UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Ciencias

USO DE COMPONENTES PRINCIPALES EN LA DETECCION INFORMAL DE OBSERVACIONES ABERRANTES

T E S I S

que para obtener el título de

A C T U A R I O

p r e s e n t a

OLGA VICTORIA SERRANO SANCHEZ

MEXICO, D. F. 1981

I N D I C E

P R O L O G O

El propósito de este trabajo, es probar los métodos gráficos en la detección informal de observaciones aberrantes multivariadas, enfocado primordialmente al análisis de componentes principales, ya que ha sido uno de los métodos gráficos más empleados. Las diferentes -- propuestas que existen acerca de este método, tales como las gráficas individuales y por parejas de las componentes principales, se prueban a lo largo del presente estudio. Se prueban además, otros dos méto-- dos gráficos: las muestras marginales de cada una de las variables y las raíces cuadradas de las distancias ordenadas.

De esta forma, en el Capítulo I, se desarrolla el concepto analítico del método de análisis de componentes principales, su historia, significado geométrico y las diferentes pruebas de significancia que existen.

En el Capítulo 2, se presenta el problema de observaciones abe-- rrantes multivariadas y los diferentes métodos que han sido propues-- tos para detectarlos, clasificándose éstos en pruebas de hipótesis y métodos informales. Para probar los métodos propuestos, se generaron gráficas para cada uno de ellos con diferente tamaño de muestra y diferente número de aberrantes, presentándose un resumen del análisis - de los mismos en el Capítulo 3.

Por último, en los anexos se presentan las diferentes gráficas - con sus respectivos comentarios, los que pueden servir para darnos una idea de la ayuda que los métodos gráficos nos pueden proporcionar así como la limitación de los mismos.

C A P I T U L O 1

ANALISIS DE COMPONENTES PRINCIPALES

INTRODUCCION

En este capítulo, presentamos una descripción detallada del método de componentes principales. Esta técnica es usada cuando se trabaja con muestras multivariadas, principalmente para reducir la dimensión de los datos y así facilitar el manejo de los mismos. En la sección 1.1 se discute el concepto del método en una forma más amplia. El desarrollo histórico se presenta en la sección 1.2. Posteriormente, las secciones 1.3 y 1.4 muestran el desarrollo analítico, así como la interpretación geométrica del método. Por último, las diferentes pruebas de significancia que han sido desarrolladas en relación a componentes principales, se presentan en la sección 1.5.

1.1 DESCRIPCION DE LA TECNICA DE COMPONENTES PRINCIPALES

La idea básica del análisis de componentes principales es reducir el conjunto de variables originales X's correlacionadas, en otro conjunto de variables no correlacionadas Y's de menor dimensión, de tal manera que se retenga la mayor información esencial de la muestra. Es decir, el método de componentes principales consiste en transformar un conjunto de variables $X_1, X_2, \ldots, X_p$ en un nuevo conjunto $Y_1, Y_2, \ldots, Y_p$ con las siguientes propiedades:

a) Cada Y es una combinación lineal de las X's, digamos:

$$Y_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p$$

$$j = 1, 2, \ldots, p$$

b) La suma de cuadrados de los coeficientes a_{ij}, $j = 1, 2, \ldots p$ es igual a la unidad.

c) De todas las posibles combinaciones lineales de este tipo, Y_1 tiene varianza mayor.

d) De todas las combinaciones lineales de este tipo, no correlacionadas con Y_1, Y_2 describe la máxima variación de las observaciones. Similarmente, Y_3 es la combinación lineal no correlacionada con Y_1 y Y_2, que describe la máxima variación de las observaciones. Así sucesivamente, hasta que $Y_1, Y_2, \ldots, Y_p$ son determinadas.

Generalmente, en la práctica sólo se consideran las primeras combinaciones lineales que son las que explican la mayor variabilidad de los datos y así descartar los "sobrantes" y reducir el número de variables a considerar. Esta reducción de variables, es de gran utilidad práctica, ya que, en primer lugar, facilita el manejo matemático y presentación gráfica de los datos. En segundo lugar, en el supuesto caso de que exista redundancia en las observaciones originales, tal que las variables están linealmente relacionadas, se dificulta el análisis numérico y esto se puede evitar reduciendo el número de variables. Por último, dicha reducción no ocasiona obstáculo

alguno en la interpretación de los resultados finales, ya que las variables transformadas pueden conducir ellas mismas a esta interpretación en términos de las variables originales.

Es importante mencionar que el método de componentes principales puede ser aplicado, entre otros, para los siguientes propósitos:

a) Métodos gráficos y análisis de conglomerados (Cluster Analysis).

En análisis de conglomerados, no es tan fácil definir criterios mediante los cuales decidir, usando métodos numéricos, si hay una justificación para dividir conjuntos de observaciones en grupos. Si se grafican las componentes principales puede ayudar al análisis de conglomerados de diferentes maneras. Por ejemplo, si hay grupos bien definidos y separados, un método analítico no es necesario. Por otro lado, en el caso de que ninguna prueba de significancia fuera posible, el graficar las componentes principales puede al menos confirmar que un grupo sugerido se ve razonable y es realmente indicado por las observaciones.

b) Redundancia.

La mayoría de las técnicas del análisis multivariado consideran matrices de orden igual al número de variables. Cuando éstas son demasiadas, los cálculos llegan a ser tediosos

y si las correlaciones entre las observaciones son altas, - tal que las matrices de dispersión son casi singulares o -- singulares, las dificultades se incrementan. Una manera de prevenir este problema, es hacer un análisis de componentes principales preliminar, y descartar las componentes más grandes, ya que en muchos casos éstas contienen más "ruido" que "información" y poco o nada se pierde con eliminarlas. Entonces el análisis puede ser llevado a cabo sobre las componentes restantes.

c) Detección de observaciones aberrantes.

Un problema muy frecuente cuando se obtienen datos de muestras multivariadas, es el de detectar aquellas observaciones que se encuentran "alejadas" del conjunto de datos. Un análisis de componentes principales y métodos gráficos puede ser de gran utilidad en la detección de estas observaciones. Esto se verá con más detalle en el capítulo 2.

1.2 HISTORIA Y APLICACIONES

Existe literatura del análisis de componentes principales desde antes de 1904, pero Karl Perason fue el primero que dió las bases matemáticas, describiendo la línea de mejor ajuste en un grupo elipsoidal de puntos. En 1933, Harold Hotelling formuló la definición moderna de componentes principales como los ejes que describen una máxima variación en los datos.

Las aplicaciones del análisis de componentes principales han sido bastante amplias en las diferentes ramas de la ciencia. Uno de los primeros artículos aparecidos fue publicado por Jolicoeur y Mossiman (1960), aplicando el análisis de componentes principales en morfométrica, en un estudio sobre una variedad de tortugas (painted turtle), para evaluar la variación de tamaño y forma en grupos de organismos vivientes. Para este estudio, contaron con 48 tortugas, 24 machos y 24 hembras, y midieron la longitud (X_1), el ancho máximo (X_2) y la altura (X_3) del carapacho. Como resultado obtuvieron los valores de los coeficientes a_{ij} (i, j, = 1, 2, 3) que se muestran en la tabla I.

T A B L A I

MACHOS			HEMBRAS		
.84012	- .48811	- .23654	.81263	- .54537	- .20540
.49190	.86938	- .04690	.49549	.83213	- .24907
.2284	- .0769	.97049	.30676	.10062	.94645

Se calcularon por separado las componentes principales para cada sexo. La primer columna de la tabla, muestra los coeficientes de la primera componente principal.

Así, la primera combinación lineal es:

$$Y_1 = .84012\ X_1 + .49190\ X_2 + .2284\ X_3 \quad \text{(machos)}$$

$$Y_1 = .81263\ X_1 + .49549\ X_2 + .30676\ X_3 \quad \text{(hembras)}$$

la cual tiene todos los coeficientes positivos. Esto implica que una unidad que incrementa a Y_1, produce un incremento en X_1, X_2 y X_3 (longitud, ancho y alto del carapacho). Por esta razón, y puesto que el crecimiento generalmente es definido como el incremento en tamaño de un organismo, Y_1 fue interpretado como una "tendencia" de crecimiento.

La segunda columna de la tabla I da los coeficientes de la segunda componente principal. Esto es:

$$Y_2 = -.48811\ X_1 + .86938\ X_2 - .07690\ X_3 \quad \text{(machos)}$$
$$Y_2 = -.54537\ X_1 + .83213\ X_2 + .10062\ X_3 \quad \text{(hembras)}$$

Como puede observarse, Y_2 tiene algunos coeficientes positivos y otros negativos, motivo por el cual lo interpretaron como una tendencia de "variación de forma", ya que un incremento en Y_2 incrementa alguna X_i y decrementa otras.

Reyment y Sandberg[1)] aplicaron el análisis de componentes principales en un estudio en el cual midieron cuatro características de fósiles de caracoles en una escala logarítmica. Como resultado obtuvieron que la primera componente principal explicó el 90% de la variación total.

1) De Blackith y Reyment (1971)

Mosser y Scott[1)] usaron el análisis de componentes principales para estudiar la intercorrelación entre 57 variables socioeconómicas medidas en 157 pueblos británicos. Los resultados que obtuvieron -- fueron que las primeras cuatro componentes juntas explicaron un 60% de la varianza total. Estas cuatro componentes fueron identificadas como clase social, cambio de población 1931-1951, cambio de pobla--- ción 1951-1958 y sobrepoblación.

Amtman,[1)] en una investigación sobre los cráneos de ardillas, - interpretó la primera componente principal como una medida de tamaño y la segunda como una medida de forma, sin dar ninguna razón apriori.

1.3 DESARROLLO ANALITICO

Supongamos que las variables originales X_1, X_2,...X_p tienen una distribución multivariada con vector de medias $\underline{\mu}$ y matriz de varian zas y covarianzas V. De esta población se selecciona una muestra de N vectores de observaciones independientes.

Entonces la primera componente principal es la combinación li-- neal de p variables correlacionadas:

$$Y_1 = a_{11} X_1 + a_{21} X_2 + \ldots + a_{p1} X_p = \underline{a}'_1 \underline{X} \quad 2)$$

1) De Blackith y Reyment

(2) $\underline{a}'_1$ denota el vector transpuesto de $\underline{a}_1$.

La varianza estimada de Y_1 puede ser expresada como una forma - cuadrática en términos de las varianzas y covarianzas de las varia-- bles originales. Esto es:

$$\hat{V}(Y_1) = \underline{a}'_1 S \underline{a}_1 = \sum_{i=1}^{P} \sum_{j=1}^{P} a_{i1} a_{j1} s_{ij}$$

donde $S = \frac{1}{N-1} \sum_{i=1}^{N} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})'$ es la matriz de varian- zas y covarianzas muestral, con elementos s_{ij}.

La varianza de Y_1 es la más grande de todos los vectores norma- lizados tal que la suma de cuadrados de a_{i1} sea igual a uno. Esto - es:

$$\underline{a}'_1 \ \underline{a}_1 = 1$$

Esta restricción ha sido tomada porque el valor de $\hat{V}(Y_1)$ pudo ser incrementado sin límite con sólo multiplicar los coeficientes -- por una constante.

Ahora bien, el problema consiste en encontrar el vector $\underline{a}_1$ que maximice $\underline{a}'_1 S \underline{a}_1$ sujeto a la restricción $\underline{a}'_1 \underline{a}_1 = 1$. Para determi- nar los coeficientes, introducimos los multiplicadores de Lagrange, l_1 y derivamos con respecto a $\underline{a}_1$.

$$\frac{\partial}{\partial \underline{a}_1} \left(\underline{a}'_1 S \underline{a}_1 - l_1 (\underline{a}'_1 \underline{a}_1 - 1) \right) =$$

$$= 2 S \underline{a}_1 - l_1 (2 \underline{a}_1) = 2 (S - l_1 I) \underline{a}_1$$

Igualando a cero esta ecuación, obtenemos el sistema de p ecuaciones lineales simultáneas:

$$(S - l_1 I) \underline{a}_1 = 0$$

Los coeficientes de $\underline{a}_1$ tienen que satisfacer este sistema de ecuaciones, y si la solución es diferente a la trivial, entonces l_1 - tiene que ser seleccionado tal que el determinante:

$$| S - l_1 I | = 0$$

Así, l_1 es una raíz característica (o eigenvalor) de S y $\underline{a}_1$ es su vector característico asociado.

Si la matriz de covarianzas S es de rango completo, entonces -- tiene p raíces características $\neq 0$. Para determinar cuáles de las p raíces deben ser seleccionadas, premultiplicamos la ecuación

$$S \underline{a}_1 = l_1 \underline{a}_1$$

por $\underline{a}'_1$:

$$\underline{a}'_1 S \underline{a}_1 = l_1 \underline{a}'_1 \underline{a}_1 = l_1 = \hat{V}(Y_1)$$

ya que $\underline{a}'_1 \underline{a}_1 = 1$

Por lo tanto, l_1 es la raíz característica más grande, porque el vector $\underline{a}_1$ fue seleccionado de tal manera que maximice la varianza de Y_1.

Después de haber determinado la primera componente principal, el siguiente paso es encontrar una segunda combinación lineal normalizada y ortogonal a la primera. Esto es:

$$Y_2 = a_{12} X_1 + a_{22} X_2 + \ldots + a_{p2} X_p$$

cuya varianza $\hat{V}(Y_2) = \underline{a}'_2 S \underline{a}_2$ sea máxima sujeta a las restricciones

$$\underline{a}'_2 \underline{a}_2 = 1$$
$$\underline{a}'_1 \underline{a}_2 = 0$$

La restricción $\underline{a}'_1 \underline{a}_2 = 0$ implica que $\underline{a}_1$ y $\underline{a}_2$ son ortogonales, es decir, correlación cero entre Y_1 y Y_2 y por consiguiente covarianza cero.

Mediante el mismo procedimiento que se siguió para encontrar los coeficientes de Y_1, se determinan los coeficientes de Y_2, pero ahora introduciendo un nuevo multiplicador de Lagrange, μ, y diferenciando con respecto a $\underline{a}_2$:

$$\frac{\partial}{\partial \underline{a}_2} \left(\underline{a}'_2 S \underline{a}_2 - l_2 (\underline{a}'_2 \underline{a}_2 - 1) - \mu (\underline{a}'_1 \underline{a}_2) \right) =$$

$$= 2 S \underline{a}_2 - 2 l_2 \underline{a}_2 - \mu \underline{a}_1$$

$$= 2 (S - l_2) \underline{a}_2 - \mu \underline{a}_1$$

Igualando a cero y premultiplicando esta ecuación por $\underline{a}'_1$ resulta:

$$2 \underline{a}'_1 S \underline{a}_2 - 2 l_2 \underline{a}'_1 \underline{a}_2 - \mu \underline{a}'_1 \underline{a}_1 = 0 \Longrightarrow$$

$$2 \underline{a}'_1 S \underline{a}_2 - \mu = 0$$

Igualmente, si premultiplicamos la ecuación $(S-l_1 I)\ \underline{a}_1$ por $\underline{a}'_2$ se sigue que:

$$\underline{a}'_2 S \underline{a}_1 - l_1 \underline{a}'_2 \underline{a}_1 = 0 \Longrightarrow \underline{a}'_2 S \underline{a}_1 = 0$$

$$\Longrightarrow \mu = 0 \quad \text{ya que } 2 \underline{a}'_1 S \underline{a}_2 = \mu$$

El vector $\underline{a}_2$ tiene que satisfacer las p ecuaciones

$$(S - l_2 I)\ \underline{a}_2 = 0$$

Si se premultiplica esta ecuación por $\underline{a}'_2$, se tiene:

$$\underline{a}'_2 S \underline{a}_2 - l_2 \underline{a}'_2 \underline{a}_2 = 0$$

$$\Longrightarrow \underline{a}'_2 S \underline{a}_2 = l_2$$

Por consiguiente, el vector $\underline{a}_2$ es el vector correspondiente a - la segunda raíz característica más grande, l_2. El mismo proceso se sigue para determinar las componentes principales restantes.

Estableciendo lo anterior en forma general, entonces la j-ésima componente principal es la combinación lineal

$$Y_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p = \underline{a}'_j \underline{X}$$

tal que para cualquier valor l_j que satisface

$$|S - l_j I| = 0 \quad ,$$

le corresponde un valor de $\underline{a}_j$ para el cual

$$(S - l_j I)\, \underline{a}_j = 0, \qquad \underline{a}'_j \underline{a}_j = 1$$

y para este valor $\underline{a}'_j S \underline{a}_j = l_j$. Así l_j es la j-ésima raíz característica más grande y $\underline{a}_j$ es su vector característico asociado.

El hecho de que la matriz S sea simétrica, implica que todas -- las raíces características son reales, y el hecho de que sea positiva definida, implica que todas son positivas. Puesto que $\underline{a}_i$ y $\underline{a}_j$ -- son ortogonales, entonces A, la matriz que contiene los vectores característicos, es ortogonal. Si $l_i = l_j$, sus correspondientes componentes principales no están especialmente definidas, ya que hay una infinidad de vectores ortogonales. La existencia de una raíz cero,

implica que las variables originales son linealmente dependientes; - en este caso, una ó más componentes valen cero. Además, que A' S A= L, donde L es la matriz diagonal que contiene las raíces características, indica independencia entre las p componentes principales. Finalmente, de la ecuación

$$\left| S - l_j I \right| = 0$$

se sigue que la suma de valores l_j es igual a la suma de la diagonal de S. Esto es:

$$l_1 + l_2 + \ldots + l_p = \text{traza de } S$$

El resultado de lo expuesto anteriormente, es que el problema - de determinar las componentes principales, se puede expresar en términos de una transformación ortogonal

$$Y = A' X$$

tal que

$$A' S A = L$$

y

$$A' A = I$$

La utilidad fundamental del análisis de componentes principales, radica en que la variación total de una muestra p-dimensional sería descrita en pocas dimensiones. En el caso de que S fuera de rango 1

existe una raíz característica $\neq 0$ y p - 1 raíces iguales a cero, lo cual indica que la primera componente principal explicaría toda la - variación del sistema. Si S es de rango completo, existen entonces p raíces características $\neq 0$ y la importancia de la componente j-ésima es medida por la relación

$$\frac{l_j}{\text{traza de } S}$$

es decir, la proporción de varianza que contribuye la j-ésima componente en la varianza total de las variables originales.

El signo algebraico y la magnitud de un elemento del vector $\underline{a}_j$, expresa la dirección e importancia de una variable particular a una componente particular, es decir, la razón

$$\frac{a_{ij}\sqrt{l_j}}{\sqrt{s_{ii}}}$$

da la correlación de la i-ésima variable y la j-ésima componente, ya que

$$\text{cov}(\underline{X}, Y_j) = \text{cov}(\underline{X}, a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p) =$$

$$S\underline{a}_j$$

s_{ii} es la varianza de la i-ésima variable (i = j)

y por la relación

$$(S - l_j I)\, \underline{a}_j = 0 \Longrightarrow S\, \underline{a}_j = l_j\, \underline{a}_j$$

y la covarianza de la i-ésima variable con la j-ésima componente es - $l_j\, a_{ij}$. Por lo tanto

$$\rho_{ij} = \frac{l_j\, a_{ij}}{\sqrt{S_{ii}}\ \sqrt{l_j}} = \frac{\sqrt{l_j}\ a_{ij}}{\sqrt{S_{ii}}}$$

es la correlación de la i-ésima variable y la j-ésima componente.

Ahora bien, es frecuente que las observaciones sean medidas en unidades diferentes, y en esta situación es necesaria una estandarización preliminar. La más común es la de dividir la matriz de covarianza por la desviación estándar de cada una de las variables, reduciendo así las varianzas a la unidad y trabajar con la matriz de correlación. Entonces, si las componentes principales han sido extraídas de la matriz de correlación, el vector

$$\sqrt{l_j}\ \underline{a}_j$$

da la correlación de las variables con la j-ésima componente. Asimismo

$$\text{traza de } R = p$$

donde R es la matriz de correlación y la proporción de la varianza -

total que contribuye la j-ésima componente, está dada por la rela- - ción

$$\frac{l_j}{p}$$

Es de importancia señalar que si el rango r de la matriz S es - menor que p, entonces la varianza total de las variables originales sería explicada por menos de p componentes; pero la matriz S puede - ser escrita como

$$S = A L A' = A \sqrt{L} \cdot \sqrt{L} A'$$

Sea $K = A \sqrt{L}$

Entonces, las columnas de K reproducen a S por la relación

$$S = l_1 \underline{a}_1 \underline{a}'_1 + l_2 \underline{a}_2 \underline{a}'_2 + \ldots + l_r \underline{a}_r \underline{a}'_r = K K'$$

Por lo tanto, como las componentes han sido extraídas de S, las matrices $l_j \underline{a}_j \underline{a}'_j$ pueden formarse y su suma compararla con S para - determinar qué tan bien está siendo generada la matriz S por un número más pequeño de variables.

1.4 SIGNIFICADO GEOMETRICO

En las secciones anteriores, hemos definido el método de componentes principales como las combinaciones lineales que describen, en

orden decreciente, la mayor variabilidad de la muestra. Ahora, bajo la suposición de que la muestra $\underline{X}_1$, $\underline{X}_2$..., $\underline{X}_N$ ha sido seleccionada de una población normalmente distribuída, y además que $\underline{X}_1$,..., $\underline{X}_N$ estén correlacionadas entre sí, introduciremos la interpretación geométrica del método como las variables correspondientes a los ejes principales de una elipse (elipsoide o hiperelipsoide, $p \geq 3$).

La existencia de dichas correlaciones reflejan la tendencia de las variables a variar juntas, esto significa que los puntos tienden a estar en un grupo elipsoidal más o menos bien definido (figura 1.1 $p = 2$).

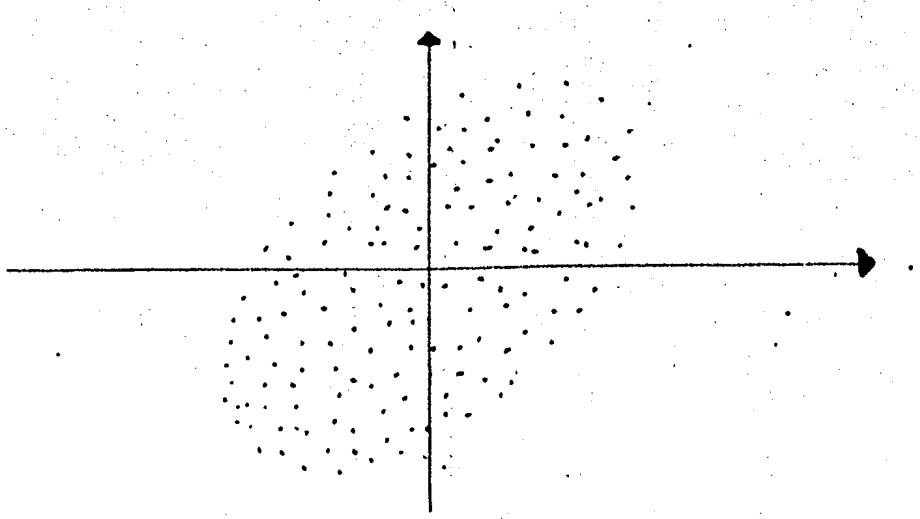


FIGURA 1.1

En términos de vectores, esto significa que los vectores tienden a estar asociados en grupos con ángulos pequeños entre ellos - - (figura 1.2).

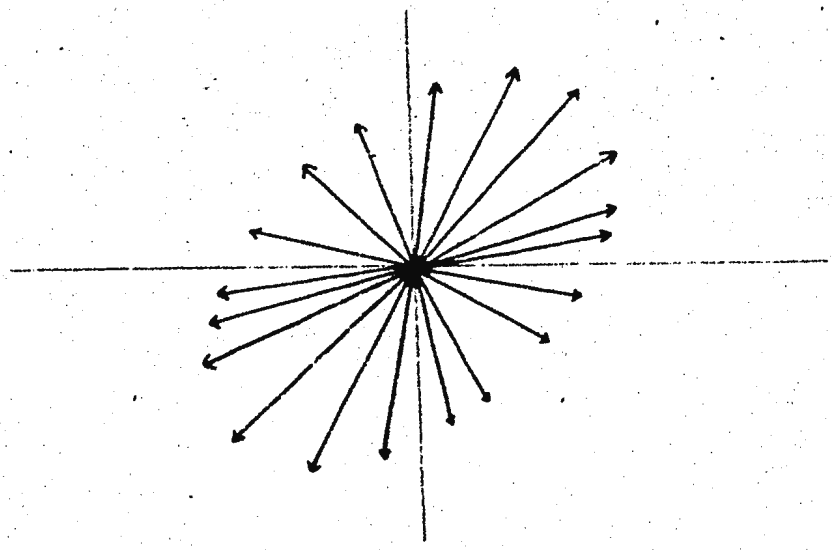


FIGURA 1.2

La existencia de más o menos correlaciones fuertes en la matriz, implica que los N puntos (o vectores) están contenidos dentro de una región que está restringida en el espacio p-dimensional. Esto permite que nuevos ejes coordenados sean colocados estratégicamente a través de las dimensiones mayores de la subregión, en la cual están actualmente los puntos. Estos puntos pueden ser descritos más eficientemente con respecto a los nuevos ejes que con los originales.

Precisamente, el análisis de componentes principales es esencialmente una técnica para introducir un sistema coordenado nuevo de esta clase, dentro de la masa de puntos. Los nuevos ejes coordenados son las componentes principales o ejes principales.

Como ya fue mencionado, las correlaciones de las variables (distribuídas normalmente) en el espacio tienden a tener una forma elipsoidal. En la figura 1.3, se muestra la efectividad de escoger como nuevos ejes de referencia aquellos que coinciden con los ejes mayores de la elipse (o hiperelipsoide en un espacio dimensional mayor). Esto en parte es porque los ejes de la elipse pueden ser arreglados tal que llegan a ser progresivamente más pequeños y, por lo tanto, menos importantes para propósitos descriptivos.

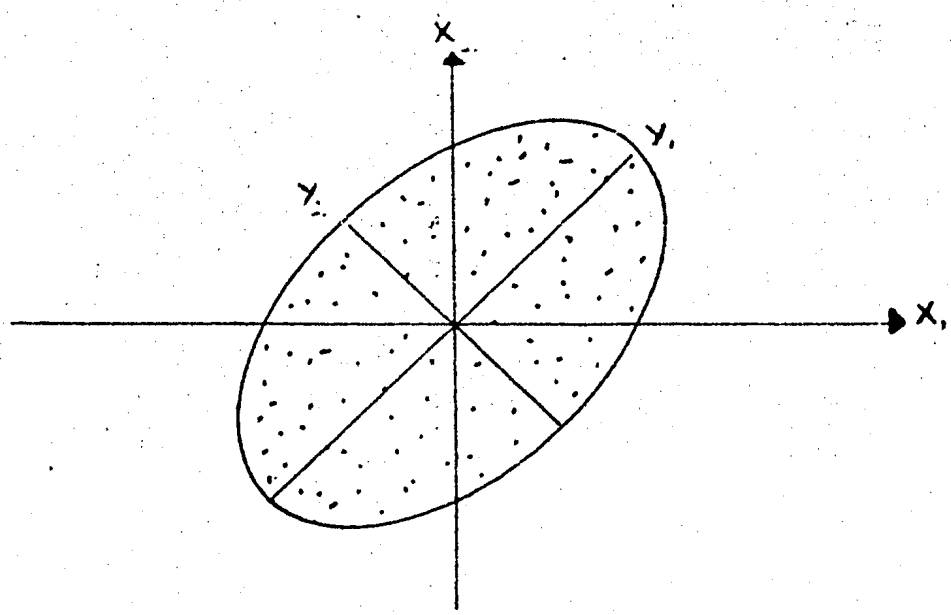


FIGURA 1.3

En el caso de no existir correlación, i.e., $\rho_{ij} = 0$ para toda i, j, la masa de puntos sería circular (o esfera o hiperesferoidal - para $p \geq 3$) y todos los ejes de referencia serían de igual longitud e importancia. En el caso de que todas las correlaciones fueran perfectas, i.e. $\rho_{ij} = \pm 1 \ \forall$ i, j, los puntos caerían en una recta --

(plano o hiperplano para $p \geq 3$) correspondiente a los ejes principales.

