Universidad Nacional Autónoma de México

Unidad Académica de los Ciclos Profesionales y de Posgrado del C.C.H.

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

INFERENCIA ESTADISTICA EN EL ANALISIS DE CORRESPONDENCIAS

T E S I S

Que para obtener el grado de:
MAESTRO EN ESTADISTICA E INVESTIGACION DE OPERACIONES

Presenta el Actuario

Eduardo Castaño Tostado

México, D. F. Febrero, 1987

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Indice.

1.Introducción.

El Análisis de Correspondencias(A.C.) es una técnica encuadrada en el enfoque del Análisis de Datos siendo su propósito fundamental la descripción gráfica, a través de elementos algebraicos y geométricos,de las relaciones más relevantes en una tabla de contingencias.Lo anterior es la idea que tradicionalmente se ha tenido en el A.C.,con lo que el carácter exploratorio del análisis tiene preponderancia sobre los aspectos inferenciales.En este trabajo se sigue la interpretación del Análisis de Correspondencias debida a la escuela de Estadística en Francia;sin embargo,el objetivo es presentar elementos inferenciales para enriquecer y validar las conclusiones del Análisis de Correspondencias.

Trabajos sobre inferencia en el Análisis de Correspondencias ya existen en la literatura , aunque sufren de algunos inconvenientes.Un primer grupo de artículos que se enfoca al tratamiento de los valores singulares,parte de la suposición de independencia entre los criterios de clasificación,supuesto que en la práctica es poco frecuente,ya que precisamente el A.C. tiene por objeto mostrar la estructura de dependencia de la

tabla;trabajos relevantes en este sentido son los de Lebart(1976) y Corsten(1976).Un segundo grupo de resultados,debidos a O'Neill(1978,1981),si bien llega a distribuciones asintóticas de los valores singulares,éstas dependen fuertemente del cumplimiento de condiciones que involucran la estructura teórica de la tabla;estas condiciones son muy complejas y resulta completamente imprâctico verificarlas,por lo que se utilizan las distribuciones sin tomarlas en cuenta.Una tercera alternativa es desarrollada por Goodman(1985),mediante un enfoque general de ajuste de modelos a una tabla de contingencias con dos criterios de clasificación, en términos de validar un sistema de pesos o calificaciones sobre los renglones y/o las columnas de la tabla bajo estudio.El Análisis de Correspondencias es un caso especial del tipo de modelos manejados por este autor;sin embargo las herramientas de análisis aún no han sido totalmente desarrolladas, siendo además un camino menos directo y más complejo que las propuestas anteriores.

Los tres grupos de tabajos mencionados anteriormente versan sobre inferencia en valores singulares del A.C.;Greenacre(1984),utilizando técnicas de remuestreo analiza la

estabilidad de la representación gráfica del A.C.,pero que desde la perspectiva de este trabajo,pueden mejorarse substancialmente. Es propósito de esta tesis presentar propuestas de análisis de tipo inferencial sobre aspectos relevantes en la interpretación de resultados del A.C.;para ello se combinan ideas ya desarrolladas en los trabajos mencionados anteriormente con las propiedades de mayor alcance de los métodos de remuestreo.

La estructura de la presentación es como sigue;en el capitulo dos se presentan los aspectos algebraicos y geométricos del Análisis de Corresponedencias.En el capitulo tres se muestran los resultados referentes a los trabajos sobre inferencia a través de teoria asintótica,es decir los desarrollados por Lebart(1976),Corsten(1976),O'Neill(1978,1981) y Goodman(1985).En el cuarto capitulo se hacen nuevas propuestas inferenciales ,no sólo sobre los valores singulares sino en variables involucradas en la representación gráfica.En el capitulo cinco se presenta una aplicación del A.C. al análisis sensorial en tecnologia de alimentos.Por último se dan conclusiones y se discute lo desarrollado en todo el trabajo.

2. Análisis de Correspondencias.una introducción.

El Análisis de Correspondencias es resultado de desarrollos geométricos y algebraicos.Existen distintos enfoques del A.C. en su utilización;uno de estos enfoques debido a la escuela francesa, y que se sigue en este trabajo,se debe primordialmente a los trabajos de Jean Paul Benzécri.Esta técnica se sitúa en el análisis multivariado de datos y su objetivo principal es la representación gráfica de las relaciones más relevates que se hayan en forma subyacente en una tabla de contigencia de dos o más criterios de clasificación.

Es propósito de esta sección el presentar los detalles más importantes tanto algebraicos como geométricos del A.C. y para ello se sigue la notación de Greenacre(1984).La presentación por sencillez se hará para el caso de dos criterios de clasificación en la tabla de contigencias,aunque la generalización al caso multicriterio es inmediata,ya que a fin de cuentas las conclusiones suceptibles de obtenerse mediante el A.C. son sólo sobre las interacciones de orden uno entre los criterios de clasificación de la tabla bajo estudio(Greenacre,1984).

2.1 Notación.

Sea N una tabla de contingencias con pq celdas y con frecuencias n_{ij} $i=1,\ldots,p+1; j=1,\ldots,q+1$ $(p>q)$, y denote por n a

$$n=\sum_i \sum_j n_{ij} \qquad (2.1)$$

Llame $\hat{P}$ a la matriz de frecuencias relativas

$$\hat{P}=(1/n)N \qquad (2.2)$$

y a

$$\underline{r}=\hat{P}\underline{1} \text{ y } \underline{c}=\hat{P}'\underline{1}$$

donde $\underline{1}$ es un vector de unos con las dimensiones adecuadas al caso, $\underline{r}$ representa el vector de sumas por renglón de $\hat{P}$ y $\underline{c}$ el vector de sumas por columna de la misma matriz. Denote a los elementos de estos dos vectores como $p_{i.}, i=1,\ldots,p+1; p_{.j}, j=1,\ldots,q+1$ respectivamente.

Se construyen las siguientes matrices:

$$R=D_r^{-1}\hat{P} \quad , C=D_c^{-1}\hat{P}', \qquad (2.3)$$

donde

$$D_r=\text{diag}(\underline{r}), D_c=\text{diag}(\underline{c}) \qquad (2.4)$$

Los renglones de la matriz R, denotados por $(\underline{r}_i, i=1,\ldots,p+1)$ son llamados los perfiles renglón de la matriz $\hat{P}$; analogamente, los renglones de la matriz $C(\underline{c}_j, j=1,\ldots,q)$ son nombrados perfiles columna. Así, el i-ésimo perfil renglón es un vector cuyas entradas representan la distribución discreta marginal en el i-ésimo

renglòn(i=1,...,p+1)de la matriz P;analogamente en el caso de los perfiles columna.En vista de lo anterior la estructura relativa de la tabla representada en dos formas por estos conjuntos de perfiles es invariante ante el tamaño total de la tabla n.

Ya que los perfiles renglòn como los perfiles columna pueden considerarse como vectores de q+1 y p+1 entradas respectivamente,el objetivo del A.C. es buscar un subespacio de rango menor en los espacios respectivos,en los que las proyeccciones de los corespondientes perfiles representen las asociaciones màs relevantes que se dan en los espacio q+1-variado y p+1-variado respectivamente(figura 2.1).Posteriormente a la obtenciòn de estos dos subespacios el A.C. obtiene una representaciòn gràfica conjunta de ambos subespacios.

Si se denota por $d_m(\underline{y},\underline{w})$ a la distancia de $\underline{y}$ a $\underline{w}$ bajo la mètrica m,un criterio para encontrar el subespacio(S*)en el espacio de perfiles renglòn es

$$\min_{S^*} \sum_{i=1} w_i \, d_m(\underline{r}_i, \underline{s}_i) \quad \underline{s}_i \in S^* \qquad (2.5)$$

donde w_i es un peso diferencial por perfil y $\underline{s}_i$ la proyecciòn ortogonal de $\underline{r}_i$ en S*.

En el A.C. se escoge $w_i = p_{i.}$ y a la mètrica m se le asigna la

matriz D_c^{-1};es decir que los perfiles se ponderan en función de su frecuencia relativa total y las q+i coordenadas de cada uno de estos por su frecuencia relativa acumulada por estas en todos los perfiles. Así,(2.5) queda expresado como

$$\min \|R-S\$\|_{r,c} = \min \sum_i p_{i.} (r_i - s_i)' D_c^{-1} (r_i - s_i) \qquad (2.6)$$

Analogamente para los perfiles columna,es decir,los renglones de la matriz C,se tiene que el criterio utilizado por el A.C. para encontrar el subespacio de mejor ajuste(T\$)es

$$\min \|C-T\$\|_{r,c} = \min \sum_j p_{.j} (\underline{c}_j - \underline{t}_j)' D_r^{-1} (\underline{c}_j - \underline{t}_j) \qquad (2.7)$$

con $t \in T\$$.

Los centroides de los perfiles renglón y de los perfiles columna son respectivamene $\underline{c} = R\underline{r}'$ y $\underline{r} = C'\underline{c}$.Se puede demostrar que estos están contenidos en los planos S\$ y T\$ por lo que se parte sin pérdida de generalidad de los perfiles centrados(Greenacre,1984)

$$R-\underline{1}\underline{c}', C-\underline{1}\underline{r}'. \qquad (2.8)$$

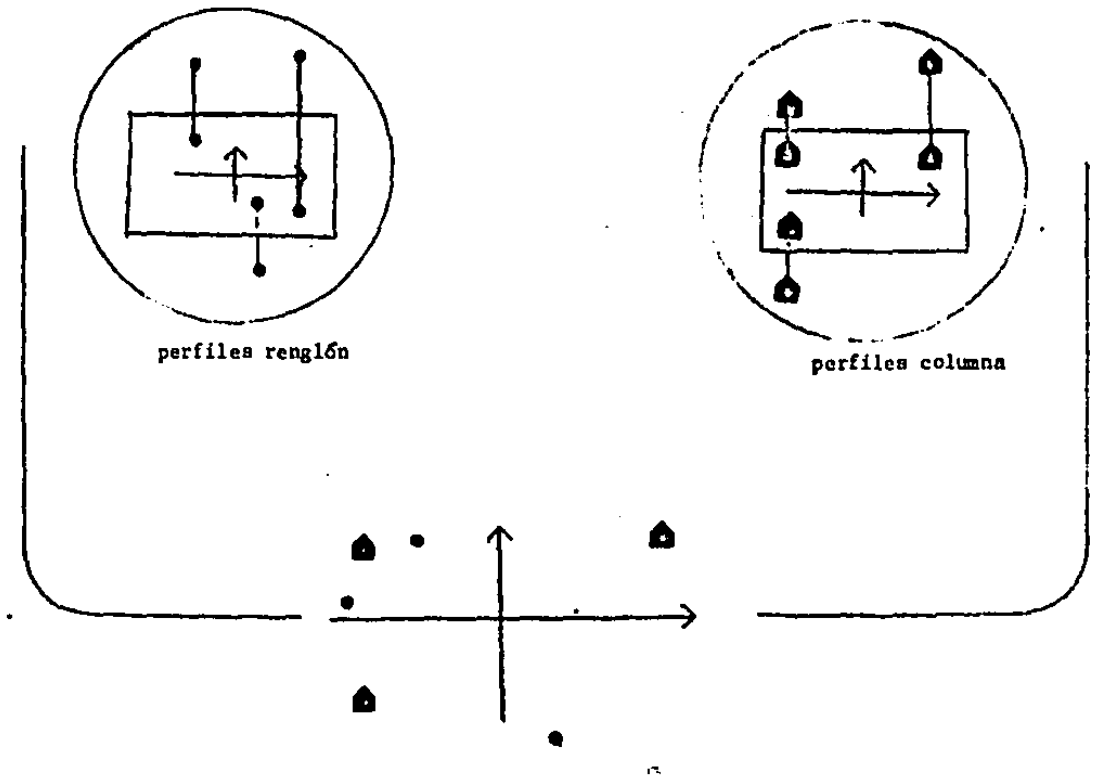
2.2 Descomposición en Valor Singular y el B.G.

Suponga que $R-\underline{1}\underline{c}'$ es de rango K ($K<=\min(p,q)$);se pueden encontrar K vectores ortogonales tales que(Green y Carroll,1976):

$$R-\underline{1}\underline{c}' = FB' \qquad (2.9)$$

donde $B=(\underline{b}_1, \underline{b}_2, \ldots \underline{b}_k)$ son los ejes principales de los perfiles

Figura 2.1 Objetivo del Análisis de Correspondencias.



renglòn y F las coordenadas de los perfiles renglòn centrados,respecto a B .Analogamente

$$C-\mathbf{1}\underline{c}' = GA' \qquad (2.10)$$

con $A=(\underline{a}_1,\underline{a}_2,\ldots,\underline{a}_K)$ejes principales de los perfiles columna y G las coordenadas respectivas.Ahora,

$$D_r(R-\mathbf{1}\underline{c}')=\hat{P}-\underline{r}\underline{c}'$$
$$D_c(C-\mathbf{1}\underline{r}')=\hat{P}'-\underline{c}\underline{r}' \qquad (2.11)$$

con lo que encontrar A y B son problemas interrelacionados.
La descomposiciòn en valor singular(vèase apèndice A),da la pauta para encontrar estos dos conjuntos de vectores.Esta herramienta algebraica provee las matrices L ,M y $D_{\hat{u}}$(K x K) tales que

$$\hat{P}-\underline{r}\underline{c}'=LD_{\hat{u}}M' \qquad (2.12)$$

sujeto a que

$$L'D_r^{-1}L=M'D_c^{-1}M=I(K\times K) \qquad (2.13)$$

Las columnas de la matriz L son K vectores ortogonales bajo la mètrica D_r^{-1},y constituyen una base ortonormal para los renglones de $\hat{P}-\underline{r}\underline{c}'$, mientras que las columnas de M son K vectores ortogonales bajo la mètrica D_c^{-1} ,siendo una base ortonormal para las columnas de la matriz $\hat{P}'-\underline{c}\underline{r}'$.Por ùltimo,$D_{\hat{u}}$ es una matriz diagonal con elementos $(\hat{u}_1>\hat{u}_2>=\ldots>=\hat{u}_K>0)$ llamados valores singulares.En vista de (2.12)-(2.13) se tiene que

$$\hat{P}-\underline{r}\underline{c}' = \sum_{k=1}^{K} \hat{u}_k \underline{m}_k \underline{l}_k' \qquad (2.14)$$

Dado el carácter aditivo en (2.14) se puede aproximar la matriz(2.11) por K*<=K de los vectores de M y L .Si K*=2 se tiene una matriz de rango 2 que aproxima a la matriz de rango K.Esta forma de aproximar a (2.12) es óptima bajo el criterio

$$\min\|A-X\| = \min\sum_i p_{i.}(\underline{a}_i-\underline{x}_i)'D_c^{-1}(\underline{a}_i-\underline{x}_i) \qquad (2.15)$$

con $A=\hat{P}-\underline{r}\underline{c}'$ (Green y Carroll,1976).

