Revista Colombiana de Estadística
Universidad Nacional de Colombia
revcoles_fcbog@unal.edu.co
ISSN (Versión impresa): 0120-1751
COLOMBIA

2005
Guillermo Ramírez / Maura Vasquez / Alberto Camardiel / Betty Perez / Purificación Galindo
DETECCIÓN GRÁFICA DE LA MULTICOLINEALIDAD MEDIANTE EL H-PLOT DE LA INVERSA DE LA MATRIZ DE CORRELACIONES
Revista Colombiana de Estadística, diciembre, año/vol. 28, número 002
Universidad Nacional de Colombia
Bogotá, Colombia
pp. 207- 219

Revista Colombiana de Estadística
Volumen 28 N^{o} 2. pp. 207 a 219. Diciembre 2005

Detección gráfica de la multicolinealidad mediante el h-plot de la inversa de la matriz de correlaciones

Multicolinearity Detection by Means of the h-Plot of the Correlation Matrix Inverse

Guillermo Ramírez[*]
Maura Vasquez[**]
Alberto Camardiel[***]
Betty Perez[****]
Purificacion Galindo[*****]

Resumen

La multicolinealidad origina imprecisión en los estimadores de los coeficientes de un modelo lineal. En este trabajo proponemos un gráfico basado en la representación h-plot de la inversa de la matriz de correlaciones, que permite visualizar con cierto grado de aproximación las relaciones lineales entre las variables predictoras. En este dispositivo se obtienen representaciones aproximadas de los coeficientes de inflación de varianza de cada variable y de las correlaciones parciales entre ellas. Con el objeto de ilustrar el método, éste se aplicó en una investigación sobre la caracterización morfológica de jóvenes nadadores venezolanos.

Palabras Claves: Multicolinealidad, h-plot, correlación parcial, coeficiente de inflación de varianza.

Abstract

Multicollinearity generates imprecision in the estimates of the coefficients in linear models. We propose to use the h-plot of the inverse of the correlation matrix to obtain a representation of the linear relations between the

[*]Profesor Titular, Postgrado en estadística, Universidad Central de Venezuela (UCV), E-Mail: guiram@cantv.net
[**]Profesor Titular, Postgrado en estadística, Universidad Central de Venezuela, (UCV), E-Mail: mauravasquez@cantv.net
[***]Profesor Titular, Postgrado en estadística, Universidad Central de Venezuela (UCV), E-Mail: acamar@reacciun.ve
[****]Profesor Titular, Instituto de Investigaciones Económicas y Sociales, UCV, E-Mail: mariusa@telcel.net
[*****]Profesor Titular, Departamento de Estadística y Matemática Aplicadas, Universidad de Salamanca, E-Mail: pgalindo@aida.usal.es

predictor variables. In the resulting plot the variance inflation factor of each variable and the partial correlation between them area roughly displayed. In order to illustrate the method it was applied in an anthropometric study of young Venezuelan swimmers.

Keywords: Multicollinearity, h-plot, partial correlation, variance inflation factor.

1. Introducción

1.1. El Problema

Multicolinealidad es el término usualmente utilizado para referirse a la existencia de relaciones lineales o cuasilineales entre las variables predictoras en un modelo lineal, lo que indica que parte sustancial de la información en una o más de estas variables es redundante. Mandell (1982) señala que una de las principales dificultades en el uso de estimaciones mínimo cuadráticas es la presencia de este fenómeno que, aún cuando no afecta la capacidad predictiva del modelo, representa un problema grave si su propósito fundamental es evaluar la contribución individual de las variables explicativas. Esto es debido a que en presencia de multicolinealidad los coeficientes b_j tienden a ser inestables, es decir sus errores estándar presentan magnitudes indebidamente grandes. Esta falta de precisión afecta los contrastes parciales diseñados para evaluar la contribución individual de cada variable explicativa, corriéndose un alto riesgo de no encontrar significación en variables que realmente la tengan. Jackson (1991) subraya además que bajo condiciones de colinealidad resulta imposible distinguir los efectos individuales de cada variable predictora, debido a que la fuerza de la correlación entre ellas produce relaciones lineales de similar magnitud entre los coeficientes.