En efecto, la ecuación

$$(\underline{X} - \underline{\bar{X}})' \; S^{-1} \; (\underline{X} - \underline{\bar{X}}) = K$$

define una familia de elipsoides concéntricos centradas en la media $\underline{\bar{X}}$ de las observaciones. Si las X's son variables con una distribución normal multivariada, estas elipsoides son los contornos de densidad - de igual probabilidad (fig. 1.4 para $p = 2$).

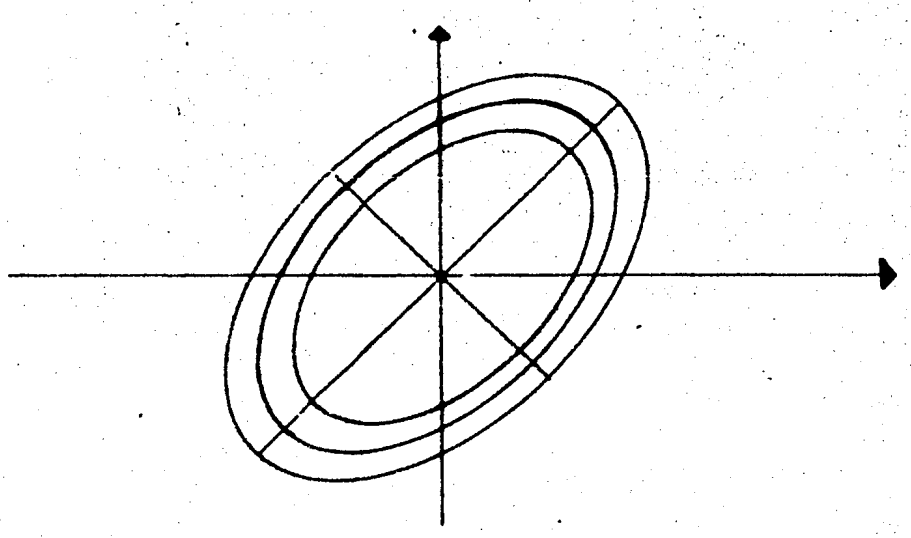


FIGURA 1.4

La transformación a componentes principales de los datos, es -- precisamente la proyección de las observaciones en los ejes principales de esta familia, y es equivalente a determinar estos ejes en orden de magnitud.

Consideremos el caso de 3 variables y N observaciones. El origen de las variables ha sido tomado como las medias de éstas. Desea mos que nuestro primer eje transformado, Y_1, esté en una orientación correspondiente a la variabilidad más grande; el eje Y_2, en ángulo - recto al primero, va a estar orientado en dirección de la siguiente variabilidad más grande; y así para Y_3.

Denotemos por α_1, α_2, y α_3, los ángulos que forma el eje Y_1 con X_1, X_2 y X_3 respectivamente (fig. 1.5).

Si Y_1 pasa por el punto de medias, su orientación es determinada por los cosenos directores $a_{11} = \cos\alpha_1$, $a_{21} = \cos\alpha_2$, $a_{31} = \cos\alpha_3$, donde $\cos^2\alpha_1 + \cos^2\alpha_2 + \cos^2\alpha_3 = 1$.

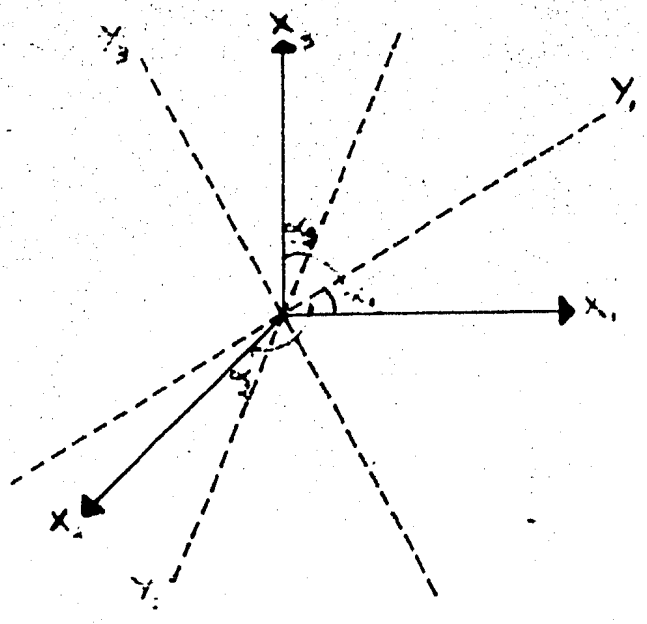


FIGURA 1.5

El valor de las observaciones en el nuevo eje coordenado Y_1, será expresado en términos de los cosenos directores

$$Y_{i1} = a_{11} (X_{i1} - \bar{X}_1) + a_{21} (X_{i2} - \bar{X}_2) + a_{31} (X_{i3} - \bar{X}_3)$$

La media de Y_1 es

$$\bar{Y}_1 = \sum_{i=1}^{N} \frac{1}{N} \cdot Y_{i1} =$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{P} a_{j1} (X_{ij} - \bar{X}_j) =$$

$$= \frac{1}{N} \sum_{i=1}^{N} (a_{11} (X_{i1} - \bar{X}_1) + a_{21} (X_{i2} - \bar{X}_2) + a_{31}(X_{i3} - \bar{X}_3) =$$

$$= a_{11} \bar{X}_1 - a_{11} \bar{X}_1 + a_{21} \bar{X}_2 - a_{21} \bar{X}_2 + a_{31} \bar{X}_3 - a_{31} \bar{X}_3 = 0$$

Por lo tanto, la varianza estimada de Y_1 es:

$$\hat{V}(Y_1) = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i1} - \bar{Y}_1)^2 = \frac{1}{N-1} \sum Y_{i1}^2 =$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left(\sum_{j=1}^{3} a_{j1} (X_{ij} - \bar{X}_j) \right)^2$$

Los ángulos de Y_1 serían determinados diferenciando esta ecua--

ción con respecto a a_{j1}, los cuales hagan las derivadas cero. La solución sería el vector característico de la raíz característica más grande de la matriz de covarianza muestral y Y_1 sería la primera componente principal.

Probemos lo anterior para el caso general de p variables, escribiendo los cosenos directores del primer eje componente principal como

$$\underline{a}'_1 = (a_{11}, a_{21}, \ldots, a_{p1})$$

y la restricción $\underline{a}'_1 \underline{a}_1 = 1$ tiene que satisfacerse. Entonces, la varianza de las proyecciones sobre el eje Y_1 será

$$\hat{V}(Y_1) = \frac{1}{N-1} \sum_{i=1}^{N} Y_i^2 =$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left(\sum_{j=1}^{P} a_{j1} (X_{ij} - \bar{X}_j) \right)^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left((\underline{X}_i - \bar{\underline{X}})' \underline{a}_1 \right)^2 =$$

$$= \frac{1}{N-1} \sum^{N} \left(\underline{a}'_1 (\underline{X}_i - \bar{\underline{X}}) (\underline{X}_i - \bar{\underline{X}})' \underline{a}_1 \right)$$

$$= \underline{a}_1 S \underline{a}_1$$

Considerando la restricción $\underline{a}'_1 \underline{a}_1 = 1$ e introduciendo el multiplicador de Lagrange l_1 para maximizar, resulta

$$\underline{a}'_1 S \underline{a}_1 + l_1 (1 - \underline{a}'_1 \underline{a}_1)$$

que es la ecuación que se obtuvo en el desarrollo analítico. Los cosenos directores del primer eje principal son los elementos del primer vector característico de S, y la varianza máxima $\hat{V}(Y_1)$, es la raíz característica más grande. Los vectores y raíces característicos restantes de S determinan las orientaciones y longitudes de los demás ejes componentes. En el caso de que dos raíces sucesivas l_j y l_{j+1} sean iguales, la configuración de los puntos sería más bien circular que elíptica, y una infinidad de vectores ortogonales pueden ser seleccionados, y la existencia de una raíz cero, significa que la elipsoide es degenerada y pudo ser representada en menos de p dimensiones.

Así, geométricamente se puede decir que las componentes principales de N observaciones y p variables, son las nuevas variables especificadas por los ejes de una rotación ortogonal del sistema coordenado de variables originales en una orientación correspondiente en las direcciones de máxima varianza. Los cosenos directores de los nuevos ejes son los vectores característicos normalizados correspondientes a las raíces características de la matriz de covarianza muestral. Si las componentes son calculadas de la matriz de correlación, la misma interpretación geométrica se mantiene, aunque el sistema coordenado de variables es expresado en unidades de medias cero y varianzas unitarias.

1.5 PRUEBAS DE SIGNIFICANCIA

La sección anterior fue desarrollada bajo la suposición de po--

blaciones normalmente distribuídas. Para el presente, la misma suposición será requerida.

Supongamos que N observaciones independientes han sido seleccionadas de una población multinormal con distribución N ($\underline{\mu}$, V). V -- tiene las raíces características

$$\lambda_1, \lambda_2, \ldots, \lambda_p > 0$$

con vectores característicos asociados

$$\alpha_1, \ldots, \alpha_p$$

El primer resultado que presentaremos, es para probar si el cálculo de componentes principales es necesario. Bartlett demuestra -- que la cantidad

$$\chi^2 = -N \ln \left(|S| \left(\frac{\text{tr}(S)}{p} \right)^p \right)$$

se distribuye aproximadamente como una chi cuadrada con 1/2(p(p+1)) grados de libertad. Si esta prueba no da un resultado significativo entonces no tiene caso el cálculo de componentes principales.

Una prueba similar es aplicada a la matriz de correlación R, -- que prueba si las variables son independientes. Bartlett demuestra que la cantidad

$$\chi^2 = -N \ln |R|$$

se distribuye aproximadamente como chi-cuadrada con 1/2 (p(p-1)) -- grados de libertad.

Anderson trabajó la prueba de las hipótesis

$$\text{Ho: } \lambda_{q+1} = \ldots = \lambda_{q+r}$$

que r de las raíces características intermedias de la matriz de covarianza poblacional V, son iguales. Las q raíces más grandes y las - p - q - r más chicas, no son restringidas ni a sus valores ni multiplicidades. H_1, la hipótesis alternativa, es que alguna de las raíces intermedias son distintas. Por el criterio de razón de verosimilitud, tenemos la estadística

$$\chi^2 = -(N-1) \sum_j \ln l_j + (N-1)\, r \ln \sum_j \frac{l_j}{r}$$

$$j = q+1, q+2, \ldots, q+r$$

Cuando Ho es verdadera, la estadística se distribuye como chi-cuadrada con 1/2(r(r + 1)) - 1 grados de libertad para N grande. Un caso especial de la hipótesis, es cuando q + r = p, o cuando la variación en las r dimensiones es esférica.

En el caso de raíces múltiples, un intervalo de confianza asimptótico al 100(1 - α) por ciento sería

$$\frac{\bar{l}_i}{1 + z_{1/2\alpha}\sqrt{\frac{2}{Nr}}} \leq \lambda_i \leq \frac{\bar{l}_i}{1 - z_{1/2\alpha}\sqrt{\frac{2}{Nr}}}$$

donde

r es la multiplicidad de λ_i

$z_{1/2\alpha}$ denota el 50α por ciento de los puntos de la distribución normal

$$Ti = \frac{1}{r} (l_{q+1} + \ldots + l_{q+r})$$

También existen algunas pruebas asociadas con los vectores característicos. Por ejemplo, la hipótesis

$$Ho: \alpha_i = \alpha_{i_o}$$

que el vector característico asociado con la raíz distinta λ_i de V es igual a algún vector específico α_{i_o}. Anderson obtuvo que la prueba estadística

$$\chi^2 = N(l_i \alpha'_{i_o} S^{-1} \alpha_{i_o} + \frac{1}{T_i} \alpha'_{i_o} S \alpha_{i_o} - 2)$$

se distribuye asimptóticamente como una chi-cuadrada con p-1 grados de libertad cuando Ho es verdadera.

Una prueba para la igualdad de las últimas p-1 raíces características de la matriz de correlación, fue propuesta por Lawley. Esta es equivalente a probar la igualdad de todas las 1/2 (p) (p-1) correlaciones. Bajo la suposición de que la muestra de tamaño N ha sido extraída de una población multinormal N ($\underline{\mu}$, V) y si el i j-ésimo

elemento de V es $\sigma_{ij} = \rho_{ij} \sigma_i$ σ_i, la hipótesis Ho es

$$Ho: \rho_{ij} = \rho$$

$$\forall \quad i \neq j$$

Lawley encontró que como N tiende a infinito, la prueba estadística

$$\chi^2 = \frac{n}{\hat{\lambda}^2} \left(\sum_i \sum_j (r_{ij} - \bar{r})^2 - \sum_{K=1} (\bar{r}_k - \bar{r})^2 \right)$$

para Ho se distribuye como una chi-cuadrada con 1/2 (p+1) (p-2) grados de libertad, donde

$n = N - 1$

$\lambda = 1 - \rho$ es la segunda raíz característica de la matriz de correlación poblacional, cuando Ho es verdadera

$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik}$ es la correlación promedio de la k-ésima variable con las otras variables

r_{ij} correlación de la iésima y j-ésima variable.

$\bar{r} = \frac{2}{p(p-1)} \sum_i \sum_j r_{ij}$ es la gran media de las correlaciones

$\hat{\lambda} = 1 - \bar{r}$ estimador de λ.

Otro caso importante es el de la matriz de correlación con las dos raíces diferentes de multiplicidades q_1 y q_2 respectivamente.

Como la matriz tiene 1's en la diagonal, esto implica que --- $\lambda_1 q_1 + \lambda_2 q_2 = p$.

Sean $l_1 > l_2 > \ldots > l_p$ las raíces características de la - matriz de correlación muestral R y sean los estimadores de λ_1 y - λ_2 los promedios

$$T_1 = \frac{1}{q_1} \sum_{i=1}^{q_1} l_i$$

$$T_2 = \frac{1}{q_2} \sum_{i=q_1+1}^{p} l_i$$

Será suficiente considerar solamente uno de estos estimadores, ya que

$$T_2 = \frac{(p - q_1 - l_1)}{q_2}$$

Anderson encontró que la esperanza asimptótica de T_2 es λ_2 y su varianza asimptótica es:

$$\frac{2\lambda_2^2 \quad (p - q_2 \lambda_2)^2}{N p q_1 q_2}$$

y que cuando $n \longrightarrow \infty$, la cantidad

$$\frac{\bar{T}_2 - \lambda_2}{\lambda_2 (p - q_2 \lambda_2)} \sqrt{\frac{N_D \, q_1 \, q_2}{2}}$$

se distribuye como una normal con media 0 y varianza unitaria. De - aquí se puede construir el intervalo para la raíz poblacional. Esto es, el intervalo al $100(1 - \alpha)$ % de la raíz λ_2 es:

$$\frac{k\,p + 1 - \sqrt{(k\,p + 1)^2 - 4kq_2 \bar{T}_2}}{2\,kq_2} \leq \lambda_2 \leq \frac{k\,p - 1 + \sqrt{(kp - 1)^2 + 4kq_2 \bar{T}_2}}{2\,kq_2}$$

con $k = z_{1/2\alpha} \sqrt{2/\, n\, p\, q_1\, q_2}$

C A P I T U L O 2

OBSERVACIONES ABERRANTES

INTRODUCCION

En este capítulo, se trata el problema de las observaciones aberrantes, es decir, aquellas observaciones que se encuentran "alejadas" del resto de los datos. En la sección 2.1, se presenta en una forma general el concepto de aberrantes, enfocado principalmente al caso multivariado, así como sus posibles causas. Los diferentes métodos que se han propuesto para detectar estas observaciones, son presentados en la sección 2.2.

2.1 ABERRANTES MULTIVARIADOS

Al extraer una muestra de una población, es importante analizar que todas las observaciones cumplan la condición de igual distribución. Esto se debe, principalmente, a que pueden aparecer observaciones inconsistentes (aberrantes), es decir, datos que estén marcadamente desviados de la "masa" de los mismos. Algunos autores clasifican las causas de estas desviaciones de la siguiente manera:

1.- Errores de medición.

Este tipo de errores es ejecutado con algún instrumento de medición.

2.- Errores de ejecución.

Este error se comete en el momento de efectuar cálculos o registrar los valores numéricos.

3.- Variabilidad inherente.

Este tipo de errores es natural e intrínseco de los datos.

Cualquiera que sea la causa de los aberrantes, es necesario contar con criterios objetivos para probar la inconsistencia de los mismos, esto es, detectar las observaciones aberrantes, problema que será tratado en la sección 2.2.

Ahora bien, el concepto de un aberrante univariado como aquella observación que se encuentra alejada del resto de las observaciones, es similar al caso multivariado, es decir, un aberrante multivariado es un valor extremo que se encuentra alejado en el espacio p-dimensional de las demás observaciones en la muestra. Este problema ha sido tratado tanto para muestras univariadas como multivariadas. Para el primer caso, existen numerosos métodos de detección de aberrantes, basados generalmente en estadísticas de orden y propuestos de una manera intuitiva. Asimismo, dichos métodos se manejan, en la mayoría de los casos, bajo la suposición de que la muestra proviene de una población normal y que una o más observaciones han sufrido un cambio en los parámetros de localización y escala (media y varianza).

En el caso de muestras multivariadas, el encontrarse con aberrantes no es un problema fácil de tratar, ya que, primero, un aberran

te multivariado, además de afectar los parámetros de localización y escala, también puede afectar los de orientación (correlación); en segundo lugar, un aberrante univariado puede pensarse como un valor extremo que se encuentra alejado del resto de las observaciones, pero esta idea no es tan sencilla en dimensiones mayores. Por último, es la variedad de tipos de aberrantes multivariados que puede haber, esto es, un vector respuesta puede ser un aberrante a causa de una falla en una de sus componentes, o bien, a causa de un error sistemático en cada una de ellas. En el presente trabajo, sólo serán presentados los métodos para detectar aberrantes multivariados.

2.2 METODOS DE DETECCION.

La noción básica de un aberrante como aquella observación que se aleja del conjunto de datos, involucra una forma de ordenamiento de los mismos. Barnett y Lewis (1978), proponen y definen algunas formas de subordenamiento y afirman que no existen principios de ordenamiento total.

Así, un vector de observaciones $\underline{X}$, se puede representar por medio de alguna métrica o medida de distancia, por ejemplo:

$$R(\underline{X};\ \underline{Xo},\ \ ) = (\underline{X} - \underline{Xo})' \quad (\underline{X} - \underline{Xo})$$

donde $\underline{Xo}$ puede ser el vector de medias y la matriz de varianzas y covarianzas. Entonces, si la muestra $\underline{X}_1, \underline{X}_2 \ldots, \underline{X}_N$ se ordena de acuerdo a los valores ordenados de $R_i\ (\underline{Xo},\ \ ) = R(\underline{X}_i;\ \underline{Xo},\ \ )$, es llamado

ordenamiento reducido, y esto puede ser usado como un marco de trabajo sobre el cual expresar la "extremidad" de ciertos elementos de la muestra.

Los métodos para detectar aberrantes los hemos clasificado en pruebas de hipótesis y métodos informales. Las pruebas que se han propuesto para declarar una observación como aberrante (discordante), han surgido, en la mayoría de los casos, de bases intiutivamente razonables. Es por esto que tenemos que considerar el grado de arbitrariedad con que se cuenta al tener que emplear el sub-ordenamiento en lugar del ordenamiento total de la muestra.

2.2.1 PRUEBAS DE HIPOTESIS

En esta sección presentaremos algunos resultados que se han propuesto para la detección de aberrantes de muestras normales.

Supongamos que $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N$ es una muestra de N observaciones que provienen de una población normal p-dimensional $N(\underline{\mu}, V)$, donde $\underline{\mu}$ es el vector de medias y V es la matriz de covarianza.

Entonces, un modelo alternativo especificando las hipótesis alternativas, sería

Modelo A:

$$E(\underline{X}_i) = \underline{\mu} + \underline{a} \quad \text{(alguna i)}$$

$$E(\underline{X}_j) = \underline{\mu} \qquad (j \neq i)$$

con matriz de covarianza $V(\underline{X}_j) = V$

$$j = 1, 2, \ldots, N$$

Modelo B:

$$V(\underline{X}_i) = bV \qquad (\text{alguna } i \text{ y } b > 1)$$

$$V(\underline{X}_j) = V \qquad (j \neq i)$$

con vector de medias $E(\underline{X}_j) = \underline{\mu}$.

$$j = 1, 2, \ldots, N$$

Bajo la hipótesis alternativa del modelo A y en la suposición de V conocida, declararemos como el aberrante $\underline{X}_{(N)}$ aquella observación - $\underline{X}_i$ para el cual $R_i(\underline{\bar{X}}, V) = (\underline{X}_i - \underline{\bar{X}})' V^{-1} (\underline{X}_i - \underline{\bar{X}})$ es un máximo, tal que las observaciones han sido ordenadas en términos de - - - -- $R(\underline{X}; \underline{\bar{X}}, V)$. Declararemos a $\underline{X}_{(N)}$ como un aberrante discordante si

$$R_{(N)}(\underline{\bar{X}}, V) = (\underline{X}_{(N)} - \underline{\bar{X}})' V^{-1} (\underline{X}_{(N)} - \underline{\bar{X}}) =$$

$$\text{máximo } R_j(\underline{\bar{X}}, V)$$
$$j = 1, 2, \ldots, N$$

es significativamente grande.

La distribución nula de $R_{(N)}$ $(\underline{\bar{X}}, V)$ no ha sido determinada en -- forma exacta. Siotani[1] (1959) estudia los problemas en determinar -- puntos de porcentaje de $R_{(N)}$ $(\underline{Xo}, \Gamma)$ cuando $\Gamma = V$ y $\underline{Xo}$ es $\underline{0}$, $\underline{\mu}$ o $\underline{\bar{X}}$.

La suposición de V conocida, en general no es real. Entonces, bajo la hipótesis alternativa del modelo A con V desconocida, tenemos lo siguiente:

$$\text{Sea } A = \sum_{j=1}^{N} (\underline{X}_j - \underline{\bar{X}})(\underline{X}_j - \underline{\bar{X}})'$$

$|A^{(i)}|$ el determinante de la matriz obtenida por eliminar $\underline{X}_i$ e i, es escogida a maximizar

$$-\frac{N}{2} \log |A^{(i)}| - \left(-\frac{N}{2} \log |A|\right)$$

Se ordenan los $|A^{(j)}|$ y la observación correspondiente con el - valor más pequeño de $|A^{(j)}|$ es declarado un aberrante. Si denotamos

$$\mathcal{R}_{(j)} = \frac{|A^{(j)}|}{|A|}$$

las observaciones son ordenadas de acuerdo a las $\mathcal{R}_{(j)}$ ordenadas, y el aberrante es aquella observación correspondiente con la $\mathcal{R}_{(j)}$ más pequeña, $\mathcal{R}_{(1)}$, y $\mathcal{R}_{(1)}$ será discordante si su valor es significativamente bajo.

La distribución de la prueba estadística es muy complicada. Exis

1) De Barnett y Lewis (1978)

te muy poca literatura sobre la distribución conjunta de las $\mathcal{R}_{(j)}$ y del mínimo $\mathcal{R}_{(1)}$.

Wilks ha propuesto resultados para probar 2, 3 ó 4 aberrantes, considerando la razón

$$\mathcal{R}_{j_1, j_2, \ldots, j_s} = \frac{\left| A^{(j_1, j_2, \ldots, j_s)} \right|}{|A|}$$

donde $\left| A^{(j_1, j_2, \ldots, j_s)} \right|$ es el determinante de A eliminando de la muestra $X_{j_1}, X_{j_2}, \ldots, X_{j_s}$. Así, el subconjunto de observaciones que minimice $\mathcal{R}_{j_1, j_2, \ldots, j_s}$ es declarado un subconjunto de aberrantes y es declarado discordante si

$$r_s = \min \mathcal{R}_{j_1, j_2, \ldots, j_s}$$

es significativamente pequeño.

Bajo la hipótesis alternativa del modelo B, supongamos de nuevo a $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N$ una muestra aleatoria de una población normal $N(\underline{\mu}, V)$ con $\underline{\mu}$ y V desconocidas.

Así como para el modelo A, declararemos un aberrante aquella observación con distancia máxima $R_j(\overline{\underline{X}}, S)$ y será discordante si el máximo -- $R_N(\overline{\underline{X}}, S)$ es suficientemente grande, donde $\overline{\underline{X}}$ y S son los estimadores - de $\underline{\mu}$, y V respectivamente.

Es importante mencionar que cuando μ y V son desconocidas, es lo mismo adoptar la formulación del modelo A o el modelo B. En -- cualquier caso, las pruebas tienen la misma forma.

Para ilustrar lo anterior, en la figura 2.1 se presenta la gráfica de las observaciones de una encuesta sobre salario y edad de ingenieros eléctricos en el Reino Unido (Barnett, 1978).

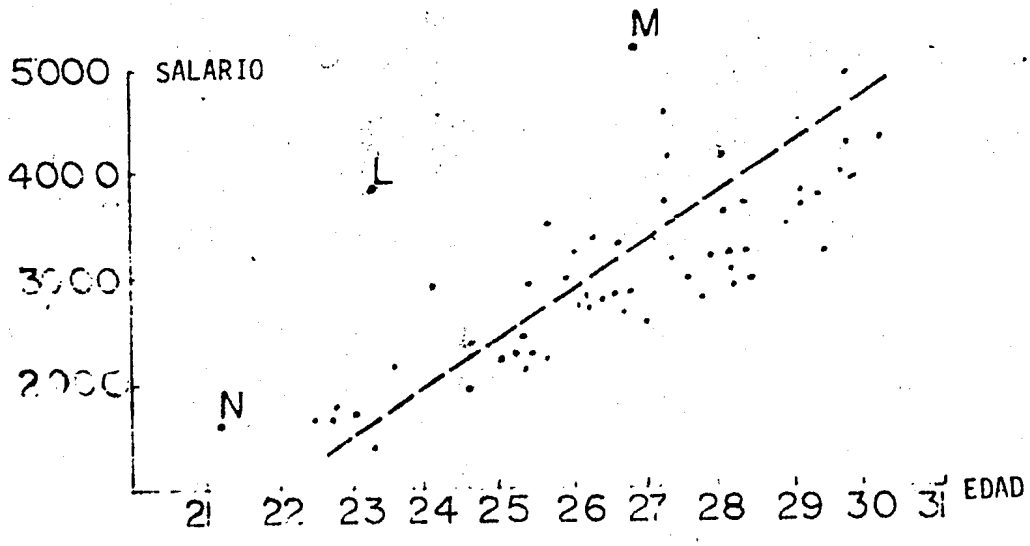


FIGURA 2.1

Las observaciones L, M y N parecen ser aberrantes, ya que éstas dan los tres valores más grandes de R_i $(\bar{X}, S)$

Para

$$M: \quad R_{(55)} \ (\bar{X}, S) = 13.7$$

L: $R_{(54)}\ (\underline{X}, S) = 9.2$

N: $R_{(53)}\ (\underline{X}, S) = 6$

De la tabla del apéndice que presenta valores críticos para -- pruebas de discordancia al 1 y 5 por ciento de un aberrante en una -- muestra multivariada, vemos que M es un aberrante discordante. Las tablas de Wilks dan valores críticos al nivel 5% y 2 1/2% de 12.52 y -- 13.58 respectivamente. Así, M es discordante al nivel 2 1/2%. En la prueba de Wilks, examinando la pareja (M,L), tenemos $\sqrt{r_2} = 0.769$ el cual es significativo al 5%.

Por último, la prueba de la abertura de Rolhf, que consiste en -- comparar aberturas grandes entre los datos, también es aplicada para - declarar la discordancia de un aberrante. La idea de este procedimiento es analizar la distancia más grande en el mínimo árbol generado por los datos en la muestra, y una observación es declarada aberrante discordante si ésta se encuentra aislada y conectada solamente a un punto en el mínimo árbol generado por una distancia significativamente grande.

De acuerdo a lo expuesto anteriormente, en la figura 2.2, el punto A sería un aberrante.

