InnOvaciOnes de NegOciOs 4(2): 405 - 422, 2007
© 2007 UANL, Impreso en México (ISSN 1665-9627)

Análisis de correlación canónica (ACC) e investigación científica
(Canonical correlation analysis and scientific research)

Badii, M.H., J. Castillo, K. Cortez, A. Wong & P. Villalpando
UANL, San Nicolás, N.L., México, mhbadii@yahoo.com.mx

Key words: ACC, multivariate statistics, scientific application

Abstract. The concept of Analysis of Canonical Correlation (ACC) is given. The basic conditions, initial questions, and main objectives are provided. The fundamentals of ACC design and the adjustments are touched upon. Field application of ACC is highlighted. The intricacies involving the profiling, validation, and redundant variables of the method are discussed. Finally, the statistical significance and theoretical interpretation of the model are explained.

Palabras claves: ACC, aplicación científica, estadística multivariable

Resumen. Se presenta el concepto de Análisis de Correlación Canónica (ACC). Se discutan los supuestos, fundamentales, preguntas iniciales y objetivos principales de éste método. Se manejan las bases del diseño, las funciones y los ajustes del método. Se presentan las nociones del estudio de campo y la aplicación del método. Se notan asuntos relacionados con el diagnóstico y la validación de ACC cubriendo el concepto de las variables redundantes. Finalmente, se explican la significancia estadística del modelo y la forma de interpretación teórica y la visualización del mismo.

Introducción

Hasta hace pocos años, el análisis de correlación canónica era una técnica estadística relativamente desconocida (Badii et al., 2004, Badii et al., 2006, Badii & castillo, 2007, Badii et al., 2007a, Badii et al., 2007b). La disponibilidad de programas de computadora ha facilitado el aumento de su utilización en problemas de investigación. Es particularmente útil en situaciones donde se tienen múltiples variables dependientes como satisfacción, compra o volumen de ventas. Si las variables predictoras fueran

exclusivamente categóricas, se podría emplear el análisis multivariante de la varianza. Pero, ¿qué ocurre si las varianzas predictoras son métricas? La correlación canónica es la respuesta, ya que permite la valoración de la relación entre variables predictoras métrica y múltiples medidas dependientes. La correlación canónica es considerada como el modelo general en que se basan otras técnicas multivariantes, dado que se pueden emplear tanto datos métricos como no métricos para variables dependientes como independientes. Expresamos la forma general del análisis canónico como:

$$Y_1 + Y_2 + Y_3 + \ldots + Y_n = X_1 + X_2 + X_3 + \ldots + X_n$$

En este capítulo se discuten estos problemas y algunas soluciones a ellos. Se ilustran los problemas y sus soluciones basadas en la experiencia de utilizar correlación canónica en el análisis de estudio de campo de las interacciones automáticas de la tripulación de la aviación comercial. Empezamos con una breve descripción de análisis de correlación canónica (ACC), seguida de la descripción del estudio de campo y de los datos que se analizarán. Se describen cinco problemas específicos que se encontraron durante el análisis, y las soluciones propuestas a cada problema. Se concluye con una afirmación de la utilidad del ACC en el contexto del espectro de métodos analíticos para datos complejos del mundo real.

El concepto

El análisis de correlación canónica es un tipo de análisis estadístico linear de múltiples variables, descrito inicialmente por Hotelling (1935). Actualmente se usa en química, biología, meteorología, demografía, inteligencia artificial, ciencias del conocimiento, ciencias políticas, sociología, psicometría, investigaciones de educación y ciencias de administración para analizar relaciones multidimensionales entre múltiples variables independientes y múltiples variables dependientes.

Aunque el ACC esta documentado en libros de texto, y se encuentra en paquetes computacionales, existen ciertos problemas técnicos y de interpretación que impiden su uso rutinario por los investigadores. Se incluyen problemas de computación (singularidad de las matrices, tiempo de computadora), interpretación (visualización, examen de casos individuales), y

significancia estadística (niveles de significancia e intervalos de confianza para datos multidimensionales no-normales, incluyendo variables discretas).