1.2. Antecedentes

Numerosos métodos han sido desarrollados con el objeto de detectar la posible existencia de multicolinealidad y sus efectos anómalos sobre un modelo de regresión. Una primera aproximación al problema plantea analizar la matriz de correlaciones $\mathbf{R}$, procedimiento útil pero que no capta el fenómeno en toda su intensidad puesto que estudia las relaciones entre las variables dos a dos, obviando las relaciones de éstas con las otras variables predictoras. Otras propuestas alternativas se basan en el coeficiente de determinación múltiple de cada variable X_j con las restantes, y en los coeficientes de correlación parcial de las variables X_j y X_k, controlando por los efectos lineales de las restantes. En particular, Kendall (1957) enfoca el abordaje práctico de la colinealidad a través de un procedimiento diseñado en función de los autovalores y autovectores de la matriz de correlaciones. Silvey (1969) plantea como una forma de superar el problema, agregar nuevos valores de las variables explicativas que eliminen la colinealidad, obtenidos como función de los autovectores asociados a autovalores nulos. Mandell (1982) demues-

tra que el error estándar del j-ésimo coeficiente de regresión puede expresarse como el producto del error estándar residual de la regresión por el factor de inflación de varianza (VIF), ampliamente utilizado para detectar multicolinealidad; en particular demuestra que el VIF se ve severamente afectado por los autovalores más pequeños de la matriz $\mathbf{R}$. Este coeficiente mide el incremento que se produce en la varianza de b_j respecto del valor mínimo que se alcanzaría en ausencia total de colinealidad de la correspondiente variable X_j respecto de las restantes variables predictoras (Glantz & Slinker 2001). Adicionalmente, la expresión VIF(j) - 1 coincide, excepto por los grados de libertad, con el estadístico que contrasta la bondad del ajuste de la regresión de X_j como función de las restantes variables explicativas. Mason, Gunst & Webster (1975) proponen el índice conocido como condition number, definido como el cociente entre el mayor y el menor autovalor de la matriz $\mathbf{R}$. Por su parte, Gleason & Staelin (1975) proponen un índice basado en los autovalores de la matriz de correlaciones que toma el valor 0 cuando las variables son independientes ($\mathbf{R} = \mathbf{I}$) y el valor 1 cuando las variables están perfectamente correlacionadas ($\mathbf{R} = \mathbf{J}$). Raveh (1985) discute la importancia de ciertos elementos fuera de la diagonal principal de la inversa de la matriz de correlaciones para detectar predictores importantes en un análisis de regresión y como criterio para evaluar los supuestos requeridos para aplicar un análisis de factores. Whitakker (1990) resalta la utilidad de la inversa de la matriz de correlaciones para establecer relaciones de dependencia entre variables y propone su representación gráfica mediante los denominados grafos de independencia condicional. Belsley (1991) refiere las ventajas y debilidades de los VIF como medida para el diagnóstico de la colinealidad. Yu (1998) desarrolla un programa multimedia para ilustrar visualmente la forma como un modelo de regresión puede colapsar cuando las variables predictoras están intercorrelacionadas.

2. Fundamentación teórica

La propuesta que presentaremos en este trabajo está basada en la técnica h-plot aplicada a la inversa de la matriz de correlaciones (Si $\mathbf{R}$ fuese singular, la técnica se aplicaría sobre la inversa generalizada $\mathbf{R}^g$). En este apartado describiremos los fundamentos de este método y centraremos nuestra atención en los elementos genéricos de la matriz $\mathbf{R}^{-1}$.

2.1. Método h-plot

El h-plot es un procedimiento introducido por Cornsten & Gabriel (1976) para obtener representaciones gráficas de la información contenida en una matriz de varianzas y covarianzas $\mathbf{S}_{pxp}$ de rango r, sobre espacios reducidos de baja dimensión. La selección adecuada de los vectores, denominados marcadores por sus autores, garantiza que en su representación sobre el primer plano h-plot se cumple que:

(a) El producto escalar entre dos marcadores aproxima la covarianza entre las variables correspondientes,

(b) La longitud de los marcadores aproxima la desviación estándar de las variables,

(c) El coseno del ángulo entre dos marcadores aproxima la correlación entre las variables correspondientes, y,

(d) El plano proporciona la mejor representación bidimensional aproximada, desde el punto de vista de los mínimos cuadrados, de las relaciones entre las variables en términos de varianzas y correlaciones.

