén en T. Las variable son excluidas usando SBS hasta que la tasa de reconocimiento comience a decrecer.
- El proceso continúa hasta que SFS no se pueda ejecutar.

Ejemplos

```
> sffs(bupa,"lda")
Las variables seleccionadas son:
[1] 3 4 5
> library(class)
> sffs(bupa,"knn")
Las variables seleccionadas son:
[1] 5 3
> library(rpart)
> sffs(bupa,"rpart")
Las variables seleccionadas son:
[1] 3 5 6 2
> sffs(breastw,"lda")
Las variables seleccionadas son:
[1] 1 2 6 4
> sffs(breastw,"knn")
Las variables seleccionadas son:
[1] 6 3 7 1
> sffs(breastw,"rpart")
Las variables seleccionadas son:
[1] 6 3 2
```

Conclusiones

- Entre los métodos wrapper, SFFS se comporta mejor que SFS: menor porcentaje de variables seleccionadas y casi la misma exactitud que SFS. Cómputo rápido.
- Entre los métodos de filtrado, FINCO tiene el menor porcentaje de variables seleccionadas.
- El rendimiento de LVF y RELIEF es bastante similar, pero LVF toma más tiempo de cómputo.
- Los métodos wrapper son más efectivos que los metodos de filtardo reduciendo el error de mala clasificación.
- La velocidad de cómputo de los métodos de filtrado es afectada por el tamaño de la muestra y el número de clases.

Conclusiones (Cont.)

- SFFS y FINCO tienen menor porcentaje de variables seleccionadas .
- En LVF, un incremento en el número de iteraciones disminuye la variabilidad de las variables seleccionadas.
- En LVF y FINCO, una reducción del nivel de inconsistencia mínimo aumenta el número de variables seleccionadas.

Extracción de Variables

Consideraremos solamente problemas de Clasificación Supervisada.

Objetivo: Construir nuevas variables a partir de las variables originales tal que la precisión del clasificador construido usando estas nuevas variables originales no varíe de forma significativa.

Veremos solamente el metodo de componentes principales.

Análisis de Componentes Principales (PCA)

El objetivo del Análsis de Componentes Principales (Hotelling, 1933) es reducir la información disponible. Esto es, la información contenida en p variables $\mathbf{X}=(X_1,\ldots,X_p)$ puede ser reducida a $\mathbf{Z}=(Z_1,\ldots Z_q)$, *con* $q<p$. Las nuevas variables Z_i's , llamadas ***Componentes Principales*** y son no son correlacionadas.