La aplicación del método

El análisis de correlación canónica es el método más generalizado de la familia de las técnicas estadísticas multivariante. Se relaciona directamente con varios métodos de dependencia. Al igual que en la regresión, el objetivo de la correlación canónica es cuantificar la validez de la relación, en este caso entre los dos conjuntos de variables (dependiente e independiente). Se asemeja al análisis factorial en la creación de compuestos de variables. También se parece al análisis discriminante en su capacidad para determinar las dimensiones independientes para cada conjunto de variables que produce la correlación máxima entre las dimensiones. De esta manera, la correlación canónica identifica la estructura óptima o la dimensionalidad de cada conjunto de variables, que maximiza la relación entre los conjuntos de variables dependientes e independientes.

El análisis de correlación canónica trata con la asociación entre los conjuntos de variables múltiples dependientes e independientes. Por ello, desarrolla varias funciones canónicas que maximizan la correlación entre combinaciones lineales, también conocidas como valores teóricos canónicos, que son conjuntos de variables dependientes e independientes. Cada función canónica se basa realmente en la correlación entre dos valores teóricos canónicos, un valor teórico para las variables dependientes y otro para las variables independientes. Otra característica única de la correlación canónica es que se obtienen los valores teóricos de forma que se maximice su correlación. Además, la correlación canónica no acaba con la obtención de una relación simple entre los conjuntos de variables. En su lugar, se pueden conseguir varias funciones canónicas.

Objetivos del método

1. Determinar si dos conjuntos de variables (medidas realizadas sobre los mismos objetivos) son independientes uno de otro ó, inversamente, determinar la magnitud de las relaciones que pueden existir entre los dos conjuntos.

2. Obtener un conjunto de ponderaciones para cada conjunto de variables criterio y variables predictoras, para que las combinaciones lineales de cada conjunto estén correlacionadas de forma máxima. Las funciones lineales adicionales que maximizan la restante correlación son independientes de los conjuntos anteriores de combinaciones lineales.
3. Explicar la naturaleza de cualquiera de las relaciones existentes entre los conjuntos de variables criterio y variables predictoras, generalmente mide la contribución relativa de cada variable a las funciones canónicas.

Estudio de campo

El estudio involucra observaciones en la cabina de pilotos de las interacciones de la tripulación con el sistema de control automático del avión Boeing 757/767 durante vuelos de contratados por un carguero de EUA. Cada dato registrado caracterizó un cambio en el modo de selección, al mismo tiempo que un número de variables que describían las condiciones bajo las cuales el cambio ocurrió. Los datos iniciales usados en el ACC consistieron en más de 1500 registros, cada uno caracterizado por 75 variables. Aproximadamente, la mitad de las variables tenían que ver con la respuesta de la tripulación, esto es, su elección de volar en el modo de piloto automático. Se puede encontrar una descripción completa del estudio en Degani (1996).

Preguntas iniciales

En general, estamos interesados en caracterizar las relaciones entre las situaciones y los patrones de respuestas, esto es, entre el estado del medio ambiente operativo y la acción humana (tipo de elección). El valor de usar el ACC en este caso esta derivado de su especial adaptabilidad para encontrar patrones en grupos de datos grandes. Se tienen múltiple variables independientes que caracterizan situaciones operacionales (permisos otorgados por la torre de control, comandar el vuelo por el capitán contra el primer oficial, distancia del aeropuerto, altitud, facilidades de la torre de control, permisos concedidos, aeropuerto de salida y destino), así como múltiples variables dependientes consistentes principalmente es variables categóricas usadas para describir la elección de la tripulación de las modalidades de piloto automático. Adicionándose a patrones caracterizados

de relaciones de situaciones- respuesta, queríamos poder reconocer casos raros (atípicos), para enfocar nuestro análisis en esos casos individuales que pudieran iluminar el comportamiento inusual de la tripulación o errores de la tripulación. Finalmente, usando el ACC para este análisis inicial de reducción de datos, usamos tanto los patrones de comportamiento típicos y los casos atípicos (outliers), como puntos de partida para desarrollar modelos dinámicos de interacciones automáticas de la tripulación (Degani & Kirlik, 1995).

Diseño del método

Frecuentemente con la correlación canónica se deben de resolver cuestiones acerca del impacto del tamaño de la muestra (tanto pequeño como grande) y la necesidad de una cantidad suficiente de observaciones por variable. Los investigadores pueden tener la tentación de incluir muchas variables tanto en el conjunto de variables independientes como en el de dependientes, ignorando sus implicaciones en el tamaño muestral. Los tamaños muestrales que son muy pequeños, no representarán correlaciones adecuadamente y como consecuencia esconderá cualquier relación significante que pueda existir. Los tamaños muestrales muy grandes, tendrán una tendencia a indicar una significación estadística en todas las instancias, incluso donde la significación práctica no está indicada. Se sugiere al investigador a mantener por lo menos diez observaciones por variable para evitar el "sobreajuste" de los datos.