La representación en referencia es aplicable a cualquier matriz simétrica $\mathbf{A}_{pxp}$ de rango r, y se efectúa eligiendo vectores marcadores $(\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_p)$ para sus columnas, tales que los elementos de la matriz se obtienen a partir de operaciones de producto interno entre los marcadores:

$$\mathrm{a}_{jj} = \mathrm{h}_j^t \mathrm{h}_j \qquad \mathrm{y} \qquad \mathrm{a}_{jk} = \mathrm{h}_j^t \mathrm{h}_k$$

El procedimiento para la selección de los marcadores se basa en la descomposición espectral de la matriz $\mathbf{A}$:

$$\mathbf{A} = \mathbf{V_{(r)}} \mathbf{D_{(r)}} \mathbf{V_{(r)}^t}$$

de manera que la matriz cuyas columnas contienen los marcadores se define en la forma:

$$\mathbf{H_{(r)}^t} = \mathbf{D_{(r)}^{1/2}}\ \mathbf{V_{(r)}^t} = (\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_p)$$

siendo $\mathbf{V_{(r)}}$ una matriz cuyas columnas son los autovectores de $\mathbf{A}$ asociados con sus r autovalores no nulos, y $\mathbf{D_{(r)}}$ una matriz diagonal que contiene tales autovalores.

Además, Gabriel (1971) ha propuesto como medida de la bondad de la aproximación sobre el primer plano h-plot, al cociente:

$$\left(\frac{\lambda_1^2 + \lambda_2^2}{\sum \lambda_\alpha^2} \right)$$

siendo λ_1 y λ_2 los dos mayores autovalores de la matriz $\mathbf{A}$.

2.2. Elemento genérico de la matriz $\mathbf{R}^{-1}$

Con el objeto de detectar multicolinealidad en un conjunto de variables, construiremos un plano h-plot sobre el cual se representará la información contenida en la inversa de la matriz de correlaciones $\mathbf{R}^{-1} = (\mathrm{r}^{jk})$. Con este fin consideraremos la siguiente permutación de las columnas de la matriz de variables $\mathbf{X}_{nxp}$:

$$\mathbf{X}_* = (\mathrm{X}^j \mathrm{X}^k \mathbf{X}_{(-j,-k)})$$

donde $\mathbf{X}_{(-j,-k)}$ es una matriz cuyas columnas contienen la información de todas las variables excepto la j y la k.

La correspondiente matriz de correlaciones particionada:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{(j,k)} & \mathbf{R}_{(j,k),(-j,-k)} \\ \mathbf{R}_{(-j,-k),(j,k)} & \mathbf{R}_{(-j,-k)} \end{bmatrix}$$

siendo $\mathbf{R}_{(j,k),(-j,-k)}$ la matriz de correlaciones, de orden $2 \times (p-2)$, de X^j y X^k respectivamente con las restantes variables:

$$\mathbf{R}_{(j,k),(-j,-k)} = \begin{bmatrix} R_{j,(-j,-k)} \\ R_{k,(-j,-k)} \end{bmatrix}$$

La inversa de $\mathbf{R}$ se denotará mediante:

$$\mathbf{R}^{-1} = \begin{bmatrix} \check{\mathbf{R}}_{11} & \check{\mathbf{R}}_{12} \\ \check{\mathbf{R}}_{21} & \check{\mathbf{R}}_{22} \end{bmatrix}$$

donde $\check{\mathbf{R}}_{11} = (\mathbf{R}_{(j,k)} - \mathbf{R}_{(j,k),(-j,-k)}\ \mathbf{R}^{-1}_{(-j,-k)}\mathbf{R}_{(-j,-k),(j,k)})^{-1}$, siendo $\check{\mathbf{R}}^{-1}_{11}$ de la forma:

