ANÁLISIS ESTADÍSTICO MULTIVARIADO

Análisis Multivariado

1. Análisis de Correlaciones Canónicas
   - Propiedades
   - Estimación
   - Test de Hipótesis
   - Relación con otras técnicas
2. Detección de observaciones atípicas
   - Introducción
   - Métodos basados en proyección de los datos
   - Métodos basados en distancias
   - Referencias

Análisis Multivariado

1. Análisis de Correlaciones Canónicas
   - Propiedades
   - Estimación
   - Test de Hipótesis
   - Relación con otras técnicas

2. Detección de observaciones atípicas
   - Introducción
   - Métodos basados en proyección de los datos
   - Métodos basados en distancias
   - Referencias

Introducción

- El análisis de correlaciones canónicas es una técnica estadística que investiga la relación entre dos conjuntos de variables aleatorias.
- Sea p el número de variables en el primer grupo y q el número de variables en el segundo grupo:

$$x = (x_1, x_2, \ldots, x_p)$$
$$y = (y_1, y_2, \ldots, y_q)$$

- Con μ_x identificamos al vector de medias de x mientras que μ_y identifica al vector de medias de y.
- Llamaremos con Σ_{11} a la matriz $(p \times p)$ de variancias y covariancias del vector x, Σ_{22} a la matriz $(q \times q)$ de variancias y covariancias del vector y. Además, $\Sigma_{12} = \Sigma'_{21}$ es la matriz de covariancias entre x e y.

Introducción

- Analizaremos el caso simétrico donde no se da preferencia a ninguno de los dos conjuntos de variables para explicar el otro, se desea investigar la relación global entre ambos conjuntos de variables.
- Se plantea la búsqueda de dos variables de resumen, una para cada grupo de variables, que tengan correlación máxima:

$$x^* = \alpha' x = \sum_{i=1}^{p} \alpha_i x_i, \qquad y^* = \beta' y = \sum_{i=1}^{q} \beta_i y_i$$

- El problema es hallar α y β tal que la correlación entre x^* e y^* sea máxima.

Correlaciones canónicas

- Suponemos que $x \sim N_p(0, \Sigma_{11})$, $y \sim N_q(0, \Sigma_{22})$, de manera que las variables estan medidas en desviaciones a la media.
- El coeficiente de correlación entre dos combinaciones lineales de las variables de cada grupo x^* e y^* es:

$$\rho(x^*, y^*) = \frac{E[\alpha' x y' \beta]}{\sqrt{E[\alpha' x x' \alpha] E[\beta' y y' \beta]}}$$

que puede escribirse como:

$$\rho(x^*, y^*) = \frac{\alpha' \Sigma_{12} \beta}{\sqrt{[\alpha' \Sigma_{11} \alpha][\beta' \Sigma_{22} \beta]}}$$

Correlaciones canónicas

- Como nos interesa la magnitud de la relación y no el signo vamos a maximizar el cuadrado de la correlación entre x^* e y^* con respecto a α y β. Impondremos la condición de variancias unitarias:

$$Var(x^*) = \alpha' \Sigma_{11} \alpha = 1, \qquad Var(y^*) = \beta' \Sigma_{22} \beta = 1$$

y la función objetivo a maximizar es:

$$\rho^2 = \frac{(\alpha' \Sigma_{12} \beta)^2}{[\alpha' \Sigma_{11} \alpha][\beta' \Sigma_{22} \beta]}$$

sujeta a las restricciones de variancias unitarias.

Correlaciones canónicas

- Si utilizamos multiplicadores de Lagrange, la función a maximizar es:

$$M = (\alpha' \Sigma_{12} \beta)^2 - \lambda_1(\alpha' \Sigma_{11} \alpha - 1) - \lambda_2(\beta' \Sigma_{22} \beta - 1)$$

Derivando respecto a los vectores de coeficientes:

$$\begin{aligned} \frac{\partial M}{\partial \alpha} &= 2\Sigma_{12}\beta - 2\lambda_1 \Sigma_{11}\alpha = 0 \\ \frac{\partial M}{\partial \beta} &= 2\Sigma_{21}\alpha - 2\lambda_2 \Sigma_{22}\beta = 0 \end{aligned}$$

Se obtiene:

$$\begin{aligned} \Sigma_{12}\beta &= \lambda_1 \Sigma_{11}\alpha \\ \Sigma_{21}\alpha &= \lambda_2 \Sigma_{22}\beta \end{aligned}$$

Correlaciones canónicas

- Premultiplicando por α' la primera ecuación y por β' la segunda:

$$\begin{aligned} \alpha'\Sigma_{12}\beta &= \lambda_1\alpha'\Sigma_{11}\alpha = \lambda_1 \\ \beta'\Sigma_{21}\alpha &= \lambda_2\beta'\Sigma_{22}\beta = \lambda_2 \end{aligned}$$

  por lo que $\lambda_1 = \lambda_2$ y concluimos:

$$\begin{aligned} \Sigma_{12}\beta &= \lambda\Sigma_{11}\alpha \\ \Sigma_{21}\alpha &= \lambda\Sigma_{22}\beta \end{aligned}$$