La clasificación de las variables tanto dependientes o independientes tiene poca importancia en la estimación estadística de las funciones canónicas, ya que el análisis de correlación canónica pondera ambos valores teóricos para maximizar la correlación y no establece ningún énfasis particular en alguno de los valores teóricos. Aunque dado que la técnica produce valores teóricos que maximizan la correlación entre ellos, un valor teórico en cualquier conjunto relaciona a todas las otras variables en ambos conjuntos. Con ello se permite la incorporación o la supresión de una sola variable que afecte a la solución total, particularmente el otro valor teórico. La composición de cada valor teórico, ya sea dependiente o independiente, llega a ser muy importante. El investigador, antes de aplicar el análisis de correlación canónica, debe relacionar conceptualmente los dos conjuntos de variables. De esta forma, la especificación de los valores teóricos

dependientes frente a los independientes es esencial para establecer una base conceptual fuerte para las variables.
El ACC es una herramienta potencialmente valiosa en las investigaciones de factores humanos que tienen 1) una clara distinción entre variables independientes y dependientes, 2) múltiples variables dependientes, y 3) el potencial para relaciones multidimensionales entre estos dos grupos de variables. Por ejemplo, estas condiciones generalmente aparecen en estudios de campo de toma de decisiones y acciones, pruebas de campo de productos o sistemas de utilidad, estudios simulados de actuación profesional de parte o de toda una misión, y datos de actuación en línea tales como registro de datos de vuelo.

Las ecuaciones generales para realizar una correlación canónica son relativamente simples. Primero, se hace una matriz de correlación (R). Esto se compone de: correlaciones entre VDs (R_{yy}), correlaciones entre VIs (R_{xx}), y correlaciones entre VDs y VIs (R_{xy}).

$$R = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$$

Para el análisis canónico se resuelve la ecuación anterior para eigenvalores y eigenvectores de la matriz R. Los eigenvalores consolidan la varianza de la matriz, redistribuyendo la varianza original en unas pocas variantes compuestas. Los eigenvectores, transformados a coeficientes, se usan para combinar las variables originales con las compuestas. Los eigenvalores están relacionados en la correlación canónica por la siguiente ecuación:

$$\lambda_i = r_{ci}^2$$

Esto es, cada eigenvalor es igual al cuadrado de la correlación canónica para cada par de variantes.

La prueba de significancia utiliza la siguiente fórmula, y sigue una distribución de chi-cuadrada:

$$\chi^2 = -\left[N - 1 - \left(\frac{k_x + k_y + 1}{2}\right)\right] \ln \Lambda_m$$

$$\Lambda_m = \prod_{i=1}^{m} (1 - \lambda_i)$$

con
N= número de casos
k_x= número de variables en el grupo de VI
$\mathbf{k_y}$= número de variables en el grupo de VD
DF = $(k_x)(k_y)$
m = número de correlaciones canónicas

Para probar la significancia de una correlación canónica, se utiliza la prueba de Bartlett de la lambda de Wilks. La lambda varía de 0 a 1 y muestra la varianza del error, la varianza no contabilizada por las variables independientes. Entonces se interpreta de forma opuesta al cuadrado de la correlación múltiple R^2. Obtener un 1.0 significa que las variables independientes no están contabilizando nada de la varianza en la variable dependiente, y un 0 significa que las variables independientes están contabilizando toda la varianza. Para una lambda menor, una varianza mayor. $1-\Lambda$ será equivalente a R^2. La prueba de chi cuadrada se usa para probar la significancia de lambda.

La redundancia de las variables se menciona frecuentemente cuando se tienen resultados de correlación canónica. A mayor redundancia, o correlación entre un grupo de variables, mejor será la habilidad para predecir de un grupo a otro.