$$\begin{aligned} \check{\mathbf{R}}^{-1}_{11} &= \begin{bmatrix} 1 - R_{j,(-j,-k)}\mathbf{R}^{-1}_{(-j,-k)}R_{(-j,-k),j} & \mathrm{r}_{jk} - R_{j,(-j,-k)}\mathbf{R}^{-1}_{(-j,-k)}R_{(-j,-k),k} \\ \mathrm{r}_{jk} - R_{k,(-j,-k)}\mathbf{R}^{-1}_{(-j,-k)}R_{(-j,-k),j} & 1 - R_{k,(-j,-k)}\mathbf{R}^{-1}_{(-j,-k)}R_{(-j,-k),k} \end{bmatrix} \\ &= \begin{bmatrix} 1 - R^2_{Xj.\{X(-j,-k)\}} & \mathrm{r}_{jk} - R_{j,(-j,-k)}R^{-1}_{(-j,-k)}R_{(-j,-k),k} \\ \mathrm{r}_{jk} - R_{k,(-j,-k)}R^{-1}_{(-j,-k)}R_{(-j,-k),j} & 1 - R^2_{Xk.\{X(-j,-k)\}} \end{bmatrix} \end{aligned}$$

matriz de orden 2×2 que puede escribirse como:

$$\begin{bmatrix} Tol_{Xj.\{X(-j,-k)\}} & \mathcal{T} \\ \mathcal{T} & Tol_{Xk.\{X(-j,-k)\}} \end{bmatrix}$$

donde $\mathcal{T} = (Tol_{Xj.\{X(-j,-k)\}})^{1/2}(Tol_{Xk.\{X(-j,-k)\}})^{1/2}r_{jk.X(-j,-k)}$ y Tol denota el índice conocido como Tolerancia, definido como uno menos el coeficiente de determinación múltiple correspondiente.

El término $(1,2)$ de la matriz anterior $(\check{\mathbf{R}}^{-1}_{11})$, normalizado por la raíz del producto de los elementos correspondientes en la diagonal, es el coeficiente de correlación parcial entre X_j y X_k. Este término coincide, excepto por el signo, con el correspondiente término normalizado en la matriz inversa $(\check{\mathbf{R}}_{\mathbf{11}})$:

$$\frac{\mathrm{r}^{jk}}{\sqrt{\mathrm{r}^{jj}}\sqrt{\mathrm{r}^{kk}}} = -\mathrm{r}_{jk.\{-j,-k\}}$$

y el término $(1,1)$ de la matriz $\check{\mathbf{R}}_{\mathbf{11}}$ (inversa de $\check{\mathbf{R}}^{-1}_{11}$) es de la forma:

$$\mathrm{r}^{jj} = \frac{1}{(1 - \mathrm{r}^2_{jk.(-j,-k)})}\frac{1}{Tol_{Xj.\{X(-j,-k)\}}}$$

Se demuestra además que este último término es igual a:

$$\frac{1}{Tol_{Xj.\{X(-j)\}}}$$

expresión que coincide con $\mathrm{VIF}(j)$. Farrar & Glauber (1967) señala que una inspección de los r^{jj} puede dar importantes pistas acerca de la severidad de las redundancias en el modelo.

3. Propuesta

En este trabajo proponemos un dispositivo gráfico cuyo propósito es obtener una representación aproximada de las relaciones de dependencia lineal que se producen entre un conjunto de variables. Específicamente, en dicha representación se visualizan los VIF para cada variable y los coeficientes de correlación parcial.

En este caso la descomposición espectral de rango r de la matriz $\mathbf{R}^{-1}$ toma la forma:

$$\mathbf{R}^{-1} = (\mathbf{r}_{ij}) = \mathbf{V}_{(\mathbf{r})}\mathbf{D}^{-1}_{(\mathbf{r})}\mathbf{V}^{t}_{(\mathbf{r})}$$

siendo $\mathbf{V}_{(\mathbf{r})}$ la matriz cuyas columnas son los autovectores de $\mathbf{R}^{-1}$ asociados con sus r autovalores no nulos, organizados sobre la matriz diagonal $\mathbf{D}^{-1}_{(\mathbf{r})}$.