- Despejando β de la segunda ecuación,

$$\beta = \lambda^{-1}\Sigma_{22}^{-1}\Sigma_{21}\alpha$$

Correlaciones canónicas

- Sustituyendo β en la primera:

$$\begin{aligned} \Sigma_{12}\beta &= \lambda\Sigma_{11}\alpha \\ \lambda^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha &= \lambda\Sigma_{11}\alpha \\ (\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\alpha &= \lambda^2\alpha \end{aligned}$$

- Por lo tanto α es el autovector asociado al autovalor λ^2 de la matriz $A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
- De manera análoga, β se obtiene como el autovalor ligado al autovector λ^2 de la matriz $B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Correlaciones canónicas

- Observemos que $\lambda_1^2 = \lambda_2^2 = \rho^2$ es el cuadrado de la correlación entre las variables canónicas x^* e y^*, por lo que tendremos que tomar el autovector asociado al mayor autovalor de las matrices A o B.
- De las expresiones anteriores resulta que:

$$\begin{aligned} \alpha &= \lambda_1^{-1}\Sigma_{11}^{-1}\Sigma_{12}\beta \\ \beta &= \lambda_2^{-1}\Sigma_{22}^{-1}\Sigma_{21}\alpha \end{aligned}$$

Por lo que solo necesitamos obtener autovectores de una sola de las matrices A o B, y de manera análoga, al conocer α podemos obtener β.

Correlaciones canónicas

- Por otra parte podemos observar que

$$\lambda^2 = \alpha' \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \alpha$$

  lo que indica que el coeficiente de correlación canónica λ^2 es el cuadrado del coeficiente de correlación múltiple entre la variable x^* y el vector de variables y.
- Las covariancias entre la variable x^* y el vector de variables y vienen dadas por el vector $\Sigma_{21}\alpha$.
- Las correlaciones entre la variable x^* y el vector de variables y vienen dadas por $D_{22}^{-1/2}\Sigma_{21}\alpha$, donde D_{22} es una matriz diagonal que contiene las variancias de las variables y.

Correlaciones canónicas

- Es posible que una vez encontrada la primera relación entre estas dos variables indicadoras no exista más relación entre ambos conjuntos de variables y entonces decimos que toda la relación entre ambos conjuntos se puede resumir en una dimensión.
- Para comprobar si esto es asi podemos buscar una segunda variable indicadora (o canónica) del primer conjunto de variables x que este no correlacionada con $x^* = \alpha' x$ y que tenga máxima correlación con una segunda variable indicadora definida a partir del segundo conjunto de variables y.
- Si procedemos de esta manera podemos hallar $r = min(p, q)$ relaciones entre variables indicadoras de ambos grupos de variables.

Correlaciones canónicas

- El proceso para obtener las $2r$ combinaciones lineales que llamaremos variables canónicas $(x_1^*, x_2^*, \ldots, x_r^*)$, $(y_1^*, y_2^*, \ldots, y_r^*)$ consiste en hallar los autovalores y autovectores de la matriz A o B que definimos antes, donde ambas matrices tienen rango r.
- Los autovalores y autovectores de las matrices A o B nos permiten formar las variables canonicas para ambos grupos, que satisfacen lo siguiente:
  - Tienen correlación máxima cuando las variables canónicas provienen del mismo autovalor: $\rho(x_j^*, y_j^*) = max$.
  - Serán ortogonales dentro de cada grupo: $\rho(x_j^*, x_k^*) = 0$.
  - Serán ortogonales si corresponden a autovalores distintos: $\rho(x_j^*, y_k^*) = 0$, $\forall j \neq k$.

Propiedades

- Las variables canónicas son indicadoras de los dos conjuntos de variables que se definen por pares, con la condición de correlación máxima.
- Los coeficientes de las variables canónicas son los autovectores ligados al mismo autovalor de las matrices $\Sigma_{ii}^{-1}\Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}$ para $i = 1, 2$ e $i \neq j$.
- Si $x^* = \alpha' x$ es una variable canónica también lo es $-\alpha' x$ y los signos de las variables canónicas suelen tomarse de manera tal que las correlaciones entre las variables canónicas $\alpha' x$ y $\beta' y$ sean positivas.
- Las correlaciones canónicas λ_i^2 son el cuadrado del coeficiente de correlación entre las dos variables canónicas correspondientes.