Dos grupos de coeficientes canónicos son necesarios para cada correlación canónica – uno para combinar las VDs y otro para combinar las VIs.
Para las VDs la ecuación es:

$$B_y = \left(R_{yy}^{-1/2}\right)' \hat{B}_y$$

Correlación canónica

Para las VIs:

$$B_x = R_{xx}^{-1/2} R_{xy} B_y^*$$

donde
B_y = matriz normalizada de eigenvectores
R = matriz de correlaciones

Las dos matrices de coeficientes canónicos se utilizan para estimar el puntaje en las variantes canónicas:

$X = Z_x B_x$
$Y = Z_y B_y$

Los puntajes en las variantes canónicas (X, Y) son el producto de los puntajes de las variantes originales y los coeficientes canónicos usados para ponderarlas. La suma de los puntajes canónicos para cada variante es igual a cero. El llenado de las matrices (A) se realiza por la multiplicación de la matriz de las correlaciones entre variables con la matriz de coeficientes canónicos. Estas matrices A son usadas para interpretar las variantes canónicas.

$A_x = R_{xx} B_x$
$A_y = R_{yy} B_y$

¿Qué tanta varianza explica cada variante canónica? La proporción de la varianza para VIs es:

$$pv_{xc} = \sum_{i=1}^{k_x} \frac{a_{ixc}^2}{k_x} \qquad pv_{yc} = \sum_{i=1}^{k_y} \frac{a_{iyc}^2}{k_y}$$

a = correlaciones llenas
k = número de variables en un grupo

Supuestos básicos

La generalidad del análisis de correlación canónica también se extiende a sus supuestos estadísticos básicos. El supuesto de linealidad afecta a dos aspectos de los resultados de la correlación canónica, primero, el coeficiente de correlación entre cualquiera de dos variables esta basado en una relación lineal. Si la relación no es lineal, entonces se debe transformar una ó ambas variables si esto fuera posible. Segundo, la correlación canónica es la relación lineal entre los valores teóricos.

Si los valores teóricos se relacionan de una manera no lineal, la relación no será reflejada por la correlación canónica. De esta manera, aunque el análisis de correlación canónica es el método multivariante más extendido, esta restringido ha la identificación de relaciones lineales.

El análisis de correlación canónica puede emplear cualquier variable métrica sin que cumpla el estricto supuesto de normalidad. La normalidad es deseable porque estandariza una distribución que nos permite una mayor correlación entre las variables. Pero en un estricto sentido, el análisis de correlación canónica puede utilizar incluso variables no normales si la forma de las distribuciones altamente simétrica no disminuye la correlación con otras variables ficticias, también sin embargo, se requiere normalidad multivariante para los contrastes de significación de inferencia estadística de cada función canónica. Dado que los contrastes de normalidad multivariantes no se están disponibles fácilmente, la línea a seguir que prevalece es asegurar que cada variable presenta una normalidad univariante. De este modo, aunque estrictamente no se requiere normalidad, es altamente recomendable que se compruebe la normalidad de todas las variables y que se transformen si fuese necesario.

Funciones y ajustes

Esta sección corresponde a la obtención de una ó más funciones canónicas. Cada función está formada por un par de valores teóricos, uno que representa las variables independientes y el otro que representa las variables independientes. El número de variables es igual al número de variables que hay en el conjunto de datos menor, ya sea dependiente o independiente. Por ejemplo, cuando en un problema de investigación incluye

cinco variables independientes y tres variables dependientes, el máximo número de funciones canónicas que puede obtener es tres.

La obtención de sucesivos valores teóricos canónicos es similar al empleado en el análisis factorial sin rotación. El primer factor extraído explica la máxima cantidad de varianza en el conjunto de variables. Después se calcula el segundo factor para que explique lo más posible la varianza no explicada por el primer factor, y así sucesivamente, hasta que todos los factores hayan sido considerados. Por tanto, los posteriores factores se calculan a partir de los residuos o de la varianza restante de los primeros factores. El análisis de correlación canónica sigue un procedimiento similar, pero centrándose en la explicación de la cantidad máxima de relación entre los dos conjuntos de variables, en lugar de en un solo conjunto. El resultado es que el primer par de valores teóricos se calcula con el fin de obtener la mayor intercorrelación posible entre dos conjuntos de variables. El segundo par de valores teóricos canónicos es obtenido después para que represente la máxima relación entre los dos conjuntos de variables que no han sido explicados por el primer par de valores teóricos. En resumen, los sucesivos pares de valores teóricos canónicos están basados en la varianza residual y sus respectivas correlaciones canónicas disminuyen a medida que se calculan funciones adicionales, es decir el primer par de valores teóricos canónicos refleja la mayor intercorrelación, el siguiente par la segunda intercorrelación, y así, sucesivamente.