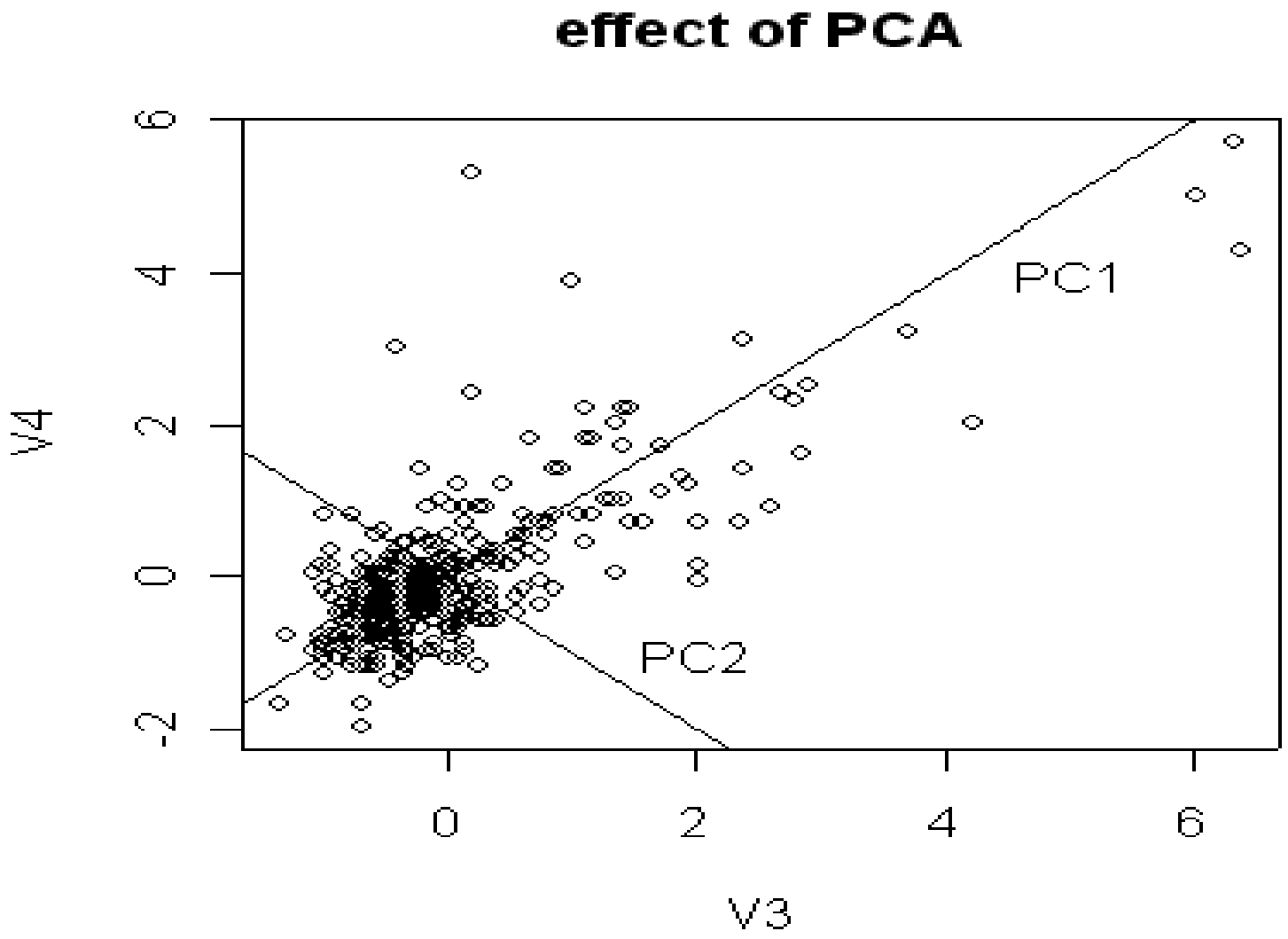
Los componentes principales de un vector **X** son los elementos de una transformación lineal ortogonal de **X**