De (2.12) y (2.11) se tiene que las columnas de M representan una base para los renglones de la matriz

$$R-\underline{1}\underline{c}';$$

si se denota por $\hat{F}$ las coordenadas respectivas,es decir,

$$\hat{F}M = (R-\underline{1}\underline{c}') \qquad (2.16)$$

por (2.13) se tiene que

$$\hat{F} = (R-\underline{1}\underline{c}')D_c^{-1}M \qquad (2.17)$$

Analogamente,si G denota las coordenadas de los renglones de

$$(C-\underline{1}\underline{r}')$$

respecto a L ,

$$\hat{G} = (C-\underline{1}\underline{r}')D_r^{-1}L\ . \qquad (2.18)$$

Si K*=2 se tendrá

$$\hat{F}2 = (R-\underline{1}\underline{c}')D_c^{-1}M2 \qquad (2.19)$$

$$\hat{G}2 = (C-\underline{1}\underline{r}')D_r^{-1}L2 \qquad (2.20)$$

serán las coordenadas de los perfiles renglón y de los perfiles columna en el subespacio de rango dos que mejor aproxima las relaciones en los espacios q y p variados respectivamente,bajo el criterio (2.15),donde M2 y L2 representan los dos primeros ejes principales respectivos.Con estos resultados se tendrán las representaciones gráficas de los perfiles renglón y de perfiles columna.Debe remarcarse que estas representaciones son en dos espacios diferentes;el objetivo final del A.C. es la representación simultánea de ambos conjuntos,que formalmente no es posible realizar,pero que,a pesar de lo anterior,en la siguiente sección se muestran las expresiones que permitirán establecer tal representación conjunta ad hoc.

2.3 Fórmulas de Transición.

Las expresiones (2.17)-(2.18) están relacionadas mediante las llamadas fórmulas de transición.Para mostrar lo anterior,primero de (2.17) y (2.11) se tiene que

$$\hat{F}=D_r^{-1}(\hat{P}-cc')D_c^{-1}M \ . \qquad (2.21)$$

Ahora de (2.12),

$$(\hat{P}-cc')D_c^{-1}M=LD\Omega$$

con lo que

$$\hat{F}=D_r^{-1}LD_{\hat{\Delta}}. \qquad (2.22)$$

En forma anàloga,

$$G=D_c^{-1}MD_{\hat{\Delta}}. \qquad (2.23)$$

De (2.21),(2.23) y (2.12) se tiene que

$$\hat{F}=R\hat{G}D_{\hat{\Delta}}^{-1} \qquad (2.24)$$
$$\hat{G}=C\hat{F}D_{\hat{\Delta}}^{-1}$$

llamadas las fórmulas de transiciónjestas expresiones significan que las coordenadas de un conjunto de perfiles se puede expresar en función de las coordenadas del otro conjunto de perfiles.Para explicar las implicaciones de estas expresiones,considere la coordenada del i-ésimo perfil renglón respecto al k-ésimo eje principal respectivo denotada por $\hat{f}_{ik}$;de acuerdo con (2.24),

$$\hat{f}_{ik}=\sum_{j=1} r_{ij}\hat{g}_{jk}/\hat{\Delta}_k \qquad (2.25)$$

donde r_{ij} es la entrada j del perfil i,$\hat{g}_{jk}$ la coordenada del perfil columna j en su eje principal k y $\hat{\Delta}_k$ el valor singular respectivo.En vista de (2.25),se tiene que $\hat{f}_{ik}$ es una combinación de las coordenadas de los perfiles columna con respecto a su eje k.Geométricamente,un perfil renglón tenderá a una posición que corresponderá a las coordenadas de los perfiles columna más importantes respecto a aquel.Las fórmulas de transición permiten así la superposición de gráficas de los dos tipos de perfiles

para su interpretación conjunta.

2.4 Interpretación de los Valores Singulares .

La variación estimada de los perfiles renglón en el eje k es

$$\sum_i p_{i.}(\hat{f}_{ik}-\bar{f}_{.k})^2 \qquad k=1,\ldots,K \qquad (2.26)$$

$$\bar{f}_{.k}=\sum_i p_{i.}\hat{f}_{ik} \qquad k=1,\ldots,K \qquad (2.27)$$

Puede mostrarse que

$$(\bar{f}_{.1},\bar{f}_{.2},\ldots,\bar{f}_{.k})=0$$

con lo que (2.26) es

$$\sum_i p_{i.}\hat{f}_{ik}^2 \quad . \qquad (2.28)$$

La expresión (2.28) puede verse como

$$(\hat{f}_{1k},\hat{f}_{2k},\ldots,\hat{f}_{pk})\begin{bmatrix} p_{1.} & 0 & 0\ldots.0 \\ 0 & p_{2.} & 0\ldots0 \\ 0 & 0 & 0\ldots p_{K.} \end{bmatrix}\begin{bmatrix} \hat{f}_{1k} \\ \hat{f}_{2k} \\ \hat{f}_{pk} \end{bmatrix} \qquad (2.29)$$

el producto del primer renglón de $\hat{F}$ por sí mismo ponderado por la matriz D_r, es decir, como parte del producto de matrices $\hat{F}'D_r\hat{F}$.

De (2.22), la expresión (2.28) desarrollándola se llega a

$$\hat{F}'D_r\hat{F}=D_{\hat{U}}L'D_r'D_rD_r'LD_{\hat{U}}=D_{\hat{U}}^2. \qquad (2.30)$$

y

$$\text{traza}(\hat{F}'D_r\hat{F})=\text{traza}(D_{\hat{U}}^2).$$

El sumando correspondiente a $\hat{\alpha}_k$ es precisamente (2.28),con lo que $\hat{\alpha}_k^2$ es el estimador de la varianza de los perfiles renglòn respecto al eje principal $\mathbf{m}_k$,k=1,...,K. Si se recuerda los valores singulares estàn ordenados,por lo que en vista de la expresiòn (2.30),$\mathbf{m}_1$ serà el eje de mayor variaciòn en las coordenadas de los perfiles renglòn y asi hasta el eje $\mathbf{m}_K$.

La expresiòn (2.6),pero ahora considerando la variaciòn de los perfiles renglòn respecto a su centroide es

$$\sum_i p_{i.}(\mathbf{r}_i-\mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{r}_i-\mathbf{c}) \qquad (2.31)$$

que es igual a

$$\text{traza}[\mathbf{D}_r(\mathbf{R}-\mathbf{1c}')\mathbf{D}_c^{-1}(\mathbf{R}-\mathbf{1c}')'] \; . \qquad (2.32)$$

Ahora,por (2.17),(2.30) y (2.11) se tiene que

$$\text{traza}[\mathbf{D}_{\hat{\alpha}}^2]=\text{traza}[\hat{\mathbf{P}}'\mathbf{D}_r\hat{\mathbf{P}}]=\text{traza}[\mathbf{D}_r\hat{\mathbf{P}}\hat{\mathbf{P}}']$$

$$=\text{traza}[\mathbf{D}_r\mathbf{D}_r^{-1}(\hat{\mathbf{P}}-\mathbf{rc}')\mathbf{D}_c^{-1}\mathbf{M}\mathbf{M}'\mathbf{D}_c^{-1}(\hat{\mathbf{P}}-\mathbf{rc}')'\mathbf{D}_r^{-1}] \qquad (2.33)$$

$$=\text{traza}[\mathbf{D}_r(\mathbf{R}-\mathbf{1c}')\mathbf{D}_c^{-1}\mathbf{M}\mathbf{M}'\mathbf{D}_c^{-1}(\mathbf{R}-\mathbf{1c}')']$$

pero

$$\mathbf{D}_c^{-1}\mathbf{M}\mathbf{M}'\mathbf{D}_c^{-1}=\mathbf{D}_c^{-1}$$

ya que

$$\mathbf{M}'\mathbf{D}_c^{-1}\mathbf{M}=\mathbf{I}$$

por lo que

$$\text{traza}[\mathbf{D}_{\hat{\alpha}}^2]=\text{traza}[\mathbf{D}_r(\mathbf{R}-\mathbf{1c}')\mathbf{D}_c^{-1}(\mathbf{R}-\mathbf{1c}')'] \qquad (2.34)$$

En vista de (2.34) la variaciòn total original es recuperada

integramente en la descomposición en valor singular, en la variación de las coordenadas F.

Analogamente respecto a las coordenadas de los perfiles columna

$$\sum u_k^2 = \text{traza}[\hat{G}' D_c \hat{G}] = \sum_j p.j (g_j - g)' D_r^{-1} (g_j - g) \quad .(2.35)$$

Por último,el A.C. provee de tres medidas para cuantificar la calidad de la representación gráfica;estas tres medidas son:

i)la calidad global de la representación en l dimensiones,reprentada por

$$(\hat{u}_1^2 + \ldots + \hat{u}_l^2) / \sum_k^K \hat{u}_k^2 \qquad l<=K.$$

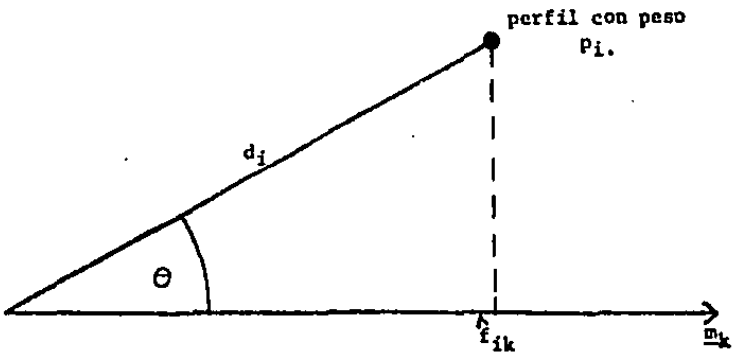
Usualmente se buscará que con l=1 ó l=2 se tenga la representación gráfica.Se puede interpretar como la proporción de la variación global que representan l dimensiones.

ii)La contribución relativa,que permite saber el ángulo de un perfil con respecto a uno de los ejes principales;de la figura 2.2,en términos de la definición del coseno de un ángulo se tiene que

$$(p_{i.} \hat{f}_{ik}^2) / (p_i . d_i^2) = (\hat{f}_{ik} / d_i)^2 = \cos^2 \theta$$

es decir,si $\cos^2 \theta$ es cercano a 1 el ángulo del perfil i será pequeño con respecto al eje en cuestión.

Figura 2.2 Contribuciones relativa(correlación) y absoluta.



Correlación$=(\hat{f}_{ik}/d_i)^2$

Contribución
Absoluta $=(p_i.\hat{f}_{ik})^2/\hat{G}_k^2$.

iii)La contribución absoluta,que cuantifica la aportación de un perfil en la variación total de un eje principal

$$(p_{i.} \hat{f}_{ik}^{2})/\hat{\lambda}_{k}^{2}.$$

En la práctica la utilidad de estas medidas es muy importante;respecto a la calidad global,lo que generalmente se desea es que con l=1 ó l=2 se tenga aglutinada la mayor parte de la variación original en la nube de perfiles multivariada ya que la aproximación de las relaciones a través de la gráfica unidimensional o bidimensional será mejor.

En el caso en que la calidad global no sea del todo convincente,es decir que oscile entre el 10% y 20%,las otras dos medidas mencionadas arriba entran en juego;la contribución absoluta identifica a aquellos perfiles que pesaron más en la orientación de cada uno de los ejes principales,mientras que la contribución relativa medirá la correlación de cada perfil con respecto a cada uno de los ejes principales,dando así una idea de que tan bien o que tan mal estuvieron representados.Los puntos que aparecen cercanos al origen de la gráfica estarán mal representados,con excepción de aquellos que coincidan con su centroide respectivo.

Asi,mientras la calidad global es suficiente para cuando agrupa la mayor parte de la varianza original,las otras dos medidas calificaràn en forma individual a cada perfil en tèrminos de su aporte en la redistribuciòn de la variabilidad en cada uno de los ejes generados por la DVS asi como su correlaciòn con èstos.

3. Inferencia: enfoque Asintótico.

Es propósito de esta capítulo el presentar los resultados fundamentales hasta ahora encontrados en la literatura respecto a la inferencia estadística en el A.C. Estos resultados son presentados cronológicamente comenzando con los de Lebart(1976) y Corsten(1976) que operan bajo la suposición de independencia entre los criterios de clasificación de la tabla; posteriormente son mostrados para el caso de dependencia las distribuciones sobre los valores singulares, debidas a O'Neill(1978,1981). Como tercer punto se establecen los elementos principales del enfoque de ajustes de modelos a tablas de contingencias debido a Goodman(1985). Finalmente se hacen comentarios generales sobre los desarrollos presentados y a partir de ellos se mencionan las necesidades a cubrir, que se atacarán con las herramientas propuestas en el capítulo cuatro.

3.1 Inferencia en caso de Independencia.

Bajo independencia entre los dos criterios de clasificación, Lebart(1976) mostró que la distribución de los valores singulares estimados mediante la descomposición en valor

singular en el A.C. puede ser aproximada por la distribución de las raíces correspondientes a una matriz Wishart central,con matriz de covarianzas la identidad de orden p-1 y grados de libertad q-1.Trabajando con el vector cuyos elementos son

$$\sqrt{n}(n_{ij}-n_{i.}n_{.j})/\sqrt{n_{i.}n_{.j}} \qquad (3.1)$$

y utilizando la aproximación a la distribución Normal,además de aplicar transformaciones ortogonales a (3.1),llega a la aproximación deseada,es decir que si

$$\hat{\alpha}_1<,\ldots,<\hat{\alpha}_p$$

son los valores singulares estimados,la distribución asintótica conjunta de estos es

$$f(\hat{\alpha}_1,\ldots,\hat{\alpha}_p)=c\prod_{i<j}n(\hat{\alpha}_i^2-\hat{\alpha}_j^2)\prod_{i=1}^{p}(n\hat{\alpha}_i^2)^{1/2(q-p+1)}\exp((1/2)\sum n\hat{\alpha}_i^2)$$

con (3.2)

$$c=\pi^{p/2}2^{(-1/2)pq}\prod_{i=1}^{p}(\Gamma((1/2)[p+1-i])\Gamma((1/2)[q+1-i]))^{-1}$$

Para detalles en la obtención de (3.2) véase Anderson(1958).

El resultado de Lebart también fue obtenido por Corsten(1976) según Greenacre(1984).

Por otra parte,Lebart(1976) mostró que el porcentaje de variación explicada hasta el l-ésimo eje principal,definido por

$$\sum_{i=1}^{l}\hat{\alpha}_i^2/(\sum_{i=1}^{p}\hat{\alpha}_i^2) \qquad (3.3)$$

es independiente de

$$\sum_{i=1}^{p} \delta_i^2). \qquad (3.4)$$

Este resultado es importante para el A.C. ya que aún en el caso de independencia,la representación gráfica producida podrá manejarse.En este sentido Lebart para ciertas dimensiones de una tabla de contingencias,generó por simulación tablas de significancia tanto para los porcentajes como para los 5 valores más grandes.

De esta manera,bajo independencia el problema inferencial sobre los valores singulares está resuelto.Si bien esto es un avance importante,en la práctica generalmente el caso de interés es cuando se presenta dependencia entre los criterios de clasificación,por lo que las conclusiones estadísticas de las herramientas propuestas son de un uso restringido constituyendo sólo puntos de referencia.