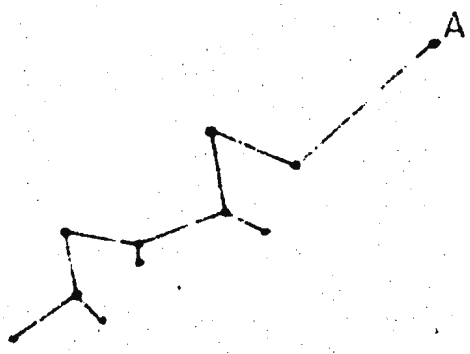


FIGURA 2.2

Podemos resumir el procedimiento anterior de la siguiente manera:

a) Deben estimarse las desviaciones estándar de cada componente del vector aleatorio, eliminando, si es posible, aberrantes - marginales.

b) Se debe estandarizar cada componente del vector aleatorio, -- con el propósito de igualar el efecto de las variables con varianza diferente

$$X^{*}_{ki} = \frac{X_{ki}}{P_{k}}$$

$k = 1, 2, \ldots, p$

$i = 1, 2, \ldots, N$

c) Encontrar el mínimo árbol generado por los datos en la mues-- tra, usando

$$d_{ij} = \left(\sum_{k=1}^{P} X^{*}_{ki} - X^{*}_{kj} \right)^2 / p \Big)^{1/2}$$

como una medida de proximidad entre cualesquiera dos puntos y tomar

$$d_i = \min_{j \neq i} d_{ij}$$

$$i = 1, 2, \ldots, N-1$$

como la longitud de los k arcos en el árbol generado.

d) Probar si

$$G = \frac{\max (d_i^2)}{\bar{d}^2}$$

donde $\bar{d}^2 = \frac{1}{N-1} \sum_{i=1}^{N-1} d_i^2$

es significativamente grande y tomar el punto correspondiente como un aberrante discordante.

Bajo la hipótesis Ho, Rolhf encontró que las distancias al cuadrado d_i^2 tienen una distribución aproximadamente gamma. Asimismo, encontró una tabla de cotas superiores para los valores críticos de la distribución de G, la cual resulta aproximadamente una beta.

2.2.2 DETECCION INFORMAL

La detección informal de aberrantes es considerada más bien como una ayuda intuitiva en la detección de los mismos, ya que no se considera ningún tipo de pruebas de discordancia.

En muestras univariadas, la observación "sospechosa" a ser aberrante puede ser obvia sobre simple inspección de los datos. Pero como mencionamos anteriormente, el concepto de "extremidad" no es tan sencillo en el caso multivariado. Para esto, existen algunas técnicas que sirven como ayuda en la detección de aberrantes, tales como - el estudio de componentes marginales individuales de las observaciones, reducción de las observaciones multivariadas como combinaciones lineales de componentes, cambios en las bases coordenadas de las observaciones y métodos de representación gráfica apropiados. A continuación se presenta una breve descripción de algunos de estos métodos.

ABERRANTES MARGINALES.

Las muestras marginales son las muestras univariadas de cada valor componente de los datos multivariados, y tienen una gran importancia en la detección de aberrantes, primero porque tenemos facilidades para probar la discordancia de aberrantes univariadas para diferentes modelos básicos y podemos adoptar modelos para explicar los aberrantes, y segundo, porque es posible esperar aberrantes que estén dentro de - componentes específicos. Esto es verdadero cuando los aberrantes son originados por errores de medición, donde casi inevitablemente una com

ponente de las observaciones multivariadas será afectada, sin afectar la medición de otras componentes.

RESTRICCIONES LINEALES.

Si establecemos una relación entre las componentes multivariadas o entre los valores esperados de las componentes, podemos facilitarnos la detección de aberrantes. Un ejemplo, (Barnett 1978), pueden ser - los 3 ángulos del triángulo de una encuesta geográfica. En este caso, la suma de los ángulos, aparte de los errores de medición, debe ser 180°, y esto puede servir como indicadores de aberrantes.

METODOS GRAFICOS.

Una gran variedad de métodos gráficos han sido propuestos como ayuda en la detección de aberrantes. Uno de los métodos más usados es graficar 2 de las p variables. En el ejemplo de los datos obtenidos en la encuesta de edad y salario del Reino Unido, notamos que las observaciones L y M están separadas del conjunto de datos (fig. 2.1) y N parece ser sospechosa. La observación L puede tener mayor efecto en reducir la correlación entre edad y salario y N parece que produce una variación más grande en edad y salario. Este efecto de N es más notorio si proyectamos las observaciones sobre la recta AB (fig. 2.3) mientras que L y M no serían aberrantes. Si proyectamos las observaciones sobre la perpendicular a AB, (figura 2.4), L y M serían aberrantes y N no.

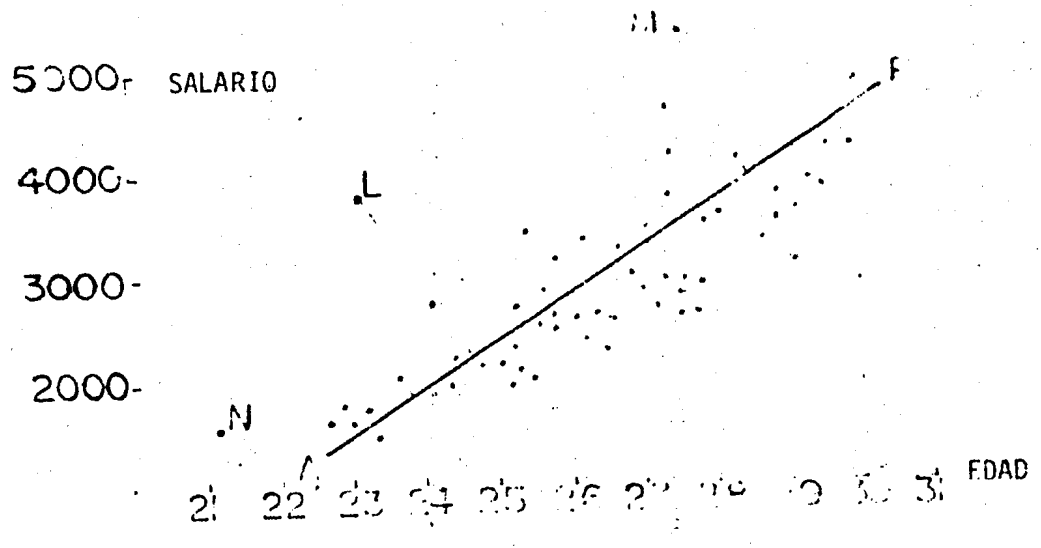


FIGURA 2.3

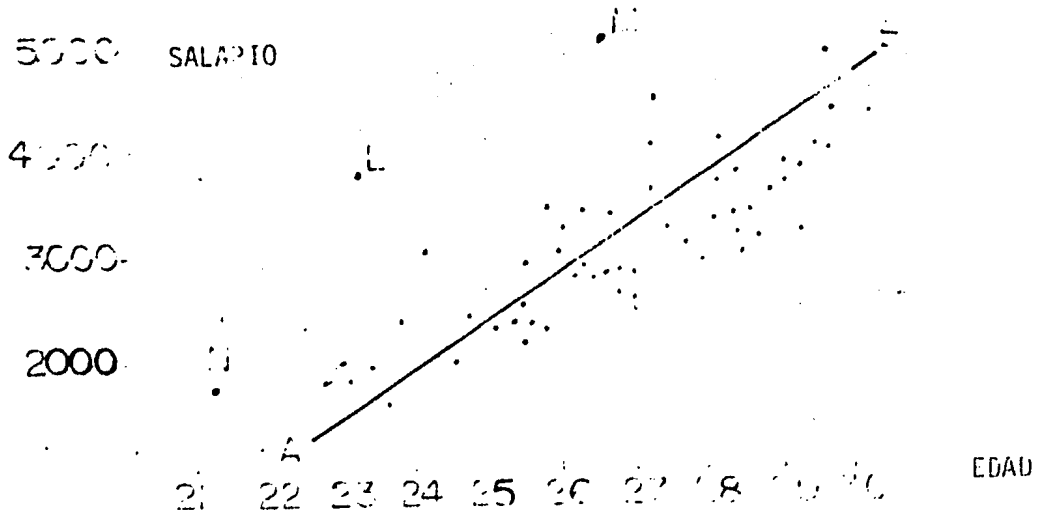


FIGURA 2.4

Otro método gráfico que podría sernos útil, sería el de graficar las $R_j(\underline{\mu},V)$ ordenadas, esto es $R_1(\underline{\mu}, V)$, $R_2(\underline{\mu}, V),\ldots, R_N(\underline{\mu}, V)$ - contra los valores esperados de las estadísticas de orden de una muestra de tamaño N de una χ_p^2.

Para el caso bivariado, la distribución χ^2 tiene 2 grados de libertad y es una distribución exponencial con media 2 (Healy, 1968). La esperanza de las estadísticas de orden de una muestra de tamaño N de esta distribución son dadas por

$$\frac{2}{N}, \frac{2}{N}+\frac{2}{N-1}, \frac{2}{N}+\frac{2}{N-1}+\frac{2}{N-2}, \ldots, \frac{2}{N}+\frac{2}{N-1}+\ldots+\frac{2}{1}$$

Es importante mencionar que para un número grande de grados de libertad, la $\sqrt[2]{\chi^2}$ y $\sqrt[3]{\chi^2}$ pueden ser usadas como transformaciones normalizantes. Este procedimiento parece ser adecuado en la detección de aberrantes, y será ilustrado con el siguiente ejemplo.

En una muestra de 39 pacientes de un hospital se miden el consumo de grasa y niveles de colesterol (en logaritmos), para observar el efecto de la dieta en la sangre (Healy, 1968). En las figuras 2.5 y 2.6 se grafican las distancias ordenadas $\sqrt{R_j(\underline{\mu}, V)}$ contra las estadísticas de orden ya mencionadas. En la figura 2.5, se observan cuatro aberrantes. En la figura 2.6, se omiten estos cuatro valores, y se puede observar un razonable ajuste lineal.

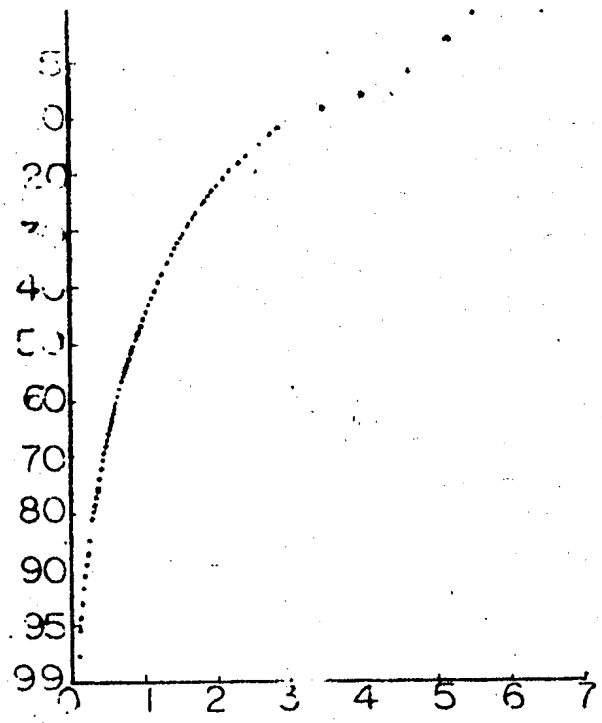


FIGURA 2.5

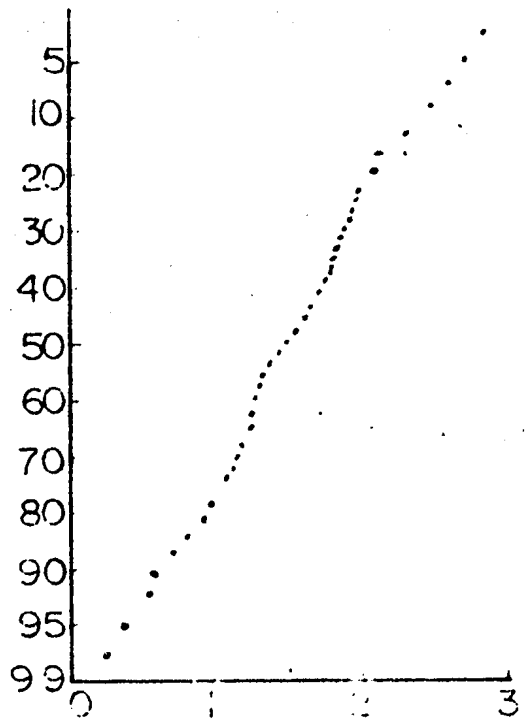


FIGURA 2.6

Los métodos gráficos tienen varias ventajas:

1a. La misma gráfica puede indicar el aberrante como las observaciones que están en la periferia de la "masa" de datos.

2a. Si alguna medida agregada, tal como el coeficiente de correlación, es alterada, puede revelar la presencia de aberrantes (en la figura 2.1, L y M).

3a. Un cambio de bases coordenadas, y la representación de los datos sobre las nuevas bases, puede revelar la presencia de aberrantes.

4a. La rotación de los ejes coordenados en dirección de su perpendicular, puede ayudar a identificar aberrantes.

Por último, cabe hacer mención que el método propuesto por Rolhf, que ya fue descrito en la Sección 2.2.1, puede ayudar a visualizar gráficamente a los aberrantes, pero solamente para el caso bivariado. Una posible aplicación de dicho método a datos multivariados, sería graficar las componentes principales, y en base al sistema coordenado generado por éstas, graficar el mínimo árbol generado de las observaciones.

DISTANCIAS GENERALIZADAS.

Otro procedimiento informal que puede ayudar a identificar aberrantes, es construir medidas univariadas reducidas.

Gnanadesiken y Kettenring (1972)[1], proponen las medidas:

$$\text{I:}\quad (\underline{X}_j - \underline{\bar{X}})' \; S^b \; (\underline{X}_j - \underline{\bar{X}})$$

$$\text{II:}\quad (\underline{X}_j - \underline{\bar{X}})' \; S^b \; (\underline{X}_j - \underline{\bar{X}}) \;/\; \left((\underline{X}_j - \underline{\bar{X}})' \; (\underline{X}_j - \underline{\bar{X}}) \right)$$

y para la clase I consideran los casos b = 0, -1 y 1, que denotaron - por q_j^2, d_j^2 y t_j^2 respectivamente, y llegaron a los sig. resultados:

1) $q_j^2 = (\underline{X}_j - \underline{\bar{X}})' \; (\underline{X}_j - \underline{\bar{X}})$

Esta forma de distancia es útil para detectar observaciones que - afectan el parámetro de escala.

2) $t_j^2 = (\underline{X}_j - \underline{\bar{X}})' \; S \; (\underline{X}_j - \underline{\bar{X}})$

Esta medida es útil para detectar aberrantes que afectan la orien tación y escala de las primeras componentes principales de S.

3) $d_j^2 = (\underline{X}_j - \underline{\bar{X}})' \; S^{-1} \; (\underline{X}_j - \underline{\bar{X}})$

Es conveniente usarla para observaciones que estén lejos de la ma- sa general de puntos.

Representación Tipo Fourier:

Andrews sugiere que $\underline{X}_j = (X_{1j}, X_{2j}, \ldots, X_{pj})'$ debe ser represen- tado por la función

1) Barnett y Lewis (1978).

$$FX_j(t) = \frac{X_{1j}}{\sqrt{2}} + X_{2j} \operatorname{sen} t + X_{3j} \cos t + X_{4j} \operatorname{sen} 2t + X_{5j} \cos 2t \ldots$$

sobre el rango de $(-\pi, \pi)$ para t. Cada observación aparece en el espacio muestral como una curva sobre todos los valores de t. Esto -- puede revelar ciertos rasgos distintivos en los datos.

METODO DE ANALISIS DE COMPONENTES PRINCIPALES.

La ecuación $Y = AX$ define la transformación a componentes principales lineales de los datos, en términos de los eigenvectores de la matriz de covarianza muestral S. Cada columna de A nos da una coordenada componente principal y cada renglón de Y nos da la proyección sobre las coordenadas componentes principales de la desviación de las N observaciones originales alrededor de $\bar{X}$.

Gnanadisikan dice: "Cuando el análisis de componentes principales es visto como un método de ajuste de subespacios lineales, o como una técnica estadística para detectar posibles singularidades lineales en los datos, el interés yace especialmente en las proyecciones de los datos sobre las coordenadas componentes principales correspondientes a los eigenvalores más pequeños".

Así, por ejemplo, para p = 2, esto puede ilustrarse en la figura 2.7. La recta de mejor ajuste a los datos es el eje Y_1.

El vector $\overrightarrow{QP}$ es el residual ortogonal del punto P, y $\overrightarrow{QP}$ es equivalente al vector $\overrightarrow{O'P'}$ y P' es la proyección de P en el eje Y_2, -

la segunda componente principal.

Con datos p dimensionales, la proyección sobre la componente principal más pequeña, sería relevante para estudiar la desviación de una observación de un hiperplano ajustado, mientras que las proyecciones sobre las q componentes principales más pequeñas sería relevante para estudiar la desviación de una observación de un subespacio lineal ajustado de dimensión (p - q).

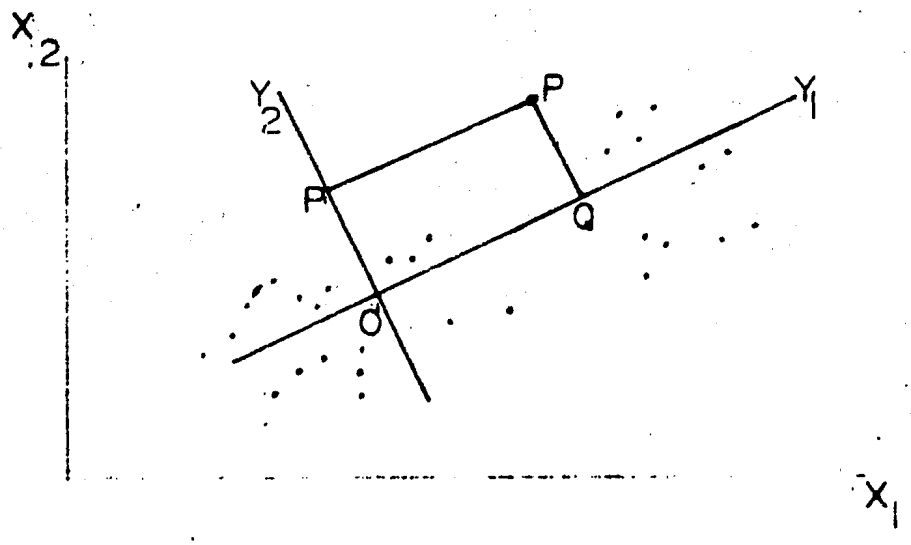


FIGURA 2.7

Rao sugiere un método para detectar la falta de ajuste de observaciones individuales, que consiste en analizar la suma de longitudes cuadradas de las proyecciones de las observaciones sobre las últimas q coordenadas componentes principales. Para cada observación $\underline{X}_i$, se calcula

$$d_i^2 = \sum_{j=p-q+1}^{p} (\underline{a}_j' (\underline{X}_i - \overline{\underline{X}}))^2 =$$

$$= (\underline{X}_i - \overline{\underline{X}})' (\underline{X}_i - \overline{\underline{X}}) - \sum_{j=1}^{p-q} (\underline{a}_j' (\underline{X}_i - \overline{\underline{X}}))^2$$

y considerar los valores grandes de d_i^2 como posibles aberrantes.

Puede pensarse también en las primeras componentes principales como un medio de investigar la presencia de aberrantes. La construcción de diagramas de parejas de Y_i ya sean las primeras o las últimas, pueden gráficamente exhibir aberrantes. Además las pruebas de aberrantes univariadas pueden ser aplicadas a las Y_i individuales. Si p es razonablemente grande, la transformación lineal involucrada en el análisis de componentes principales puede conducir a que las Y_i sean muestras de distribuciones normales. En tal caso, graficando probabilidades -- normales, en el cual el j-ésimo valor ordenado de Y_i es graficado contra α_j, donde

$$\alpha_j = E(U_j)$$

y U_j es la j-ésima estadística de orden de la distribución normal -- N(0, 1), puede revelar la presencia de aberrantes, los cuales se encuentran como puntos extremos de la relación lineal de la gráfica.

Siguiendo el ejemplo de la encuesta de edad y salario de ingenieros eléctricos del Reino Unido, esto último es ilustrado en las figuras 2.8 y 2.9, graficando primero Y_1 contra α_j y Y_2 contra α_j.

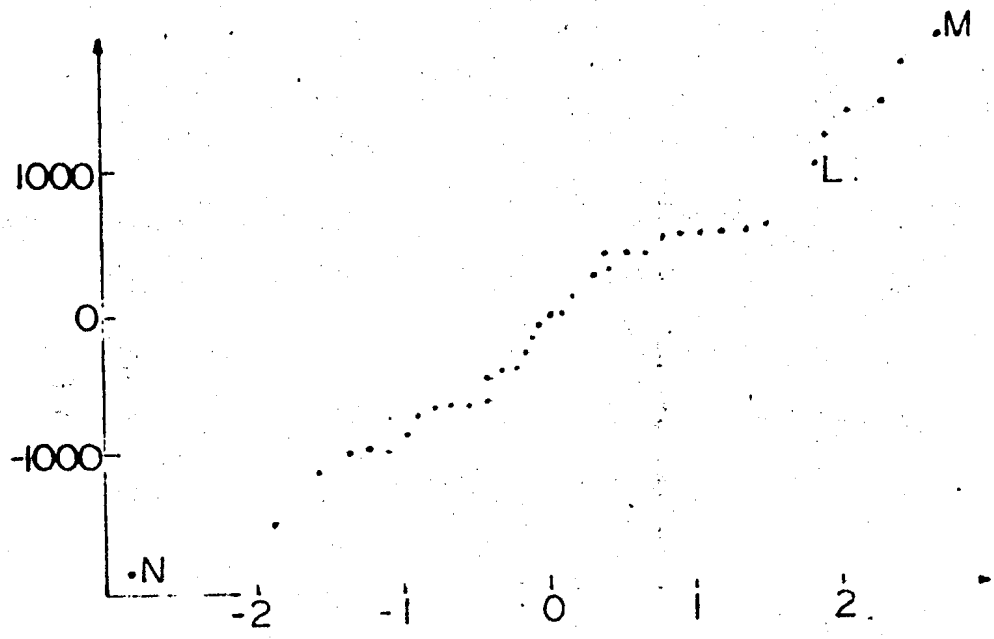


FIGURA 2.8

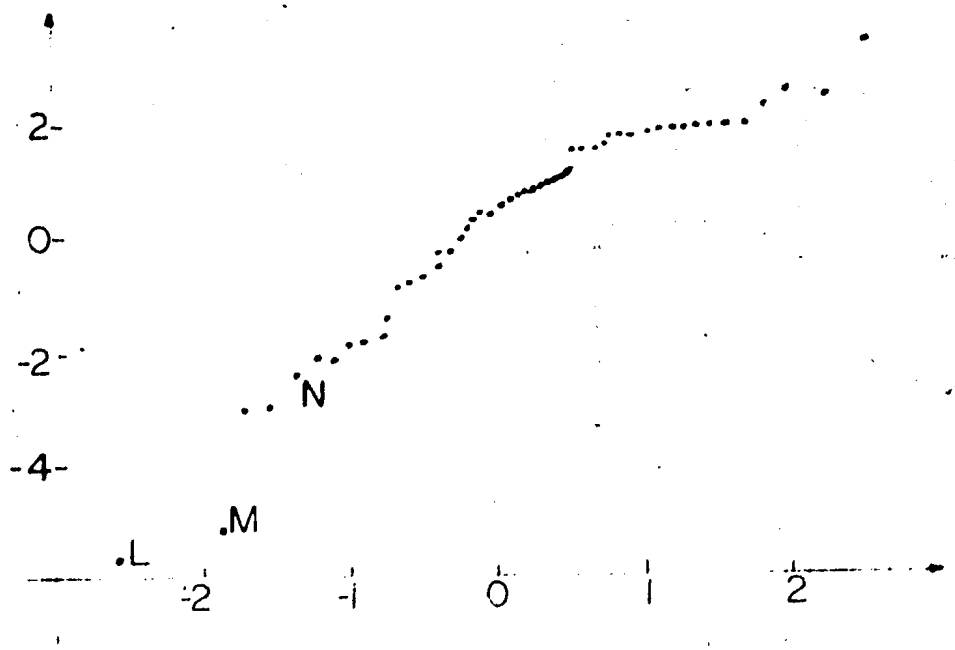


FIGURA 2.9

En la figura 2.8, los aberrantes N y M son notoriamente valores - extremos y L no lo es. En cambio, en las proyecciones de la segunda - componente principal (fig 2.9), L y M son valores extremos y N no, y - están por debajo de la relación lineal, lo cual indica discordancia.

C A P I T U L O 3

RESULTADOS DE LOS METODOS GRAFICOS

INTRODUCCION

En este capítulo trataremos de exponer y probar algunos de los métodos gráficos que fueron propuestos en el capítulo 2. Para ilustrar - lo anterior, utilizando el paquete de sub-rutinas del IMSL (que calcula componentes principales, descompone matrices,etc.), se generaron mues-- tras multivariadas de cinco variables, con datos alejados de la masa de observaciones, de tal manera que alguna prueba como la de Wilks, las declare como aberrantes. El número de estos varió de acuerdo al tamaño - de muestra, lo cual se puede apreciar en la siguiente tabla, marcada - con una cruz.

Tamaño de muestra \ No. de aberrantes	1	5	10
50	X	X	
100	X	X	X

A su vez, en cada una de las muestras, los aberrantes fueron generados en dos desviaciones d_1 y d_2 de tal forma que si $X \sim N(\mu, V)$, ahora -

$$X \sim N(\mu + A\, d_i,\ V) \qquad i = 1,2$$

donde

A matriz triangular superior que resulta de descomponer V de tal forma que $V = A\ A'$

$$d_1 = \begin{bmatrix} 4.2426406 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 3 \\ 0 \\ 0 \\ 3 \\ 0 \end{bmatrix}$$

que forman un ángulo de 45° entre sí y son de la misma magnitud.

Asimismo, las mismas muestras fueron graficadas también sin aberrantes, con el propósito de ilustrar las posibles diferencias entre unas y otras: En las secciones siguientes, se presentan los resultados que se obtuvieron al probar los métodos de componentes principales, gráficas de los marginales, y la raíz cuadrada de la distancia $(\underline{X}_j - \overline{\underline{X}})' \ S^{-1} (\underline{X}_j - \overline{\underline{X}})$.

3.1 COMPONENTES PRINCIPALES

Para probar este método, se calcularon las coordenadas componentes principales y se graficaron estos valores ordenados de Y_j contra α_j, donde $\alpha_j = E(U_j)$ y U_j es la j - ésima estadística de orden de la distribución normal N(0,1). También se graficó por parejas la primera componente principal más grande contra la segunda, así como la quinta componente principal más chica contra la cuarta.

Los resultados de analizar las gráficas variaron de acuerdo al número de aberrantes en las muestras, así como también al tipo de desvia-

ción con la que fueron generadas dichas observaciones (ver Anexo 1). Por ejemplo, se pudo apreciar que en las muestras con un aberrante, y bajo la desviación d_1 , éste se detectó fácilmente en la primera componente principal, debido a que el dato se localizó bastante alejado de la masa de observaciones y fuera de la tendencia lineal, lo cual no sucedió bajo la desviación d_2 , ya que en ésta última el aberrante no se pudo detectar porque se encontró alejado pero dentro del patrón lineal.

En lo que respecta a las gráficas con 5 y 10 aberrantes, el comportamiento de las misma fue similar a las de un aberrante, no obstante que éstos, en la mayoría de los casos, no pudieron ser identificados en su totalidad; por ejemplo, veamos el caso para N = 50 y bajo la desviación d_1 , los cuatro datos extremos coinciden con los aberrantes, ya que éstos se encuentran por debajo de la tendencia lineal, lo cual no sucede para d_2 , debido a que aquí solo una observación fue identificada. Algo similar sucede en las muestras de tamaño 100, ya que también aquí solo bajo d_1 y en la primera componente principal se pudo identificar a dos aberrantes, mientras que la configuración de los datos bajo la desviación d_2 , siguió una tendencia lineal. Por otro lado, las gráficas sin aberrantes, en la mayoría de los casos siguieron un comportamiento diferente a los anteriores, lo que nos da una idea de la información que la primera componente nos puede proporcionar.

Ahora bien, cabe precisar que las muestras con 10 aberrantes, no presentaron resultados de importancia, debido a que las gráficas bajo d_1 y d_2 no exhibieron irregularidad alguna en su comportamiento.