Al igual que cualquier investigación que utiliza otras técnicas estadísticas, la práctica más común es analizar las funciones cuyos coeficientes de correlación canónica son estadísticamente significativos para un nivel, normalmente 0.05 o mayor. Si se consideran no significativas otras funciones independientes, estas relaciones entre las variable no se interpretan. La interpretación de los valores teóricos canónicos en una función significativa está basada en la premisa de que las variables de cada conjunto, que contribuyen fuertemente a las varianzas compartidas por estas funciones, son consideradas como relacionadas unas con otras. El uso de un único criterio como el nivel de significación es demasiado superficial. En lugar de esto, se recomiendan que sean empleados tres criterios que son: **1.** El nivel de significación estadística de las funciones. **2.** La magnitud de la correlación canónica. **3.** La medida de la redundancia por el porcentaje de varianza explicado por los dos conjuntos de datos.

El nivel de significación de una correlación canónica, que generalmente se considera como el mínimo aceptable para la interpretación, es el nivel 0.05, que se ha llegado a convertir junto con el de 0.01 como los más habitualmente aceptados para considerar que un coeficiente de correlación es estadísticamente significativo. Para superar el sesgo y la incertidumbre propios del empleo de raíces canónicas (correlaciones al cuadrado) como una medida de la varianza compartida se ha propuesto un índice de redundancia. Este es el equivalente de calcular el coeficiente de correlación múltiple al cuadrado entre el conjunto predictor total entre cada una de las variables criterio, y después promediar estos coeficientes al cuadrado para obtener un R^2 medio. Proporciona una medida que resume de la capacidad del conjunto de las variables predictoras, para explicar la variación de la variable criterio como tal, la medida de redundancia es perfectamente análoga al estadístico R^2 de la regresión múltiple, y su valor como índice es similar.

Interpretación del valor teórico canónico

Si la relación canónica resulta estadísticamente significativa y las magnitudes de la raíz canónica y del índice de redundancia son aceptables, el investigador aún necesita realizar interpretaciones de los resultados. La realización de estas interpretaciones comprende el examen de las funciones canónicas para determinar la importancia relativa de cada uno de las variables originales en las relaciones canónicas. Se han propuesto tres métodos. **1.** Ponderaciones canónicas. **2.** Cargas canónicas. **3.** Cargas cruzadas canónicas.

El enfoque tradicional para interpretar las funciones canónicas comprende el examen del signo y la magnitud de la ponderación canónica asociada a cada variable en su valor teórico canónico. Las variables con ponderaciones relativamente mayores contribuyen más al valor teórico y viceversa. Igualmente, las variables cuyas ponderaciones tienen signos contrarios presentar una relación directa, sin embargo, la interpretación de la importancia o contribución relativa de una variable por su ponderación canónica esta sujeta a las mismas críticas asociadas con la interpretación de los coeficientes beta de las técnicas de regresión.

El empleo de las cargas canónicas ha sustituido al uso de ponderaciones canónicas como base de interpretación, debido a las

deficiencias inherentes a estas últimas. Las cargas canónicas, también denominadas correlaciones de estructura canónica, miden la correlación lineal simple entre una variable original observada del conjunto dependiente o independiente y el valor teórico canónico del conjunto. Las cargas canónicas reflejan la varianza que la variable observada compare con el valor teórico canónico, y puede ser interpretada cono una carga factorial para valorar la contribución relativa de cada variable a cada función canónica. Se considera cada función canónica independiente de forma separada, y se calcula la correlación dentro del conjunto entre variables y valores teóricos. Cuanto mayor sea el coeficiente, mayor es la importancia que tiene para calcular el valor teórico canónico. Los criterios para determinar la significación de las correlaciones de estructura canónica también son los mismos que con las cargas factoriales. Las cargas canónicas se consideran más válidas que las ponderaciones canónicas.