3.2 Inferencia en el caso de dependencia.

O'Neill(1978) obtuvo las distribuciones asintóticas de los valores singulares estimados en el caso de dependencia;sus resultados se presentan a continuación.

Denote por $\hat{\mu}$ el vector de p entradas cuyos elementos son

$\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_p$. Particione a $\hat{\mu}$ en dos subvectores $\hat{\mu}_1$ de orden K=p-s y $\hat{\mu}_2$ de orden s, donde $\hat{\mu}_1$ corresponde a los valores estimados cuyos valores teóricos son distintos cero.

Lancaster(1958) mostró que bajo un esquema de correlación · las probabilidades teóricas de la tabla $P_{ij}, i=1,\ldots,p; j=1,\ldots,q$ son tales que

$$P_{ij} = P_{i.} P_{.j} (1 + \sum_{k}^{K} u_k x_{ki} y_{kj})$$

donde

$$\sum_i x_{wi} x_{vi} P_{i.} = \delta_{wv} \qquad (3.5)$$

$$\sum_i \sum_j x_{wi} y_{vjk} P_{ij} = u_w \delta_{wv}.$$

A la vista de esta parametrización llamada la fórmula de reconstitución, los resultados de O'Neill(1978,1981) son los siguientes:

i) $n \hat{\mu}_2' \; I \hat{\mu}_2$ se distribuye asintóticamente como lo hacen las raíces de una matriz Wishart $W_{q-p+s}(I,s)$ si y solo si para cada i,j,k,l>p-s

$$\sum_{w=1} u_w (\sum_m x_{im} x_{km} x_{wm} P_{m.}) (\sum_n y_{jn} y_{ln} y_{wn} P_{.n}) = 0. \qquad (3.6)$$

ii) $n(\hat{\mu}_1 - \mu_1) \sim AN(0, \Delta)$ (3.7)

donde Δ depende en su cálculo explícito de la estructura teórica de la tabla(expresiones 10 y 11 de O'Neill,1981).

Asi en el caso de dependencia con valores singulares teòricos de multiplicidad 1,estas distribuciones seràn de utilidad si se conoce la parametrizaciòn especifica que sustente la tabla bajo estudio.

3.3 Modelos de Asociaciòn y Modelos de Correlaciòn.

Goodman(1985) muestra otras posibilidades de anàlisis en una tabla de contingencias de dos criterios de clasificaciòn.Su propuesta se orienta a la validaciòn de un sistema de calificaciones(scoring system),para los renglones de la tabla y/o para las columnas de la misma.Los modelos propuestos son distintas parametrizaciones de las probabilidades teòricas por celda;estas parametrizaciones son divididas por el autor en dos clases:

- modelos de correlaciòn
- modelos de asociaciòn

A contiunuaciòn se presentan los modelos de correlaciòn ya que sòlo en este caso se encontraron elementos importantes a reportarse desde la perspectiva del A.C.

3.3.1 Modelos de Correlación.

Este tipo de modelos parte de la parametrización llamada RC de correlación

$$P_{ij}=P_{i.}P_{.j}(1+ux_iy_j) \qquad (3.8)$$

donde x_i,y_j son calificaciones de renglón y de columna respectivamente tales que

$$\begin{aligned}\sum_i x_i P_{i.} &= \sum_j y_j P_{.j} = 0\\ \sum_i x_i^2 P_{i.} &= \sum_j y_j^2 P_{.j} = 1\end{aligned} \qquad (3.9)$$

Dadas estas condiciones

$$u=\sum_i\sum_j x_i y_j P_{ij} \qquad (3.10)$$

con lo que este parámetro representa la correlación entre las x's y y's .De este hecho se desprende el nombre de estos modelos.Por otra parte

$$(P_{ij}-P_{i.}P_{.j})/(P_{i.}P_{.j})=ux_iy_j \qquad (3.11)$$

es decir que u muestra la importancia de x_iy_j en la explicación de la dependencia;por último

$$\begin{aligned}\sum_i\sum_j (P_{ij}-P_{i.}P_{.j})^2/(P_{i.}P_{.j}) &= u^2\\ \sum_j y_j(P_{ij}/P_{i.}) &= ux_i\\ \sum_i x_i(P_{ij}/P_{.j}) &= uy_j\end{aligned} \qquad (3.12)$$

Del modelo RC general de correlación, cuando los renglones y columnas tienen un orden especifico se tiene

Modelo U $\quad x_i - x_{i+1} = A_i \,, y_j - y_{j+1} = A_j$

Modelo R $\quad y_j - y_{j+1} = A_j$ (3.13)

Modelo C $\quad x_i - x_{i+1} = A_i$

Siendo el objetivo el seleccionar uno de los modelos propuestos en el sentido de encontrar un sistema de calificaciones (x_i) (y_j) que mejor ajuste a los datos en la tabla de contingencias, en el ajuste de este tipo de modelos se utilizan la estadistica ji-cuadrada de bondad de ajuste y el cociente de verosimilitudes para datos categóricos en el caso asintótico.

Por otra parte, para establecer la relación de este tipo de modelos con el A.C. se define el modelo RC de correlación saturado como

$$P_{ij} = P_{i.} P_{.j} \left(1 + \sum_{k}^{K} u_k x_{ik} y_{jk}\right) \qquad (3.14)$$

$$\sum_i x_{ik} P_{i.} = \sum_j y_{jk} P_{.j} = 0$$

$$\sum_i x_{ik}^2 P_{i.} = \sum_j y_{jk}^2 P_{.j} = 1$$

$$\sum_i x_{ik} x_{imk} P_{i.} = \sum_j y_{jk} y_{jk'} P_{.j} = 0 \quad k \neq k'$$

con lo que

$$u_k = \sum_i \sum_j x_{ik} y_{jk} P_{ij} \quad k = 1, \ldots, K. \qquad (3.15)$$

La parametrización que maneja el A.C. es

$$P_{ij}=P_{i.}P_{.j}(1+\sum_{k}^{\delta}x'_{ik}y'_{jk}/u_k)$$

con (3.16)

$$x'_{ik}=x_{ik}u_k$$

$$y'_{jk}=y_{jk}u_k$$

Así las restricciones sobre las calificaciones se modifican a

$$\sum_i x'^2_{ik}P_{i.}=u^2_k$$

$$\sum_j y'^2_{jk}P_{.j}=u^2_k. \qquad (3.17)$$

Con esto,el Análisis de Correspondencias es un caso especial del modelo RC de correlación saturado.En vista de lo anterior si las herramientas utilizadas para el caso del modelo RC no saturado con K=1 son generalizadas para K$(1<=K$<=K),el problema inferencial sobre los valores singulares del A.C. podrá ser atacado mediante este camino.Sin embargo,sobre la generalización Goodman sólo la señala como una posibilidad citando referencias para el caso;por otra parte la comparación entre los resultados del A.C. y los arrojados por el ajuste de sus modelos de correlación se hace en una tabla en la que el modelo no saturado (3.8) es suficiente estadísticamente,quedando por ver un caso en que al menos sean necesarios dos valores singulares en la explicación aceptable de la tabla.Como ya se mencionó, el objetivo del ajuste de los los modelos de Goodman es la

validación de un sistema de calificaciones, por lo que al considerar los modelos no saturados con una sola dimensión sólo se supone que hay un único sistema;sin embargo en tablas en las que no baste una sola dimensión para su ajuste satisfactorio será necesario considerar la posibilidad de varios sistemas de calificaciones producidos no necesariamente por el mismo modelo.Para poder determinar que las últimas s raices son iguales a cero estadisticamente,habria que ajustar el modelo saturado con p y con p-s,para posteriormente por medio del cociente de razón de verosimilitudes, saber cual es el adecuado de los dos.Sin embargo en términos de las restricciones en (3.14) a cumplir por las calificaciones y por las raices,esto resulta ser muy complejo por el momento.

Una linea interesante de investigación consiste en la concreción de la generalización de las herramientas manejadas por Goodman,para enfrentar la situación señalada además de enfrentar el problema inferencial propio del A.C. en lo referente a la significancia de los valores singulares.

3.4 Comentarios.

Como se ha visto en lo expuesto en este tercer capitulo,el

problema de la inferencia en el A.C. ha sido visualizado en dos perspectivas.Se tiene en primer lugar la obtenciòn de las distribuciones asintòticas de los valores singulares como lo hacen los trabajos de Corsten(1976),Lebart(1976) y O'Neill(1978,1981).La segunda perspectiva ataca el problema inferencial respecto a estos valores singulares desde un enfoque de mayor generalidad y por ende de mayor estructura,es decir,el trabajo de Goodman(1985).

Respecto al primer enfoque,especialmente respecto a los resultados de Lebart y Corsten,el hecho de que se parta de la suposiciòn de independencia entre los criterios de clasificaciòn limita fuertemente su uso en la pràctica, en donde el caso de dependencia es el de interès en tèrminos generales;los trabajos de O'Neill si bien suplen esta carencia al proveer de distribuciones en el caso de dependencia,su uso estarà condicionado al cumplimiento de condiciones que en la pràctica no es posible verificar.En cuanto al tratamiento de Goodman,se tiene el inconveniente de que aùn no se han generalizado las herramientas de ajuste,con lo que no se ha tenido la posibilidad de comparar resultados en situaciones màs generales como las que enfrenta el A.C.,por lo que en el anàlisis de tablas que

requieran para su representación de más de una dimensión no se conoce con certeza el comportamiento de los modelos propuestos por este autor.

En el siguiente capítulo se presenta una propuesta de análisis inferencial sobre los valores singulares a partir de las ideas de O'Neill(1978) y combinándolas con elementos de técnicas de remuestreo.Se decidió trabajar en este sentido porque se le consideró más directo y menos complicado que el enfoque de Goodman.Sin embargo los comentarios sobre las posibilidades de generalización de este último debe considerarse como una línea futura de investigación.

Por otra parte,debe resaltarse el hecho de que los resultados presentados sólo han versado sobre los valores singulares.El A.C. si bien considera como parte importante el reconocimiento del número de dimensiones estadísticamente significativas,contiene otros elementos en los que sería interesante contar con herramientas de tipo inferencial.En el capítulo cuatro se presenta una estadística que permitirá realizar inferencias sobre los posibles agrupamientos de los perfiles de una misma característica,ya una vez proyectados sobre sus planos de mejor

ajuste.Nuevamente esto se hace a través de remuestreo y resultados de teoria asintòtica.

4. Inferencia: enfoque de remuestreo.

Como se observó en el capitulo tres de este trabajo, los métodos de inferencia presentados sufren de incovenientes ya que o se parte de supuestos poco realistas o aún no se han desarrollado a plenitud las posibilidades de su uso especifico en el A.C.. Otro aspecto a remarcar es que los esfuerzos han sido dirigidos a realizar inferencias sólo sobre la dimensionalidad de la representación, es decir, sobre la determinación estadistica de cuales valores singulares se puedan considerar igual a cero, sin tomar en cuenta caracteristicas de la representacion grafica. En este capitulo se persiguen dos objetivos; el primero respecto a proponer un algoritmo para probar dimensionalidad del A.C., mientras que el segundo consiste en realizar inferencias sobre la representación gráfica producida.

Así, en primer lugar se comenta sobre los fundamentos generales del método bootstrap desarrollados por Efron(1977), como las condiciones especificas a cumplir en los casos trabajados más adelante. En segundo lugar, aprovechando algunos resultados de O'Neill(1978) y combinándolos con las propiedades de la metodologia de remuestreo bootstrap, se presenta un algoritmo

estadistico para probar dimensionalidad sobre los valores singulares.En el mismo sentido también se reportan los resultados debidos a Beran(1985),respecto al uso del método bootstrap en inferencia respecto a porcentajes de variación recuperada con un número de valores singulares seleccionado de antemano.

En tercer lugar,como se mencionó en el capítulo tres,es necesario desarrollar elementos inferenciales en otros aspectos de relevancia práctica en el A.C.;en este sentido se presenta una estadística que permitirá probar estadísticamente si dos perfiles, ya una vez proyectados en el plano de mejor ajuste,se pueden considerar iguales o no.En el desarrollo de esta estadística se utilizan resultados de teoría asintótica y elementos de remuestreo tipo bootstrap.

4.1 El Método Bootstrap.

Suponga que se tiene una muestra aleatoria de tamaño n de una función de distribución de probabilidad F no especificada,

$$(\underline{X}=\underline{x})=(X_i=x_i, X_i \sim F \text{ iid}, i=1,\ldots,n) \qquad (4.1)$$

Dada una función $R(\underline{X},F)$ el problema es estimar la distribución muestral de R en base a la observación de una muestra $\underline{x}$.El método bootstrap en este caso es el siguiente(Efron,1977):

i)Construya la distribución de probabilidad muestral F_n asignando (1/n) de probabilidad a cada x_i, i=1,...,n.

ii)Dada F_n, genere una muestra de tamaño n

$$(\underline{X}^*=\underline{x}^*)=(X_1^*=x_1^*, X_i^* \sim F, iid, i=1,...,n)$$

llamada la muestra bootstrap.

iii)Aproxime la distribución de R por la distribución bootstrap de $R(\underline{X}^*, F_n)$.

Para el cálculo de la distribución de $R(\underline{X}^*, F_n)$ se tienen tres alternativas.

i)Cálculo teórico directo.
ii)Aproximación Monte Carlo de la distribución bootstrap.
iii)Métodos de expansión por series de Taylor.

En el trabajo desarrollado en esta sección se usó ii), es decir, usando masivamente la computadora.

Por otra parte, un aspecto importante a señalar es que este método de remuestreo es aplicable también en el caso en que se conozca la familia paramétrica a la que pertenezca F sólo desconociendo los valores de los parámetros que la indexan (Efron, 1977, observación K). Este es el caso de las aplicaciones de este método en los desarrollos presentados más adelante, ya que se está trabajando con tablas de contingencias que son gobernadas por la distribución Multinomial; así al considerar F se debe

pensar en un miembro de la familia Multinomial,es decir,si $\underline{X}$=vec(N) vector de L entradas entonces

$$P[\underline{X}=\underline{x}]=(n!/\prod_{l=1}^{L} x_l!)\prod_{l=1}^{L} P_l^{x_l}$$

donde

$$L=(p+1)(q+1),\ n=\sum_{l=1}^{L} x_l$$
$$\underline{x}=(n_1,\ldots,n_L),\ n_l=\#(X_i=l) \qquad (4.2)$$
$$l=1,\ldots,L;\ i=1,\ldots,n,$$
$$P_l=Prob[X_i=l]$$

denotada por $M(n,(P_l,l=1,\ldots,L))$.

En este caso se tendrán que estimar (P_l) ,con lo que

$$F_n=M(n,(\hat{P}_l,l=1,\ldots,L)$$

con

$$\hat{P}_l=n_l/n \qquad l=1,\ldots,L$$

los estimadores máximo verosimil correspondientes (Lindgren,1976).