Por otro lado, en lo que se refiere a la segunda componente principal de la cual se esperaba también que mostrara a los aberrantes, sólo en dos casos las gráficas exhibieron a uno de éstos, lo cual sucedió para N = 100 con 10 aberrantes y N = 50 con 5 aberrantes respectivamente. Como puede verse, la información que la segunda componente principal nos proporciona, es mínima.

Es importante mencionar que de las gráficas por parejas de las dos primeras componentes principales, se podría obtener resultados de mayor importancia, ya que en casos donde las gráficas individuales no aportaron información, en estas últimas se pudo apreciar que los datos se encontraron bastante alejados de la masa de obseravciones. Veamos el caso N = 100 con 10 aberrantes, las dos gráficas bajo d_1 y d_2 muestran a 5 observaciones alejadas, las cuales efectivamente son aberrantes. Más aún, el comportamiento de las gráficas sin aberrantes de las dos primeras componentes fue muy diferente a las de con aberrantes, lo cual nos puede hacer notar la ventaja de graficar las dos primeras componentes.

Un comentario importante acerca de este método, es que tanto las gráficas individuales como por parejas de las últimas componentes principales, en ningún momento mostraron ser sensibles a los aberrantes, por lo que, la propuesta de Gnanadesikan de graficar las últimas componentes principales para detectar posibles anormalidades de los datos, no resultó en los casos estudiados en el presente trabajo.

De los resultados obtenidos, podemos concluír que, graficando la primera componente individualmente, o bien, por parejas con la segunda componente principal, contamos con un método gráfico como ayuda en la detección de aberrantes multivariadas.

3.2 MARGINALES

En la presente sección, expondremos los resultados obtenidos en la detección de aberrantes mediante las gráficas marginales de las muestras multivariadas (anexo 2), esto es, los valores extremos que se encuentran en cada una de estas muestras, ya que, como se mencionó anteriormente, es posible esperar aberrantes dentro de componentes específicas, lo cual sucede generalmente a causa de errores de medida que afectan á una componente particular. Aunque no necesariamente las gráficas marginales exhiban a los aberrantes, creemos que quizá alguna anomalía de los datos multivariados, se podrían reflejar en dichas gráficas. Así, se graficaron cada una de las cinco variables contra el valor esperado de las estadísticas de orden, y los resultados, que fueron muy variados, se mencionan a conitnuación.

En las gráficas del anexo 2 se puede apreciar que en las muestras con un aberrante, para N = 50 y bajo d_1, en cuatro casos se identificó el aberrante, y bajo d_2, en 3 gráficas el dato se encontró alejado del conjunto de datos. Cabe hacer mención que las muestras de tamaño 100 y con un aberrante, sólo en un caso exhibieron irregularidad en los da--

tos, lo cual, quizá nos podría hacer pensar en la influencia del tamaño de muestra en la detección de aberrantes.

Por otro lado, en las muestras con 5 aberrantes, en ningún momento estos fueron identificados totalmente, pudiéndose apreciar a lo más, 4 datos alejados, y dudándose, en algunos casos, en declararlos como aberrantes, debido a que se encontraron como extremos pero dentro del patrón lineal.

Ahora bien, para N= 100, con 10 aberrantes, las muestras bajo d_1, en tres gráficas exhibieron a los 10 aberrantes con bastante claridad, lo cual no sucede bajo d_2, ya que, en este caso, los datos se localizaron dentro de la tendencia lineal.

Por lo expuesto anteriormente, podemos decir que, graficando cada una de las muestras marginales, contamos con otro método útil en la identificación de observaciones aberrantes.

3.3 DISTANCIAS ORDENADAS

Otro método que fue probado, es el de las raíces cuadradas de las distancias ordenadas R_j (, V) = $(\underline{X}_j - \bar{\underline{X}})' S^{-1} (\underline{X}_j - \bar{\underline{X}})$. Esto es, se graficaron cada uno de los valores ordenados $\sqrt{R_1}$ (, V), $\sqrt{R_2}$(, V), ... $\sqrt{R_N}$ (, V), contra los valores esperados de las estadísticas de orden de las muestras de tamaño 100 y 50 de una χ^2.

Los resultados obtenidos en las graficas se muestran en el anexo 3 De estos podemos decir que no fueron muy satisfactorios, debido a que - los aberrantes, en general, no pudieron ser identificados, ya que la con figuración de los datos exhibió a los mismos dentro de la masa de observaciones.

Veamos el caso para N = 100. En las muestras con 10 aberrantes y bajo d_1, sólo dos aberrantes fueron detectados, lo que nos puede dar una idea de la utilidad del método. En las muestras con 5 aberrantes, solamente bajo la desviación d_2, se encontró a una observación como valor extremo y fuera de la tendencia lineal.

En lo que respecta a las gráficas con un aberrante, se pudo observar que para N = 100, y bajo d_1, el aberrante se localizó fuera del patrón lineal y como extremo, y en los casos restantes, los datos no mostraron irregularidad en su comportamiento.

Por consiguiente, podemos concluír que el método de la raíz cuadrada de las distancias ordenadas, al menos en los casos aquí estudiados, - no resultó ser sensible a las observaciones aberrantes.

C O N C L U S I O N E S

El problema de detección de observaciones aberrantes multivariadas mediante métodos gráficos, lo podríamos considerar como ayuda en la detección"formal" de los mismos (prueba de hipótesis). Pero creemos que esta ayuda es limitada ya que el concepto de una observación como valor extremo, depende del sujeto que analiza las gráficas.

En el presente trabajo, se estudiaron 3 de los métodos gráficos - que han sido los más usados en la detección de aberrantes, y los comentarios acerca de cada uno de éstos, se pueden resumir en lo siguiente:

1.- Análisis de Componentes Principales.

De los casos estudiados en este método, que fueron las gráficas individuales y por parejas de las componentes principales, podemos decir que graficando la primera componente principal contra la segunda, es posible encontrar resultados importantes, ya que en situaciones donde las gráficas individuales no exhibieron a los aberrantes, aquéllas mostraron a los mismos alejados de la masa de observaciones. Asimismo, de las gráficas individuales de las componentes principales, solamente recomendaría la gráfica de la primera componente, debido a que las gráficas restantes no presentaron irregularidad alguna en el comportamiento de los datos.

2.- Marginales.

Las gráficas de cada una de las muestras marginales bajo la desviación d_1, mostraron a los aberrantes bastante alejados y fuera del patrón lineal, en contraposición con las gráficas bajo la desvia

ción d_2, en las que se detectaron solamente una minoría de éstos. Este método es recomendable cuando los aberrantes son causados -- principalmente por errores de medición.

3.- Raíz cuadrada de las Distancias Ordenadas.

De los métodos que fueron probados, creemos que éste es el menos útil ya que, en muy pocos casos, las gráficas exhibieron a los aberrantes como extremos, sobre todo en las muestras con 5 y 10 aberrantes.

En resumen, creemos que lo más conveniente de los métodos probados es, por un lado, graficar la primera componente principal contra la segunda, así como también las muestras marginales. Por otro lado, creemos que el procedimiento que se menciona a continuación, sería recomendable en la detección de aberrantes:

a) Graficar todas las observaciones.
b) Una vez detectados los aberrantes, eliminarlos.
c) Volver a graficar.

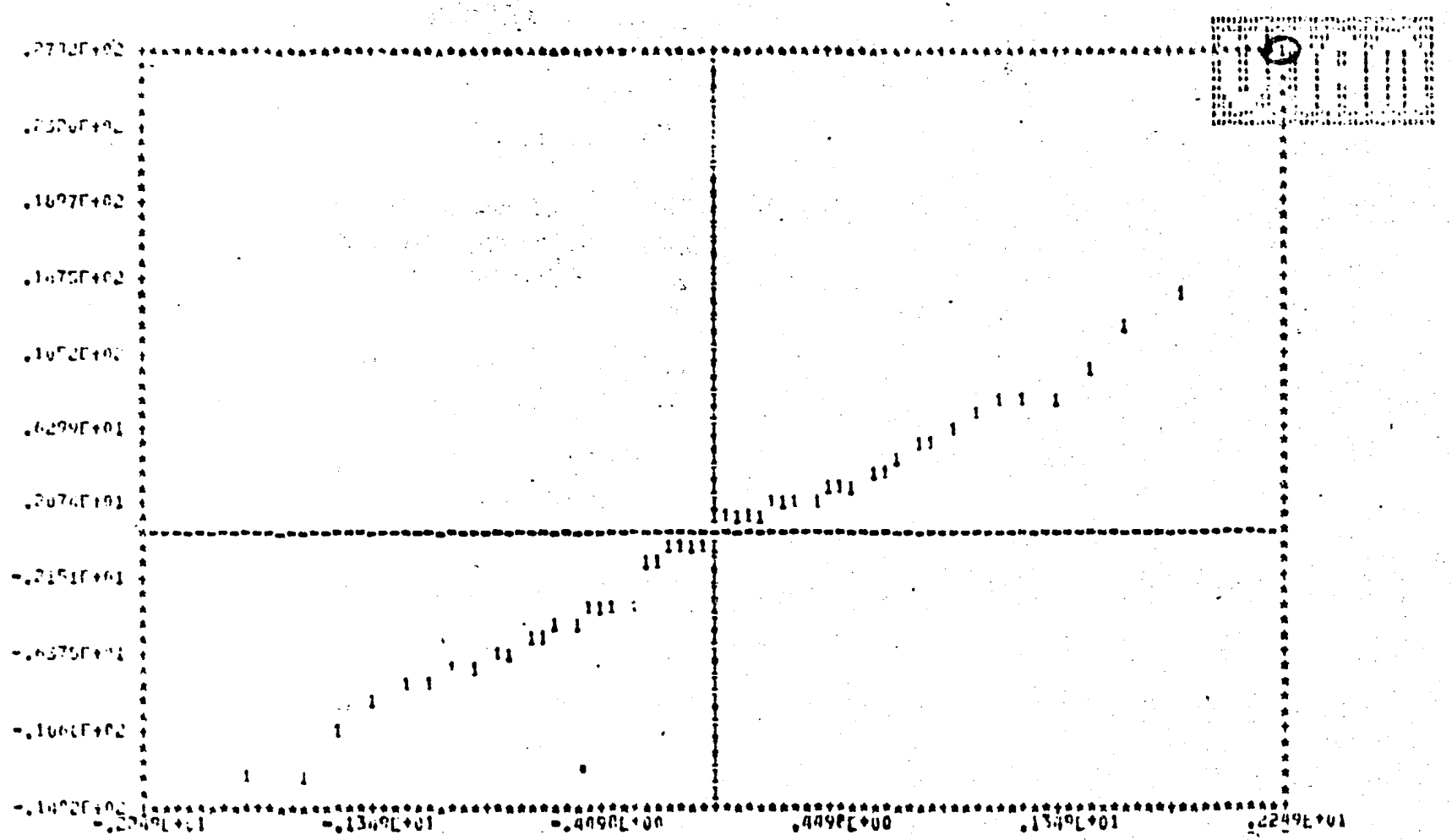
Esto se sugiere para evitar el problema de enmascaramiento de observaciones aberrantes, esto es, que la presencia de algunos aberrantes no permitan la identificación de otros.

A N E X O 1

En este anexo se presentan las gráficas con aberrantes de cada uno de los componentes principales contra C_j, ordenados de mayor a menor. Asímismo, al final de cada grupo de las gráficas individuales se muestran las gráficas por parejas de la primera componente principal más grande contra la segunda, así como la quinta componente principal más chica contra la cuarta. También, en algunos casos de interés, se presentan las gráficas sin aberrantes, para hacer notar la diferencia entre unas y otras.

Las observaciones aberrantes fueron encerradas en un círculo, y con una cruz se marcaron los datos que están alejados de la masa de observaciones a causa de la variabilidad intrínseca de las mismas.

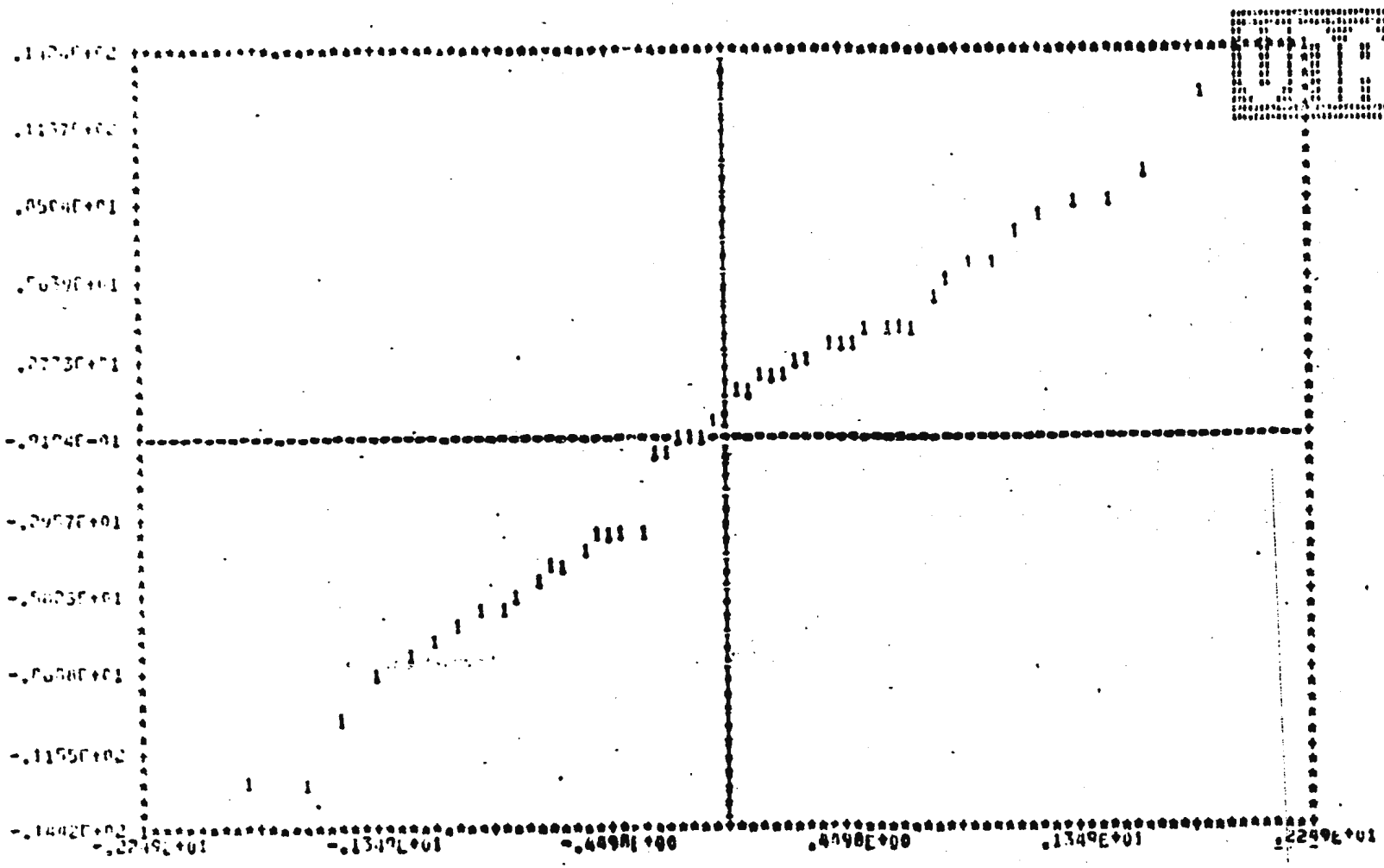
Aquí cabe hacer la aclaración que, debido a la subjetividad de los métodos gráficos, una persona con diferente preparación a la del autor podría quizás detectar más o probablemente menos aberrantes.



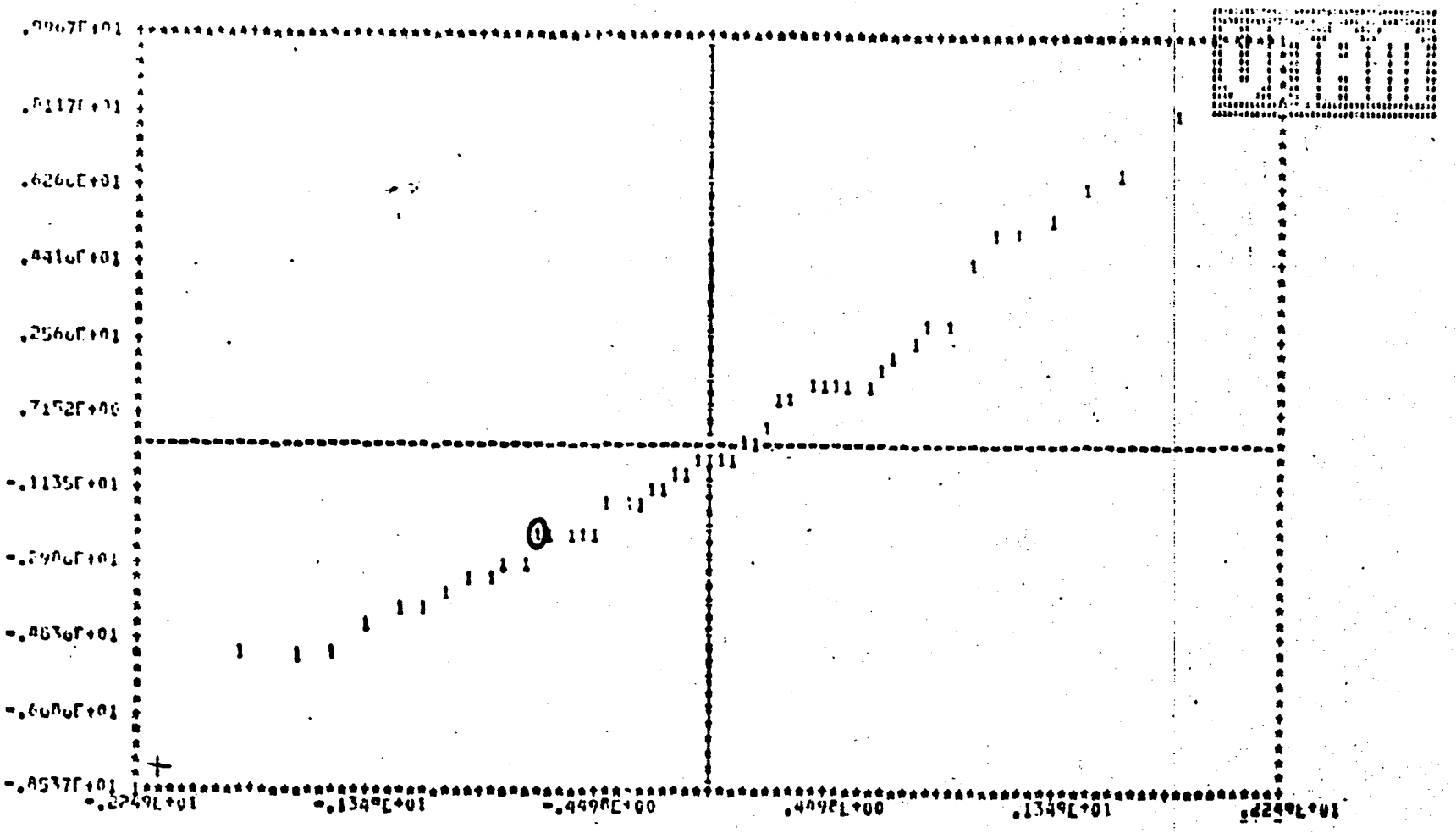
Primera componente principal contra α_j

N = 50, 1 aberrante, bajo d_1

El aberrante como punto extremo y fuera del patrón lineal.



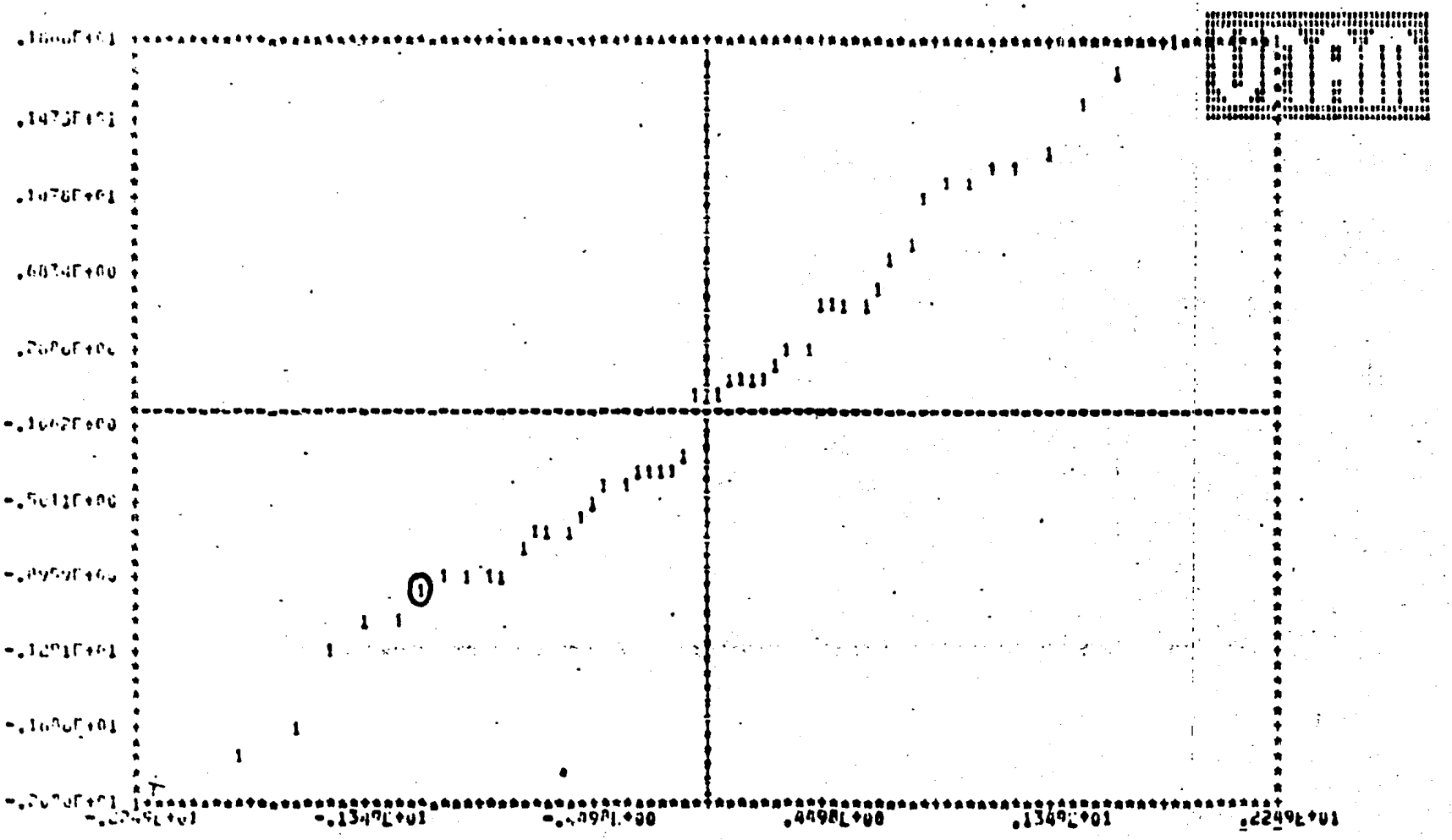
Primera componente principal contra α_j
N=50, sin aberrantes, bajo d_1
No aparece alguna observación sospechosa a ser aberrante