Al definir la matriz de marcadores como $\mathbf{H}^{\mathbf{t}}_{(\mathbf{r})} = \mathbf{D}^{-1/2}_{(\mathbf{r})}\mathbf{V}^{t}_{(\mathbf{r})} = (\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_p)$, su representación gráfica sobre el primer plano h-plot garantiza que:

1.- El coeficiente de inflación de varianza de la variable X_j es aproximado por el cuadrado de la longitud del marcador correspondiente:

$$\mathrm{h}^{t}_{j(2)}\mathrm{h}_{j(2)} \simeq \mathrm{VIF}(j) \qquad \forall j = 1, 2, \ldots, p$$

2.- La correlación parcial entre las variables X_j y X_k es aproximada, excepto por el signo, a través del coseno del ángulo entre sus marcadores:

$$\frac{\mathrm{h}^{t}_{j}\mathrm{h}_k}{\sqrt{\mathrm{h}^{t}_{j}\mathrm{h}_j}\sqrt{\mathrm{h}^{t}_{k}\mathrm{h}_k}} = -\mathrm{r}_{jk.X(-j,-k)}$$

4. Información captada por la representación $\mathbf{R}^{-1}\mathbf{h}-plot$

Tomando en cuenta que en un modelo de regresión el $\mathrm{VIF}(j)$ se interpreta como el incremento en la varianza del coeficiente b_j, debido a la multicolinealidad de X_j con las restantes variables explicativas, es posible definir los siguientes indicadores:

Imprecisión global. Este indicador se define como traza(R-1) y se interpreta como una medida de la imprecisión global de los coeficientes de regresión debido a la multicolinealidad, ya que:

$$\mathrm{traza}(\mathbf{R}^{-1}) = \sum \mathrm{VIF}(j) = \sum \lambda_\alpha$$

siendo λ_α el α-ésimo autovalor de $\mathbf{R}^{-1}$.

Imprecisión captada por un eje. Este indicador se define como el cociente $\lambda_\alpha/\mathrm{traza}(\mathbf{R}^{-1})$ y se interpreta como la proporción de la imprecisión global que es captada por el α-ésimo eje h-plot.

Contribución de cada variable a la imprecisión captada por un eje. Dado que:

$$\sum \mathrm{h}_{j\alpha}^2 = \sum \left(\sqrt{\lambda_\alpha} v_{j\alpha}\right)^2 = \lambda_\alpha$$

se define como contribución de la variable X_j a la imprecisión captada por el eje α al cociente:

$$\mathrm{CV}_j\mathrm{F}_\alpha = \frac{\mathrm{h}_{j\alpha}^2}{\lambda_\alpha}$$

Contribución de cada eje a la imprecisión del coeficiente asociado a una variable. Dado que:

$$\sum \mathrm{h}_{j\alpha}^2 = \mathrm{h}_j^t \mathrm{h}_j = \mathrm{VIF}_{(j)}$$

se define como contribución del eje α a la imprecisión del coeficiente de regresión b_j al cociente:

$$\mathrm{CF}_\alpha\mathrm{V}_j = \frac{\mathrm{h}_{j\alpha}^2}{\mathrm{VIF}_{(j)}}$$

5. Ilustración

Con el objeto de validar el dispositivo propuesto, éste ha sido aplicado en una investigación sobre la caracterización morfológica de jóvenes nadadores venezolanos (Pérez, Vásquez, Tomei, Landaeta & Ramírez 2004). El objetivo principal de este estudio consistió en identificar las variables antropométricas con mayor capacidad predictiva para clasificar correctamente un atleta según su estatus de maduración sexual (prepúber, púber inicial o púber avanzado). El procedimiento estadístico utilizado para construir la regla de clasificación es el Análisis Lineal Discriminante, del cual se conoce que resulta sensible a la presencia de multicolinealidad en las variables predictoras. Específicamente se ilustra la aplicación del método en la evaluación de la posible existencia de relaciones lineales en el conjunto de variables que describen el patrón de distribución de la grasa. Las variables consideradas aquí son los pliegues (panículos adiposos) medidos en milímetros, en diferentes partes del cuerpo: triceps (trice), subescapular (subes), biceps (bicep), cresta ilíaca (iliac), supraespinal (supra), abdomen (abdom), muslo (muslo) y pantorrilla media (panto), en el grupo de nadadores prepúberes.

5.1. Resultados

5.1.1. Matriz de correlaciones R

En esta matriz se evidencian las fuertes relaciones lineales positivas entre los niveles de grasa medidos en los ocho puntos considerados.