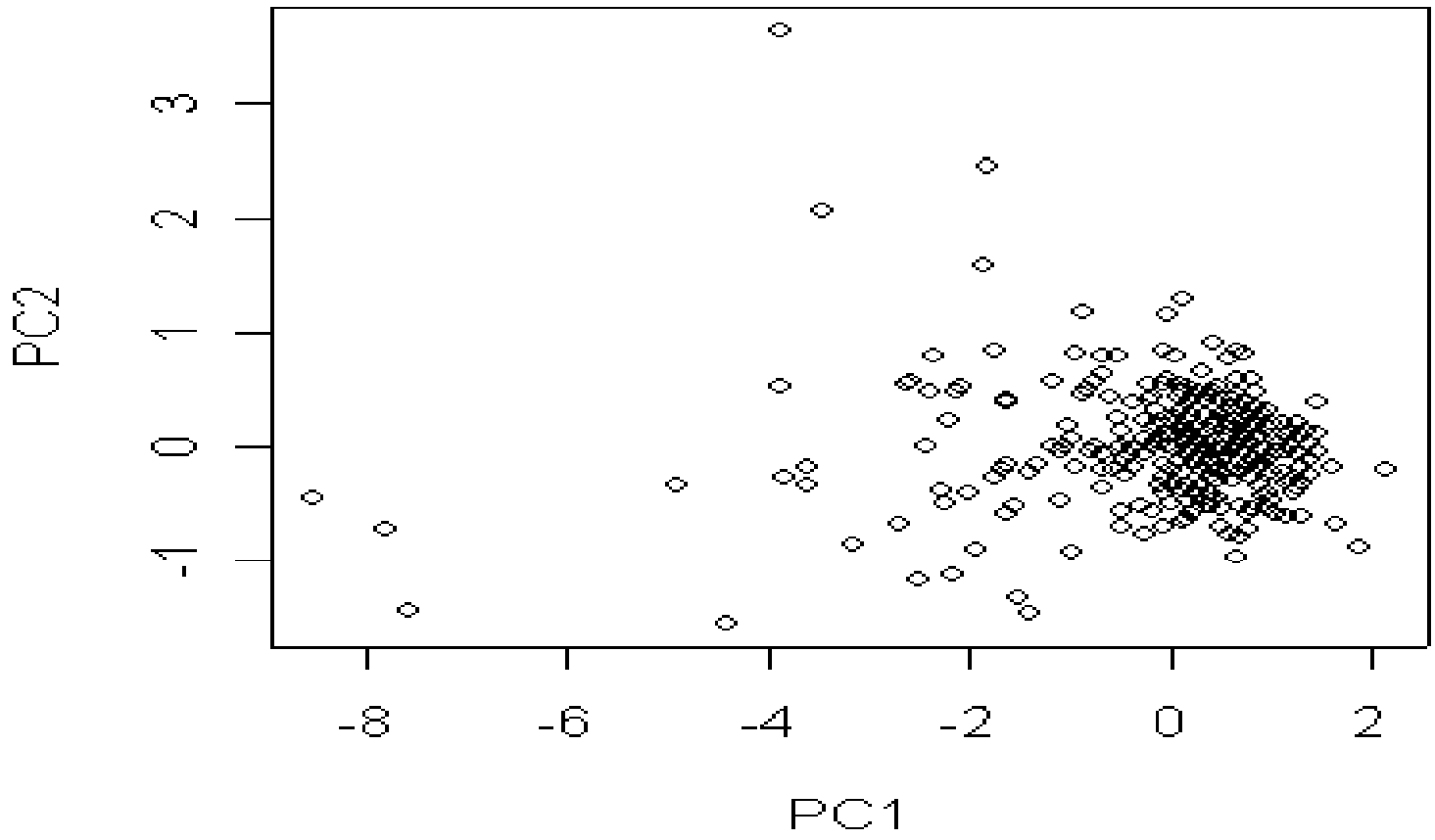
Desde un punto de vista geométrico, aplicar los componentes principales es equivalente a aplicar una rotación de los ejes coordenados.

Ejemplo: Bupa (p=q=2)

```
> bupapc=prcomp(bupa[,c(3,4)],scale=T,retx=T)
> print(bupapc)
Desviación estándar:
[1] 1.3189673 0.5102207

Rotation:
         PC1        PC2
V3 -0.7071068 -0.7071068
V4 -0.7071068  0.7071068
```


effect of PCA
PC1
PC2
V4
V3
6
4
2
0
-2
0
2
4
6



Note que PC1 y PC2 no están correlacionadas

Hallando los Componentes Principales

Para determinar los Componentes Principales Z, debemos encontrar una matriz ortogonal V tal que

i) $Z=X^*V$,

donde X* es obtenida de normalizar cada columna de X. Se puede demostrar que $VV'=V'V=I$