En otro sentido,para usar el método bootstrap se requiere del cumplimiento de ciertas condiciones de regularidad de la función R(.,.).Asi,Efron(1977,observación G),en el caso de que el espacio muestral sea finito con lo que la función F puede representarse como una tabla de probabilidades cuyos elementos son los correspondientes al vector L-dimensional P_v=vec(P),define

$$D(\hat{P}_v,P_v)=R(\underline{X},F_n). \qquad (4.3)$$

Ahora dado que

$$\hat{P}_v \mid P_v \sim M_L(n, P_v)$$

y (4.4)

$$P^*_v \mid P_v \sim M_L(n, P_v)$$

para muestras grandes $\hat{P}_v$ estará cerca de P_v con lo que la distribución de (4.3) será razonablemente aproximada por aquella generada por la muestra bootstrap.En cuanto a la validez asintótica de esta aproximación se tienen que cumplir con las condiciones de regularidad (4.5):

i. $Q(P_v, P_v)=0$

ii. $u(P_v, \hat{P}_v)=(\partial / \partial P^*_{v,1}[Q(P^*_v, \hat{P}_v)])$
exista continuamente para $(P^*_v, \hat{P}_v)$ en una vecindad abierta de (P_v, P_v). (4.5)

iii. $u=u(P_v, P_v)\neq 0$

iv. $Q(.,.)$ no dependa de de n.

Bajo estas condiciones y aplicando el teorema de Taylor(Serfling,1980),y el hecho de que P^*_v y $\hat{P}_v$ converge con probabilidad uno a P_v se tiene que

$$Q(\hat{P}_v, P_v)=(\hat{P}_v - P_v)(u+e_n)$$

y (4.6)

$$Q(P^*_v, \hat{P}_v)=(P^*_v - \hat{P}_v)(u+\hat{e}_n),$$

$e_n, \hat{e}_n$ convergen a cero con probabilidad uno.Ahora dado (4.4) y el

hecho de que $\hat{P}_v$ converge a P_v con probabilidad uno,

$$n^{(1/2)}(\hat{P}_v - P_v) \mid P_v \sim AN(0, \Omega_P) \tag{4.7}$$

y

$$n^{(1/2)}(P_v^* - \hat{P}_v) \mid P_v \sim AN(0, \Omega_P)$$

con Ω_P matriz de covarianzas con elementos $p_{v,l}(\delta_{lm} - p_{v,m})$, $\delta_{lm} = 1(0)$ si $l = m (l \neq m)$. De (4.6) y (4.7)

$$\sqrt{n}\, d(P_v^*, \hat{P}_v) \mid \hat{P}_v \tag{4.8}$$

es asintóticamente equivalente a

$$\sqrt{n}\, d(\hat{P}_v, P_v) \mid P_v \tag{4.9}$$

ambas siendo asintóticamente Normales

$$N(0, u' \Omega_P u). \tag{4.10}$$

Así, bajo el cumplimiento de las condiciones (4.5), el método bootstrap podrá utilizarse. En las secciones siguientes se hacen dos aplicaciones del enfoque bootstrap en el A.C.; en cada caso se hacen los señalamientos necesarios para satisfacer (4.5).

4.2 Valores Singulares y el Método Bootstrap.

En el capítulo 3 se presentaron los resultados de la teoría asintótica en el caso de interés práctico, es decir, cuando existe dependencia entre los criterios de clasificación y se suponen distintos los valores singulares no cero. Sin embargo, se hizo

énfasis que el uso de estos resultados depende del conocimiento de la estructura teórica de la tabla considerada,lo cual es irreal en las aplicaciones.

Aun así,los resultados asintóticos pueden combinarse con el enfoque bootstrap,para dar respuestas útiles en la práctica.

Como se mencionó en el capítulo tres,los estimadores de los valores singulares centrados (3.7) son variables aleatorias cuya distribución asintótica es Normal en el caso de que los valores teóricos correspondientes sean distintos de cero.Basándose en este resultado,al generar

$$P^{*}_{b} \quad b=1,\ldots,B \qquad (4.11)$$

matrices de frecuencias relativas bootstrap,y que a partir de estas se obtenga

$$(\hat{u}^{b}_{k}, k=1,\ldots,p) \quad b=1,\ldots,B \qquad (4.12)$$

si para k fija,la distribución empírica corresponde estadísticamente a la distribución Normal se tendrá un elemento para afirmar que el correspondiente u_k es distinto de cero.

Este procedimiento será aplicable si las funciones Q_k cumplen con las condiciones de regularidad (4.5) presentadas en la introducción de este capítulo.Así,en forma trivial se tiene que

$$Q_k(u_k,u_k)=0 \quad k=1,\ldots,K$$

cumpliéndose la primera condición de regularidad;respecto a que las funciones Q_k sean continuamente diferenciables respecto a P^{*}_{vl},resulta del hecho de que los valores singulares pueden verse como raíces de polinomios,que por teoría en esta materia (Lancaster,1969),resultan cumplir con la condición requerida.El cumplimiento de la tercera condición se da ya que de otra manera los estimadeores centrados de los valores singulares no serían asintóticamente Normales.Por último estandarizando por $n^{1/2}$, se cumple la cuarta condición.

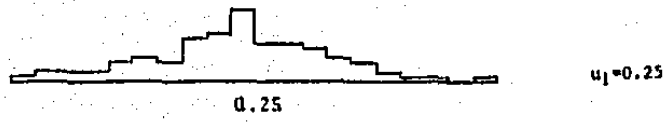
De esta manera la estadística a aplicarle el remuestreo será

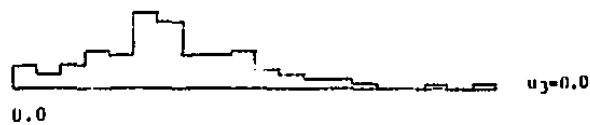
$$(\hat{u}_k - u_k) \qquad k=1,\ldots,p. \qquad (4.13)$$

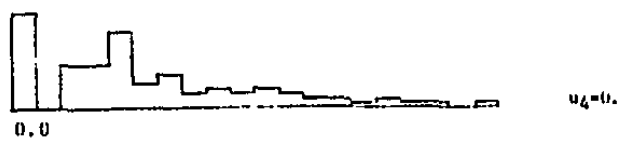
Este planteamiento fue originalmente sugerido por la observación de los histogramas generados para cada valor singular a partir de un experimento (O'Neill,1978,ejemplo 2) con 191 ensayos Monte Carlo(figura 4.1),los histogramas correspondientes a valores teóricos cero muestran un sesgo mientras que el correspondiente al valor teórico distinto de cero da una apariencia simétrica.Esta situación combinada con los resultados asintóticos ha hecho que específicamente se proponga el siguiente algoritmo:

i)A partir de la tabla de datos originales genérense B repeticiones bootstrap de la variable aleatoria

Figura 4.1 Histogramas con 191 ensayos Monte Carlo sobre Valores Singulares.







Multinomial tomando a P como el verdadero valor del parámetro, generándose las matrices (4.11).

ii)Aplique el A.C. a cada una de las tablas produciendo (4.12).

iii)En orden ascendente,a partir del último valor singular estimado distinto de cero numbricamente hablando, aplique pruebas de bondad de ajuste para determinar si la distribución bootstrap es Normal.Deténgase en el momento en que rechace la normalidad para algún valor singular,y a partir de éste hacia arriba considere que todos son distintos de cero estadisticamente.

Este algoritmo fue probado con el siguiente ejemplo:

Para una tabla 5x5 (p=q=4),y un tamaño total de 4096 individuos a contabilizar,se parte de las fórmulas de reconstitución (3.5) de la sección 3.2,

$$P_{ij}=P_{i.}P_{.j}(1+\sum_{k=1}^{K}u_{ik}x_{ik}y_{jk})$$

con las condiciones (4.14)-(4.15) a continuación:

$$P_{i.}=(1/16,4/16,6/16,4/16,1/16)=P_{.j}, \quad i,j=1,\ldots,5. \qquad (4.14)$$

$$x^{(1)}=(6,0,-2,0,6)/\sqrt{6} \qquad y^{(1)}=(0,1,0,-1,0)\sqrt{2}$$
$$x^{(2)}=(-2,-1,0,1,2) \qquad y^{(2)}=(1,0,0,0,-1)\sqrt{8} \quad (4.15)$$
$$x^{(3)}=(2,1,0,1,-2) \qquad y^{(3)}=(-4,1,0,1,-4)2/\sqrt{10}$$

y

$$u_1=0.25, u_2=u_3=u_4=0.$$

es decir con K=1.

Esta tabla fue definida por O'Neill(1978) en un experimento Monte Carlo para corroborar sus resultados asintòticos con los de muestra finita;debido a tiempo de màquina sòlo generò 191 ensayos Monte Carlo.Partiendo de las condiciones (4.14)-(4.15),se generaron 191 ensayos MonteCarlo con B=50 repeticiones bootstrap;esto fue realizado mediante programaciòn en las màquina Borroughs B7800 de la U.N.A.M. en el compilador fortran 4 utilizando el generador de nùmeros pseudo aleatorios propio de esta màquina.A cada una de las matrices P^*_b se realizò el A.C. correspondiente obtenièndose (4.12) y para cada k se aplicò la prueba Anderson-Darling caso tres cuya descripciòn es:

> Sea $X_1,\ldots,X_n$ una muestra aleatoria de la cual se requiere saber si estadisticamente proviene de una distribuciòn Normal con parÂmetros desconocidos μ, σ^2. Calcule la media y la varianza mustral.Para cada X_i

calcule

$$Z_i = F(X_i | \bar{X}, S^2) \quad i=1,\ldots,n.$$

donde F es la dist.acumulativa Normal.

Obtenga las estadísticas de orden de las Z_i's para calcular

$$A2=-n-(1/n)\sum_i (2i-1)[\log(Z(i))+\log(1-Z(n+1-i))].$$

Si A_2 es menor o igual que el cuantil $A_{2,\alpha}$ entonces no se rechazará que la muestra provenga de la distribución Normal.

Si se recuerda,K denota al número de raíces mayores a cero;denótese por K a la variable aleatoria que contabiliza el número de raíces ordenadas distintas de cero detectadas por el algoritmo propuesto.Los resultados de la simulación se muestran en el cuadro 4.1.En esta simulación K=1,y de acuerdo a los resultados

$$P[K>=3]=\begin{cases} 0.95 & \text{significancia } 0.05 \\ 1.0 & \text{significancia } 0.01 \end{cases}$$

con lo que K resulta ser una cota superior de K;sin embargo,vía el algoritmo no se tomó como cero a algún valor singular que en realidad fuera positivo.Estos resultados aunque no son del todo

Cuadro 4.1. Probabilidades estimadas del número de raíces distintas de cero por medio del experimento Montecarlo con B=50 replicaciones bootstrap en cada uno de los 191 ensayos.

$\hat{P}[\hat{k}=k]$

k	α =0.05	α = 0.01
0	0	0
1	.01	0
2	.04	0
3	.75	.49
4	.20	.51

satisfactorios,si sugieren que el algoritmo trabaja en forma conservadora evitando errores graves consistentes en no considerar a ejes principales que tengan algo que aportar en la explicación de la tabla.Tratando de explotar al máximo la información generada por la simulación, para determinar las causas en el comportamiento del algoritmo,se calcularon en forma marginal,es decir cada una de las raices por separado sin tomar en cuenta a las restantes,los porcentajes en que se rechazó la hipótesis de que la raiz sea distinta de cero cuando lo es(α),y la proporción de veces en que se rechazó la hipótesis de que una raiz es cero cuando no lo es (β).Estos porcentajes fueron calculados en base a dos niveles de significancia(0.05 y 0.01),mostrándose los resultados en la cuadro 4.2;a la vista de los resultados en este cuadro el algoritmo aplicado en forma marginal muestra tener un nivel de significancia adecuado aunque una potencia pobre para detectar ceros en el caso de raices intermedias.Debe mencionarse que se aplicaron otras pruebas de bondad de ajuste al caso Normal y los resultados fueron menos satisfactorios que los arrojados por la prueba de Anderson-Darling.

Entre las posibles causas de los malos resultados en u_2 y

Cuadro 4.2.Errores tipo 1 y tipo 2,Exp.Montecarlo 191 ensayos.

Valor Singular	$\alpha = 0.05$	$\alpha = 0.01$
$u_1 = 0.25$	$\hat{\alpha} = 0.03$	$\hat{\alpha} = 0.01$
$u_2 = 0$	$\hat{\beta} = 0.92$	$\hat{\beta} = 0.96$
$u_3 = 0$	$\hat{\beta} = 0.94$	$\hat{\beta} = 0.98$
$u_4 = 0$	$\hat{\beta} = 0.19$	$\hat{\beta} = 0.51$

u_3,siguiendo las ideas de Efron y Tibshirani(1986),se puede argumentar que el nûmero de repeticiones B=50 es insuficiente para probar una hipôtesis.En su trabajo,êstos autores afirman que en los casos en los que no sôlo se quiere tener una idea de la precisiôn en la estimaciôn de un parâmetro,sino que se requiere construir por ejemplo un intervalo de confianza,se tendrâ que incrementar sustancialmente a B para tener resultados confiables. En el presente trabajo para investigar la influencia del tamaño de B,se trabajaron dos casos;el primero fue un ensayo MonteCarlo de los generados por (4.14)-(4.15) y el segundo partiendo de la misma estructura pero considerando que el primer valor singular agrupaba el 90% de la variaciôn total y el segundo el 10% restante,es decir,u_1=0.2236068,u_2=0.0707107,u_3=u_4=0.

En ambos ejemplos se generaron 200,500 y 1000 repeticiones bootstrap,construyêndose en cada caso histogramas(figuras 4.2 y 4.3),y se aplicô el algoritmo propuesto inicialmente,cuyos resultados se muestran en la cuadro 4.3.

En el primer ejemplo tanto B=500 como B=1000 dan resultados aceptables,si bien α=0.01 se presenta la tendencia a dar una cota superior.En el segundo ejemplo sôlo con 1000 repeticiones se tienen resultados satisfactorios tanto en α=0.01 y

0.05.Observando los histogramas,se evidencia que los histogramas tienden a sesgarse cuando el valor teórico es cero;esto es marcado en los casos con B=1000.Aparentemente en estos ejemplos el número B si tiene influencia.

En vista de lo anterior,será recomendable para aplicar el algoritmo realizar B>=1000 repeticiones bootstrap.

Esta afirmación deberá ser corroborada en base a un experimento Monte Carlo con un número considerable de ensayos,experimento que no fue realizado por limitaciones en tiempo.Sin embargo los resultados apoyan los comentarios de Efron y Tibshirani en el sentido de mayor precisión a mayor B.

Una última propuesta del uso del enfoque Bootstrap se debe a Beran(1985);él propone trabajar con

$$r_k = \hat{\Omega}^2_k / \sum_k^p \hat{\Omega}^2_k \qquad (4.16)$$

es decir con el porcentaje de variación explicado por el eje k.Esta estadística si bien no es utilizable en términos de saber estadísticamente si un valor singular es cero,permite establecer cuantiles de la proporción r_k en base a la aproximación de su distribución bootstrap obtenida por simulación.