Se ha sugerido el cálculo de las cargas cruzadas canónicas como una alternativa a las cargas convencionales. Este consiste en correlacionar cada una de las variables dependientes originales observadas directamente con el valor teórico canónico independiente, y viceversa. Es parecido a la regresión múltiple pero difiere en que cada variable independiente, por ejemplo, esta correlacionado con el valor teórico dependiente en lugar de con una única variable dependiente. De esta manera las cargas cruzadas proporcionan una medida más directa de las relaciones entre las variables dependientes e independientes eliminando un paso intermedio incluido en las cargas convencionales.

Validación y diagnostico

Al igual que cualquier otra técnica mutivariante, el análisis de correlación canónica debe estar sujeto a métodos de validación que aseguren que los resultados no son solamente específicos de los datos de la muestra y que pueden ser generalizados a la población. El procedimiento más directo es crear dos submuestras de los datos y llevar a cabo el análisis en cada submuestra de forma separada. Después, los resultados se pueden comparar para buscar la igualdad de las funciones canónicas, las cargas de los valores teóricos, y demás aspectos. Si se encuentran importantes diferencias, el investigador debe considerar él realiza una investigación adicional para asegurar que los resultados finales son representativos de los

valores poblacionales y no solamente de una única muestra. Aunque existen pocos procedimientos de diagnostico desarrollados específicamente para el análisis de correlación canónica, el investigador debe observar los resultaos teniendo en cuenta las limitaciones de la técnica. Entre las limitaciones que pueden tener un mayor impacto sobre los resultados y su interpretación están los siguientes:

La correlación canónica refleja la varianza compartida por las combinaciones lineales de los conjuntos de variables y no la varianza extraída de las variables. Las ponderaciones canónicas obtenidas para calcular las funciones canónicas están sujetas a una gran inestabilidad.

Las ponderaciones canónicas son obtenidas para maximizar la correlación entre las combinaciones lineales, no para la varianza extraída. La interpretación de los valores teóricos canónicos puede ser difícil ya que estos se calculan para maximizar la relación, y no existen ayudas para la interpretación como puede ser la rotación de los valores teóricos, como se vio en el análisis factorial. Es difícil identificar una relación con significado entre los subconjuntos de variables dependientes e independientes dado que aún no se han desarrollado estadísticos precisos para interpretar el análisis canónico, y debemos utilizar medidas inadecuadas como las cargas cruzadas. Sin embargo, estas limitaciones no deben desanimar a la hora de utilizar la correlación canónica. Al contrario, se menciona para aumentar la efectividad de la correlación canónica como una herramienta de investigación**.**

Variables redundantes

El primer problema que encontramos fue que la matriz de correlación para las variables originales era singular. Esto es un problema común cuando el número de variables es grande, simplemente quiere decir que algunas variables son redundantes. Es difícil, sin embargo, determinar por la sola inspección cuales variables son redundantes. Usamos varios métodos para atacar este problema, pero la mejor solución de todas fue usar un análisis de celdas anterior al ACC.

El análisis de celdas (Jardine & Sibson, 1971) es una forma de analizar una matriz de correlación que es complementaria al ACC. Donde el ACC enfatiza los patrones globales, el análisis de celdas trabaja “de abajo hacia arriba” uniendo primero los grupos más inter-correlacionados de

variables, y después yendo a otras celdas más grandes que estén menos inter-relacionadas. Como resultado, las primeras celdas identifican las fuentes más probables de redundancia. Como un beneficio colateral, las celdas grandes nos permiten revisar la fuerza de los resultados del ACC (ya que el análisis de celdas y el ACC son bastante diferentes matemáticamente hablando).

Significancia estadística

Usamos el ACC como parte de un espectro de herramientas analíticas. Por lo tanto, sirve para dirigir la atención a patrones y a las desviaciones de esos patrones. No es nuestra intención poner peso de más en la "significancia estadística" de los resultados del ACC. Sin embargo, estamos interesados en estimar la estabilidad de las correlaciones canónicas computadas, y esto requiere el cálculo de errores estándar. La teoría del muestreo para el ACC es compleja y asume normalidad multivariable, un supuesto lejos de la realidad de nuestros datos: a mayor parte de nuestras variables dependientes son discretas. Por lo tanto, nos volvimos a un método bien conocido de remuestreo, el método de jackknife para estimar errores estándar e intervalos de confianza (Efron & Tibshirani, 1993). Encontramos que el jackknife es conceptualmente recto, aunque computacionalmente demandante (ver abajo). El problema relativo de estimar el nivel de significancia de nuestras correlaciones canónicas pidió una solución relativa, el uso de pruebas randomizadas (Edgington, 1987). Los métodos de remuestreo, tales como el jackknife y la prueba de randomización, esta siendo más familiares y aceptadas; su descripción detallada esta fuera de este escrito (Simon & Bruce, 1991). Mencionaremos, sin embargo, algunos de los problemas computacionales derivados de nuestro uso de los métodos de jackknife y randomización. Para análisis complejos tales como el ACC estos métodos de remuestreo requieren de computadoras veloces y técnicas especiales, ya que necesitan la solución interactiva de cientos de factorizaciones matriciales. Nuestros programas fueron unidos de rutinas (Koelcker, 1994) e integrados usando Lenguaje Icon de Programación (Griswold & Griswold, 1996), un lenguaje de interpretación de alto nivel. Usamos también un código que necesita mucho tiempo. El análisis jackknife de 897 casos y 50 variables corrió en una laptop Pentium en un poco más de tres horas.