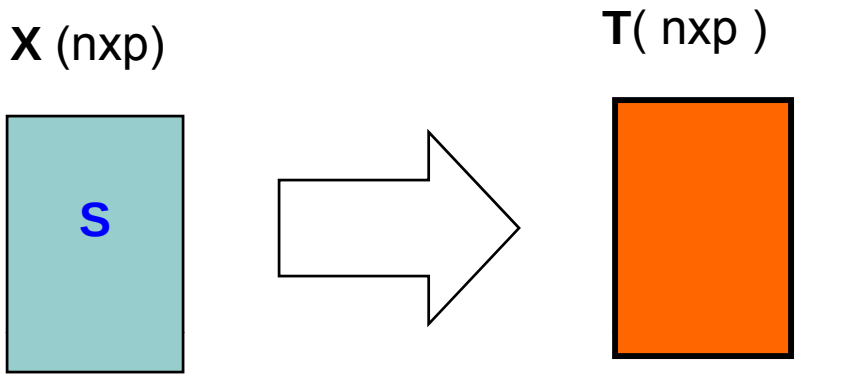
y ii) $Z'Z=(X^*V)'(X^*V) =V'X^{*'}X^*V$

$$=\text{diag}(\lambda_1,\ldots.,\lambda_p)$$

donde los λ_j's son los valores propios de la matriz de correlación $X^{*'}X^*$. V se encuentra usando descomposición de valores singulares de $X^{*'}X^*$.

La matriz V se llama matriz de cargas y contiene los coeficientes de todas las variables en cada componente principal.

PCA COMO UN PROBLEMA DE OPTIMIZACIÓN



$$\mathbf{T}_k = \text{argmax var}\ (X\gamma)$$
$$\gamma'\gamma = 1$$

Sujeto a la restricción de ortogonalidad

$$\gamma_j' \mathbf{S} \gamma_k = 0 \qquad \forall \ 1 \leq j < k$$

De (ii) el j-ésimo componente principal Z_j tiene una desviación estándar $\sqrt{\lambda_j}$ y puede ser escrito como:

$$Z_j = v_{1j} X_1^* + v_{2j} X_2^* + ... + v_{pj} X_p^*$$

Donde $v_{j1}, v_{j2}, \ldots . v_{jp}$ son los elementos de la j-ésima columna en V.

Los valores calculados del componente principal Z_j son llamados los valores rotados o simplemente los “scores”.

Selección del número de componentes principales

Existen muchas alternativas (Ferre, 1994), pero las más usadas son:

- Escoger el número de componentes con una proporción acumulada de valores propios (i.e, varianza) de por lo menos 75 por ciento.
- Escoger hasta el componente cuyo valor propio es mayor que 1. Usar "Scree Plot".