Segunda componente principal contra d_j

N=50, 1 aberrante, bajo d_1

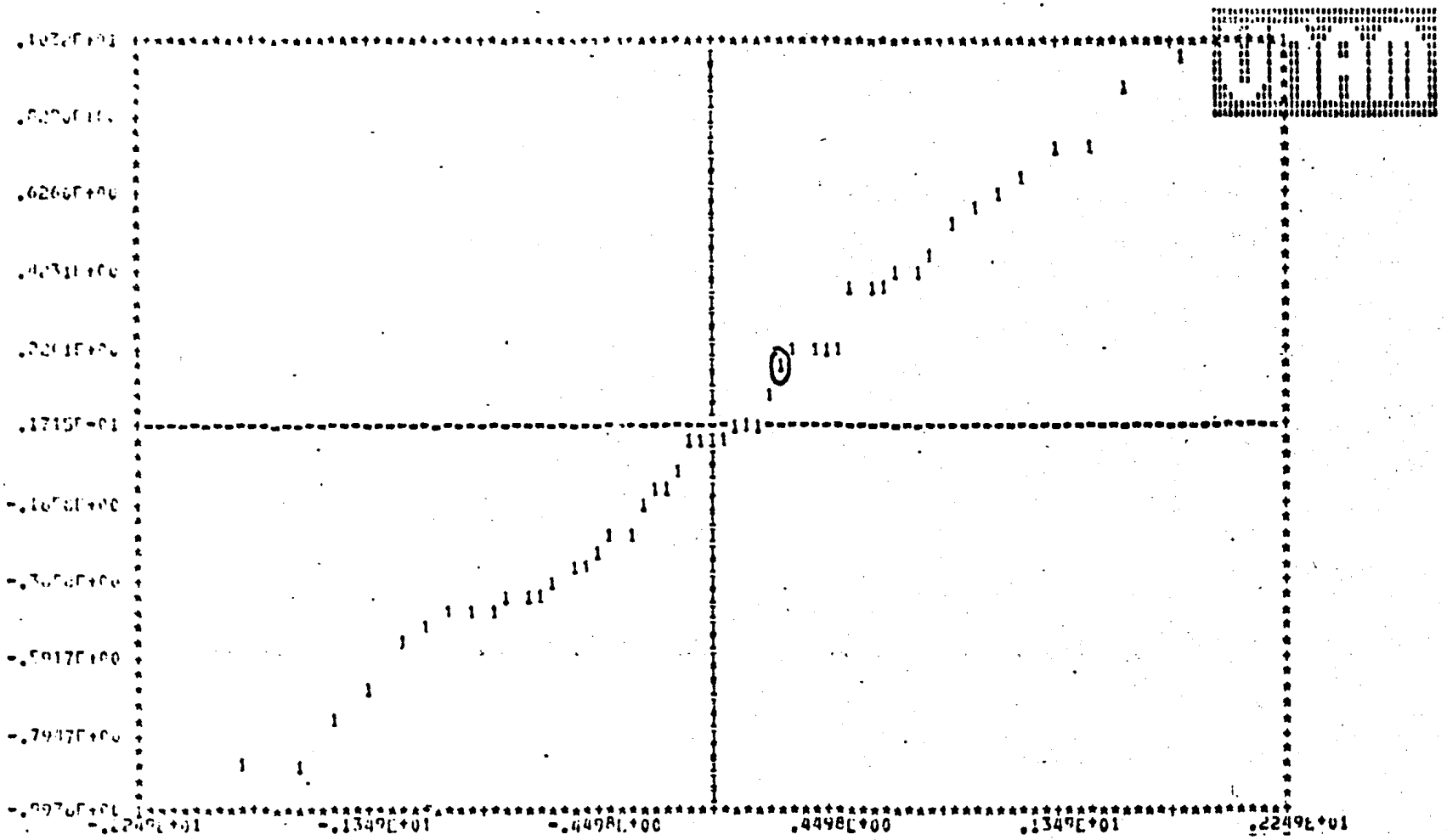
El aberrante dentro del patrón lineal



Tercera componente principal contra c_j

N= 50, 1 aberrante, bajo d_1

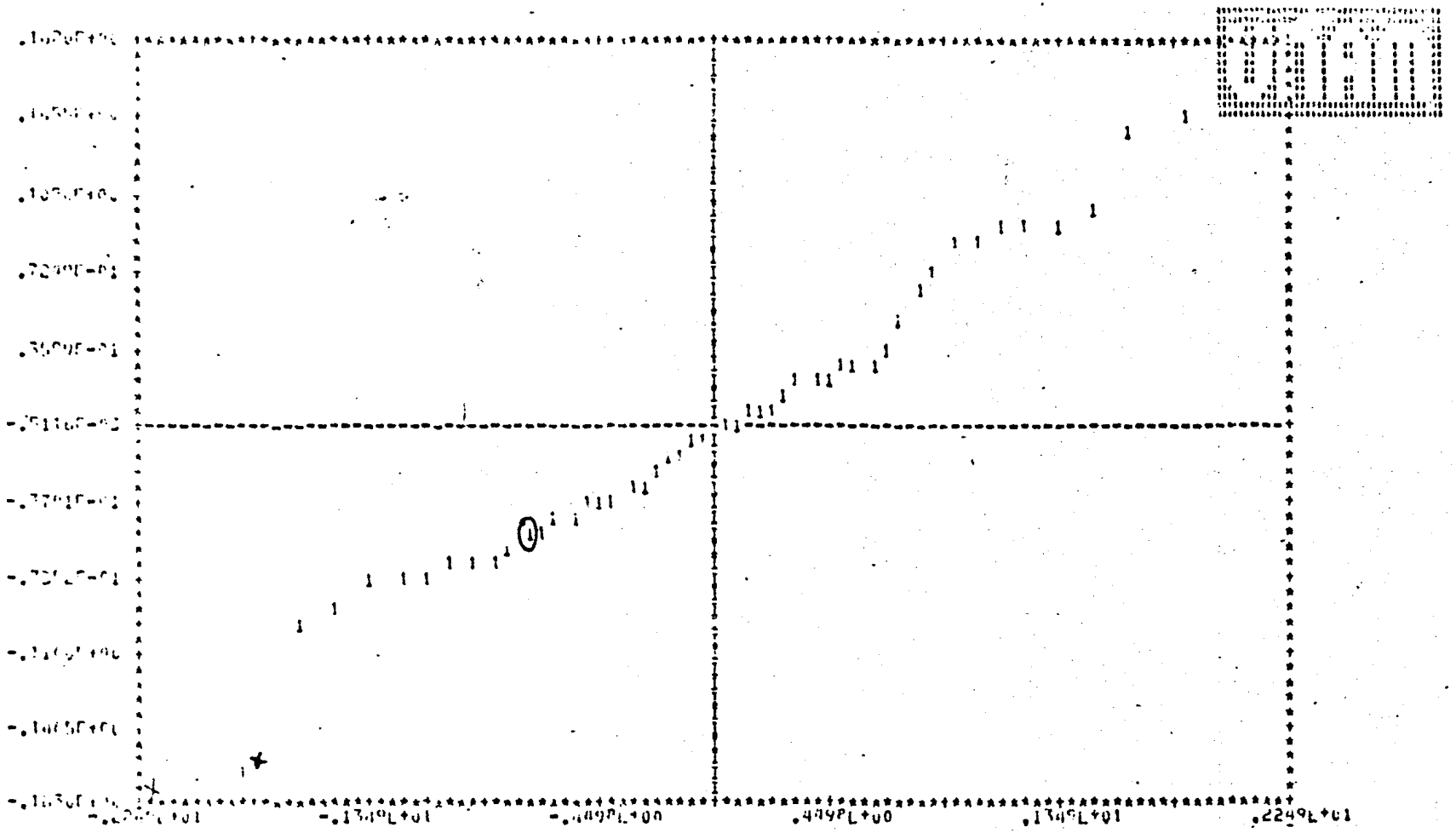
El aberrante dentro del patrón lineal



Cuarta componente principal contra c'_j

N= 50, 1 aberrante, bajo d_1

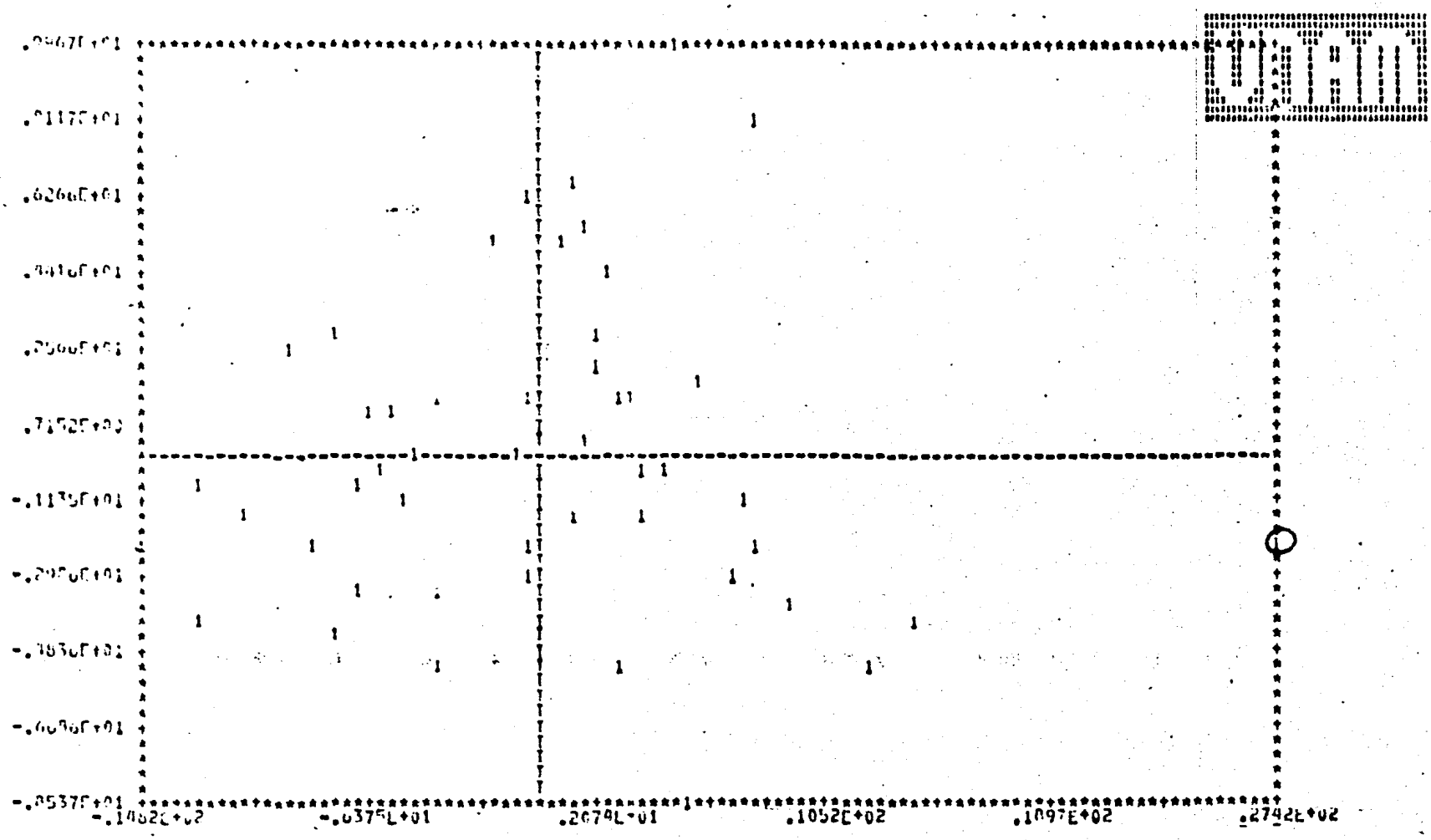
El aberrante dentro del patrón lineal



Quinta componente principal contra σ_j

N= 50, 1 aberrante, bajo d_1

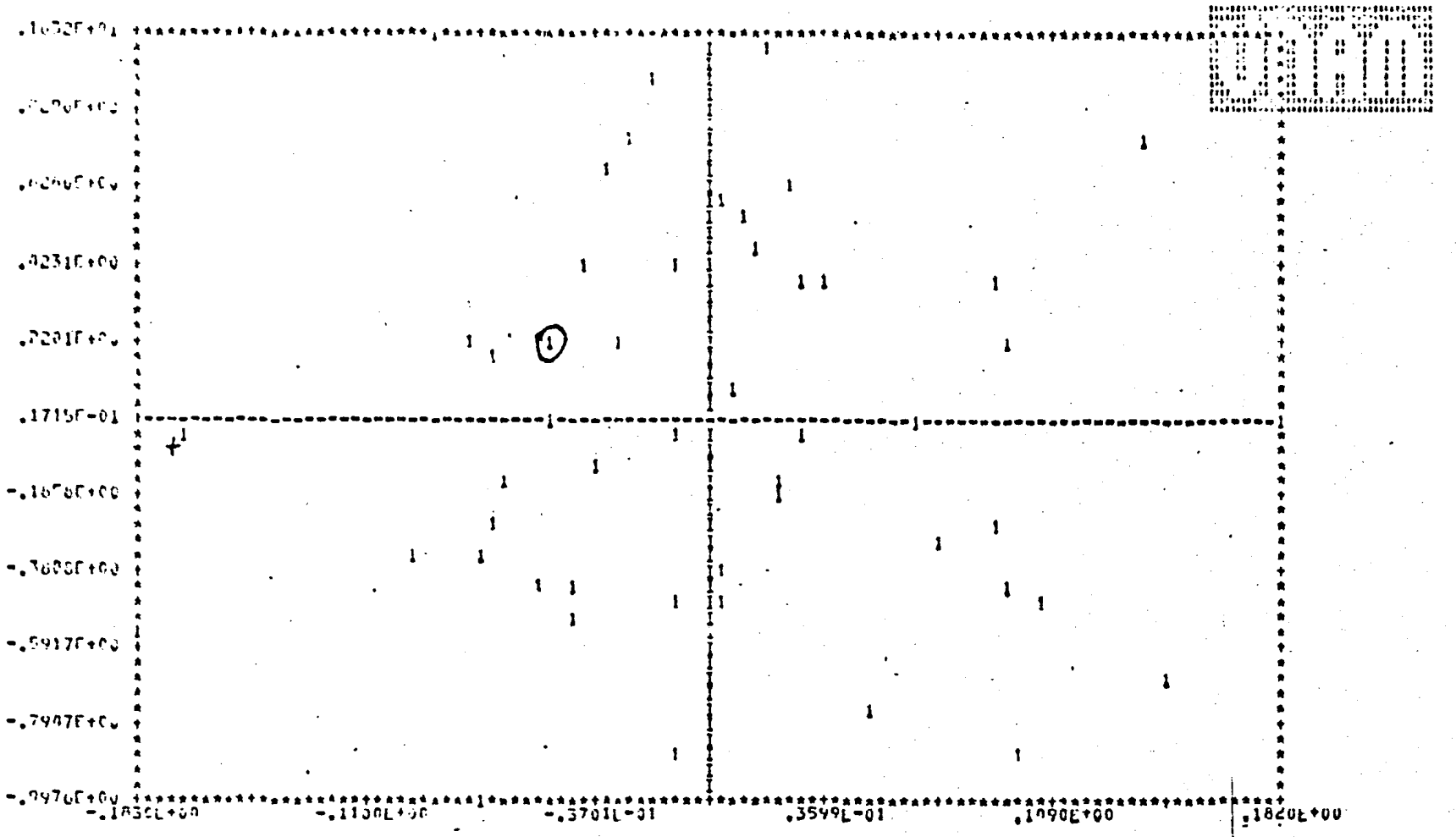
El aberrante dentro de la masa de observaciones y las dos observaciones extremas serían sospechosas a ser aberrantes



Primera componente principal contra segunda componente principal

N=50, 1 aberrante, bajo d_1

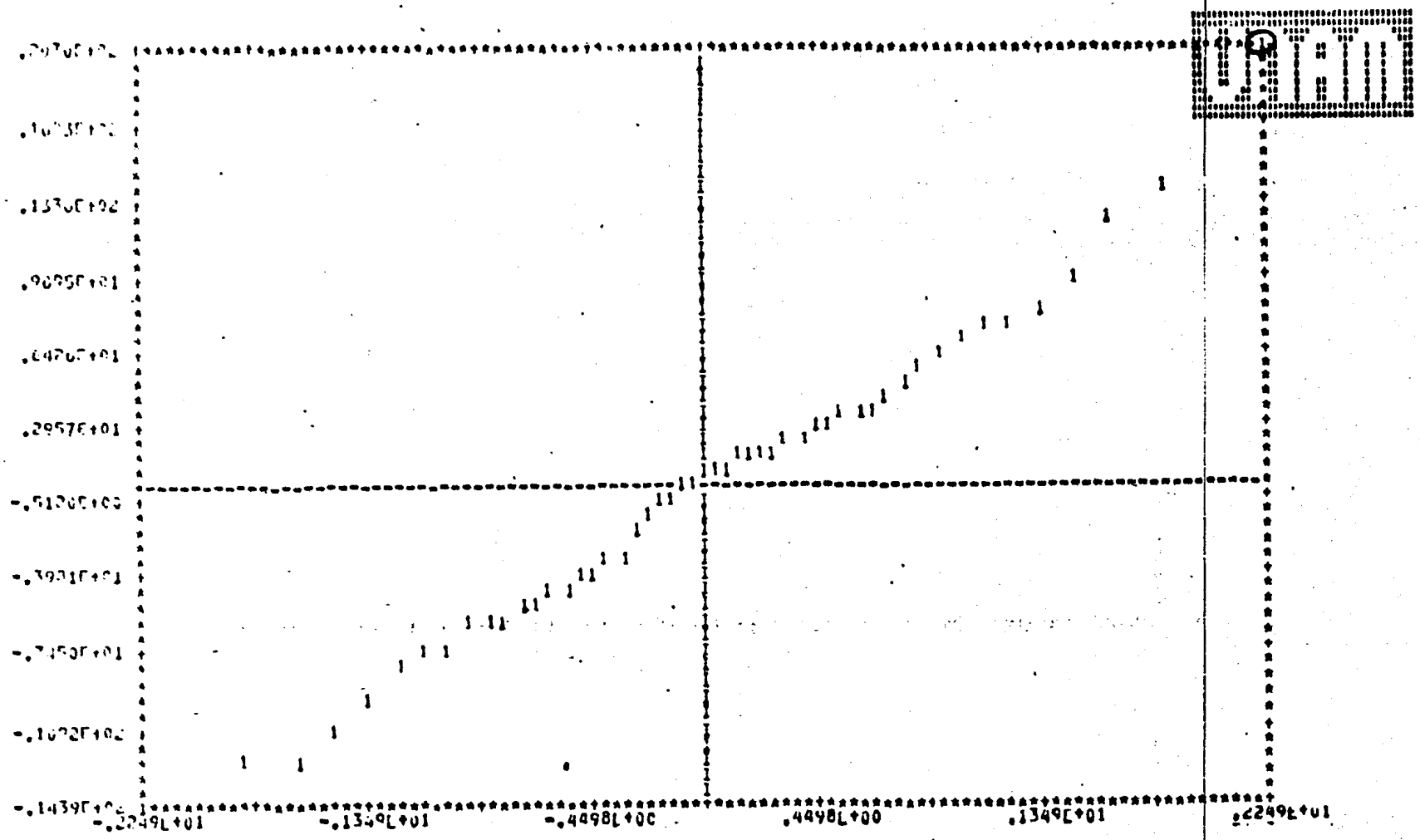
Claramente, el aberrante aparece fuera del conjunto de datos



Quinta componente principal contra cuarta componente principal

N=50, 1 aberrante, bajo d_1

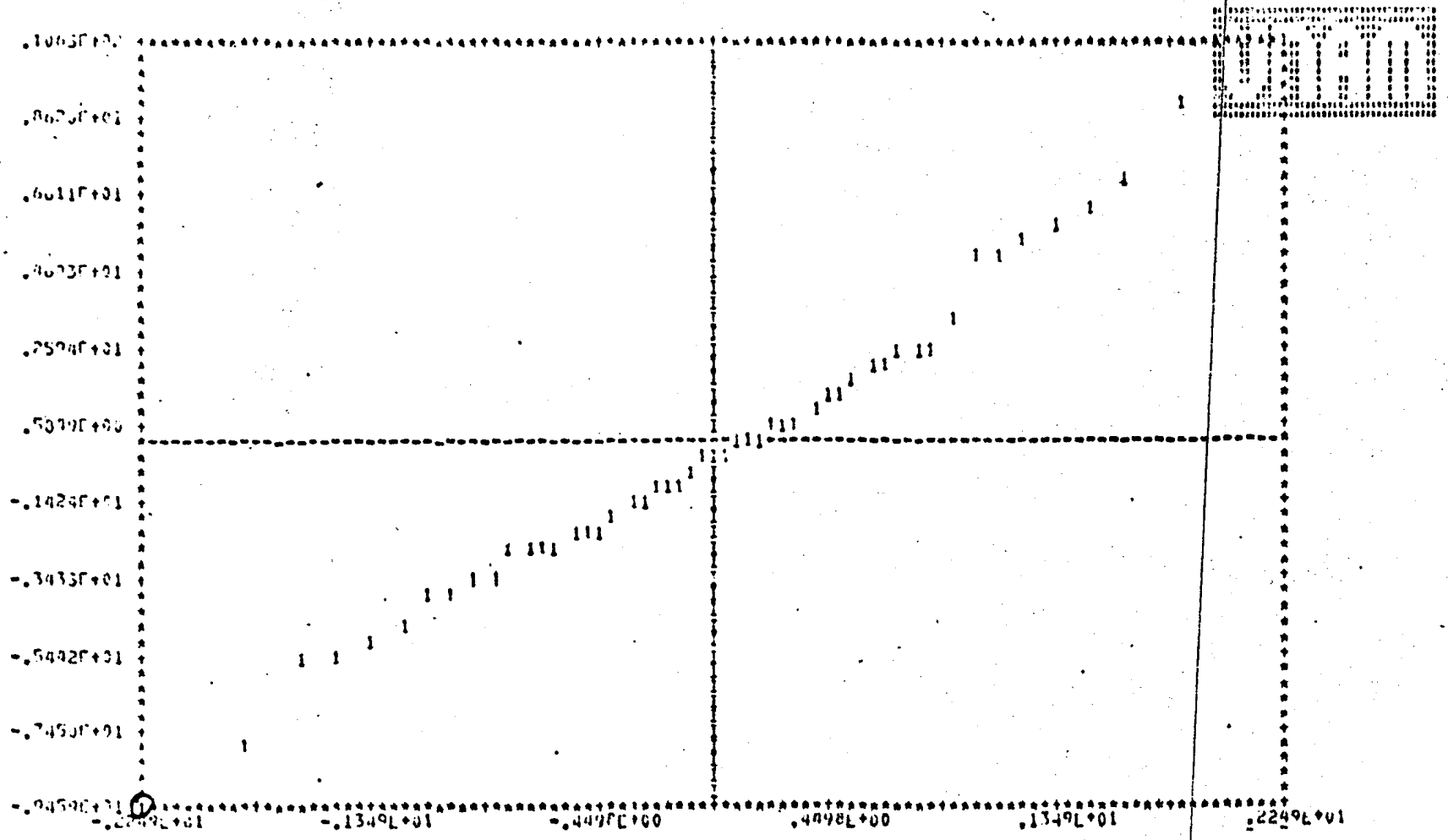
El aberrante cae dentro del conjunto de datos



Primera componente principal contra α_j

N= 50, 1 aberrante, bajo d_2

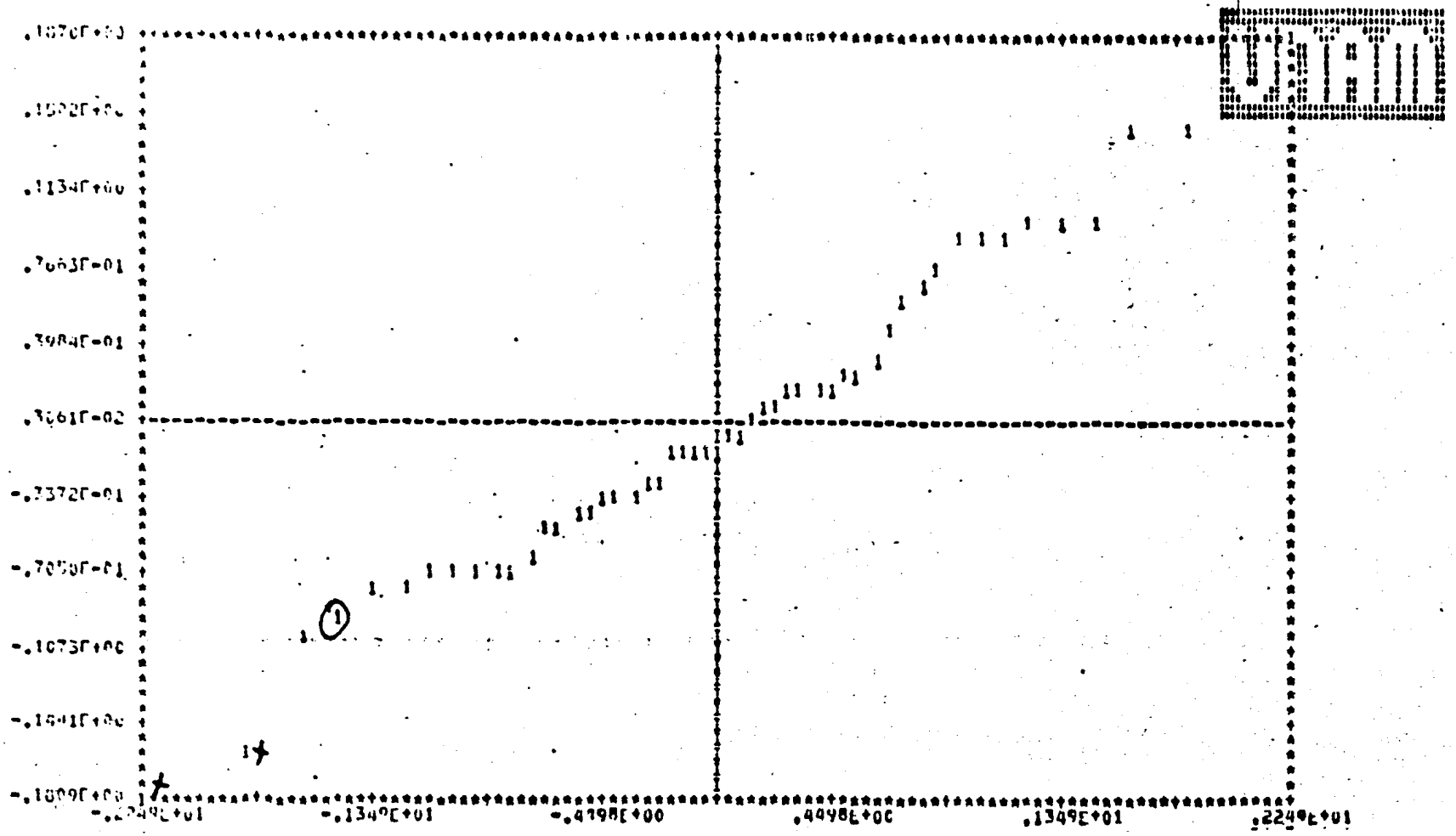
El aberrante se encuentra como valor extremo, pero dentro del patrón lineal



Segunda componente principal contra x_j

N= 50, 1 aberrante, bajo d_2

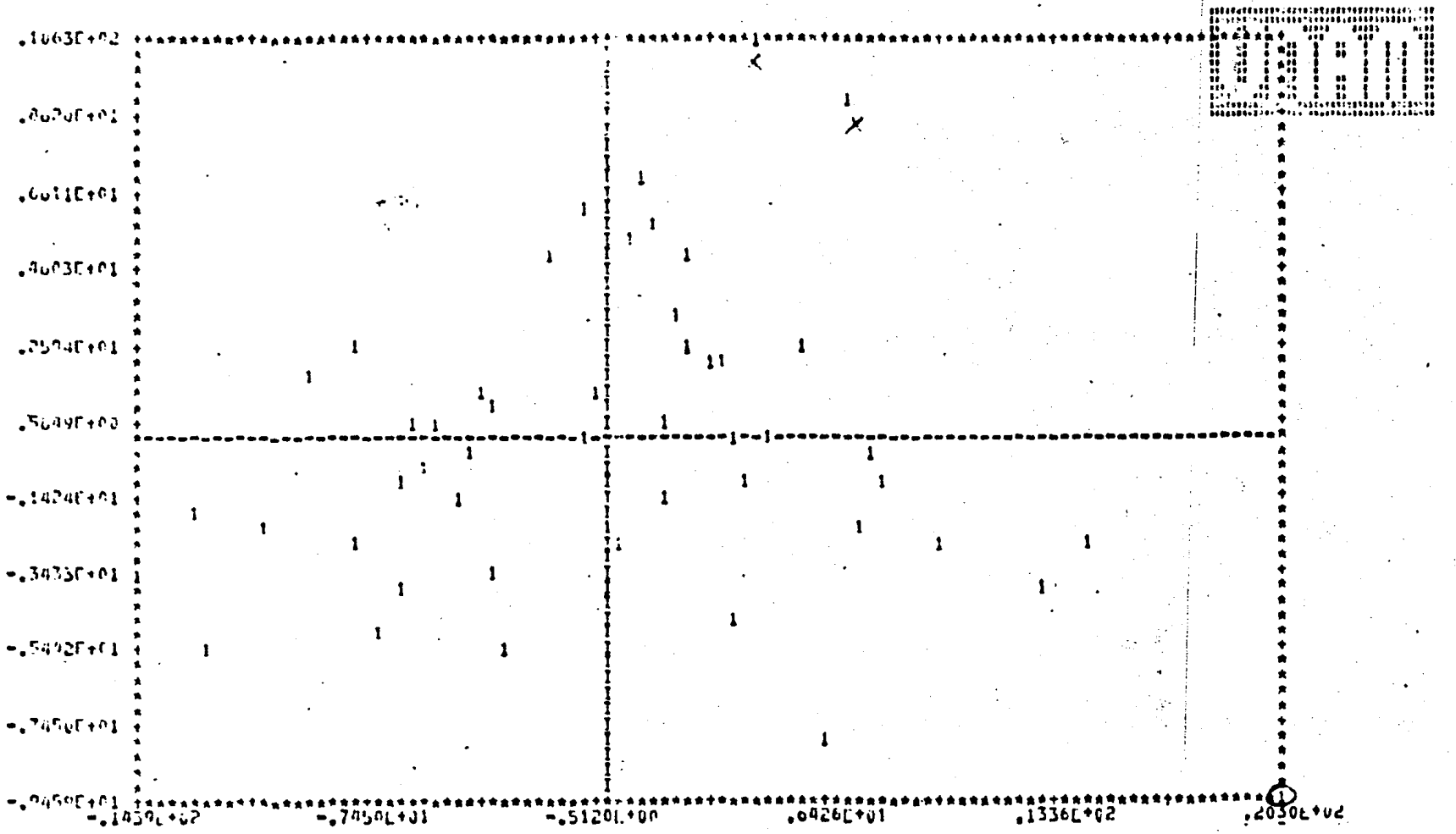
El aberrante se encuentra como valor extremo pero dentro de la tendencia lineal