Tabla 1: Matriz de Correlaciones $\mathbf{R}$

	trice	subes	bicep	iliac	supra	abdom	muslo	panto
trice	1.00	0.90	0.95	0.92	0.91	0.94	0.92	0.91
subes	0.90	1.00	0.91	0.87	0.95	0.95	0.89	0.87
bicep	0.95	0.91	1.00	0.91	0.93	0.92	0.89	0.91
iliac	0.92	0.87	0.91	1.00	0.85	0.91	0.95	0.88
supra	0.91	0.95	0.93	0.85	1.00	0.95	0.88	0.89
abdom	0.94	0.95	0.92	0.91	0.95	1.00	0.93	0.86
muslo	0.92	0.89	0.89	0.95	0.88	0.93	1.00	0.93
panto	0.91	0.87	0.91	0.88	0.89	0.86	0.93	1.00

Tabla 2: Inversa de la Matriz de Correlaciones $\mathbf{R}^{-1}$

	trice	subes	bicep	iliac	supra	abdom	muslo	panto
trice	**18.48**	-0.08	0.34	0.11	-0.34	0.50	-0.21	0.38
subes	1.25	**14.18**	0.04	0.09	0.29	0.19	0.00	0.03
bicep	-6.84	-0.77	**21.40**	0.56	0.17	0.08	-0.42	0.33
iliac	-2.12	-1.46	-11.19	**18.68**	-0.09	-0.14	0.61	-0.34
supra	8.83	-6.76	-4.94	2.28	**37.40**	0.75	-0.49	0.63
abdom	-16.81	-5.50	-3.09	4.94	-36.04	**62.21**	0.71	-0.78
muslo	6.42	-0.10	13.78	-18.98	21.29	-40.14	**51.49**	0.84
panto	-9.19	-0.67	8.59	8.43	-21.92	34.73	-34.24	**32.02**

5.1.2. Inversa de la matriz de correlaciones $\mathbf{R}^{-1}$

Nota: En el triángulo superior de esta matriz se han colocado las correlaciones parciales

En la diagonal de $\mathbf{R}^{-1}$ aparecen los coeficientes de inflación de varianza. Se destacan fundamentalmente los correspondientes a los pliegues abdominal (62.2), muslo (51.5), supraespinal (37.4) y pantorrilla (32.0), que acumulan un 71.6 % del total de la multicolinealidad existente en el sistema de variables. Se evidencian además importantes correlaciones parciales entre pantorrilla y muslo (0.84), pantorrilla y abdomen (-0.78), abdomen y muslo (0.71), y, supraespinal y abdomen (cresta ilíaca y muslo).

5.1.3. Autovalores de $\mathbf{R}^{-1}$

$$\text{Imprecisión Global} \quad 255.87 \quad \longleftarrow \text{traza}(\mathbf{R}^{-1})$$

La estructura de las relaciones de multicolinealidad entre las variables es explicada en un 78 % por el primer plano $\mathbf{R}^{-1}$ h-plot.

Tabla 3: Autovalores de $\mathbf{R}^{-1}$

Factor	Autovalor	Porcentaje de Imprecisión	Porcentaje acumulado
1	150.15	58.7	58.7
2	48.15	18.8	77.5
3	22.57	8.8	86.3
4	15.15	5.9	92.2
5	8.64	3.4	95.6
6	6.74	2.6	98.3
7	4.33	1.7	99.9
8	0.14	0.1	100.0

5.1.4. Coordenadas y VIF

Tabla 4: Coordenadas y VIF

Variable	Factor 1	Factor 2		VIF suma de cuadrados (1-8)	VIF suma de cuadrados (1-2)
trice	-1.89	1.72		18.48	6.53
subes	-0.12	-0.32		14.18	0.12
bicep	-1.06	-3.78		21.40	15.41
iliac	1.57	3.23		18.68	12.93
supra	-4.62	3.05		**37.40**	30.69
abdom	7.35	-1.74		**62.21**	57.04
muslo	-6.37	-2.75		**51.49**	48.08
panto	5.20	0.68		**32.02**	27.51
suma de cuadrados	150.15	48.15	suma	255.87	198.30

5.1.5. Contribuciones

Las variables con mayor contribución a la multicolinealidad (CVF) captada por el primer factor son abdomen (36 %), muslo (27 %), pantorrilla (18 %) y supraespinal (14 %). El segundo eje queda fundamentalmente definido por biceps (30 %), cresta ilíaca (22 %), supraespinal (19 %) y muslo (16 %). Con la excepción de triceps y subescapular, la multicolinealidad de todas las variables (VIF) es aproximada con alta calidad ($> 69\,\%$)