Cuadro 4.3. Efecto del tamaño de B en el algoritmo de dimensionalidad.

Ejemplo 1. $u_1=0.25, u_2=u_3=u_4=0$

B	0.05	0.01
200	para en u_2	para en u_3
500	para en u_1	para en u_3
1000	para en u_1	para en u_2

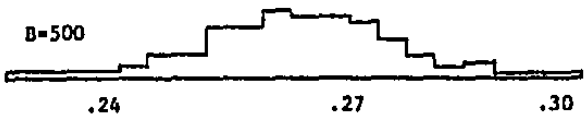
Ejemplo 2. $u_1=0.2236068, u_2=0.0707107, u_3=u_4=0$

B	0.05	0.01
200	para en u_2	para en u_3
500	para en u_1	para en u_2
1000	para en u_2	para en u_2

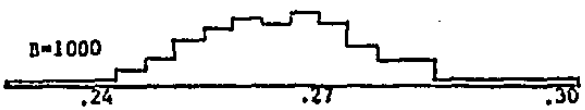
Figura 4.2 Ensayo Montecarlo 1.Sensibilidad ante cambios en B.

Valor Teórico u_1=0.25

sesgo	0.41
curtosis	2.89
$\hat{\lambda}_2$	0.84



sesgo	0.19
curtosis	2.93
$\hat{\lambda}_2$	0.52



sesgo	0.11
curtosis	2.90
$\hat{\lambda}_2$	0.37

Figura 4.2 (continuación).Ensayo MonteCarlo 1

valor teórico $u_2=0$

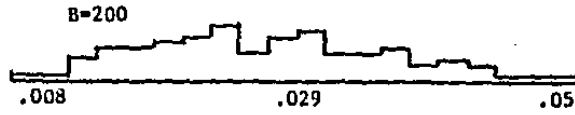
sesgo 0.44
curtosis 2.77
$\hat{A}_2$ 0.98

sesgo 0.33
curtosis 3.21
$\hat{A}_2$ 0.90

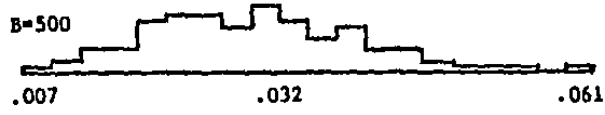
sesgo 0.28
curtosis 3.23
$\hat{A}_2$ 0.86

Figura 4.2 (continuación)

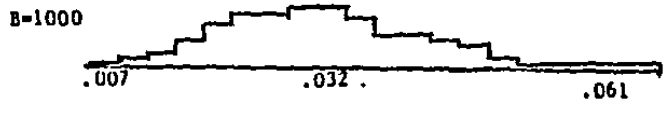
valor teórico $u_3=0.0$



sesgo	0.21
curtosis	2.30
$\hat{A}_2$	0.78



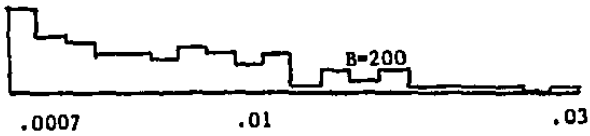
sesgo	0.25
curtosis	2.65
$\hat{A}_2$	0.97



sesgo	0.24
curtosis	2.68
$\hat{A}_2$	1.38

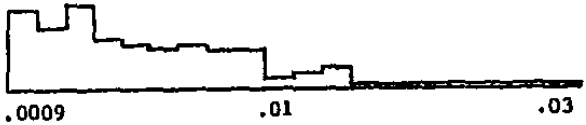
Figura 4.2 (continuación)

valor teórico u_4=0.0



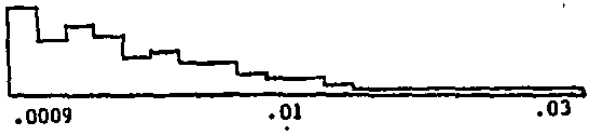
sesgo	0.63
curtosis	2.74
$\hat{A}_2$	2.96

B=500



sesgo	0.83
curtosis	3.29
$\hat{A}_2$	8.60

B=1000



sesgo	0.90
curtosis	3.47
$\hat{A}_2$	18.68

Figura 4.3 Ensayo MonteCarlo 2.Sensibilidad ante cambios en B.

valor teórico u_1=0.2236068

sesgo	0.36
curtosis	2.91
$\hat{\lambda}_2$	0.67

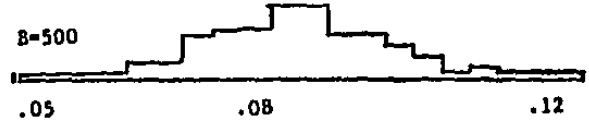
sesgo	0.23
curtosis	2.81
$\hat{\lambda}_2$	0.72

sesgo	0.17
curtosis	2.83
$\hat{\lambda}_2$	0.65

Figura 4.3 (continuación).

valor teórico u_2=0.0707107

sesgo	0.23
curtosis	3.15
$\hat{A}_2$	0.51

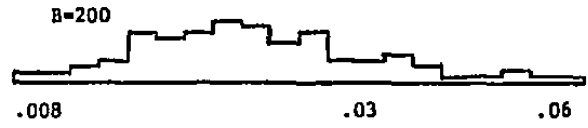


sesgo	0.23
curtosis	3.30
$\hat{A}_2$	0.80

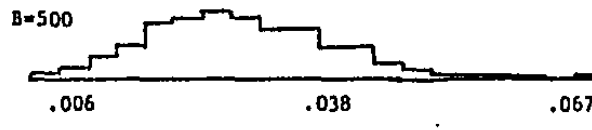
sesgo	0.04
curtosis	3.17
$\hat{A}_2$	0.60

Figura 4.3 (continuación)

valor singular u_3=0.0



sesgo	0.42
curtosis	2.83
$\hat{\lambda}_2$	0.87

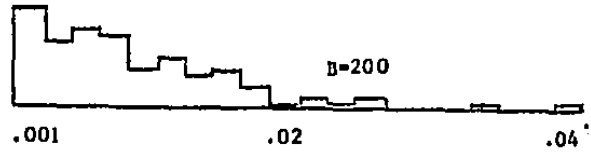


sesgo	0.50
curtosis	3.28
$\hat{\lambda}_2$	1.73

sesgo	0.46
curtosis	3.18
$\hat{\lambda}_2$	2.71

Figura 4.3 (continuación)

valor teórico $u_4=0.0$



sesgo	1.12
curtosis	4.90
$\hat{\lambda}_2$	3.52

B=500

sesgo	1.12
curtosis	5.05
$\hat{\lambda}_2$	7.99

B=1000

sesgo	1.02
curtosis	4.21
$\hat{\lambda}_2$	18.38

4.2 Agrupamientos de Perfiles en el Plano de Mejor Ajuste.

Como se mencionó en el capitulo dos de este trabajo,el objetivo primordial del A.C. es la representación gráfica de las relaciones relevantes en una tabla de contingencias con dos criterios de clasificación.En cuanto al problema inferencial la mayoría de los esfuerzos se han centrado en los valores singulares;por otra parte,en la práctica la interpretación de la cercanía en el plano de mejor ajuste de dos perfiles renglón o columna es uno de los aspectos de mayor preponderancia en el A.C..En este sentido se cree que seria útil contar con una herramienta inferencial que nos permita trabajar sobre los posibles agrupamientos de perfiles.Greenacre(1984),ha utilizado el enfoque bootstrap en lo que él llama la estabilidad de la representación en dos dimensiones.Especificamente,a partir de la tabla original genera (4.11),obteniendo en cada caso las coordenadas proyectadas sobre el plano de mejor ajuste de los perfiles originales,mediante la fórmula de transición

$$\hat{F}_b = D_r^{-1} P_b^{*} \hat{G} D_c^{-1} \qquad b=1,\ldots,B \qquad (4.17)$$

donde $\hat{F}_b$ son las coordenadas de los perfiles renglón de la repetición b,pero en el espacio ortonormal correspondiente a las

coordenadas de los perfiles columna en su espacio ortonormal original $\hat{G}$ y D_Δ la matriz diagonal de valores singulares mayores que cero;en forma anàloga para los perfiles columna se realiza el procedimiento mediante la respectiva fòrmula de transiciòn.De un anàlisis visual Greenacre obtiene conclusiones sobre la estabilidad.Para mayores detalles consùltese Greenacre(1984,cap.8).Esto presenta problemas, ya que sus conclusiones dependen fuertemente de las intersecciones entre las nubes de perfiles proyectados, hecho que puede tener implicaciones de poca claridad.

Se denota a $(\hat{f}_{i1},\hat{f}_{i2})$ como las coordenadas del i-èsimo perfil respecto al plano de mejor ajuste.Considere al vector de diferencias por coordenadas en el plano de mejor ajuste de dimensiòn dos entre el perfil renglòn i y el perfil renglòn i',es decir,

$$\hat{d}_{ii'} = (\hat{f}_{i1}-\hat{f}_{i'1},\hat{f}_{i2}-\hat{f}_{i'2}) \qquad (4.18)$$

y a la estadistica

$$\hat{d}_{ii'}\,\Phi_{ii'}^{-1}\,\hat{d}_{ii'} \qquad (4.19)$$

donde $\Phi_{ii'}$ es la matriz de covarianzas de (4.18).Esta estadistica serà utilizable en la prâctica en la medida que se conozca esta matriz.El mètodo bootstrap es utilizado para estimar

tal matriz de covarianzas, que en la pràctica no es posible conocer;para ello nòtese que via las fòrmulas de transiciòn,

$$\hat{d}_{ii'} = \hat{D}_r^{-1}\hat{P}\hat{G}\hat{D}_u^{-1}g_{ii'} \qquad (4.20)$$

donde

$$g_{ii'} = (0,0,\ldots,\underset{i}{1},0,\ldots,\underset{i'}{-1},0,\ldots 0)$$

vector de q entradas.El enfoque bootstrap como se recordarà,aproxima la distribuciòn de (4.18) via la muestra bootstrap denotada por $\hat{d}_{ii',b}$ b=1,...,B;la validez asintòtica de tal aproximaciòn,nuevamente como en la secciòn 4.1,requiere del cumplimiento de las condiciones (4.5).Para ello considere la estadistica

$$\hat{d}_{ii'} - d_{ii'},$$

donde $d_{ii'}$ es la media de la distribuciòn de (4.18);en vista de lo anterior la primera condiciòn de (4.5) se cumple.Dado (4.20),cuando se hacen las repeticiones bootstrap

$$\hat{d}_{ii',b} = D_{rb}^{-1} P_b^* \hat{G} D_u^{-1} g_{ii'} \qquad (4.21)$$
$$= H(P_b^*)\hat{G}D_u^{-1}g_{ii'}.$$

$H(P_b^*)$ es una funciòn continuamente diferenciable de P_b^* ;por otra parte, que la derivada parcial de H(.) respecto a P_{ij} evaluada en P sea distinta de cero se trabajò en el caso general de una tabla

2x2,es decir dos perfiles renglòn y dos perfiles columna.En este caso se tendrà una recta de mejor ajuste;en vista de (4.20)

$$\begin{pmatrix} \hat{f}_{11} \\ \hat{f}_{21} \end{pmatrix} \begin{pmatrix} 1/r_1 & 0 \\ 0 & 1/r_2 \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \begin{pmatrix} \hat{g}_{11}/\hat{u}_1 \\ \hat{g}_{21}/\hat{u}_1 \end{pmatrix}$$

por lo que,por ejemplo

$$\partial/\partial p_{11} \begin{pmatrix} f_{11} \\ f_{21} \end{pmatrix} = \begin{pmatrix} [p_{12}(\hat{g}_{11}-\hat{g}_{21})]/[r_1^2\hat{u}_1] \\ 0 \end{pmatrix}$$

con lo que para que sea distinta de cero se requiere de que

$$g_{11} \neq g_{21} \quad y \quad p_{12} \neq 0.$$

Para tablas de mayores dimensiones sin embargo,se espera que las condiciones no sean tan fàciles de manejar.

Por ùltimo, la cuarta condiciòn de (4.5) se cumple trivialmente.

Dado lo anterior se justifica el uso del enfoque bootstrap;aùn asi sòlo se tiene la equivalencia asintòtica de la distribuciòn de $\hat{g}_{ii'}$ con la ley de $g_{ii'b}$;sin embargo(Serfling,1980),la convergencia en distribuciòn implica convergencia en momentos de orden dos si $(g^2_{ii'b})$ es uniformemente integrable.

Para ello serà suficiente mostrar que $H(P^*_b)$es uniformemente integrable.Se tiene la siguiente demostraciòn:

Existe una vecindad cerrada alrededor de P tal que H(.) està bien

definida sobre ella.Sea $\overline{V(P)}$ tal vecindad.H(.) es continua sobre $\overline{V(P)}$ y por lo tanto es acotada.Ahora,

$$\hat{P} \dashrightarrow P \text{ c.p.1}$$

cuando el tamaño de muestra tiende a infinito y,por lo mismo,

$$P\$ \dashrightarrow \hat{P} \text{ c.p.1}$$

cuando n tiende a infinito;entonces

$$P\$ \dashrightarrow P \text{ c.p.1}$$

Si definimos

$$P^{0} = \begin{cases} P\$ & \text{si } P\$ \in \overline{V(P)} \\ P \text{ si } P\$ \notin \overline{V(P)} \end{cases}$$

entonces

$$P^{0} \in V(P)$$

siempre,y

$$n^{\delta}\,(H(P\$)-H(P^{0})) \dashrightarrow 0 \text{ c.p.1} \qquad \delta>0$$

por lo que H(P$) y $H(P^{0})$ son asintóticamente equivalentes.Por lo tanto $H(P^{0})$ converge en distribución a lo mismo que H(P$). $H(P^{0})$ es acotada,por lo tanto es uniformemente integrable(apéndice 2);también son uniformemente integrable.por teorema A sección 1.4 de Serfling(1980),los momentos de $H(P^{0})$ convergen a los momentos con respecto a la distribución límite,por la equivalencia asintótica también convergerán a lo

mismo los momentos de H(P*).

Por lo tanto, $(d^2_{ii',b})$ es uniformemente integrable.

Así el estimador bootstrap $\Phi_{ii'B}$ converge a $\Phi_{ii'}$ la matriz de covarianzas de $\hat{d}_{ii'}$.