Quinta componente principal contra a_j

N= 50, 1 aberrante, bajo d_2

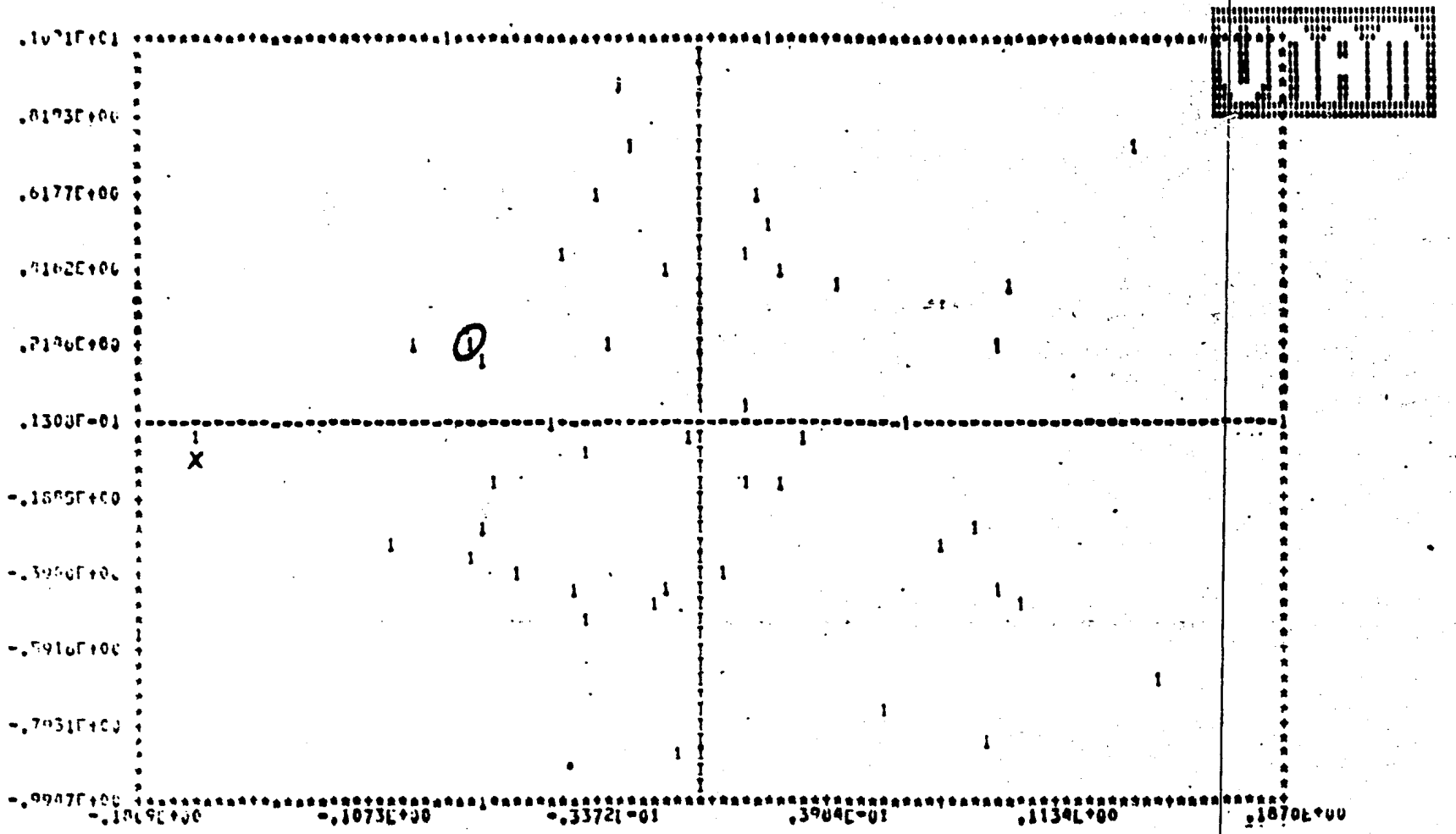
El aberrante dentro de la tendencia lineal y dentro del conjunto de datos.Los dos datos extremos serían sospechosos a ser aberrantes



Primera componente principal contra segunda componente principal

N= 50, 1 aberrante, bajo d_2

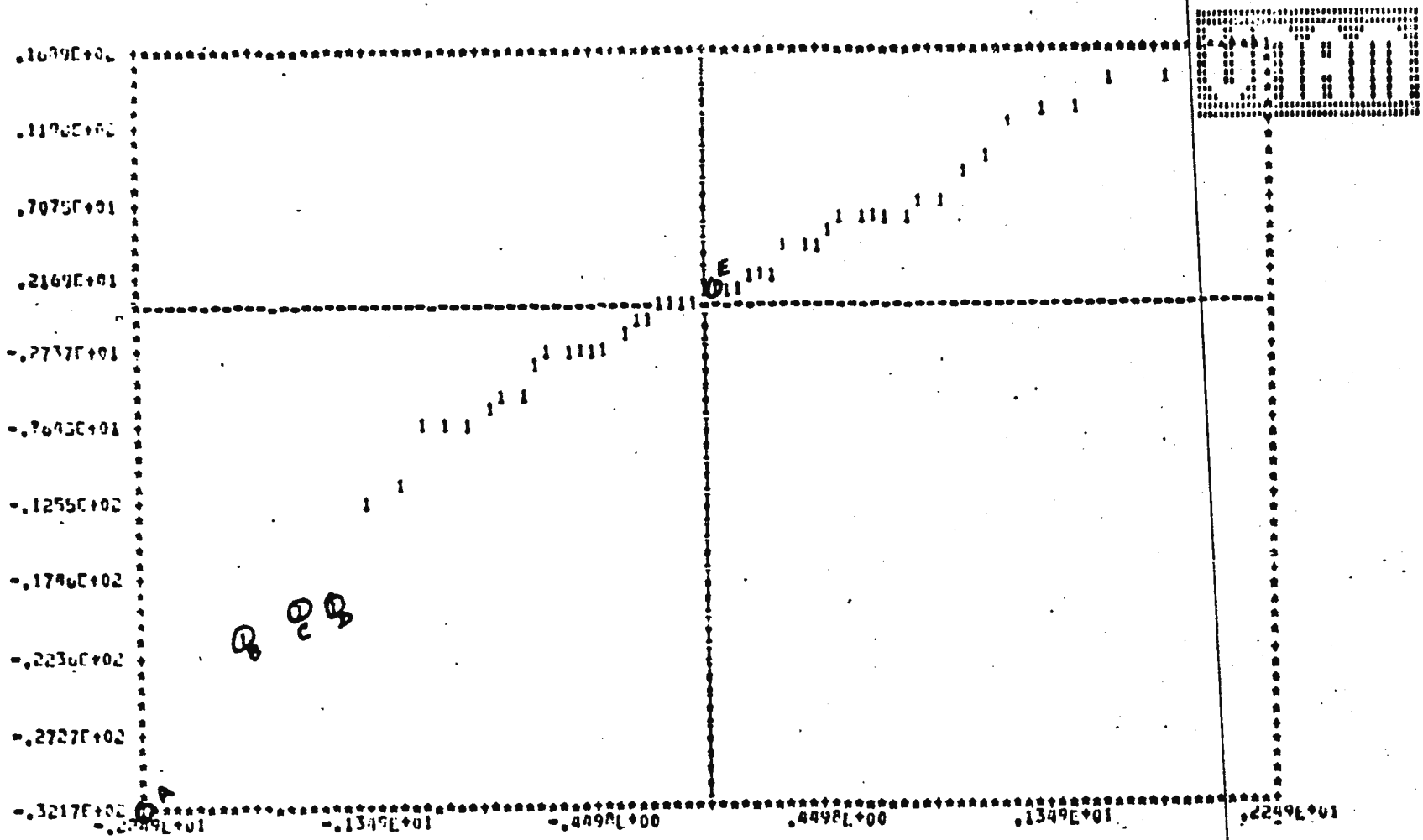
El aberrante se encuentra alejado del conjunto de datos



Quinta componente principal contra cuarta componente principal

N= 50, sin aberrantes, bajo d_2

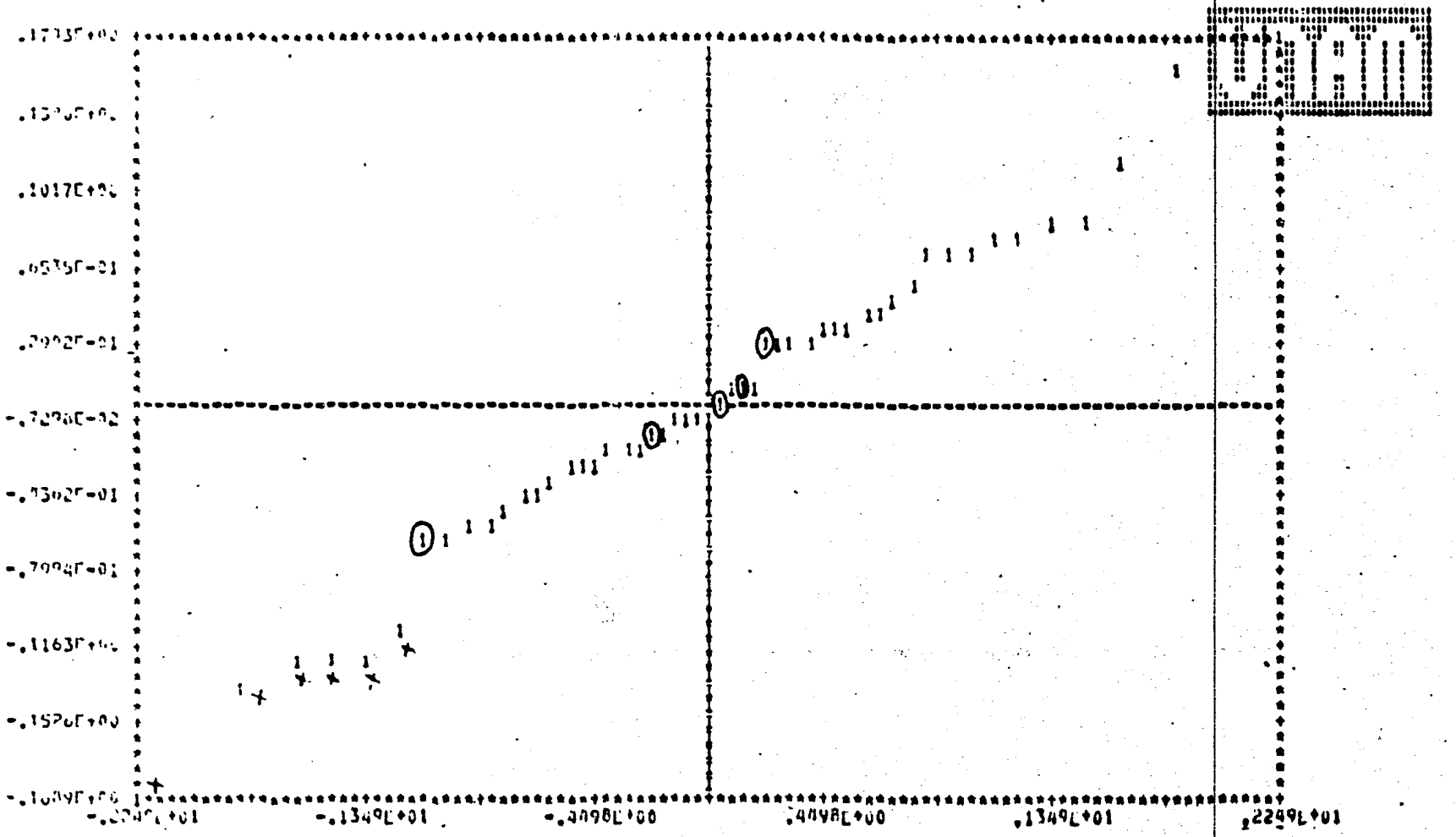
El aberrante dentro del conjunto de datos



Primera componente principal contra α_j

N= 50, 5 aberrantes, bajo d_1

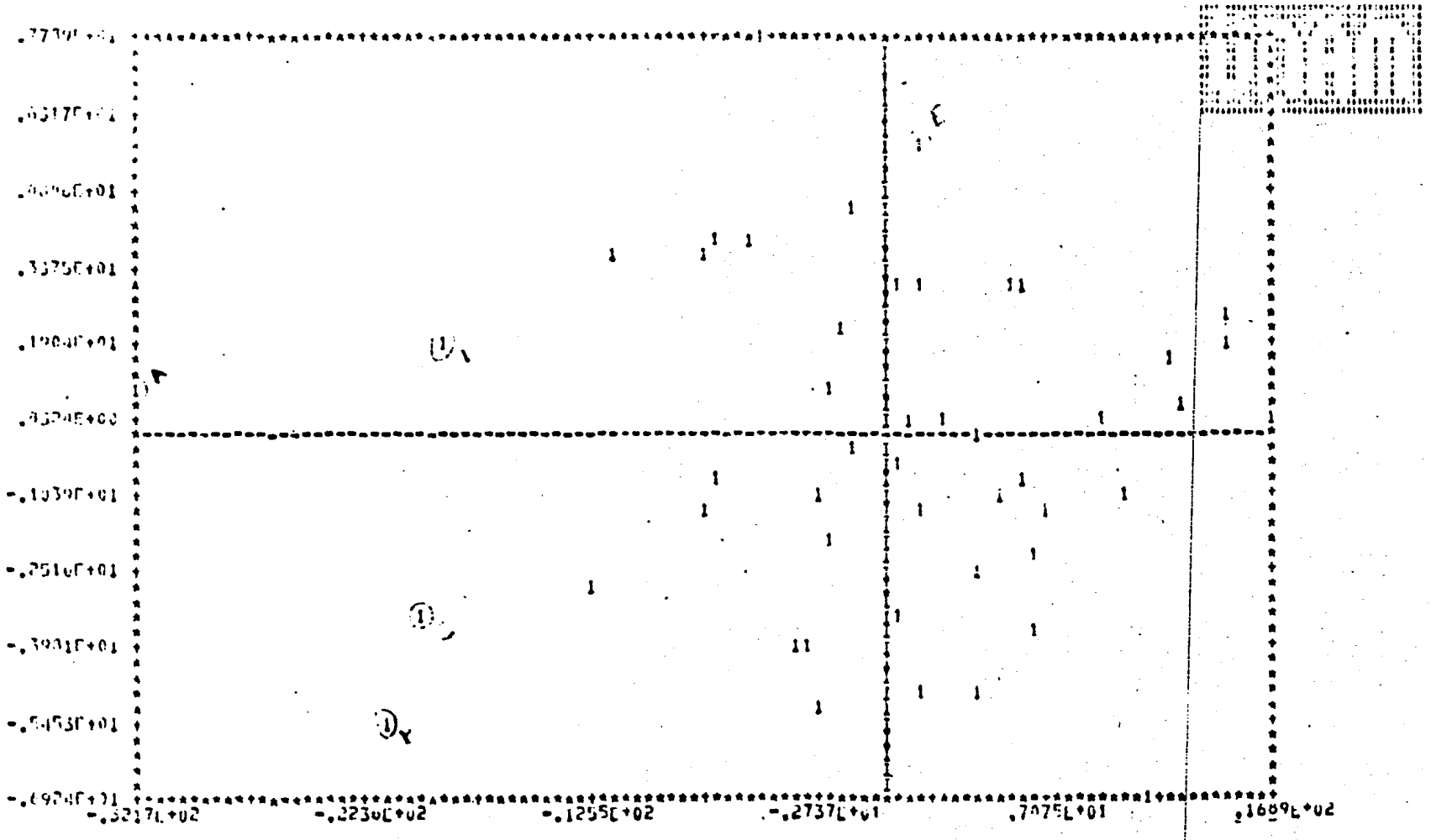
La observación E es la única que no es sospechosa a ser aberrante



Quinta componente principal contra α_j

N= 50, 5 aberrantes, bajo d_1

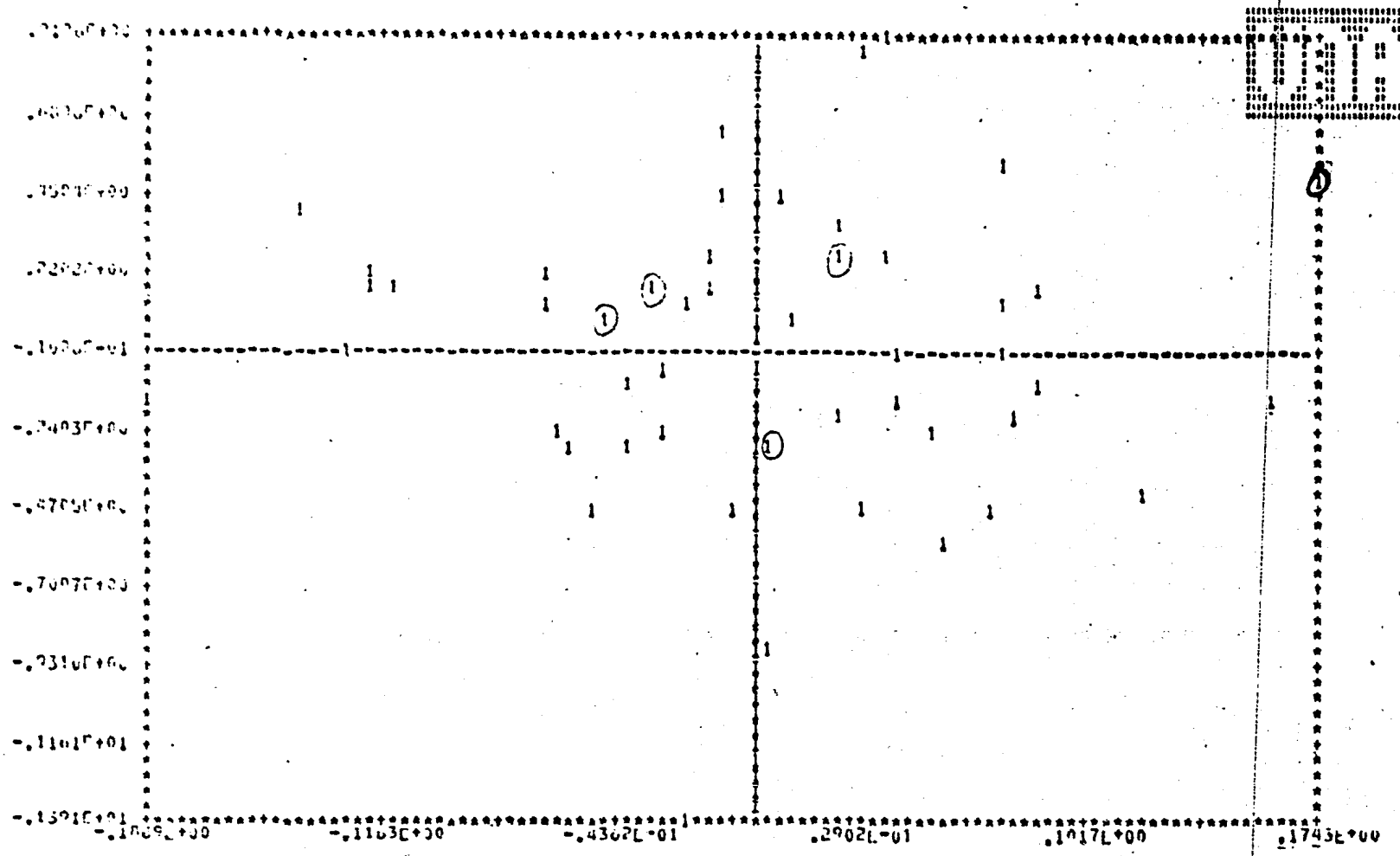
Todos los aberrantes dentro del conjunto de datos



Primera componente principal contra segunda componente principal

N= 50, 5 aberrantes, bajo d_1

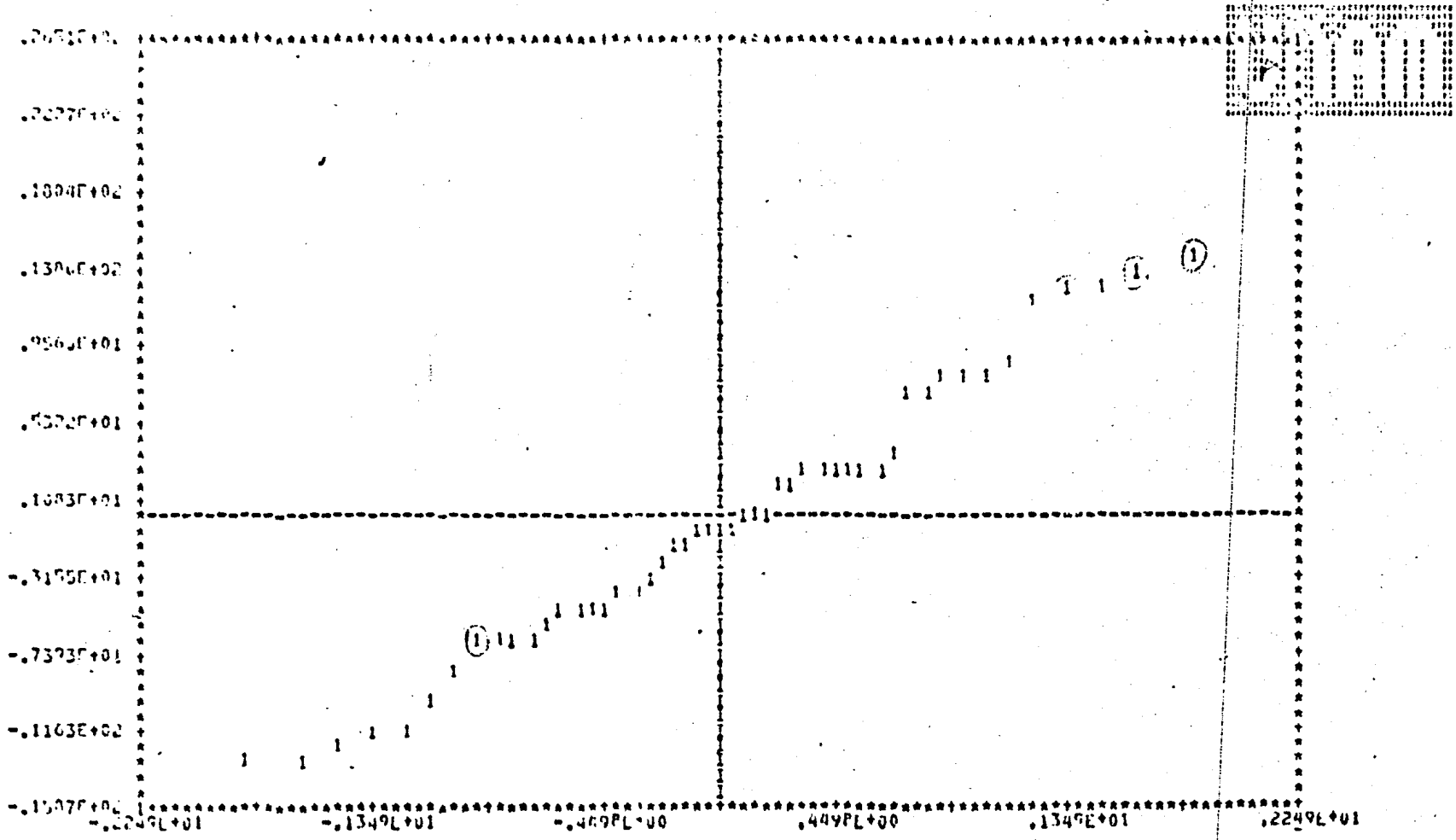
Solamente a la observación E, dudaría en declararla como aberrante



Quinta componente principal contra cuarta componente principal

N= 50, 5 aberrantes, bajo d_1

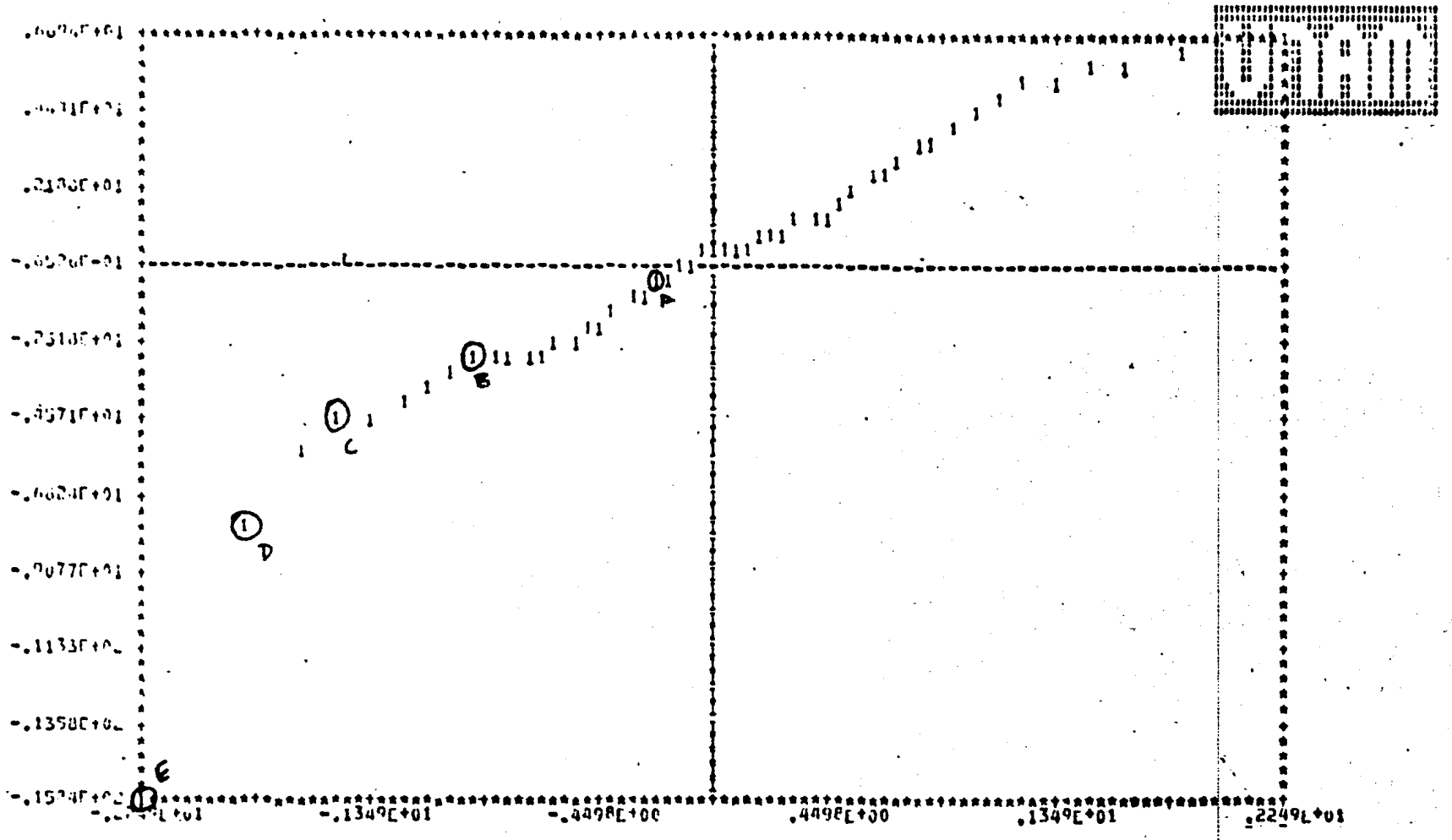
Todas las observaciones dentro del conjunto de datos



Primera componente principal contra $\hat{y}_j$

N=50, 5 aberrantes, bajo d_2

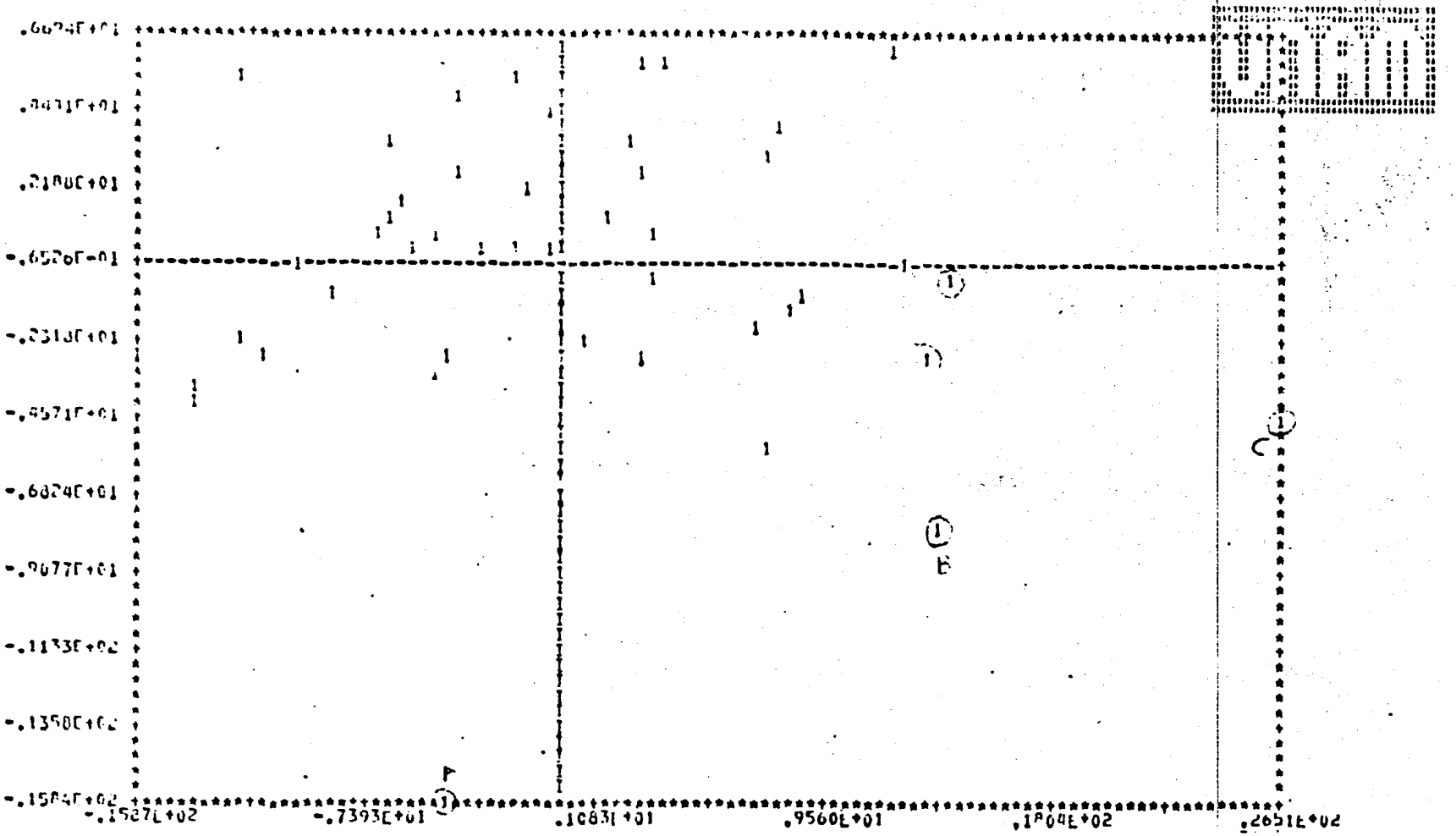
Solamente el aberrante A, aparece como extremo y fuera de la tendencia lineal.