Propiedades

- Las correlaciones canónicas son invariantes ante transformaciones lineales de las variables, son propiedades del conjunto de variables y no se modifican si sustituimos las r variables de un conjunto por r combinaciones lineales de ellas linealmente independientes.
- La primera correlación canónica λ_1^2 es mayor o igual que el mayor coeficiente de correlación simple al cuadrado entre las variables de cada conjunto.
- El coeficiente de correlación canónica λ_i^2 es el coeficiente de determinación en una regresión múltiple donde la variable respuesta es $y_i^* = \beta_i' y$ y considerando como variables explicativas al conjunto de las x y viceversa.

Estimación

- En la práctica los valores poblacionales no son conocidos por lo tanto deberemos estimarlos. Sin pérdida de generalidad supondremos que las variables de ambos conjuntos tienen promedio igual a 0.
- Las variables canónicas son funciones de la matriz de variancias y covariancias entre las variables. Bajo la hipótesis de normalidad multivariada el estimador máximo verosímil de esta matriz es S, la matriz de variancias y covariancias muestral. Por lo tanto los estimadores máximo verosímiles de las variables canónicas se obtienen al extraer los $r = min(p, q)$ autovalores y autovectores asociados a las matrices:

$$\widehat{A} = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}, \qquad \widehat{B} = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$$

Estimación

- Las matrices S_{ij} corresponden a los estimadores máximoverosímiles de Σ_{ij}, que se obtienen particionando la matriz S.
- En la práctica, suponiendo que $p \geq q$ basta obtener los autovalores de la matriz de dimensión mas chica, $\widehat{B}$.
- Si estandarizamos las variables y trabajamos con las matrices de correlación, las correlaciones canónicas no cambian. Al estandarizar las variables obtendremos:

$$\widehat{A} = A_R = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$$

donde las matrices R_{ij} son las matrices de correlación definidas como $R_{ij} = D_{ii}^{-1/2} S_{ij} D_{jj}^{-1/2}$, con $i, j = 1, 2$.

Estimación

- Las matrices D_{11} y D_{22} son diagonales y contienen las variancias de las variables de cada grupo. Vamos a comprobar que las matrices $\hat{A}$ y A_R tienen los mismos autovalores:
  Utilizando la relación $R_{ij} = D_{ii}^{-1/2} S_{ij} D_{jj}^{-1/2}$ podemos escribir:

$$A_R = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$$

$$A_R = (D_{11}^{1/2} S_{11}^{-1} D_{11}^{1/2})(D_{11}^{-1/2} S_{12} D_{22}^{-1/2})(D_{22}^{1/2} S_{22}^{-1} D_{22}^{-1/2})(D_{22}^{-1/2} S_{21} D_{11}^{-1/2})$$

$$A_R = (D_{11}^{1/2} S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} D_{11}^{-1/2})$$

$$A_R = (D_{11}^{1/2} \hat{A} D_{11}^{-1/2})$$

  La ecuación para obtener los autovalores de A_R es $|A_R - \lambda^2 I| = 0$, que puede escribirse como $|D_{11}^{1/2}||\widehat{A} - \lambda^2 I||D_{11}^{-1/2}| = 0$ por lo tanto las correlaciones canónicas son idénticas.

Estimación

- Los autovectores de A_R pueden obtenerse a partir de los autovectores de $\hat{A}$ ya que si v es un autovector de A_R resulta:

$$\begin{aligned} A_R v &= \lambda^2 v \\ D_{11}^{1/2} \hat{A} D_{11}^{-1/2} v &= \lambda^2 v \\ \hat{A} D_{11}^{-1/2} v &= \lambda^2 D_{11}^{-1/2} v \end{aligned}$$

- De este modo, la variable canónica asociada a λ^2 cuando las variables estan estandarizadas, se calcula como:

$$\hat{x}^* = (X D_{11}^{-1/2}) v$$

Mientras que si las variables no están estandarizadas se calcula como:

$$\hat{x}^* = X(D_{11}^{-1/2} v)$$

y las variables canónicas son idénticas.

Test de hipótesis

- Podemos construir un test para comprobar si los dos conjuntos de variables están no correlacionados, es decir $\Sigma_{12} = 0$ bajo la hipótesis de normalidad $x \sim N_p(0, \sigma_{11})$ e $y \sim N_p(0, \Sigma_{22})$.
- El test de no correlación entre x e y es equivalente al test de que todas las correlaciones canónicas son nulas:

$$H_0)\ \Sigma_{12} = 0 \qquad H_1)\Sigma_{12} \neq 0$$

- Bajo H_0 la función de verosimilitud conjunta se descompone como $f(x_1, \ldots, x_n, y_1, \ldots, y_n) = f(x_1, \ldots, x_n) f(y_1, \ldots, y_n)$ El cociente de verosimilitudes es:

$$RV = \frac{f(H_1)}{f(H_0)} = \frac{(2\pi)^{-n(p+q)}|S|^{-n/2}e^{-n(p+q)/2}}{(2\pi)^{-np}|S_{11}|^{-n/2}e^{-np/2}(2\pi)^{-nq}|S_{22}|^{-n/2}e^{-nq/2}}$$