Dado todo lo anterior la estadística propuesta es

$$D^2_{ii'} = \hat{d}_{ii'} \cdot \Phi^{-1}_{ii'B} \, {}^{b}\hat{d}_{ii'} \qquad (4.22)$$

donde

$$\Phi^{-1}_{ij'B} = (B-1)^{-1} \sum_{b=1}^{B} (d_{ij'b} - \bar{d}_{ij'})(d_{ij'b} - \bar{d}_{ij'})'$$

En cuanto a la distribución de ésta se tienen los siguientes argumentos:

i)vec($\hat{P}$-P)es asintóticamente Normal con media cero(Serfling,1980 pág 109).

ii)Dada la justificación del uso del método bootstrap,se cumplen las condiciones para que $\hat{d}_{ii'}$ sea una variable aleatoria con distribución asintótica Normal(Serfling,1980,pág.122 teorema A).

iii)En base a ii) y Serfling(1980,pag.130),asintóticamente

$$\hat{d}_{ii'} \cdot \Phi^{-1}_{ii'B} \hat{d}_{ii'} \sim \chi^2_{(2,gii')} \cdot \Phi^{-1}_{ii'\,gii')} \qquad (4.23)$$

iv)Bajo la hipótesis H_0 de que el perfil i es igual al perfil i' en el plano de mejor ajuste se espera que $d_{ii'}=0$ con lo que

$$\hat{g}'_{ii} \cdot \Phi^{-1}_{ii,B} \hat{g}_{ii} \sim \chi^2_{(2)} . \qquad (4.24)$$

Así, el perfil i y el perfil j son considerados estadísticamente iguales en el plano de mejor ajuste a un nivel si

$$\hat{g}'_{ii} \cdot \Phi^{-1}_{ii,B} \hat{g}_{ii} <= \chi^2_{(2),\alpha} .$$

5. Aplicaciones.

En la presente sección se presentan dos aplicaciones en el área de tecnología de alimentos, específicamente en el campo de la evaluación sensorial. Se detallan tanto el uso descriptivo de la técnica como las herramientas de corte inferencial discutidas en la sección anterior. A continuación se presenta la descripción general de la evaluación sensorial de una fruta llamada Jiotilla. Interesa el estudio de esta fruta considerada exótica y perteneciente a la familia de las cactáceas, por sus propiedades alimenticias y porque representa una fuente de ingresos adicionales para las comunidades rurales que tienen acceso a ella. Se prepararon 13 tratamientos de esta fruta:

TRATAMIENTO	IDENTIFICADOR
1. En fresco (control)	FFRE
2. Confitada en estufa	CEST
3. Confitada en sol	CSOL
4. Confitada al aire	CAIR
5. Mermelada	MERM
6. Almíbar	ALMI
7. Confitada por proceso lento	CPLE
8. Confitada por proceso rápido	CPRA
9. Glaseada por proceso lento	GPLE
10. Glaseado por proceso rápido	GPRA
11. Confitada y envasada en cloruro de polietileno	CECP
12. Confitada y envasada en celofán	CECE
13. Confitada y envasada en cloruro de polivideno.	CECV

Estos tratamientos fueron evaluados por un panel de 15 jueces no entrenados, que calificaron según una escala

organoléticas(hedónica) de nueve categorías:

CATEGORIA	IDENTIFICADOR
Gusta extremadamente	GEX
Gusta mucho	GMU
Gusta moderadamente	GMO
Gusta ligeramente	GLI
Indiferente	IND
Disgusta ligeramente	DLI
Disgusta moderadamente	DMO
Disgusta mucho	DMU
Disgusta extremadamente	DEX

Esto fue realizado en olor,sabor,color,textura y aspecto de cada uno de los trece tratamientos aplicados a la fruta;se presentan a continuación los resultados de las primeras dos evaluaciones.

5.1 OLOR.

Los datos se muestran en el cuadro 5.1.Se evidencia la nula contabilización desde DMO hasta DEX por lo que no se consideran en el análisis.La representación producida por A.C. se muestra en sus dos primeros ejes en la figura 5.1.

En forma descriptiva se observa que entre estos dos primeros ejes se tiene un 73% de la variación en los datos;respecto a la calidad de cada perfil en la representación con los dos primeros ejes, los perfiles renglón,es decir,las categorías de la escala de preferencias,en el cuadro 5.4 las contribuciones absolutas y las correlaciones.Así,en general están bien representados en estos dos ejes,ya que las contribuciones relativas son mayores al

0.5 salvo el caso de DLI.En cuanto a la contribución a los ejes,en el primero de ellos GEX y GLI son los preponderantes sucediendo esto también en el segundo eje,es decir,que ambos ejes son representantes de las posiciones relativas al gusto extremo y al gusto ligero.

En cuanto a los perfiles columna,los tratamientos,se tiene los resultados en cuanto a contribuciones absolutas y correlaciones en el cuadro 5.5.

En vista de las correlaciones,FFRE,CAIR,MERM y ALMI son tratamientos mal representados;los demàs en general no tienen mayores problemas.En cuanto a las contribuciones absolutas,CECP y GPLE son los que dan primordialmente la dirección al eje uno;respecto al segundo eje se unen al grupo anterior CEST y CECE.

En cuanto a la orientación de los trece tratamientos respecto a la escala de preferencias se tiene que GPLE descolla hacia el gusto extremo siguiendo en orden descendente CPLE,GPRA,FFRE,CPRA y CEST.Posteriormente se agrupan hacia un gusto moderado MERM,ALMI,CSOL y CAIR.Por último CECV,CECE y CECP son poco aceptados.

Utilizando las herramientas de corte inferencial comentadas en el

capitulo cuatro,en primer lugar los estimadores de los valores singulares junto con la estadistica Anderson-Darling(A^2) con 1000 repeticiones bootstrap son mostrados en el cuadro 5.2;via el algoritmo de dimensionalidad de la seccion 4.1,se tiene que los primeros cuatro valores singulares son estadisticamente distintos de cero,ya que tanto al 0.01 y al 0.05 se dio esta situación.

Con la misma informacion proporcionada por las 1000 repeticiones bootstrap,en base a(4.16) se construyeron intervalos al 95% para porcentajes de variación marginal y acumulada para cada uno de los valores singulares,mostrándose estos resultados en el cuadro 5.3,con lo que,la representación global en los dos primeros ejes oscilará entre 63 y 79 % lo que es aceptable en general y se observa que nuevamente hasta el valor singular 4 no incluye al valor cero por ciento.

Por otra parte se aplicó la estadistica D^2 para detectar agrupamientos tanto en el caso de los perfiles renglón como de los perfiles columna.Antes de mostrar algún resultado,debe mencionarse que sólo se reportan los correspondientes a los perfiles que tuvieron una correlación acumulada mayor al 0.5 con los dos primeros ejes principales;si bien lo anterior es

arbitrario,se tomó esta determinación ya que esta estadística producía agrupamientos demasiado extendidos al considerar a los perfiles mal representados.La determinación estadística de cual perfil está bien representado en un conjunto de ejes es un asunto que no ha sido atacado hasta ahora y que se cree será necesario tratar en un futuro.

A un nivel de significancia global de 0.036(cada comparación 0.001 $\binom{9}{2}$ comparaciones),los tratamientos se agrupan de la siguiente manera:

1.GPLE,CPLE,GPRA,CPRA,CEST.

2.CECV,CECE,CECP.

3.CSOL como puente entre 1 y 2.

Respecto a la escala de preferencias con un nivel global de 0.01(cada comparación con 0.001, $\binom{5}{2}$ comparaciones),se tiene que no hay agrupamiento alguno.

Conjuntando ambos grupos de resultados se tiene que:

1.GPLE,CPLE,GPRA,CPRA y CEST se orientan hacia GEX y GMU.

2.CECV,CECE y CECP se orientan hacia el GLI e IND.

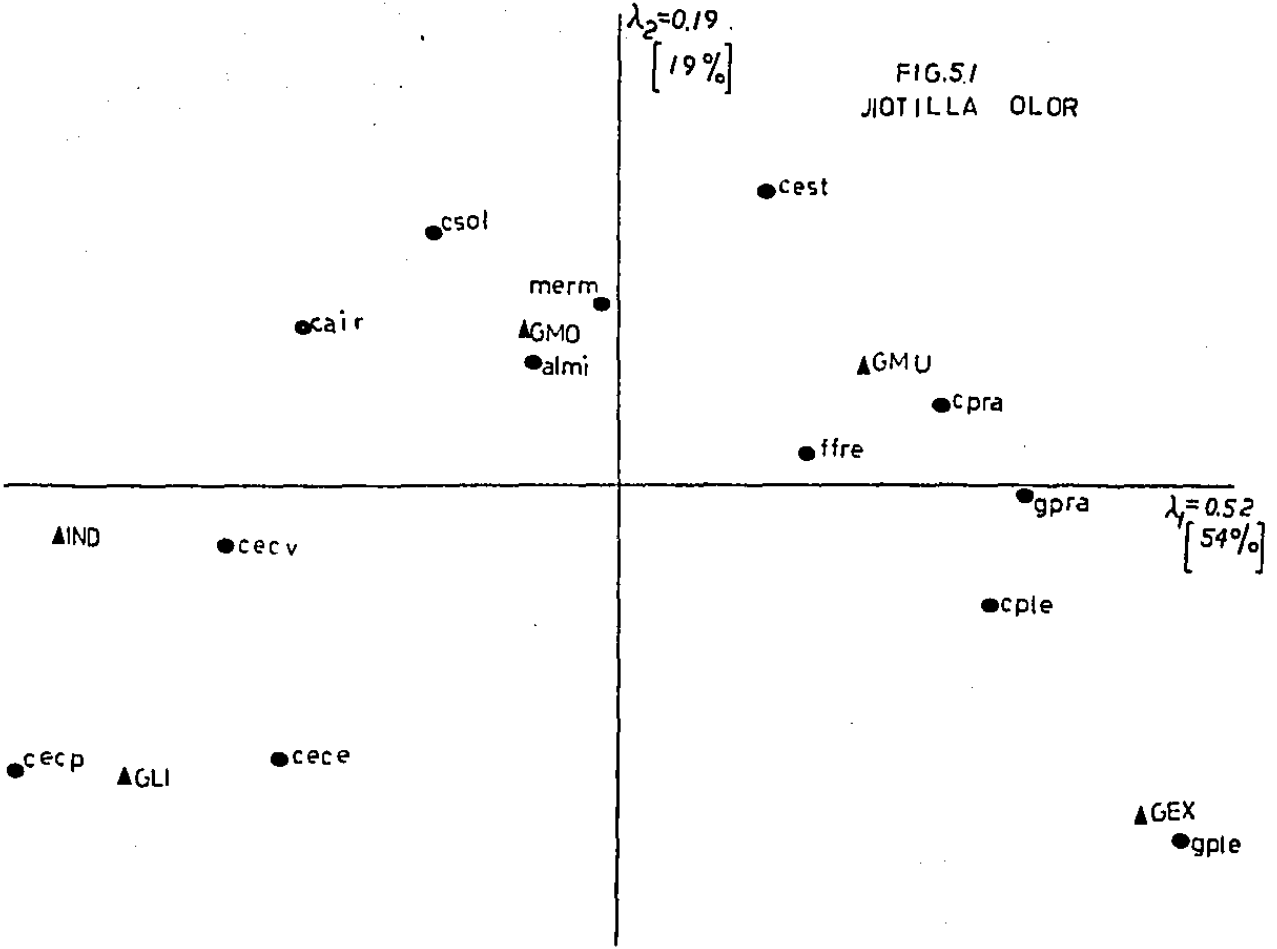
3.CSOL se sitúa hacia el gusto moderado.

Los resultados generales se muestran en el cuadro 5.6.

De lo anterior,se tiene que los glaseados y los confitados tanto

Cuadro 5.1 Tabla de contingencias,OLOR.

	FRUS	CEST	CSOL	CAIF	MLPH	ALMI	CPLE	CPRA	GPLE	GPRA	CFCP	CECE	CECV
	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.
GEY	4.	0.	0.	0.	0.	0.	4.	8.	0.	4.	1.	1.	0.
GGJ	15.	0.	0.	0.	0.	0.	14.	8.	0.	10.	1.	1.	0.
GGL	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
GLN	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
GPL	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
JEJ	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
JPE	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.


λ_2=0.19
[19%]
FIG.5.1
JIOTILLA OLOR
cest
csol
merm
cair
GMO
almi
GMU
cpra
ffre
gpra
λ_1=0.52
[54%]
IND
cecv
cple
cecp
GLI
cece
GEX
gple

Cuadro 5.2 Estadistica Anderson-Darling para los valores singulares,OLOR.

No.	valor	$\hat{A}^2$
1	0.72	0.32
2	0.46	0.31
3	0.40	0.34
4	0.26	0.19
5	0.14	67.95

Cuadro 5.3. Intervalos de confianza para porcentajes marginales y acumulados con porcentaje marginal observado.

	marginal			acumulado	
	observado	11%	1s%	11%	1s%
1	54.19	37.75	51.16	37.75	51.16
2	19.53	20.99	32.83	63.22	79.26
3	16.31	12.91	22.17	82.29	95.21
4	7.71	4.31	13.25	93.05	100.00
5	2.25	0	6.90		

Cuadro 5.4.Contribuciones Absolutas y Correlaciones.
Escala de preferencias,COLOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje 1	eje2	eje1	eje2
GEX	0.314	0.381	0.648	0.283
GMU	0.149	0.101	0.604	0.148
GMD	0.024	0.181	0.153	0.417
GLI	0.279	0.276	0.621	0.221
IND	0.223	0.006	0.558	0.005
DLI	0.011	0.055	0.098	0.184