Interpretación y visualización

Nuestro problema final es sobre la interpretación de los resultados. Tratamos de encontrar métodos gráficos que nos ayudaran a entender y explicar los patrones multidimensionales encontrados por el ACC. Estos patrones son importantes porque ayudan al analista a definir, en una forma de vista de datos, las condiciones ambientales más importantes y sus correspondientes efectos en las acciones humanas. Una de las sugerencias más útiles fue encontrada por Cliff (1987), que sugirió interpretar la estructura de las correlaciones más que las ponderaciones. Las correlaciones estructurales son las correlaciones de la variante canónica X con cada una de las variables independientes originales, y la de la variante canónica Y con cada una de las variables dependientes originales. De esta forma, algunas veces misteriosas variantes canónicas pueden ser interpretadas en términos de su correlación con las variables originales. Después usamos dos métodos gráficos para pintar el patrón de la estructura de las correlaciones y enfatizar las desviaciones del patrón y los atípicos (outliers, ver Figura 1).

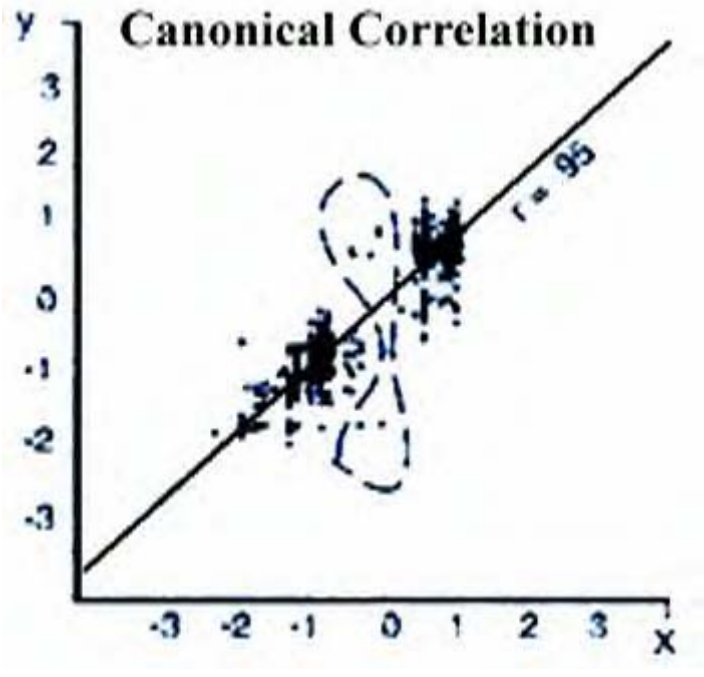

Figura 1. Descripción del patrón de la estructura de ACC.