Tabla 5: Contribuciones

Variable	CVF 1	CVF 2	CFV 1	CFV 2	suma
trice	0.02	0.06	0.19	0.16	0.35
subes	0.00	0.00	0.00	0.01	0.01
bicep	0.01	**0.30**	0.05	0.67	**0.72**
iliac	0.02	**0.22**	0.13	0.56	**0.69**
supra	**0.14**	**0.19**	0.57	0.25	**0.82**
abdom	**0.36**	0.06	**0.87**	0.05	**0.92**
muslo	**0.27**	**0.16**	**0.79**	0.15	**0.94**
panto	**0.18**	0.01	**0.84**	0.01	**0.85**

5.1.6. Primer plano factorial

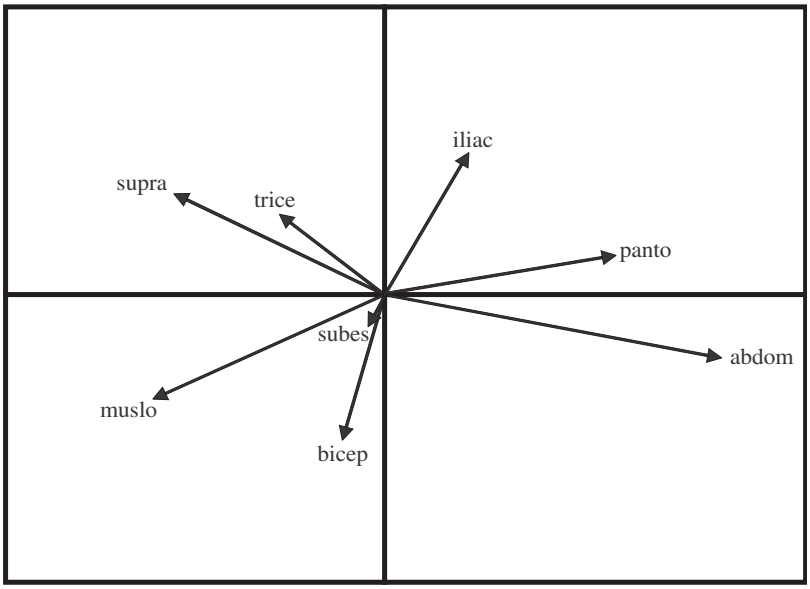

Figura 1: Plano principal 1-2

La multicolinealidad presente en las variables queda claramente reflejada por la longitud de los rayos. Se aprecia visualmente por ejemplo, la fuerte correlación parcial positiva entre muslo y pantorrilla a partir del ángulo entre ambos vectores. De igual manera, las posiciones de los vectores correspondientes a abdomen y pantorrilla indican una correlación inversa importante entre ellas, después de controlar por las restantes variables.

5.1.7. Ángulos entre algunos vectores

En la figura 2 se representan solamente cuatro de las ocho variables, indicándose además los ángulos entre los vectores correspondientes.

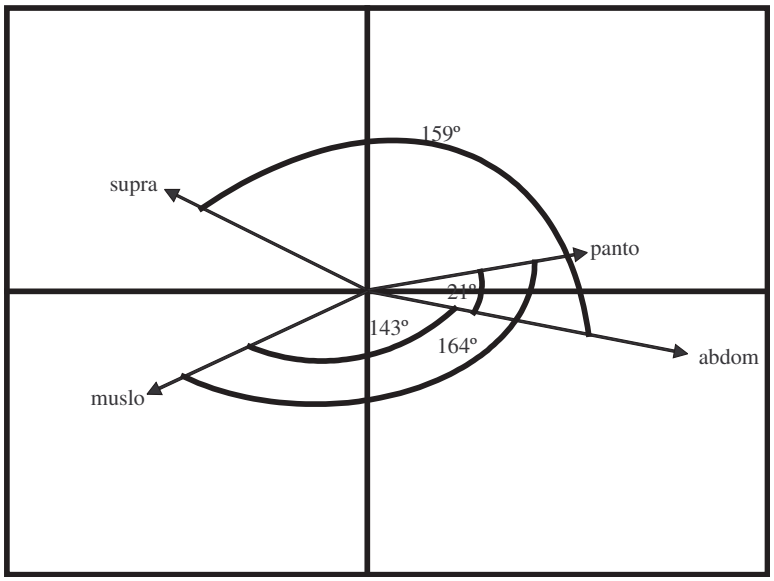

Figura 2: Ángulos entre vectores

5.1.8. Algunas correlaciones parciales

A continuación comparamos algunas de las correlaciones parciales y su aproximación a través de los cosenos de los ángulos correspondientes (multiplicados por -1) en la representación bidimensional anterior.