Test de hipótesis

- El test de razón de verosimilitudes será:

$$\lambda = 2(log(H_1) - log(H_0)) = -nlog\frac{|S|}{|S_{11}||S_{22}|}$$

Donde $|S| = |S_{11}||S_{22} - S_{21}S_{11}^{-1}S_{12}| = |S_{11}||S_{22}||I - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}|$
Por lo tanto

$$\lambda = -nlog(|I - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}|)$$

$$\lambda = -nlog(\prod_{j=1}^{r}(1 - \lambda_j^2))$$

$$\lambda = -n\sum_{j=1}^{r} log(1 - \lambda_j^2)$$

que tiene distribución asintótica χ^2 con pq grados de libertad. La aproximación mejora si se reemplaza n por $m = n - (p + q + 3)/2$.

Test de hipótesis

- Podemos utilizar este test para comprobar si las primeras s correlaciones canónicas son distintos de cero y los restantes $r - s$ son iguales a cero. Las hipótesis del test serán:

$$H_0)\lambda_i^2 > 0, i = 1, \ldots, s; \lambda_{s+1}^2 = \cdots = \lambda_r^2 = 0$$
$$H_1)\lambda_i^2 > 0 i = 1, \ldots, s; \text{ al menos algún } \lambda_j^2 > 0 j = s+1, \ldots, r$$

y el test de razón de verosimilitudes implica comparar los determinantes de la matriz de variancias y covariancias bajo H_0 y H_1 como en el caso anterior.

- El estadístico para este test es:

$$\lambda = -m \sum_{j=s+1}^{r} (1 - \lambda_j^2)$$

que bajo H_0 se distribuye como una χ^2 con $(p - s)(q - s)$ grados de libertad y se rechaza H_0 cuando λ_{obs} sea significativamente grande para el nivel de significación elegido.

Relación con otras técnicas

- El análisis de correlaciones canónicas cubre como casos particulares las técnicas de regresión y por extensión las de análisis discriminante.
- Supongamos el caso más simple en que cada uno de los conjuntos de variables contiene solo una variable. La correlación canónica entre x e y es el coeficiente de correlación al cuadrado:

  $$p = q = 1 \text{ y } R_{11} = R_{22} = 1 \text{ mientras que } R_{12} = R_{21} = r_{xy}.$$

  Entonces $R_{11}^{-1} R_{12} R_{22}^{-1} R_{21} = r_{xy}^2 = \lambda^2$.
- Si el conjunto de variables x tiene varias variables, $q = 1$ y $p > 1$, la correlación canónica entre la variable respuesta, y, y el conjunto de variables explicativas, x, es el cuadrado del coeficiente de correlación múltiple.

Relación con otras técnicas

- En efecto, siendo $S_{22} = s_y^2$, llamando S_{21} al vector de covariancias entre la variable respuesta y y las variables explicativas x y S_{11} la matriz de variancias y covariancias de las variables explicativas, obtenemos:

$$\lambda^2 = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12} = \frac{S_{21} S_{11}^{-1} S_{12}}{s_y^2}$$

y la correlación canónica es el coeficiente de determinación del modelo de regresión múltiple entre y y las variables explicativas x:

$$\lambda^2 = \frac{SCRegresion}{SCTotal} = R^2$$

Relación con otras técnicas

- El análisis discriminante también se puede abordar desde la perspectiva de correlaciones canónicas.
- Se definen $G-1$ variables explicativas binarias que indiquen a que grupo o población pertenece cada observación, que conformarán la matriz $X_{(n\times(G-1))}$. Por otro lado tendremos las p variables de la matriz $Y_{(n\times p)}$.
- El análisis de correlaciones canónicas entre las matrices de datos X e Y nos llevará al mismo resultado que el análisis discriminante.
- Se puede demostrar que llamando S a la matriz de dimensión $(p+G-1)$ de variancias y covariancias entre las variables de X e Y, obtenemos $nS_{22} = T$, $nS_{21}S_{11}^{-1}S_{12} = B$.

Relación con otras técnicas

- Las correlaciones canónicas obtenidas con la matriz $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$ serán las obtenidas con $T^{-1}B$.
- Para ver la relación entre las correlaciones canónicas y las obtenidas en análisis discriminante con la matriz $W^{-1}B$, observemos que si llamamos a_i a los vectores que definen las variables canónicas y λ_i^2 a las correlaciones canónicas, entonces:

$$\begin{aligned} T^{-1}Ba_i &= \lambda_i^2 a_i \\ (W+B)^{-1}Ba_i &= \lambda_i^2 a_i \\ (I+W^{-1}B)^{-1}W^{-1}Ba_i &= \lambda_i^2 a_i \\ W^{-1}Ba_i &= \lambda_i^2 a_i + \lambda_i^2 W^{-1}Ba_i \\ W^{-1}Ba_i &= [\lambda_i^2/(1-\lambda_i^2)]a_i \end{aligned}$$

- Los autovectores que definen las variables canónicas discriminantes son idénticas a las obtenidas por correlaciones canónicas y los autovalores se relacionan por $\phi_i = \lambda_i^2/(1-\lambda_i^2)$.