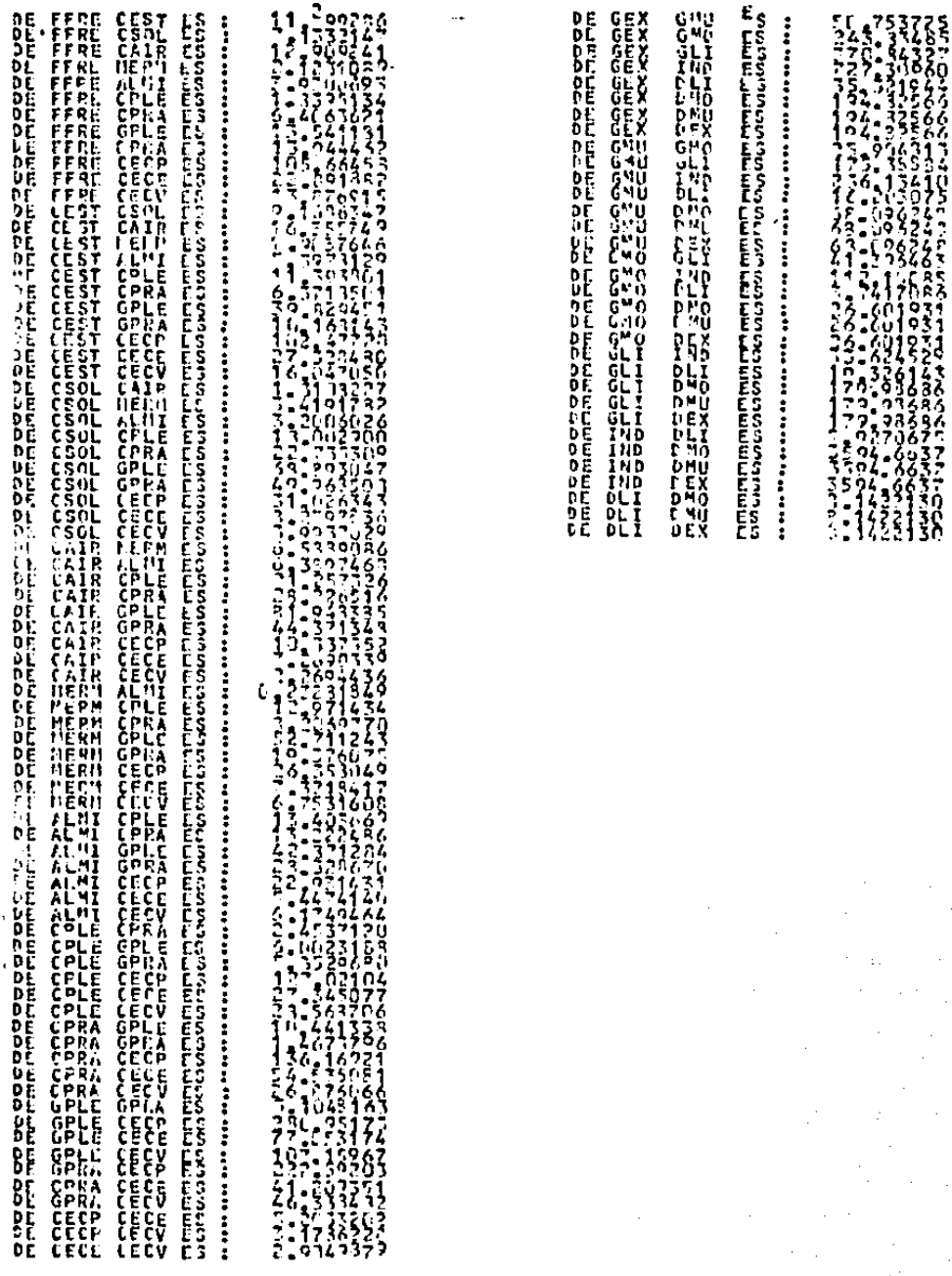
Cuadro 5.5 Contribuciones Absolutas y Correlaciones.
TRATAMIENTOS,OLOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje 1	eje2	eje1	eje2
FFRE	0.023	0.002	0.288	0.010
CEST	0.013	0.167	0.142	0.645
CSOL	0.023	0.118	0.220	0.413
CAIR	0.066	0.046	0.345	0.087
MERM	0	0.061	0.007	0.463
ALMI	0.005	0.026	0.108	0.190
CPLE	0.084	0.031	0.787	0.103
CPRA	0.063	0.010	0.789	0.045
GPLE	0.203	0.241	0.659	0.282
GPRA	0.103	0	0.762	0
CECP	0.240	0.150	0.776	0.175
CECE	0.076	0.141	0.337	0.227
CECV	0.100	0.007	0.848	0.020

Cuadro 5.6 Estadística D^2, B=1000, OLOR.

TRATAMIENTOS

ESCALA HEDONICA



por proceso lento como por proceso rápido son tratamientos de muy alta aceptación;en el mismo sentido se encuentra la jiotilla confitada por estufa.Siguiendo de este grupo de alta aceptación, se encuentra la fruta confitada al sol.Por último los tratamientos envasados se ven demeritados en su olor.

5.2. Sabor.

Los datos se muestran en el cuadro 5.7.La representación gráfica producida por el A.C. se mustra en la figura 5.2.

Descriptivamente,los dos primeros ejes contemplan el 71% de la variación en los datos;en cuanto a la representación de las categorias de la escala de preferencias se tiene en base al cuadro 5.11,los siguientes comentarios:

Se tiene que GMO y DEX tienen graves problemas en su representación.En cuanto a la contribución absoluta,en el primer eje sobresalen DLI,DMO y GMU con lo que este eje muestra la posición relativa entre el disgusto moderado y el gusta mucho;en el segundo eje nuevamente DMO pesa pero ahora superado por GLI.Respecto a los perfiles columna,es decir los tratamientos a la fruta se tiene en base al cuadro 5.12,lo siguiente:

Se tienen problemas en la representación de FFRE,CEST,CSOL,CAIR y

MERM(correlación< 0.5).En cuanto a las contribuciones absolutas CECP y CECV son los de mayor influencia en ambos ejes.en cuanto a la orientación especifica de los tratamientos a la fruta respecto a la escala de preferencias,se tiene un agrupamiento fuertemente aglutinado alrededor de GEX y GMU,estando constituido por CPLE,GPLE,GPRA y CEST siguiendo MERM,FFRE,ALMI y CSOL.Hacia GLI se tiene a CECE,CPRA,CAIR y CECV.Por último,se separa totalmente CECP hacia DMO y DMU.

En cuanto a la utilización de las herramientas inferenciales,se tiene en primer término que los seis primeros valores singulares pueden considerarse distintos de cero(véase cuadro 5.8).

Como en el caso de la evaluación del OLOR se calcularon intervalos al 95% de confianza para los porcentajes marginales y acumulados(cuadro 5.9).De esta manera,la representación en los dos primeros ejes en su calidad global es satisfactoria(del 58% al 75%).

Al utilizar la estadística propuesta en el capitulo anterior omitiendo a los mal representados,los agrupamientos más relevantes en los tratamientos a una significancia global de 0.028(cada comparación 0.001, $\binom{8}{2}$ comparaciones),son:

1.CPLE,GPLE,GPRA,ALMI.

2.CPRA,CECV.

3.CAIR.

4.CECP.

Respecto a la escala de preferencias a una significancia global de 0.021(cada comparación 0.001, $\binom{7}{2}$ comparaciones),se tiene:

1.GEX,GMU.

2.GLI,IND,DLI.

3.DMU,DMO.

Conjuntando ambos grupos de resultados se tiene que:

1.CPLE,GPLE,GPRA,ALMI se orientan hacia GEX-GMU.

2.CPRA,CECV orientados hacia el GLI-IND-DLI.

3.CAIR orientado hacia GLI-IND-DLI.

4.CECP orientado hacia DMO-DMU.

Los resultados generales se muestran en el cuadro 5.13.

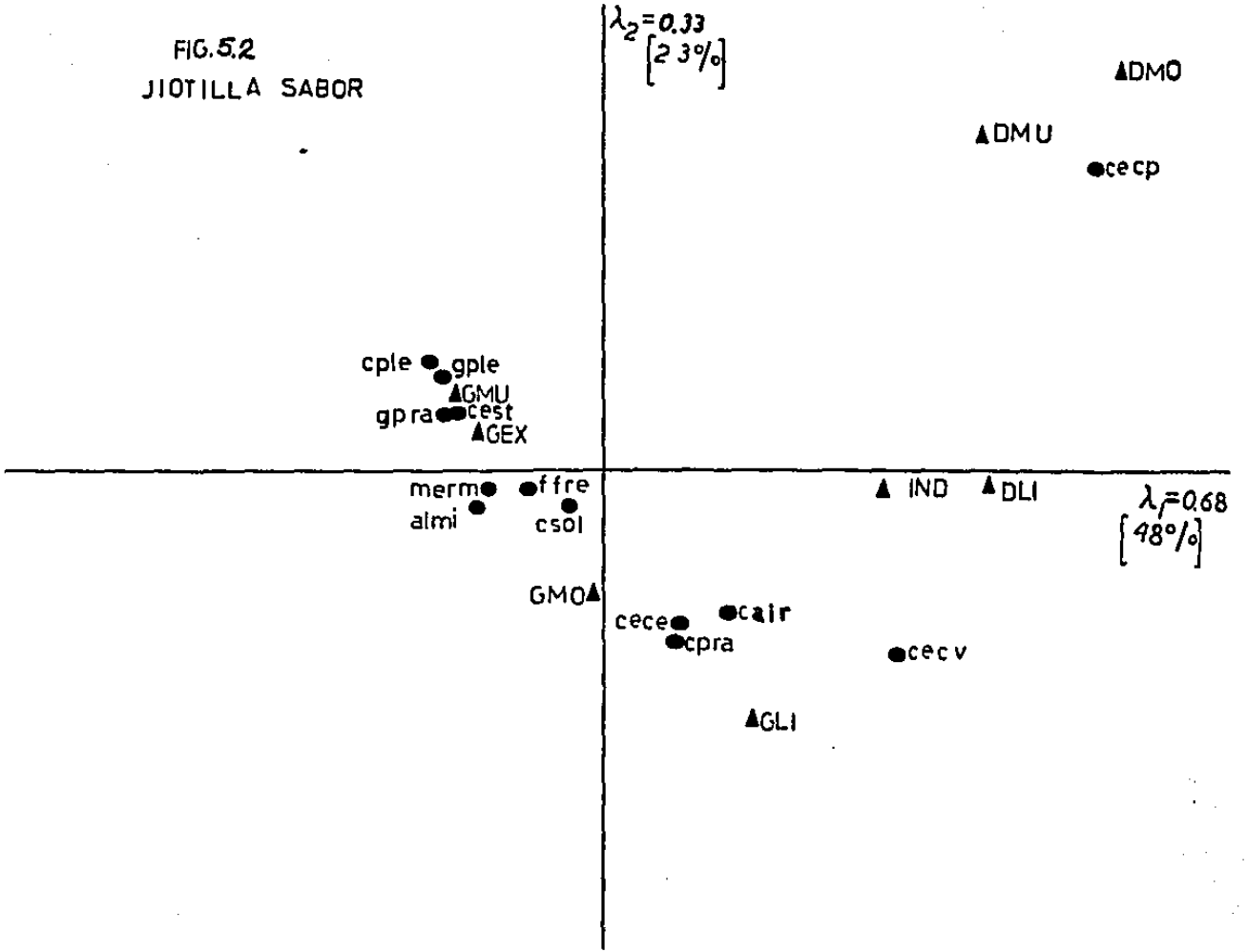
El proceso lento es muy aceptado tanto en confitado como en glaseado,cosa que no sucede con el proceso rápido dado que sólo el glaseado es muy aceptado.Así,el confitado por proceso rápido,el confitado y envasado en cloruro de polivideno y la jiotilla confitada al aire se muestran como tratamientos no aceptados.Por último,el confitado envasado en celofán se ve

fuertemente afectado en su sabor.

Cuadro 5.7.Tabla de contingencias ,SABOR

	FFRC	CFST	CSOL	CAIR	HERB	ALMI	CPLE	CPRA	GPLL	GPRA	CECP	CECE	CECV
	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.
GEX	7.	10.	7.	2.	7.	4.	6.	4.	3.	4.	.	1.	0.
GMB	3.	4.	6.	4.	3.	7.	1.	6.	7.	4.	.	2.	0.
GWB	1.	1.	6.	3.	3.	4.	8.	5.	7.	3.	.	2.	0.
GGU	1.	0.	2.	1.	0.	0.	0.	0.	0.	0.	.	0.	0.
FFC	0.	0.	1.	1.	0.	0.	0.	0.	0.	0.	.	0.	0.
DDX	0.	0.	1.	0.	0.	0.	0.	0.	0.	0.	.	0.	0.
DEV	0.	0.	1.	0.	0.	0.	0.	0.	0.	0.	.	0.	0.

FIG. 5.2
JIOTILLA SABOR



Cuadro 5.8 Estadistica Anderson-Darling para los valores singulares,SABOR.

	$\hat{\Delta}$	$\hat{A}^2$ (B=1000)
1	0.82	1.17
2	0.57	0.74
3	0.42	0.30
4	0.31	0.64
5	0.26	0.40
6	0.20	0.22
7	0.14	0.88
8	0.03	28.74

Cuadro 5.9.Porcentajes marginales y acumulados.SABOR.

k		marginal		acumulado	
	observado	li%	ls%	li%	ls%
1	48.88	36.22	48.98	36.22	48.98
2	23.01	18.42	29.02	57.76	75.28
3	12.22	9.90	18.83	73.51	86.94
4	7.19	6.06	13.09	84.88	94.40
5	5.03	3.13	9.05	92.44	98.57
6	2.16	1.08	5.56	97.08	99.85
7	1.44	0.12	2.65	99.36	100.00
8	0.07	0.00	0.61	100.00	

Cuadro 5.10 Contribuciones absolutas y Correlaciones.
Escala de preferencias,SABOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje1	eje2	eje1	eje2
GEX	0.110	0.024	0.485	0.050
GMU	0.179	0.089	0.623	0.149
GMO	0	0.147	0.001	0.464
GLI	0.054	0.350	0.214	0.666
IND	0.103	0	0.646	0.001
DLI	0.277	0	0.780	0
DMO	0.219	0.280	0.579	0.356
DMU	0.058	0.099	0.418	0.340
DEX	0	0.011	0.001	0.038

Cuadro 5.11 Contribuciones Absolutas y Correlaciones. Tratamientos,SABOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje1	eje2	eje1	eje2
FFRE	0.012	0.001	0.302	0.010
CEST	0.048	0.020	0.412	0.082
CSOL	0.003	0.03	0.018	0.008
CAIR	0.029	0.084	0.251	0.349
MERM	0.030	0.001	0.454	0.008
ALMI	0.035	0.002	0.702	0.023
CPLE	0.063	0.052	0.564	0.226
CPRA	0.008	0.126	0.084	0.610
GPLE	0.059	0.042	0.585	0.201
GPRA	0.048	0.020	0.674	0.136
CECP	0.489	0.401	0.714	0.281
CECE	0.008	0.103	0.062	0.381
CECV	0.167	0.144	0.527	0.219

Cuadro 5.12 Estadística D^2, B=1000, CADOR.