Corr(panto, abdom, restantes)	=	-0.78	$\approx$	$-\cos(164^o)$	=	0.96
Corr(muslo, panto, restantes)	=	0.84	$\approx$	$-\cos(21^o)$	=	-0.94
Corr(muslo, abdom, restantes)	=	0.71	$\approx$	$-\cos(143^o)$	=	0.80
Corr(supra, abdom, restantes)	=	0.75	$\approx$	$-\cos(159^o)$	=	0.93

5.1.9. Algunos coeficientes de inflación de varianza

En forma análoga presentamos algunos coeficientes de inflación de varianza y su aproximación a través de la norma cuadrado del vector correspondiente.

VIF(panto)	=	32.02	$\approx$	norma^2(vector)	=	27.51
VIF(abdom)	=	62.21	$\approx$	norma^2(vector)	=	57.04
VIF(muslo)	=	51.49	$\approx$	norma^2(vector)	=	48.08
VIF(supra)	=	37.40	$\approx$	norma^2(vector)	=	30.69

5.1.10. Comentarios finales

El procedimiento anterior se repitió para cada uno de los grupos de variables que definen las distintas dimensiones consideradas en el estudio (tamaño, longitudes, anchuras y pliegues), lográndose reducir el número de variables predictoras de un total de 35 a 8. Los resultados obtenidos indicaron un notable aumento en el porcentaje de la correcta clasificación al eliminar del conjunto original las variables colineales. Específicamente en la dimensión grasa, las variables predictoras que se incorporaron al modelo de predicción fueron los pliegues del triceps y del subescapular.

Recibido: 11 de Abril de 2005
Aceptado: 24 de Julio de 2005

Referencias

Belsley, D. (1991), *Conditioning Diagnostics, Collinearity and Weak Data in Regression*, Wiley, New York.

Cornsten, L. & Gabriel, K. (1976), 'Graphical exploration in comparing variance matrices', *Biometrics* **32**(851-863).

Farrar, D. & Glauber, R. (1967), 'Multicollinearity in regression analysis: The problem revisited', *Review of Economic Statistics* **49**, 92–107.

Gabriel, K. (1971), 'The biplot graphic display of matrices with application to principal component analysis', *Biometrika* **58**, 453–467.

Glantz, S. & Slinker, B. (2001), *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, New York.

Gleason, T. & Staelin, R. (1975), 'A proposal for handling missing data', *Psychometrika* **40**, 229–252.

Jackson, J. (1991), *A User's Guide to Principal Components*, Wiley, New York.

Kendall, M. (1957), *A Course in Multivariate Analysis*, Griffin, London.

Mandell, J. (1982), 'Use of the singular value decomposition in regression analysis', *The American Statistician* **36**(1), 15–24.

Mason, R., Gunst, R. & Webster, J. (1975), 'Regression analysis and the problem of multicollinearity', *Communications in Statistics* **4**, 277–292.

Pérez, B., Vásquez, M., Tomei, C., Landaeta, M. & Ramírez, G. (2004), Anthropometrics characteristics of young venezuelan male swimmers according with biological maturity status, International Congress of Auxology, Florencia, Italia.

Raveh, A. (1985), 'On the use of the inverse of the correlation matrix in multivariate data analysis', *The American Statistician* **39**(1), 39–42.

Silvey, S. (1969), 'Multicollinearity and imprecise estimation', *Journal of the Royal Statistical Society* **Series B**(31), 751–754.

Whitakker, J. (1990), *Graphical Models in Applied Multivariate Analysis*, Wiley, New York.

Yu, C. (1998), Multi-collinearity, variance inflation and orthogonalization in regression.
*http://seamonkey.ed.asu.edu/ alex/computer/sas/collinear.html