Detección de outliers

- Existen diversas definiciones de outlier en la literatura, una de ellas es la siguiente: "Se denominan outliers a una minoría de las observaciones en un conjunto de datos que tiene patrones diferentes a los de la mayoría de las observaciones".
- Se suele asumir que hay un *core* o núcleo de al menos el 50% de las observaciones que son homogéneas y las restantes observaciones presentan un patrón atípico diferente al patrón común.
- Si bien es posible identificar outliers a través de la inspección de gráficos este procedimiento no funciona para más de tres dimensiones.
- La identificación automática de outliers es complicada debido a los efectos de enmascaramiento (*masking effect*) y de confusión (*swamping effect*).
- El primero genera falsos negativos, por clasificarse como "típicos" algunos datos que no lo son, mientras que el efecto de confusión corresponde a la situación donde se clasifican como atípicas observaciones que no lo son, generando así falsos positivos.

Detección de outliers

- Existen diversos enfoques para lidiar con la presencia de outliers, muchas veces se requiere la utilización de herramientas de estimación robusta como etapa importante en la identificación de outliers.
- Los métodos robustos están diseñados para ser "resistentes" a la presencia de outliers, en el sentido de que los resultados de la estimación no se vean (muy) afectados por su presencia, y para ello en general se da menos peso a las observaciones sospechosas.
- Una vez obtenidas las estimaciones robustas de los parámetros de interés se pueden utilizar para identificar outliers. Por ejemplo, calculando distancias de Mahalanobis.
- Sea $\bar{x}$ el vector de medias muestrales y S representa la matriz de varianzas y covariancias entre las variables del vector x. La distancia de Mahalanobis (al cuadrado) entre la i-ésima observación y el centro de los datos resulta:

$$d_M^2(x_i, \bar{x}) = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) = d_i(\bar{x}, S)^2$$

Detección de outliers

- Un valor grande para la distancia $d_i(\bar{x}, S)^2$ puede indicar que la i-ésima observación es un outlier, un valor pequeño de distancia no necesariamente significa que no lo sea.
- Esto se debe a que el vector de medias $\bar{x}$ y la matriz de variancias y covariancias S no son robustas, por lo tanto tampoco lo es la distancia de Mahalanobis.
- Si reemplazamos estos parámetros por estimaciones robustas podremos calcular una distancia de Mahalanobis también robusta que resulta más efectiva para identificar outliers.
- Es importante destacar que se asume que los outliers constituyen menos del 50% de los datos y que los mismos son originados por una distribución diferente.
- Así, sea F_1 la distribución de los no-outliers y F_2 la distribución de los outliers, la distribución del conjunto total de datos será

$$F = (1 - a)F_1 + aF_2$$

donde $a \in (0, 0.5)$.

Detección de outliers

- Dentro de la literatura estadística hay dos enfoques generales para la identificación de outliers:
  - métodos calculados en la búsqueda de direcciones de proyección
  - métodos basados en el cálculo de distancias robustas.
- En el primer caso, se intenta detectar si hay una estructura particular en los datos que pasa desapercibida a simple vista. La idea es encontrar proyecciones en unas pocas direcciones que revelen información útil sobre la estructura de los datos.
- Un caso especial de este enfoque es el análisis de componentes principales, donde las direcciones de proyección de los datos se determinan de manera tal que se maximice la variancia de cada componente.
- En el caso de la detección de outliers, el resultado que se espera obtener al identificar la dirección correcta es que los outliers proyectados sobre ésta sean inmediatamente obvios.

Método basado en Kurtosis

- Peña y Prieto (2001) presentaron un método denominado Kurtosis1 que proyecta los datos en un conjunto de $2p$ direcciones, siendo p la dimensión del conjunto de datos original.
- Estas direcciones se eligen de forma tal que se minimiza y maximiza el coeficiente de kurtosis de los datos proyectados. Una vez encontradas las direcciones de proyección se determina el grado de *outlying* de los datos calculando la mediana y DMA univariada de los datos proyectados en cada dirección aplicando la fórmula de cálculo presentada antes.
- Si una observación es atípica en alguna de las direcciones de proyección se la clasifica como potencial outlier. Luego se calcula la media y dispersión de los datos a través de la muestra de observaciones no identificadas como potenciales outliers.
- Los resultados de esta estimación se pueden utilizar para calcula la distancia de Mahalanobis robusta para cada punto con respecto al centro de los datos.
- Todos aquellos puntos que superen un cierto valor crítico de la distribución χ^2 con p grados de libertad serán declarados como outliers.