Correlación canónica

Conclusiones

El ACC es el método de selección cuando se tienen variables multivariantes dependientes en un contexto de otra forma adecuado para regresión múltiple. El ACC se utiliza mejor como parte de un grupo de métodos analíticos. Todo el paquete debe incluir análisis de celdas, estado de transición (Markov) y modelos dinámicos, métodos gráficos, y otros métodos estadísticos (Degani, 1996; Degani, Shafto, & Kirlik,; Degani & Kirlik, 1995). Los métodos de premuestreo pueden ser utilizados para computar intervalos de confianza y niveles de significación de correlaciones canónicas. Las correlaciones estructurales son útiles para interpretar los resultados del ACC, y las técnicas de gráficas simples pueden ser utilizadas para entender y explicar los resultados. El ACC es capaz de describir en una forma objetiva (con datos) algunos de los patrones complejos en los datos de los estudios de campo, simulaciones, y experimentos controlados en la interacción del hombre-máquina. Dirige la atención del analista a los patrones principales de los datos, así como también a las desviaciones importantes de dichos patrones. La correlación canónica se utiliza para analizar la correlación entre dos grupos de variables cuando hay un grupo de VIs (variables independientes) y otro grupo de VDs (variables dependientes). Es un procedimiento más bien descriptivo que analítico para probar hipótesis, y existen varias formas en las que la información puede ser combinada en este procedimiento. El término “canónica” indica que la técnica se extrae de una matriz. Se extraerán tantas funciones como el menor número de variables, por ejemplo, si hay 5 variables independientes y 3 variables dependientes, se tendrá un total de 3 funciones. Cada función describe una cantidad menor de variación, por ejemplo, la primera función describirá la mayor parte de ella, después se computará otra función en la varianza residual, y así sucesivamente.

Generalmente, las funciones secundarias son de uso y valor cuestionable. Se pueden obtener y el programa lo hace, pero eso no significa que sean de utilidad o que tengan significado. Cada una tiene un coeficiente de determinación asociado a ella, y en general éste caerá rápidamente después del primero.

Son varias las preguntas que pueden ser contestadas con la Correlación Canónica. **1.** ¿Cuántos pares de variables confiables hay en el grupo de datos? **2.** ¿Qué tan fuerte es la correlación entre las variables en un

par? **3.** ¿Cómo deben ser interpretadas las dimensiones que relacionan a las variables? La Correlación Canónica esta sujeta a varias limitantes. **1.** Es matemáticamente elegante pero difícil de interpretar porque las respuestas no son únicas. **2.** La relación entre variables debe ser linear; si la información esta correlacionada de manera no-linear, entonces otros análisis serán más apropiados. **3.** Pequeños cambios en donde las variables están incluidas en el análisis pueden causar grandes diferencias en los resultados, y esto puede confundir la interpretación posterior.

Normalmente, no es necesario realizar la Correlación Canónica, pero esto aumenta el poder estadístico de una prueba. Como se mencionó antes, es esencial la relación linear entre las variables. Además, la homogeneidad (varianzas muy semejantes) aumenta la potencia de la prueba. La Correlación Canónica es muy sensible a datos faltantes en la matriz analizada y en los datos atípicos. Debe probarse que toda la información está presente y debe resolverse ese problema antes de conducir una Correlación Canónica.

Referencia

Alpert, M.I, y R.A. Peterson, 1972. On the interpretation of Canonical Analysis. Journal of marketing Research, 187.

Alpert, M.I, R.A. Peterson y W.S. Marti, 1975. Testing the significance of canonical correlations. American Marketing Association 37: 117-119.

Ashley D.A., 1996. Canonical Correlation Procedure for Spreadsheets, 27th Annual Meeting of Decision Sciences Institute USA.

Badii, M.H., A.R. Pazhakh, J.L. Abreu & R. Foroughbakhch. 2004. Fundamentos del método científico. InnOvaciOnes de NegOciOs 1(1): 89–107.

Badii, M.H., J. Castillo & A. Wong. 2006. Diseños de distribución libre. InnOvaciOnes de NegOciOs, 3(1): 141-174.

Badii, M.H. & J. Castillo (eds.). 2007. Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Badii, M.H., R. Ramírez & J. Castillo. 2007a. Papel de estadística en la investigación científica. InnOvaciOnes de NegOciOs, 4(1): 107-145.

Badii, M.H., J. Castillo, J. Rositas & G. Alarcón. 2007b. Uso de un método de pronóstico en investigación. Pp. 137-155. In: M.H. Badii & J. Castillo (eds.). Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Dillon, W.R, y M. Goldstein, 1984. Multivariate analysis: Methods and applications. New York: Wiley.

Lambert, Z., y R. Durand, 1975. Some precautions in using canonical analysis. Journal of Marketing Research 12:468-475.

Correlación canónica

Stewwart, D., y W. Love 1968. A general canonical correlation index. Psychological Bulletin 70: 160-163.
Hair J, Anderson R, Tatham R, Black W. Análisis Multivariante. Prentice Hall, 2000.