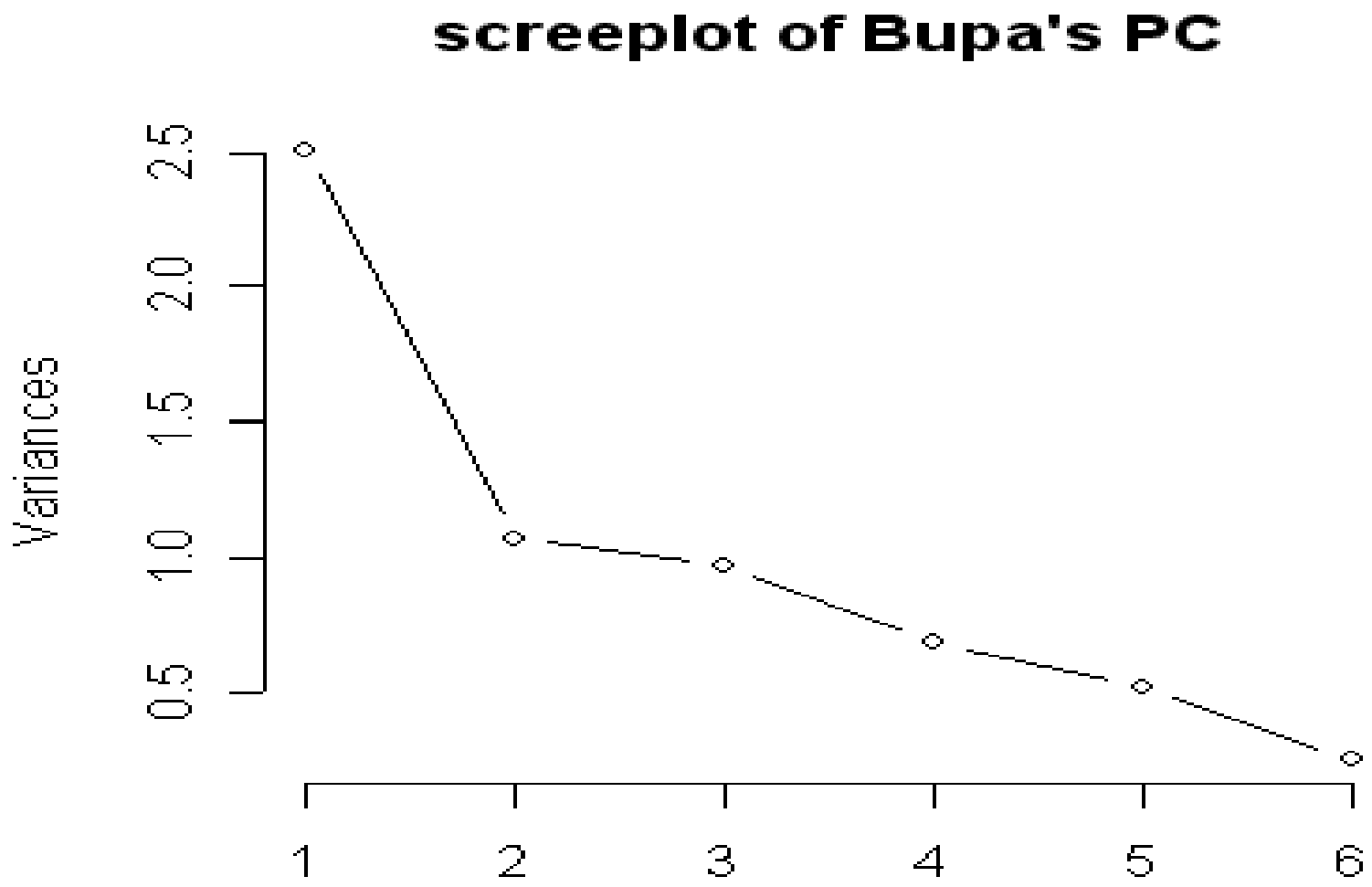
Ejemplo: Bupa

```
> a=prcomp(bupa[,-7],scale=T)
> print(a)
Desviación estándar:
[1] 1.5819918 1.0355225 0.9854934 0.8268822 0.7187226 0.5034896

Rotation:
        PC1        PC2        PC3        PC4        PC5         PC6
V1 0.2660076  0.67908900  0.17178567 -0.6619343  0.01440487
   0.014254815
V2 0.1523198  0.07160045 -0.97609467 -0.1180965 -0.03508447
   0.061102720
V3 0.5092169 -0.38370076  0.12276631 -0.1487163 -0.29177970
   0.686402469
V4 0.5352429 -0.29688378  0.03978484 -0.1013274 -0.30464653 -
   0.721606152
V5 0.4900701 -0.05236669  0.02183660  0.1675108  0.85354943
   0.002380586
V6 0.3465300  0.54369383  0.02444679  0.6981780 -0.30343047
   0.064759576
```

Ejemplo (cont)

```
> summary(a)
Importancia de componentes:
                          PC1   PC2   PC3   PC4    PC5    PC6
Desviación estándar     1.582 1.036 0.985 0.827 0.7187 0.5035
Proportion of Variance 0.417 0.179 0.162 0.114 0.0861 0.0423
Cumulative Proportion  0.417 0.596 0.758 0.872 0.9577 1.0000
>
```



El screen-plot recomienda elegir dos componentes principales

Comentarios

- Muchos estudios han demostrado que PCA no da buenas predicciones en clasificación supervisada.
- Mejores alternativas: PLS generalizados (Vega,2004) y PCA supervisados (Hastie, Tibshirani, 2004, Acuna and Porras, 2006).
- Para mejorar el rendimiento de componentes princiaples es mejor hacer primero selección de variables.