Método basado en distancias

- Como alternativa a los procedimientos anteriores se encuentra el cálculo de distancias. Al respecto, Rocke y Woodruff señalan las siguientes dos etapas:
  - Obtener estimaciones robustas del vector de medias y la matriz de varianzas y covariancias, luego calcular la distancia de Mahalanobis robusta entre cada observación y el centro (robusto) de los datos que llamaremos T. Este cálculo dará por resultado una medida de cuan alejado está cada punto del centro T de acuerdo a la escala de los datos, que llamaremos C (es la matriz de variancias y covariancias robusta de los datos).
  - Determinar una frontera de separación D tal que los puntos cuya distancia robusta d_i sean mayores a D serán declarados outliers.
- Si los datos están distribuidos según una normal y considerando los estimadores T y C de toda la muestra, la distancia de Mahalanobis tiene una distribución χ^2. Sin embargo, cuando se consideran los estimadores robustos, esta distribución podría no ser válida, como tampoco en el caso de datos altamente asimétricos.
- La determinación de D no es un problema trivial. A continuación se presentan sintéticamente dos enfoques sobre este tema.

Contornos de distribuciones bivariadas

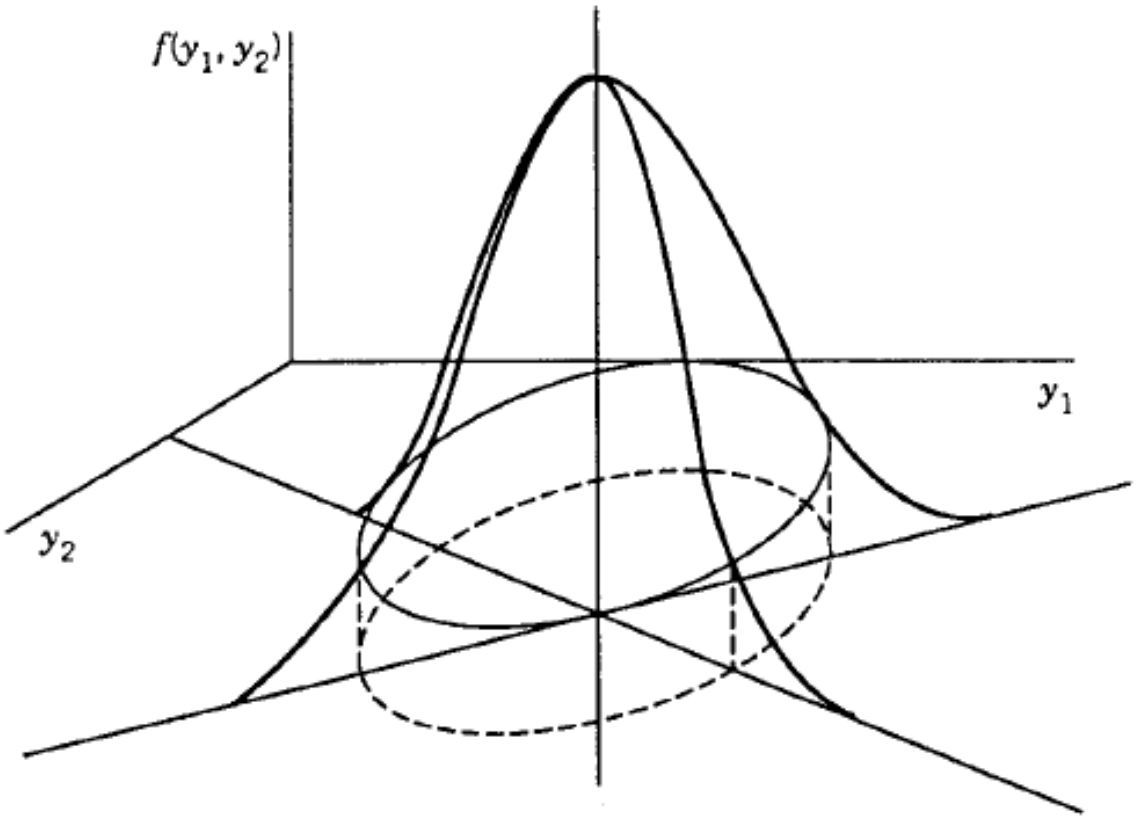

Figure 4.4. Constant density contour for bivariate normal.