Segunda componente principal contra CA_j

N= 50, 5 aberrantes, bajo d_2

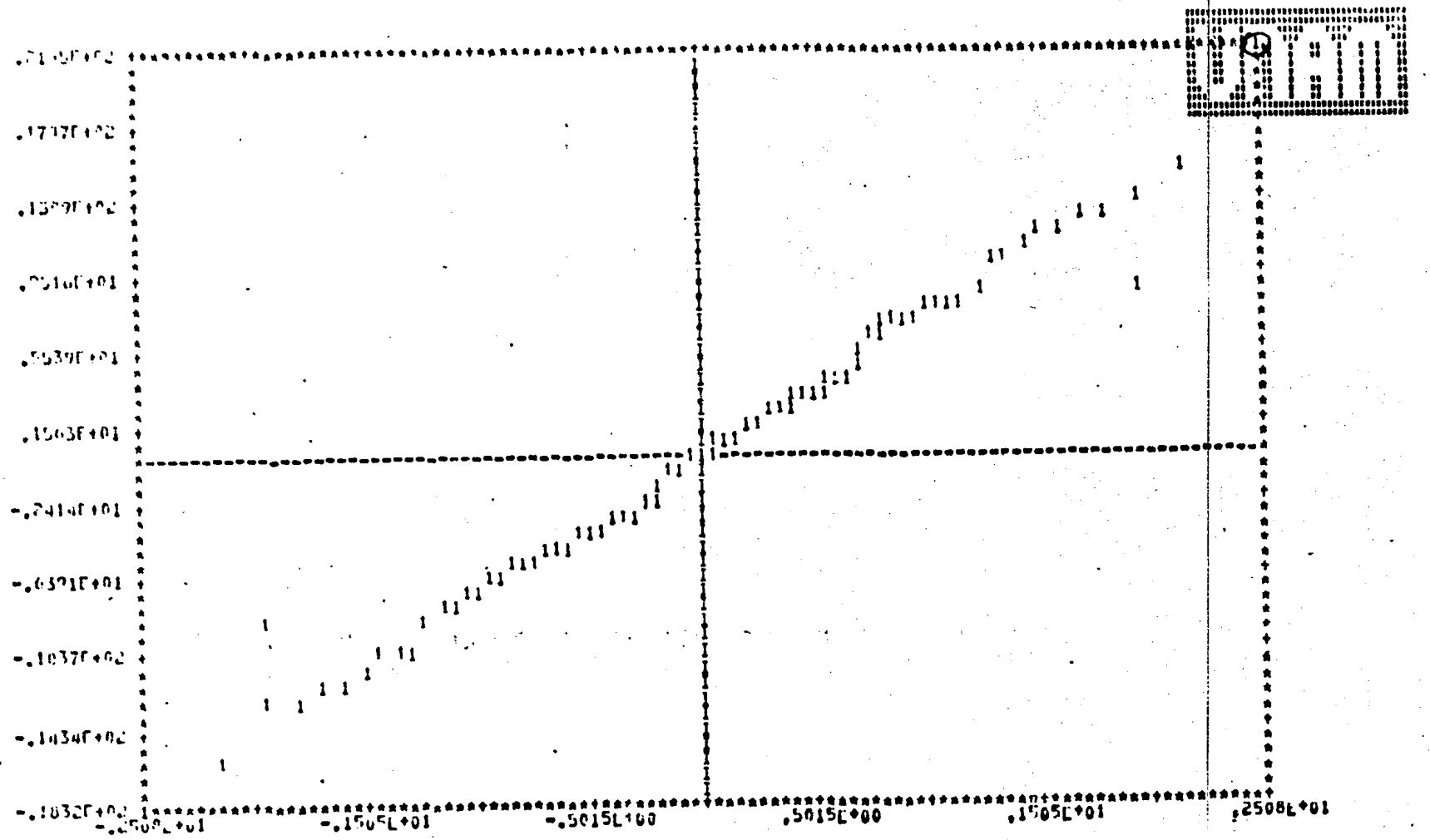
Solo el aberrante E aparece como extremo y fuera de la tendencia lineal y es uno de los casos donde la segunda componente principal es sensible a los aberrantes



Primera componente principal contra segunda componente principal

N=50, 5 aberrantes, bajo d_2

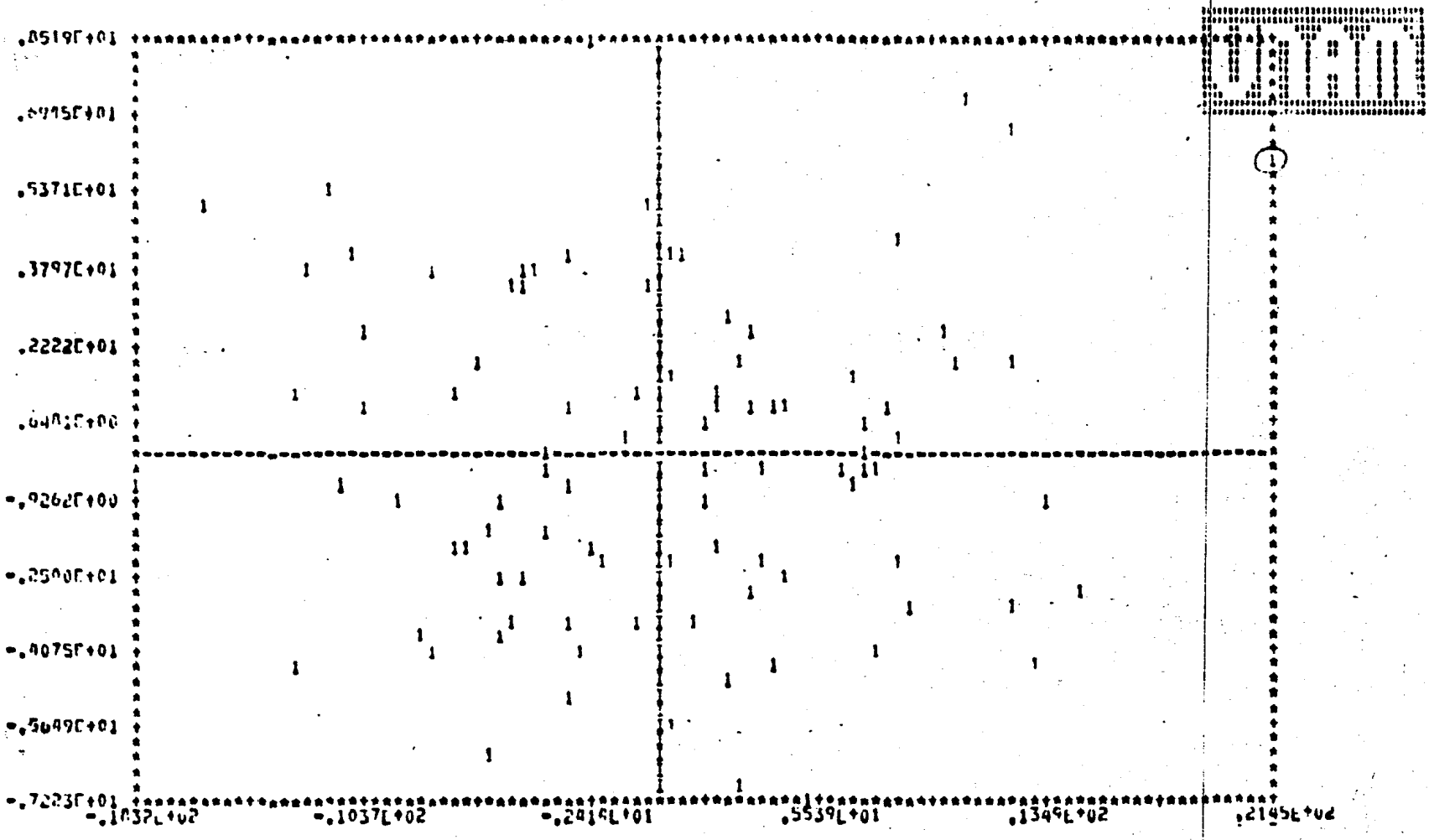
Las observaciones A, B y C son sospechosas a ser aberrantes



Primera componente principal contra α_j

N= 100, 1 aberrante, bajo d_1

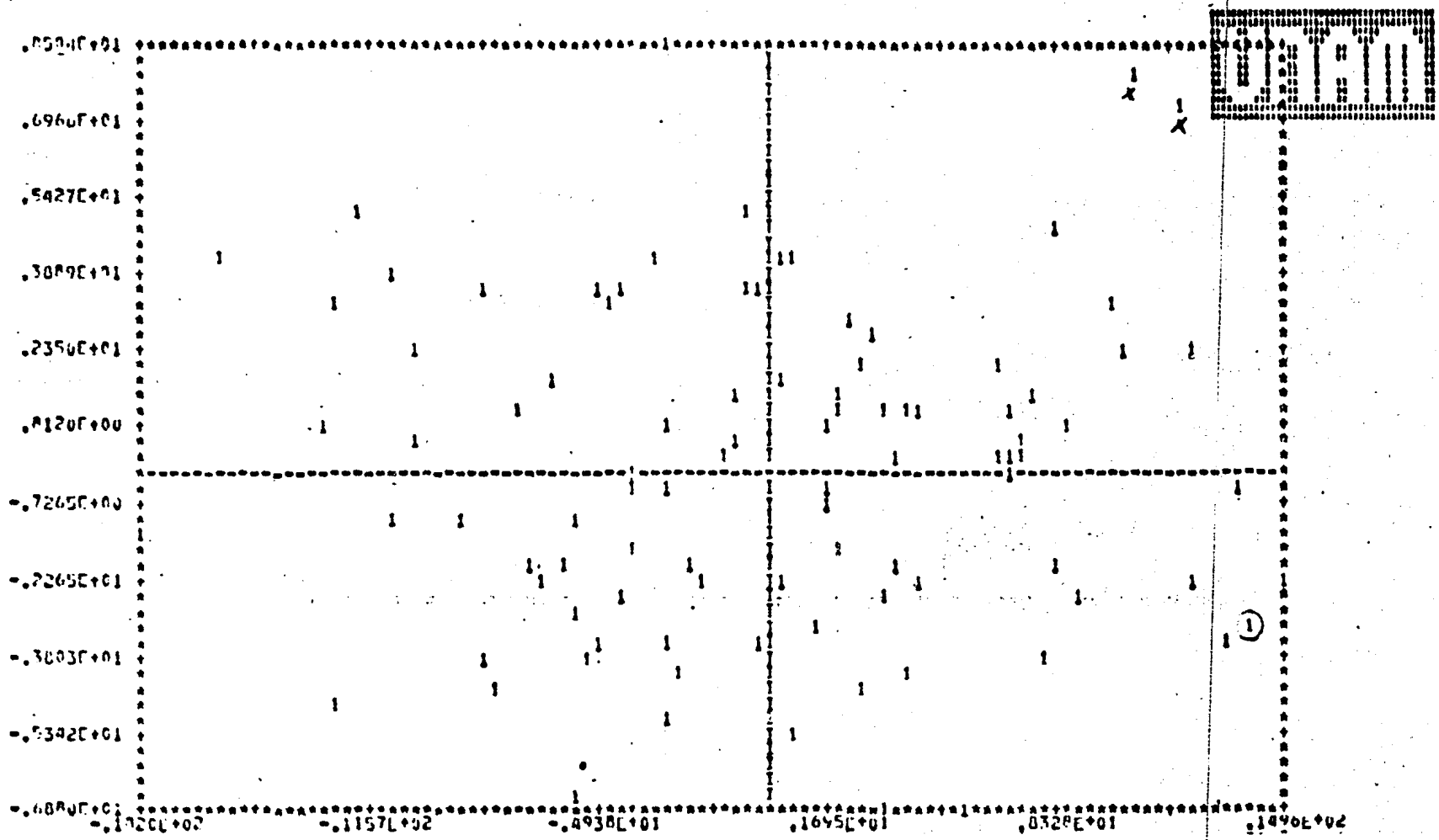
El aberrante se detecta fácilmente



Primera componente principal contra segunda componente principal

N= 100, 1 aberrante, bajo d_1

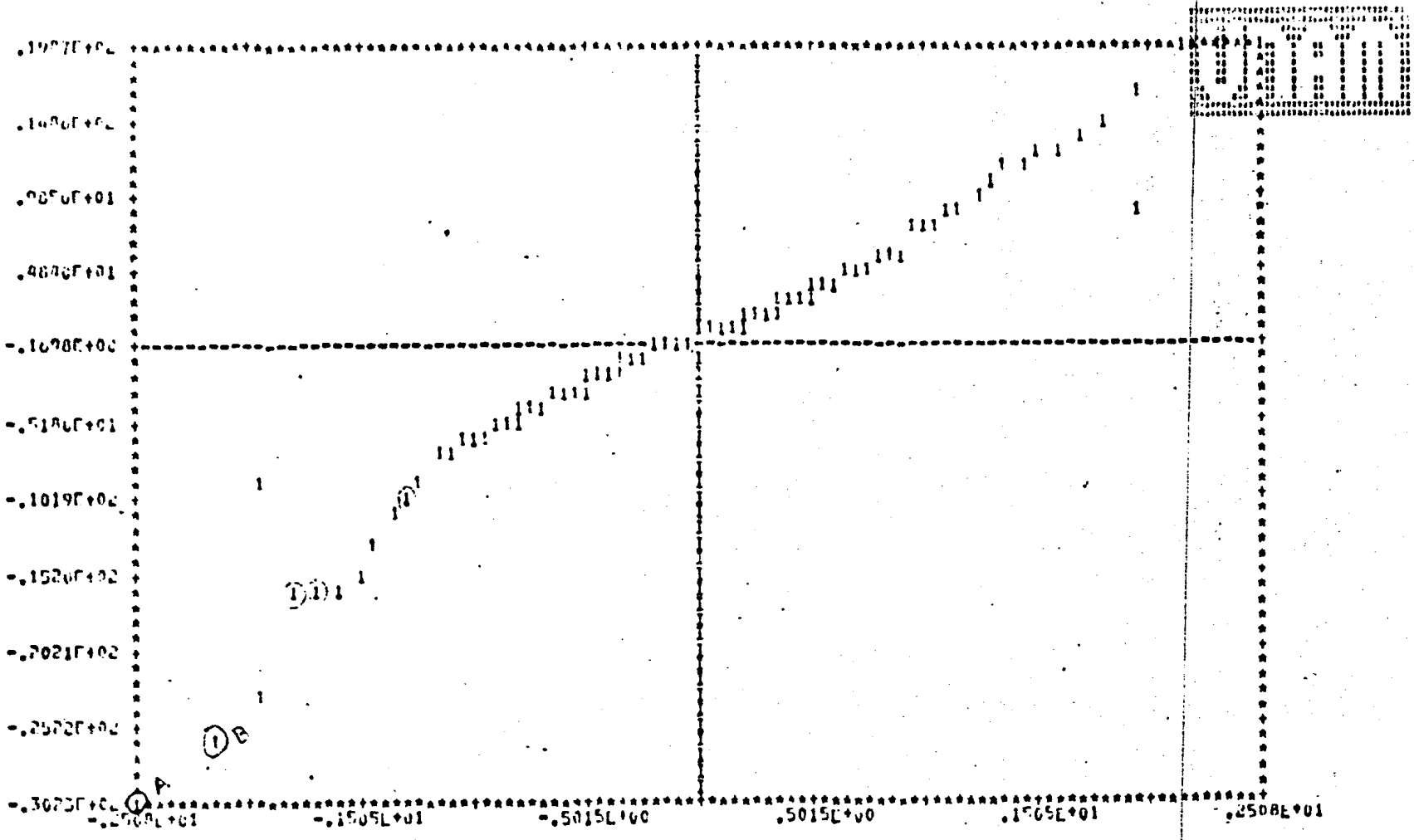
El aberrante bastante alejado del conjunto de datos



Primera componente principal contra segunda componente principal

N= 100, 1 aberrante, bajo d_2

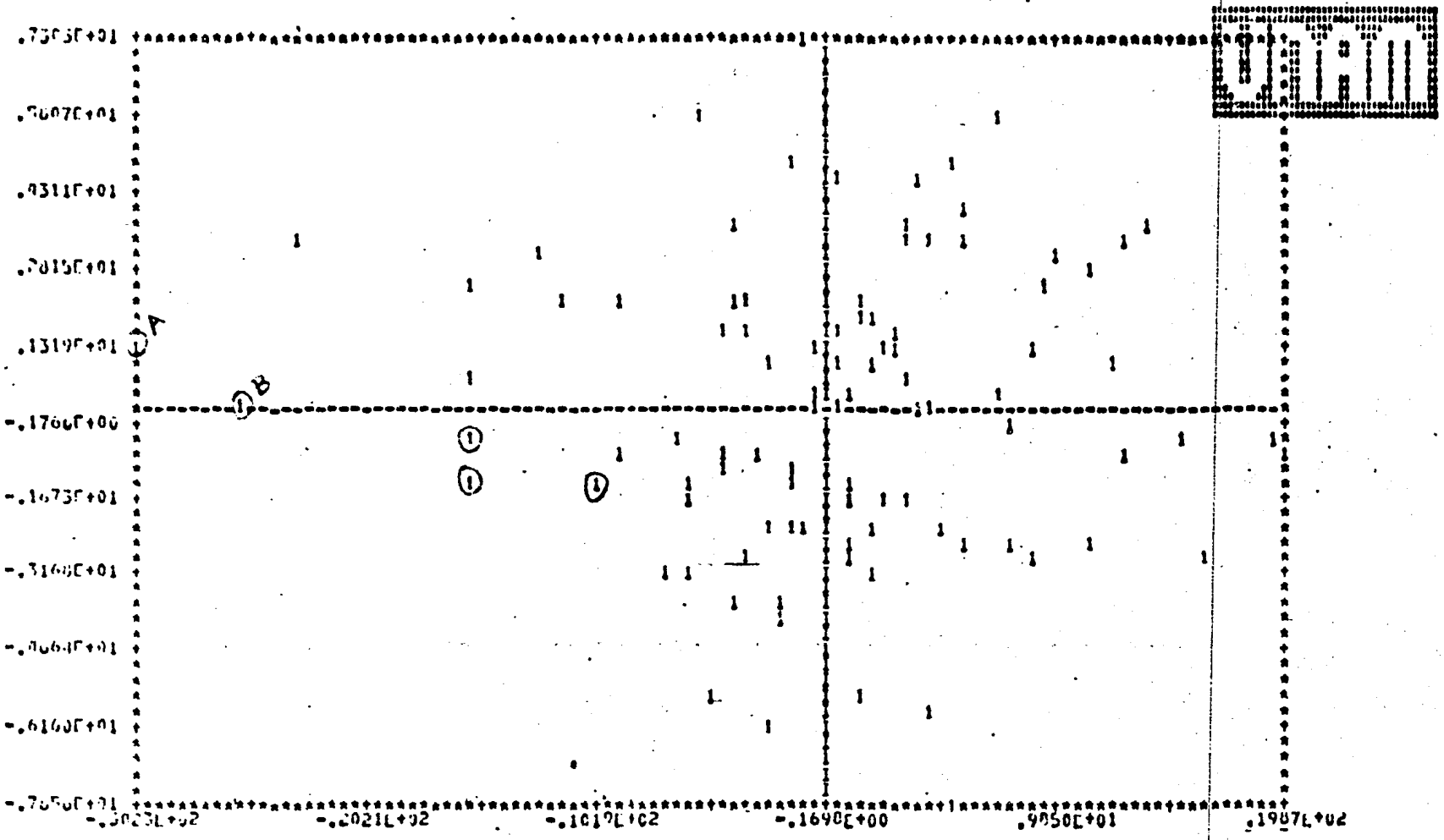
El aberrante dentro del conjunto de datos



Primera componente principal contra α_j

N= 100, 5 aberrantes, bajo d_1

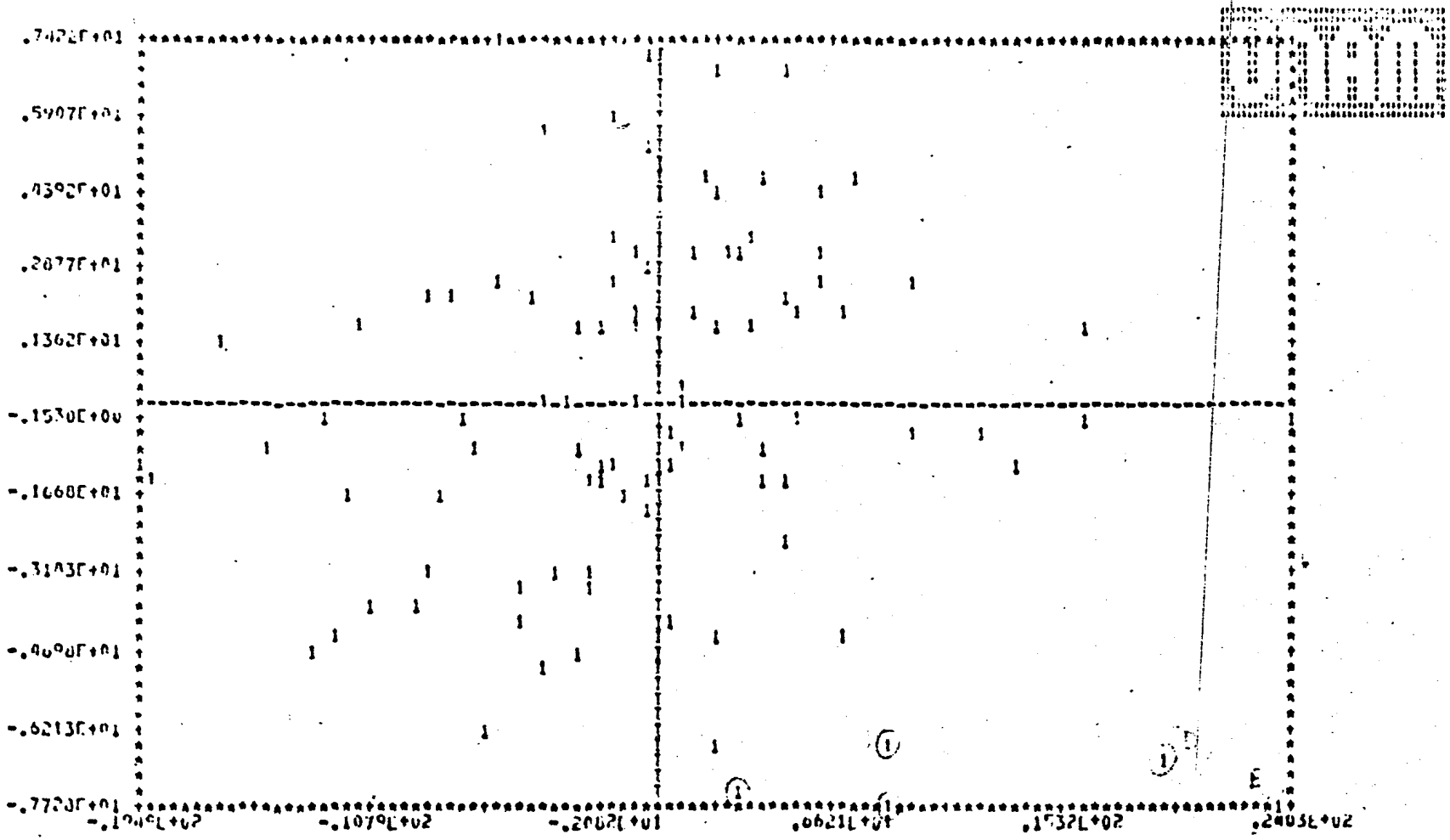
Los aberrantes A y B aparecen como extremos y fuera de la tendencia lineal



Primera componente principal contra segunoa componente principal

N= 100, 5 aberrantes, bajo d_1

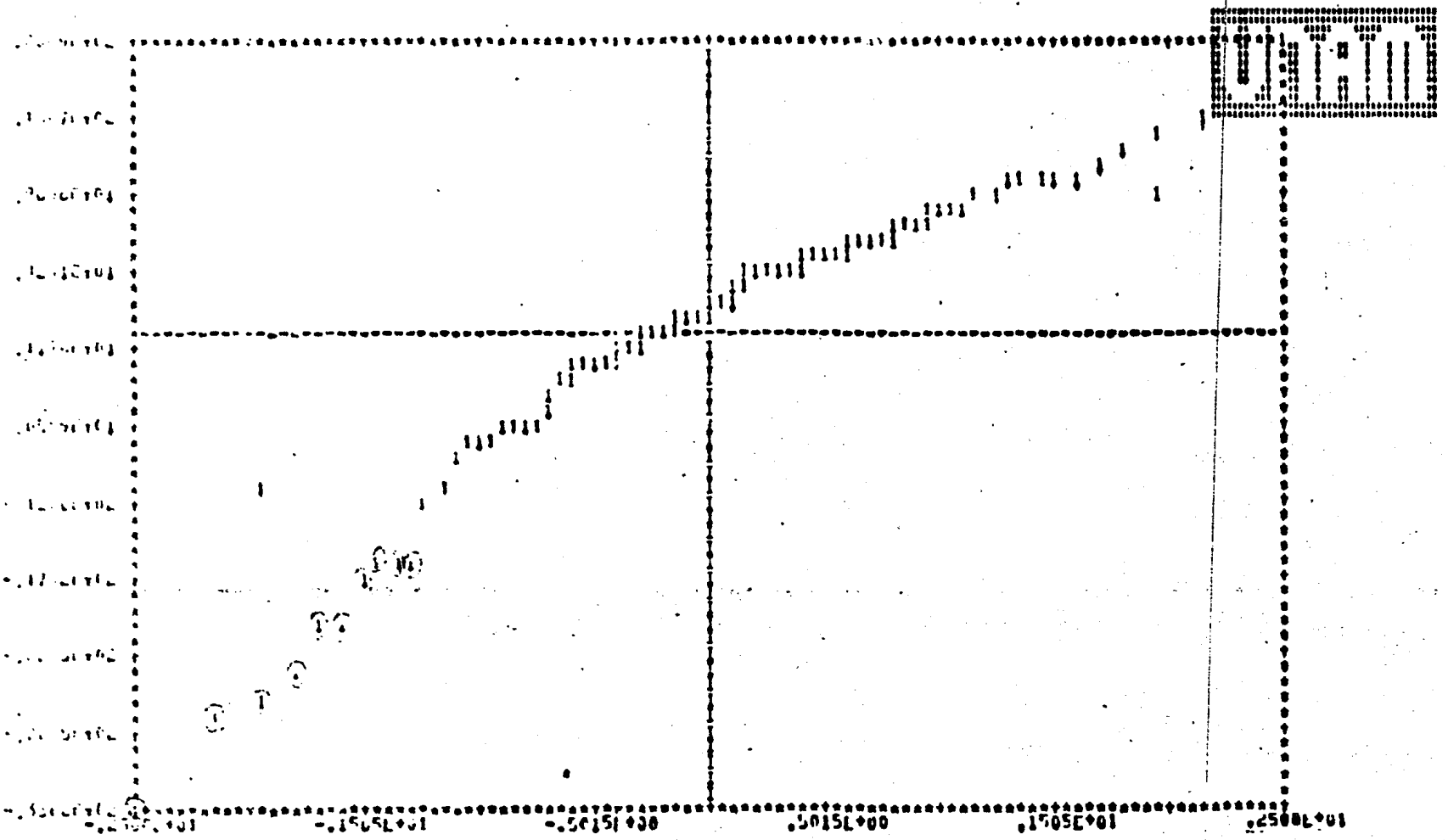
Las observaciones A y B se encuentran alejadas del resto de los datos