TRATAMIENTOS

DE	FRRE	CEST	ES :	3.3516966
DE	FRRE	CSOL	ES :	0.48231201
DE	FRRE	CAIR	ES :	8.4722063
DE	FRRE	MERM	ES :	1.0832766
DE	FRRE	ALMI	ES :	1.4303520
DE	FRRE	CPLE	ES :	4.4532252
DE	FRRE	CPRA	ES :	5.0242455
DE	FRRE	GPLE	ES :	3.7330061
DE	FRRE	GPRA	ES :	33.3121571
DE	FRRE	CECP	ES :	132.37043
DE	FRRE	CECE	ES :	6.4341651
DE	FRRE	CECV	ES :	47.6406415
DE	CEST	CSOL	ES :	13.6675394
DE	CEST	CAIR	ES :	63.5636012
DE	CEST	MERM	ES :	1.7200635
DE	CEST	ALMI	ES :	8.7624854
DE	CEST	CPLE	ES :	6.1121355
DE	CEST	CPRA	ES :	55.7342156
DE	CEST	GPLE	ES :	5.8355077
DE	CEST	GPRA	ES :	4.2708762
DE	CEST	CECP	ES :	1133.4219
DE	CEST	CECE	ES :	32.6071516
DE	CEST	CECV	ES :	45.779936
DE	CSOL	CAIR	ES :	14.5206310
DE	CSOL	MERM	ES :	1.1483135
DE	CSOL	ALMI	ES :	46.0621681
DE	CSOL	CPLE	ES :	20.2180950
DE	CSOL	CPRA	ES :	6.8157017
DE	CSOL	GPLE	ES :	17.5121991
DE	CSOL	GPRA	ES :	114.274626
DE	CSOL	CECP	ES :	316.43193
DE	CSOL	CECE	ES :	4.1921467
DE	CSOL	CECV	ES :	13.1534633
DE	CAIR	MERM	ES :	17.7598514
DE	CAIR	ALMI	ES :	12.0772711
DE	CAIR	CPLE	ES :	37.1970837
DE	CAIR	CPRA	ES :	0.65150526
DE	CAIR	GPLE	ES :	48.1951466
DE	CAIR	GPRA	ES :	33.6760381
DE	CAIR	CECP	ES :	13.9230911
DE	CAIR	CECE	ES :	0.37323745
DE	CAIR	CECV	ES :	12.3201011
DE	MERM	ALMI	ES :	1.1702847
DE	MERM	CPLE	ES :	3.6319336
DE	MERM	CPRA	ES :	0.4540389
DE	MERM	GPLE	ES :	2.3731607
DE	MERM	GPRA	ES :	1.4733069
DE	MERM	CECP	ES :	181.75670
DE	MERM	CECE	ES :	13.322173
DE	MERM	CECV	ES :	16.356630
DE	ALMI	CPLE	ES :	5.9853202
DE	ALMI	CPRA	ES :	13.9187734
DE	ALMI	GPLE	ES :	6.1207304
DE	ALMI	GPRA	ES :	0.8579097
DE	ALMI	CECP	ES :	432.4835
DE	ALMI	CECE	ES :	17.8355031
DE	ALMI	CECV	ES :	13.47897
DE	CPLE	CPRA	ES :	37.303171
DE	CPLE	GPLE	ES :	3.3342607
DE	CPLE	GPRA	ES :	2.3590724
DE	CPLE	CECP	ES :	1657.2143
DE	CPLE	CECE	ES :	48.501802
DE	CPLE	CECV	ES :	23.773514
DE	CPRA	GPLE	ES :	32.246094
DE	CPRA	GPRA	ES :	54.603050
DE	CPRA	CECP	ES :	69.279137
DE	CPRA	CECE	ES :	7.61961575E-0
DE	CPRA	CECV	ES :	10.499911
DE	GPLE	GPRA	ES :	4.3045596
DE	GPLE	CECP	ES :	1342.7526
DE	GPLE	CECE	ES :	41.854086
DE	GPLE	CECV	ES :	158.64134
DE	GPRA	CECP	ES :	379.73831
DE	GPRA	CECE	ES :	331.660583
DE	GPRA	CECV	ES :	371.10421
DI	CECP	CECE	ES :	38.546979
DE	CECP	CECV	ES :	21.103514
DE	CECE	CECV	ES :	13.659267

ESCALA HEDONICA

DE	GEX	GMU	ES :	2.9611274
DE	GEX	GMO	ES :	31.390269
DE	GEX	GLI	ES :	11.443210
DE	GEX	IND	ES :	19.558521
DE	GEX	DLI	ES :	175.251121
DE	GEX	DMO	ES :	224.165410
DE	GEX	DMU	ES :	24.175450
DE	GEX	DEX	ES :	4.2755461
DE	GMU	GMO	ES :	[illegible]
DE	GMU	GLI	ES :	[illegible]
DE	GMU	IND	ES :	[illegible]
DE	GMU	DLI	ES :	424.51066
DE	GMU	DMO	ES :	830.54635
DE	GMU	DMU	ES :	110.43107
DE	GMU	DEX	ES :	1.6351030
DE	GMO	GLI	ES :	15.587070
DE	GMO	IND	ES :	15.274019
DE	GMO	DLI	ES :	[illegible]
DE	GMO	DMO	ES :	113.52606
DE	GMO	DMU	ES :	15.106643
DE	GMO	DEX	ES :	3.5494243
DE	GLI	IND	ES :	32.7093216
DE	GLI	DLI	ES :	32.032169
DE	GLI	DMO	ES :	35.822671
DE	GLI	DMU	ES :	59.335350
DE	GLI	DEX	ES :	7.0296350
DE	IND	DLI	ES :	6.0619505
DE	IND	DMO	ES :	5.3391833
DE	IND	DMU	ES :	10.5767653
DE	IND	DEX	ES :	6.4210336
DE	DLI	DMO	ES :	[illegible]
DE	DLI	DMU	ES :	35.615205
DE	DLI	DEX	ES :	3.4513537
DE	DMO	DMU	ES :	2.1505578
DE	DMO	DEX	ES :	21.434870
DE	DMU	DEX	ES :	7.0115700

6.Conclusiones y Discusiòn.

En el desarrollo de este trabajo se realizaron primordialmente cuatro tareas;la primera consistiò en la exposiciòn de los detalles del A.C. desde un punto de vista descriptivo.La segunda tarea estuvo conformada por la revisiòn de herramientas inferenciales ya existentes en la literatura.Como tercera parte,en base a las dos primeras se propusieron procedimientos para enfrentar distintos aspectos inferenciales en el A.C.Por ùltimo,la cuarta tarea consistiò de la aplicaciòn tanto los aspectos descriptivos como los inferenciales propuestos,a un ejemplo de anàlisis sensorial proveniente del àrea de tecnologia de alimentos.

La evaluaciòn de estas tareas es como sigue;de la primera se obtuvieron los conocimientos bàsicos necesarios para la correcta aplicaciòn e interpretaciòn de los resultados descriptivos arrojados por el A.C.;respecto a la revisiòn bibliogràfica,se puede decir que en general las herramientas estudiadas tienen inconvenientes.En el caso de los trabajos de Lebart(1976) y Corsten(1976),el hecho de partir de la independencia entre los criterios de clasificaciòn de la tabla,limita su uso pràctico ya que la dependencia es,en tèrminos generales,el caso comùn y de

interès.Referente a las distribuciones asintòticas desarrolladas por O'Neill(1978,1981),como ya fue señalado en su momento,requiere del cumplimiento de condiciones que en la prâctica no es posible verificar.En cuanto a la metodogía de ajuste de modelos propuesta por Goodman(1985),si bien su objetivo primordial no es del A.C.,al menos en el caso de contar con una tabla de contingencias de rango 1,provee los elementos suficientes para realizar inferencias en forma indirecta sobre la dimensionalidad;sin embargo su desarrollo para casos màs generales està pendiente.

Respecto a la tercera tarea,se puede decir que las herramientas propuestas trabajan,a la luz de los resultados,en forma adecuada;en el caso del algoritmo para probar sobre dimensionalidad,la mezcla de una parte de la teoria asintòtica con el mètodo bootstrap,parece ser satisfactorio tanto en su concepto(utilizar teoria que si es utilizable directamente y cubrir sus inconvenientes con remuestreo),como en los resultados que arroja.Por otra parte,en cuanto a la estadistica D^2 para detectar agrupamientos de perfiles,èsta resulta ser uno de los primeros intentos para realizar inferencias en otro aspecto que

no sean los valores singulares;sus resultados son plausibles.

En cuanto a la evaluación de la cuarta tarea,se puede decir que es la primera vez que en México se realizan este tipo de aplicaciones y que de la retroalimentación con los tecnólogos en alimentos,los resultados ofrecidos por el A.C. son fàcilmente entendidos y resultan ser útiles para sus fines y objetivos de anàlisis.El A.C. es sólo una posibilidad de anàlisis entre otras posibles;así las posibilidades de aplicación son amplias.

En cuanto a líneas futuras de investigación o aspectos que quedan pendientes se tiene:

i)Diseñar y realizar un experimento Monte Carlo que considere un número mayor o igual a 1000 de repeticiones bootstrap,para tener elementos de juicio más fuertes sobre el funcionamiento del algoritmo para probar sobre la significancia de los valores singulares.Nótese que este experimento tiene una magnitud de mucha importancia.

ii)Diseñar más herramientas inferenciales en el A.C.;en este caso se pueden considerar para ello a las contribuciones absolutas y las correlaciones ,en términos de la determinación estadística de la orientación de un eje,así como la calidad de la representación de un perfil en particular .

iii)Los modelos propuestos por Goodman(1985) deben ser generalizados en forma explícita para poder ver tanto el comportamiento de estos modelos en el caso de contar con más de un sistema de calificaciones como sus implicaciones para el A.C.

iv)Profundizar en la validez la aplicación del método bootstrap en el A.C.

v)Sobre la posibilidad de que la combinación de resultados ya desarrollados con el enfoque bootstrap pueda ser explotada en otras áreas,en el caso de dimensionalidad se piensa en el tema de colinealidad en análisis de regresión.

Bibliografia.

Anderson,T.W.(1958)."An Introduction to Multivariate Statistical Theory".John Wiley and Sons.

Beran,R. y Srivastava,M.S.(1985)."Bootstrap tests and confidence regions for functions of a covariance matrix.Annals of Statistics,vol.13 no.95-115.

Corsten,L.C.A.(1976)."Matrix Approximation,a key to application of multivariate methods".Proceedings of the 9th International Conference of the Biometric Society,pp 61-77,Raleigh,North Carolina,USA.

Efron,B.(1977)."Bootstrap Methods:Another look at the Jackknife".Annals of Statistics,vol.7 no.1,1-26.

Efron,B. y Tibshirani,R.(1986)."Bootstrap Methods for Standard Errors,Confidence Intervals and other measures of Statistical Accuracy".Statistical Science,vol.1 no.1,54-77.

Green E.P. y Carroll,J.D.(1976)."Mathematical Tools for Applied Multivariate Analysis".Academic Press.

Greenacre,M.J.(1984)."Theory and Applications of Correspondance Analysis".Academic Press.

Goodman,L.A.(1985)."The Analysis of cross-classified data having

ordered and/or unordered categories:Association models,correlation models and asymetry models for contingency tables with or without missing entries". The Annals of Statistics,vol.13,no.1 pp 10-69.

Kirch,A.(1973)."Introduction to Statistics with Fortran".Holt,Rinehart and Winston,Inc.

Lancaster,H.O.(1958)."The Structure of Bivariate Distributions".Annals of Mathematical Statistics,29,719-736.

Lancaster,P(1969)."Theory of Matrices".Academic Press.

Lebart,L.(1976)."The significance of Eigenvalues Issued From Correspondence Analysis of Contingency Tables",en COMPSTAT 1976,Physica Verlag Wien,editado por J.Gordesh y P.Naeve,pp 38-45.

Lindgren,B.W.(1976)."Statistical Theory".Collier MacMillan.

O'Neill,M.E.(1981)."Asymptotic Distributions of the Canonical Correlations from Contingency Tables".Australian Journal of Statistics,20(1),75-82.

O'Neill,M.E.(1978)."Distributional Expansions for Canonical Correlations from Contingency Tables",Journal of the Royal Statistical Society Series B,40 no.3 303-312.

Rubinstein,R.Y.(1981)."Simulation and the MonteCarlo Method".John Wiley and Sons.

Serfling,R.J.(1980)."Approximation Theorems of Mathematical Statistics".John Wiley and Sons.

Tabet,N.(1973).Programa de Análisis de Correspondencias,parte de tesis Doctoral Universidad de Paris VI.

Apèndice A.Descomposiciòn en Valor Singular.

Sea A una matriz p x q real;entonces A puede expresarse como el producto de tres matrices

$$A=UD_a V' \qquad (a.1)$$

donde D_a es una matriz diagonal KxK con nùmeros positivos $a_1,a_2,..,a_K$,K el rango de A.Las matrices U pxK y V qxK tales que U'U=V'V=I,es decir ortogonales por columna.

En vista de esta descomposiciòn,la matriz A puede expresarse como

$$A=\sum_{k=1} a_k \underline{u}_k \underline{v}_k \qquad (a.2)$$

donde $\underline{u}_k,\underline{v}_k$ son las columnas de U y V.Los valores $a_1,a_2,...,a_K$ son llamados los valores singulares,mientras que$(\underline{u}_k)$ y $(\underline{v}_k)$ los vectores propios izquierdos y derechos.

Los vectores izquierdos constituyen una base ortonormal para las columnas de A mientras que los vectores propios derechos conforman una base ortonormal de los renglones de A.

La existencia de esta descomposiciòn se puede demostrar a travès de la descomposiciòn propia(eigenstructure)de A'A y de AA'.

La unicidad de la DVS serà un hecho si todos los valores singulares son distintos entre sí;en el caso de que existan empates las porciones correspondientes en U y V no serán ùnicas.Debe señalarse que D_a siempre serà ùnica.En cuanto al

problema de los empates hay que recordar sin embargo que los subespacios generados por vectores propios asociados a un mismo valor singular son iguales.

Dado (a.2),la matriz A puede aproximarse por matrices de rango menor a K,es decir,

$$A=\sum_{k=1}^{K\#} a_k \underline{u}_k \underline{v}_k' \quad 1<=K\#<=K \qquad (a.3)$$

Esta formas de aproximar es óptima bajo el criterio

$$\min_X \|A-X\| = \min_X \sum_i \sum_j (a_{ij}-x_{ij})^2$$
$$= \min \text{traza}((A-X)(A-X)') \qquad (a.4)$$

Por último,en el caso de que se desee realizar una DVS generalizada,es decir,expresar A como

$$A=NDM' \qquad (a.5)$$

sujeto a $N'\Upsilon N=M'\chi M=I$,con Υ y χ no necesariamente matrices identidad,resúelvase la DVS simple para $\Upsilon^{1/2}A\chi^{1/2}$ y con esto

$$N=\Upsilon^{-1/2}U \qquad M=\chi^{-1/2}V. \qquad (a.6)$$

En este caso la aproximación de A por matrices de rango menor será óptima bajo

$$\min_X \text{traza}(\Upsilon(A-X)\chi(A-X)')$$
$$=\min_X \sum_i w_i(\underline{a}_i-\underline{x}_i)'\chi(\underline{a}_i-\underline{x}_i) \qquad (a.7)$$

si $\Upsilon=\text{diag}(w_i)$.

Apéndice B.

1.Si (Y_n) es tal que $P[|Y_n|<=k]=1$ para toda n,entonces es uniformemente integrable,es decir,

$$\lim_{c\to\infty} \sup_n E(|Y_n| I(|Y_n|>c))=0$$

dem:

Para toda $c>k$ $E(|Y_n|>c)=0$ para toda n;por lo tanto,

$$\sup_n E(|Y_n|>c)=0 \text{ para toada } c>k.$$

De esto se tiene el resultado.

2.(Y_n) y (X_n) son asintóticamente equivalentes y acotadas.(Y_n) es uniformemente integrable y $X_n \xrightarrow{\mathcal{L}} F$;entonces,

$$Y_n \xrightarrow{\mathcal{L}} F \text{ y}$$

$$E[Xn]\longrightarrow \int x dF.$$

dem:

$Y_n - X_n \longrightarrow 0$ c.p.1;por lo tanto (Slutsky) $Y_n \xrightarrow{\mathcal{L}} F$ en distribución.Y_n al ser uniformemente integrable implica que

$$E[Y_n]\longrightarrow \int x dF$$

por teorema A sec 1.4 de Serfling(1980).

$$E[X_n]=E[X_n - Y_n]+E[Y_n]$$

Por el teorema de convergencia dominada(o acotada)

$$E[X_n - Y_n]\longrightarrow 0$$

y por lo tanto

$$E[X_n] \longrightarrow \int x dF.$$