Contornos de distribuciones bivariadas

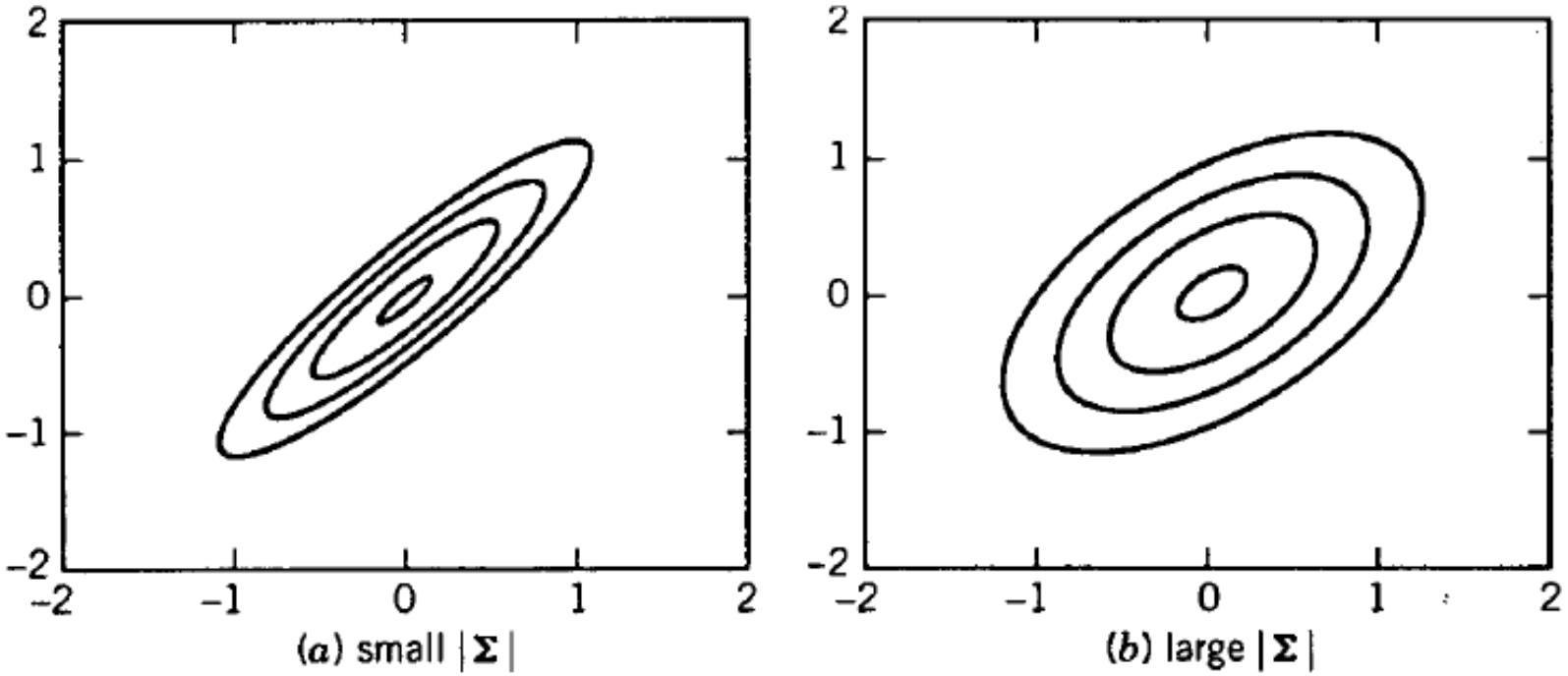

Figure 4.3. Contour plots for the distributions in Figure 4.2.

Enfoques MVE y MCD

- El estimador MVE está basado en la idea de hallar el elipsoide de volumen más chico que cubre h de las n observaciones y su utilización se volvió popular debido a su alta resistencia a la presencia de valores atípicos, que lo hacen una herramienta confiable para la detección de outliers.
- El enfoque estándar para determinar si los elementos de este conjunto de datos son homogéneos es calcular distancias de Mahalanobis para cada elemento con respecto al centro de los datos.
- Si los datos provienen de una población normal, las distancias de Mahalanobis (al cuadrado) tendrán una distribución aproximada χ^2 con p grados de libertad, por lo tanto se pueden comparar las distancias con un valor crítico de la esa distribución.
- Si el conjunto de datos es homogéneo no se esperan resultados muy por encima del valor crítico considerado, sin embargo el vector de medias y la matriz de covariancias muestrales pueden estar muy influenciados por la presencia de outliers.
- Como resultado de esto, es posible que los outliers tengan asociados valores pequeños para las distancias de Mahalanobis y por lo tanto permanezcan sin ser detectados (enmascaramiento).

Enfoques MVE y MCD

- Definición 1: Los estimadores de localización T_n y de dispersión C_n por MVE minimizan el determinante de C sujeto a la condición

$$Card\{i : (x_i - T)'C^{-1}(x_i - T) \leq c^2\} \geq h$$

  Donde la minimización se realiza sobre todos los $t \in R^p$ y $C \in PDS(p)$, que es la familia de matrices definidas positivas de dimensión p.
- El valor de c es una constante fija elegida que determina la magnitud de C_n, generalmente se elige de manera que C_n es un estimador consistente de la matriz de covariancias de los datos provenientes de una población normal multivariada, $c = (\chi^2_{p,\alpha})^{1/2}$ donde $\alpha = h/n$.
- A partir de su definición, MVE estima el centro y dispersión del conjunto de datos a partir de las h observaciones más concentradas del conjunto de datos. El usuario puede elegir el valor de h y determinar el grado de robustez de las estimaciones. La elección estándar es $h = (n+p+1)/2$.
- Las estimaciones de localización y dispersión a través de MVE son independientes de las unidades de medida de las variables como así también de traslaciones o rotaciones de los datos.

Enfoques MVE y MCD

- El enfoque MCD es similar al MVE y tiene la misma función objetivo. La diferencia está en la restricción considerada.
- El criterio MCD requiere únicamente que las estimaciones estén basadas en h puntos en lugar del elipsoide de h puntos. Actualmente el procedimiento más utilizado es el MCD ya que se comprobó que es superior en términos de convergencia y eficiencia estadística (algoritmo $FAST - MCD$).
- El corazón del algoritmo $FAST - MCD$ es el Concentration Step ($C - Step$). Dado un subconjunto H_1 que contiene h puntos, el mejor conjunto de h puntos se puede obtener al calcular la distancia de Mahalanobis para todos los n puntos basados en la media y matriz de covariancias de los h puntos anteriores y seleccionar el nuevo conjunto de elementos como aquellos h que poseen las distancias de Mahalanobis más pequeñas.
- Este procedimiento se repite hasta encontrar el conjunto final de h elementos cuando en dos pasos consecutivos de la iteración el conjunto no cambia.

Procedimiento BACON

- Este procedimiento fue desarrollado por Billor et. al. (2000) y su nombre corresponde a las siglas de *Blocked Adaptive Computationally Efficient Outlier Nominator*.
- Es un procedimiento iterativo basado en la idea de actualizar un subconjunto básico de elementos donde presumiblemente no hay outliers a partir del cual se estiman el vector de medias y la matriz de covarianzas que se utilizan luego para el cálculo de distancias robustas de Mahalanobis para todos los elementos de la matriz de datos.
- A continuación se describen a grandes rasgos los pasos del algoritmo.

Procedimiento BACON

1. Seleccionar un subconjunto inicial presumiblemente homogéneo eligiendo una de las dos opciones disponibles y calcular el vector de medias T_b y la matriz de covariancias C_b de este subconjunto.
2. Calcular la distancia robusta de Mahalanobis d_i para cada elemento del dataset utilizando T_b y C_b.
3. Determinar el nuevo subconjunto básico como aquel que contiene todos los puntos para los cuales $d_i^2 < c_{npr}^2 \chi_{p,1-\alpha/n}^2$, donde el último término representa el percentil $(1 - \alpha/n)$ de la distribución χ^2 con p grados de libertad y c_{npr}^2 es un factor de corrección para subconjuntos básicos pequeños o matrices de datos pequeñas. Este factor se aproxima a 1 a medida que se incrementa el tamaño de los subconjuntos o del conjunto de datos.
4. Se repiten los pasos 2 y 3 hasta que el tamaño del subconjunto básico se mantenga constante. En la práctica esto ocurre usualmente en 3 o 4 iteraciones.
5. Las observaciones excluidas del subconjunto básico final se "nominan" como outliers. Si es posible, graficar las distancias d_i para verificar que puntos son claramente outliers y cuales son de borde o requieren una inspección más detallada.

Algunas referencias

1. Hadi, A.S.; Rahmatullah Imon, A.H.M. y Werner, M. (2009): "Detection of outliers". John Wiley & Sons, Inc. WIREs Comp Stat 2009. Vol. 1, 57 – 70.
2. Hardin, J. y Rocke, D. (2004): "Outlier detection in multiple cluster setting using de minimum covariance determinant estimator". Computational Statistics and Data Analysis, 44: 625 – 638.
3. Hardin, J. y Rocke, D. (2005): "The distribution of robust distances". Journal of Computational and Graphical Statistics, 14, 928 – 946.
4. Peña, D. y Prieto,F. J. (2001*a*): "Robust covariance matrix estimation and multivariate outlier detection". (con discusión), Technometrics, 3, 286 – 310.
5. Peña, D. y Prieto,F. J (2001*b*). "Cluster Identification using Projections". The Journal of American Statistical Association, 96, December 2001.
6. Rocke, D. y Woodruff, D. (1996): "Identification of outliers in multivariate data". The Journal of American Statistical Association, 91, September 1996.
7. Van Aelst, S. and Rousseeuw, P. (2009): "Minimum volume ellipsoid". John Wiley & Sons, Inc. WIREs Comp Stat 2009. Vol. 1, 71 – 82.