Primera componente principal contra segunda componente principal

N= 100, 5 aberrantes, bajo d_2

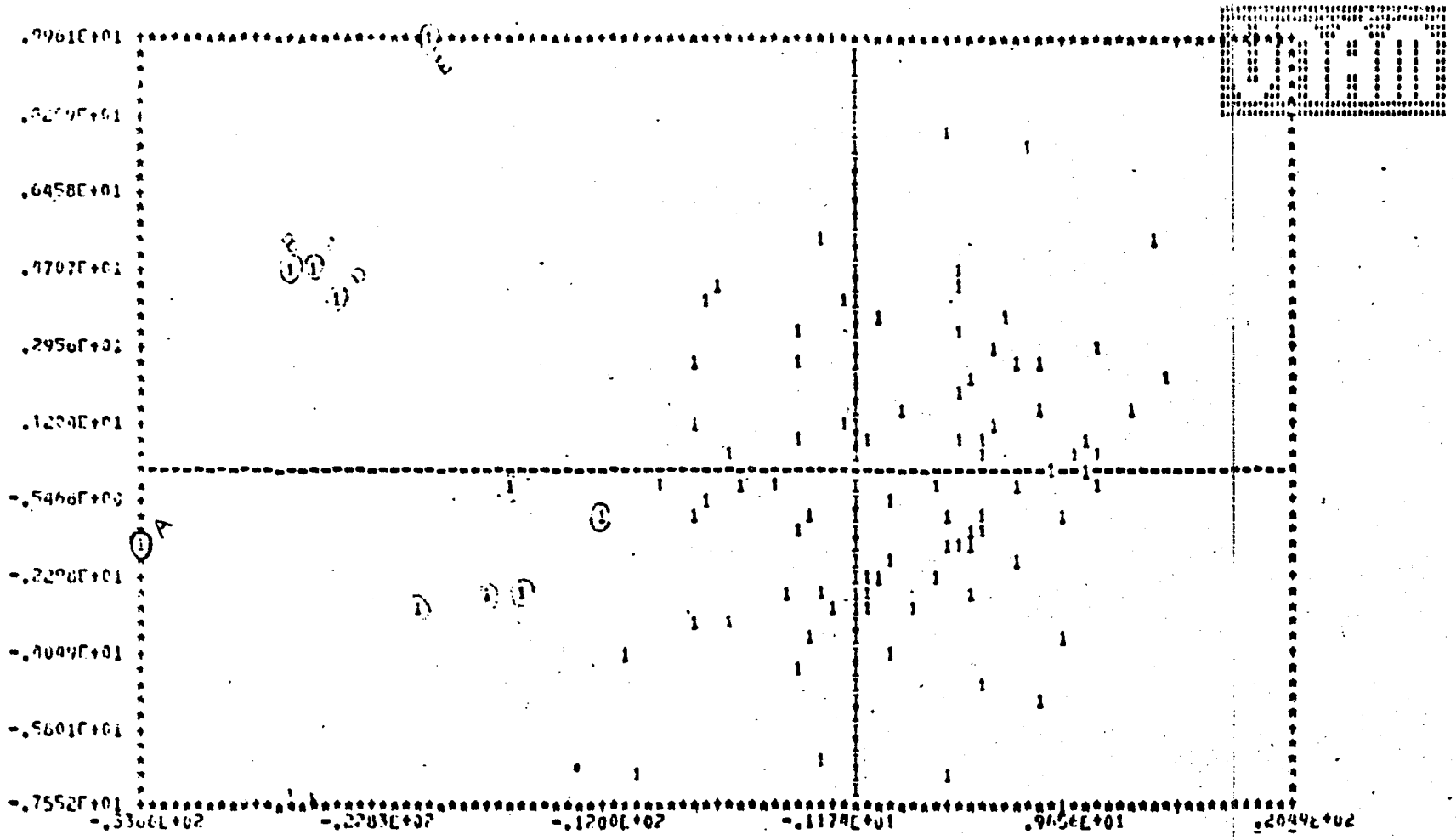
Las observaciones D y E son las mas sospechosas a ser aberrantes



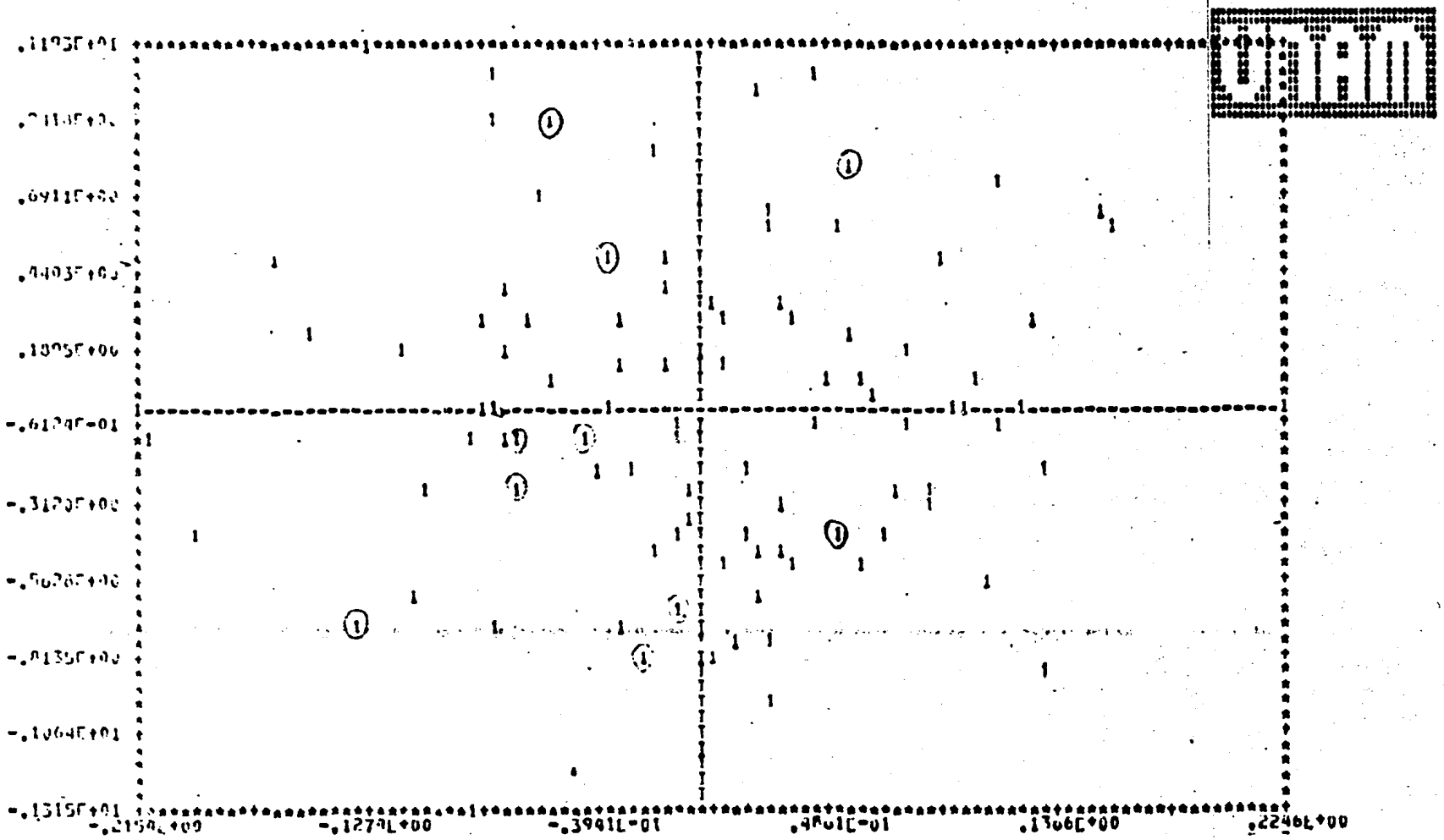
Primera componente principal contra α_j

N= 100, 10 aberrantes, bajo d_1

Todos los aberrantes dentro de la tendencia lineal



Primera componente principal contra segunda componente principal
N= 100, 10 aberrantes, bajo d_1
Las observaciones A, B, C, D y E son las mas sospechosas a ser aberrantes

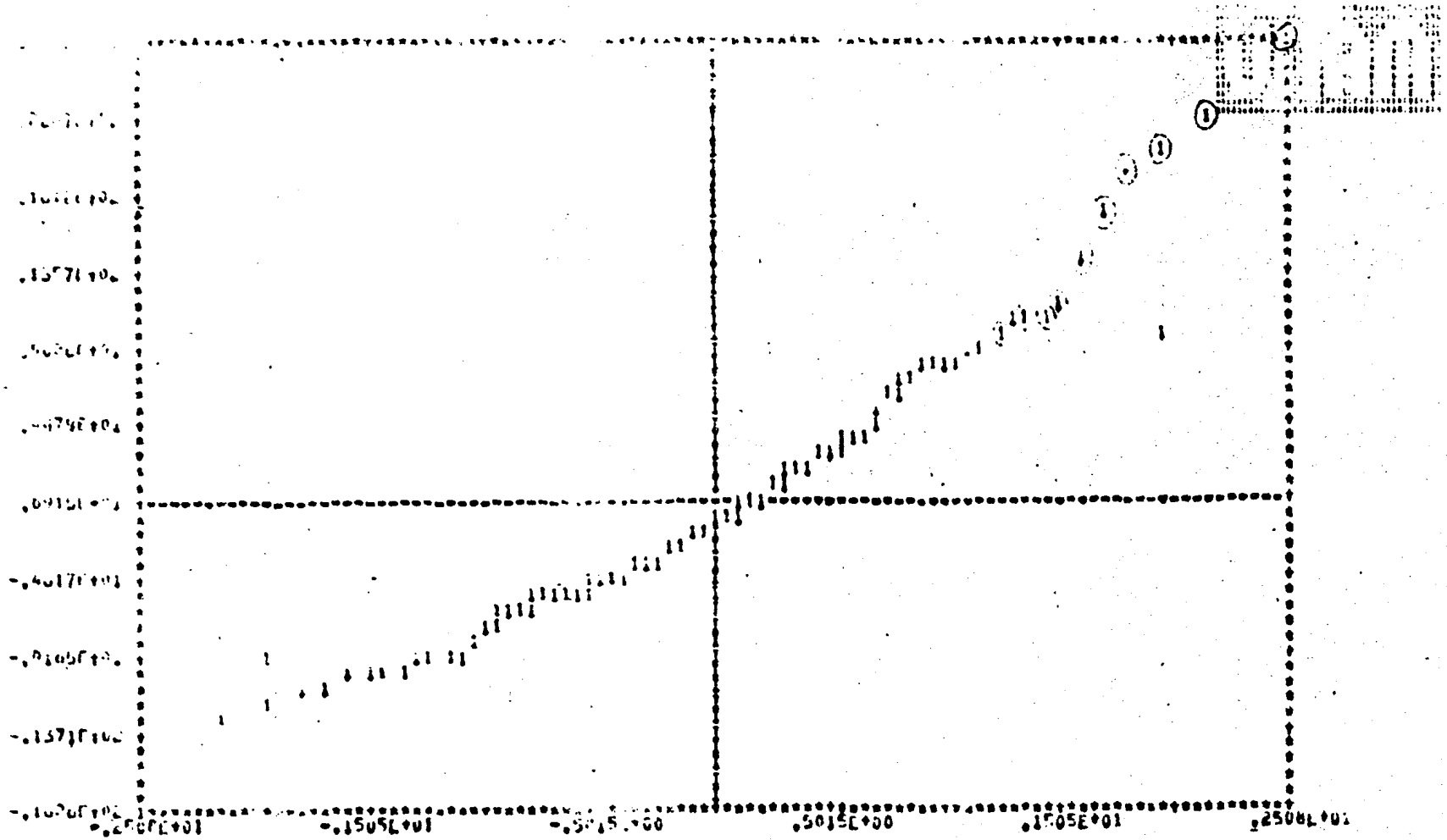


Quinta componente principal contra cuarta componente principal

N= 100, 10 aberrante, bajo d_1

Se puede notar la diferencia entre esta gráfica y la anterior

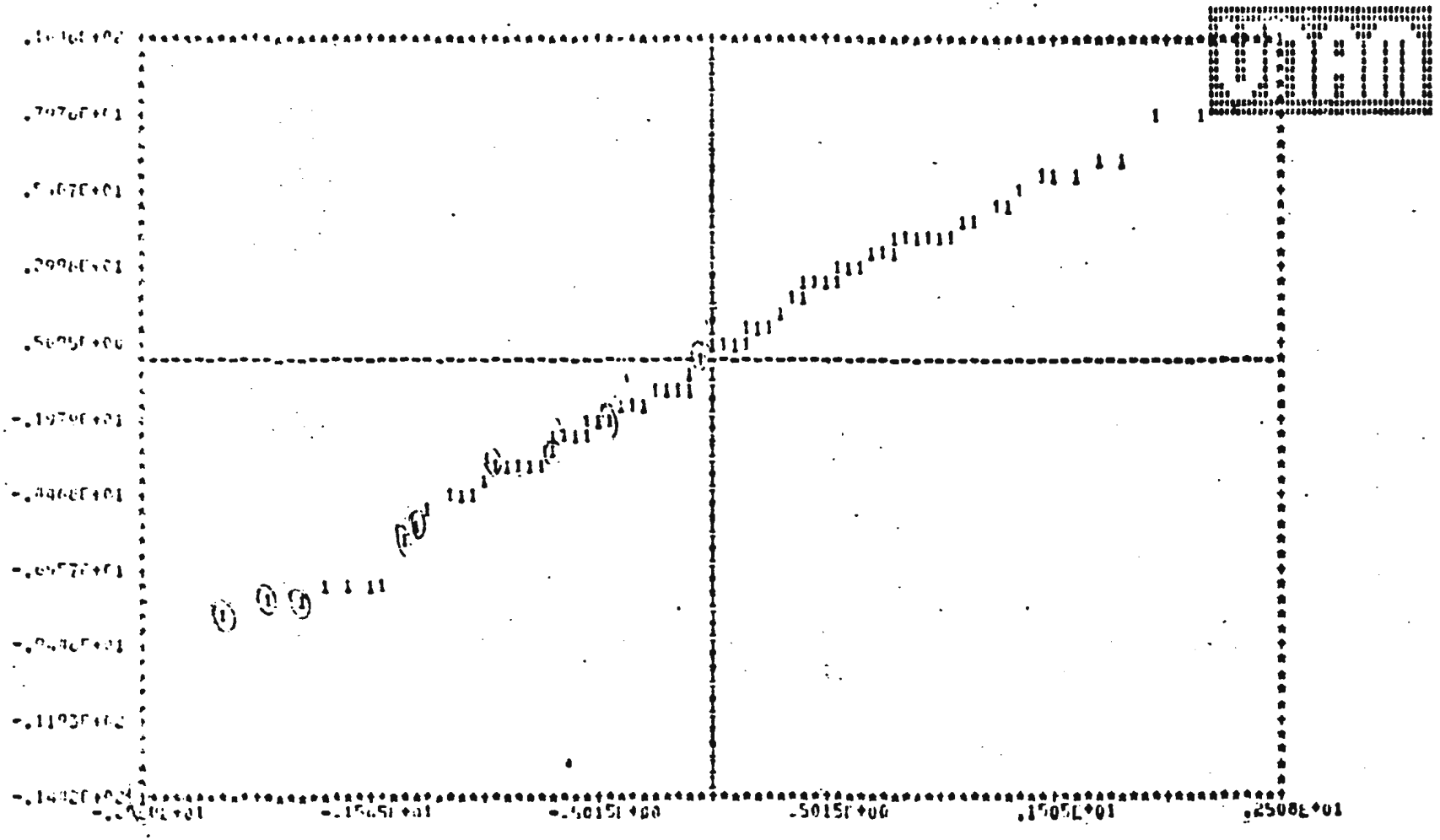
Aquí no se detecta a ningún aberrante



Primera componente principal contra d_j

N= 100, 10 aberrantes, bajo d_2

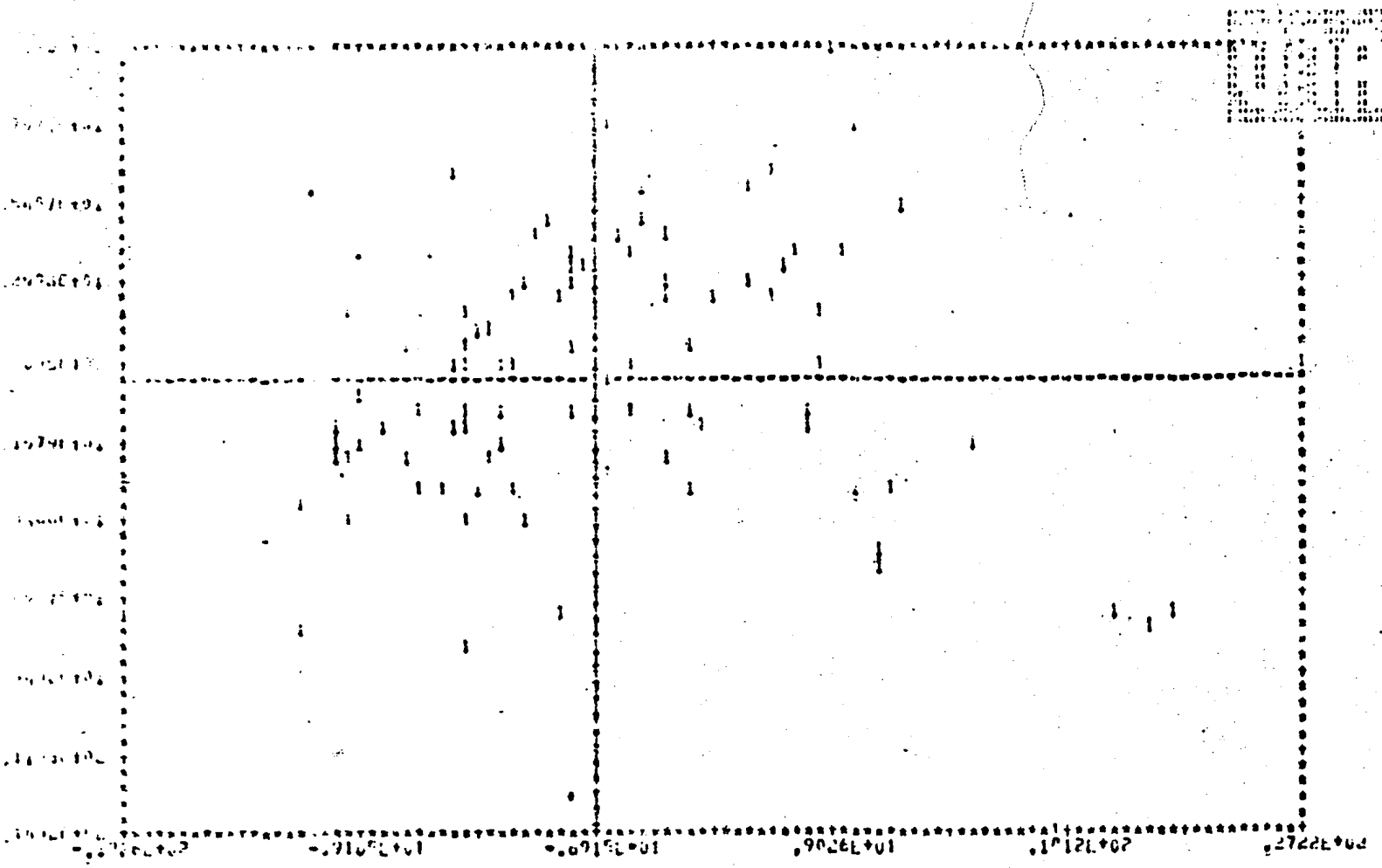
Dificilmente declararía a alguna observación como aberrante



Segunda componente principal contra x_j

N= 100, 10 aberrantes, bajo d_2

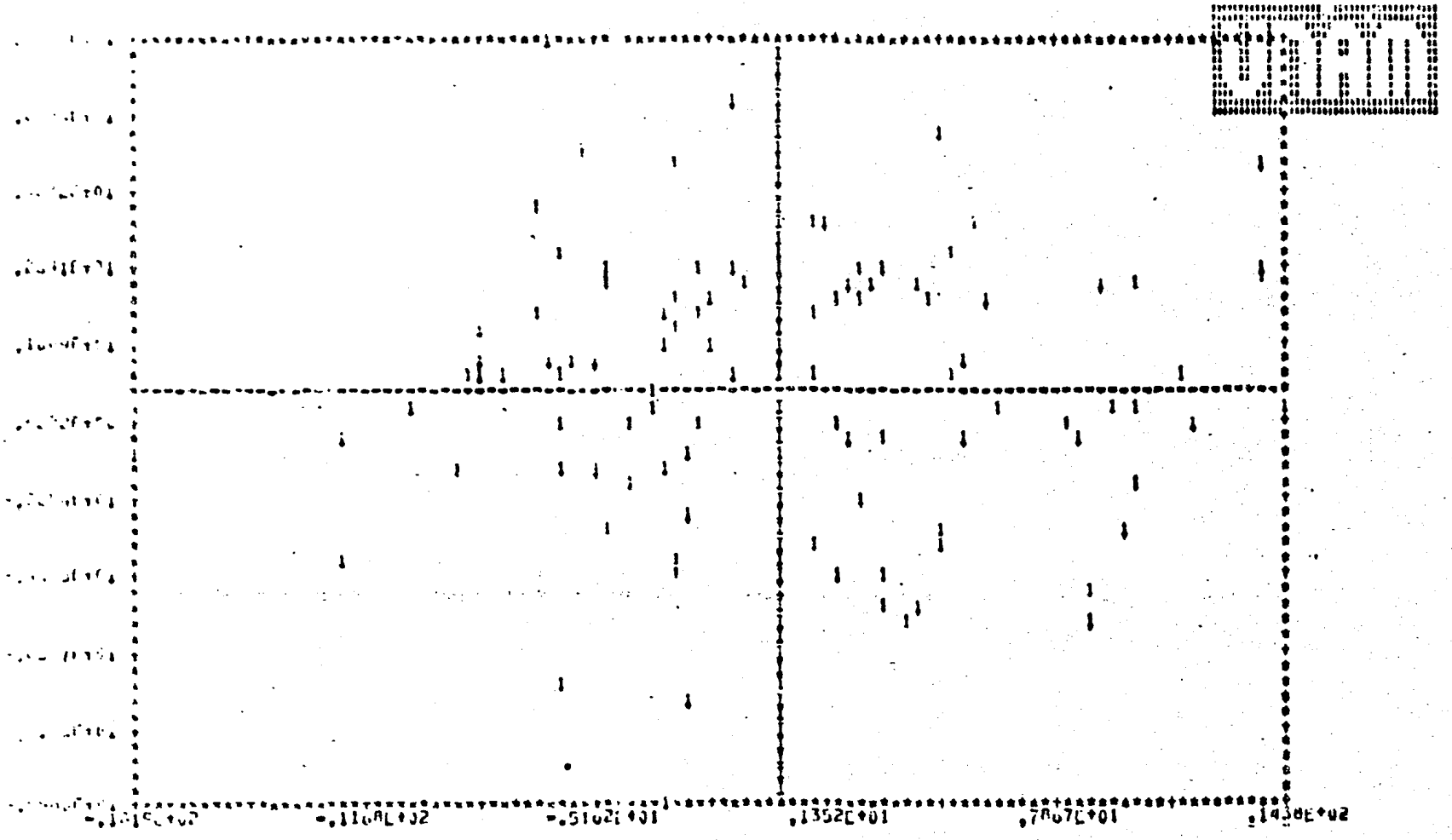
Otro caso donde la segunda componente principal muestra claramente a un aberrante



Primera componente principal contra segunda componente principal

N= 100, 10 aberrantes, bajo d_2

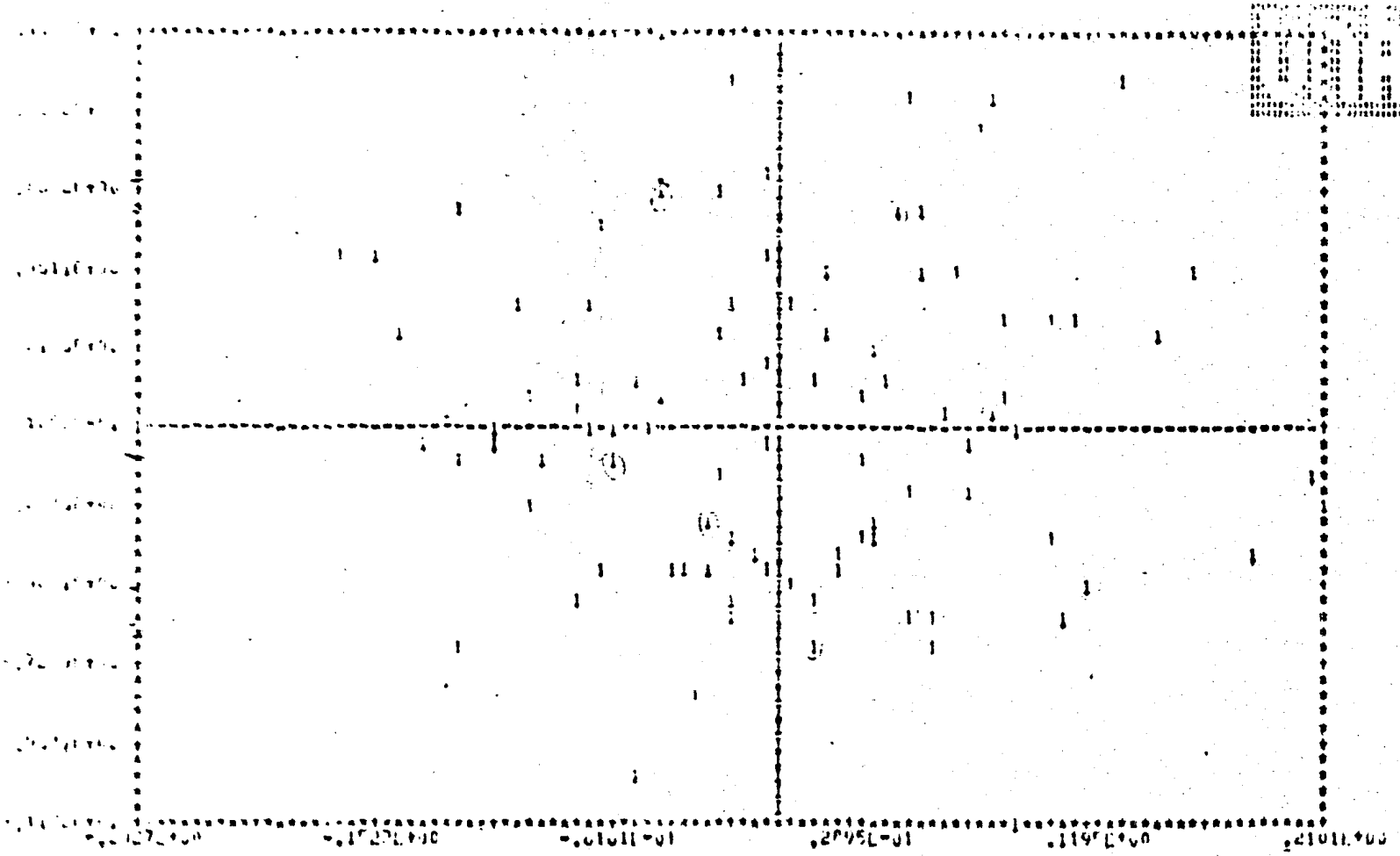
Se detecta facilmente a los aberrantes



Primera componente principal contra segunda componente principal

N= 100, sin aberrantes, bajo d_2

Se puede notar gran diferencia entre esta gráfica y la anterior



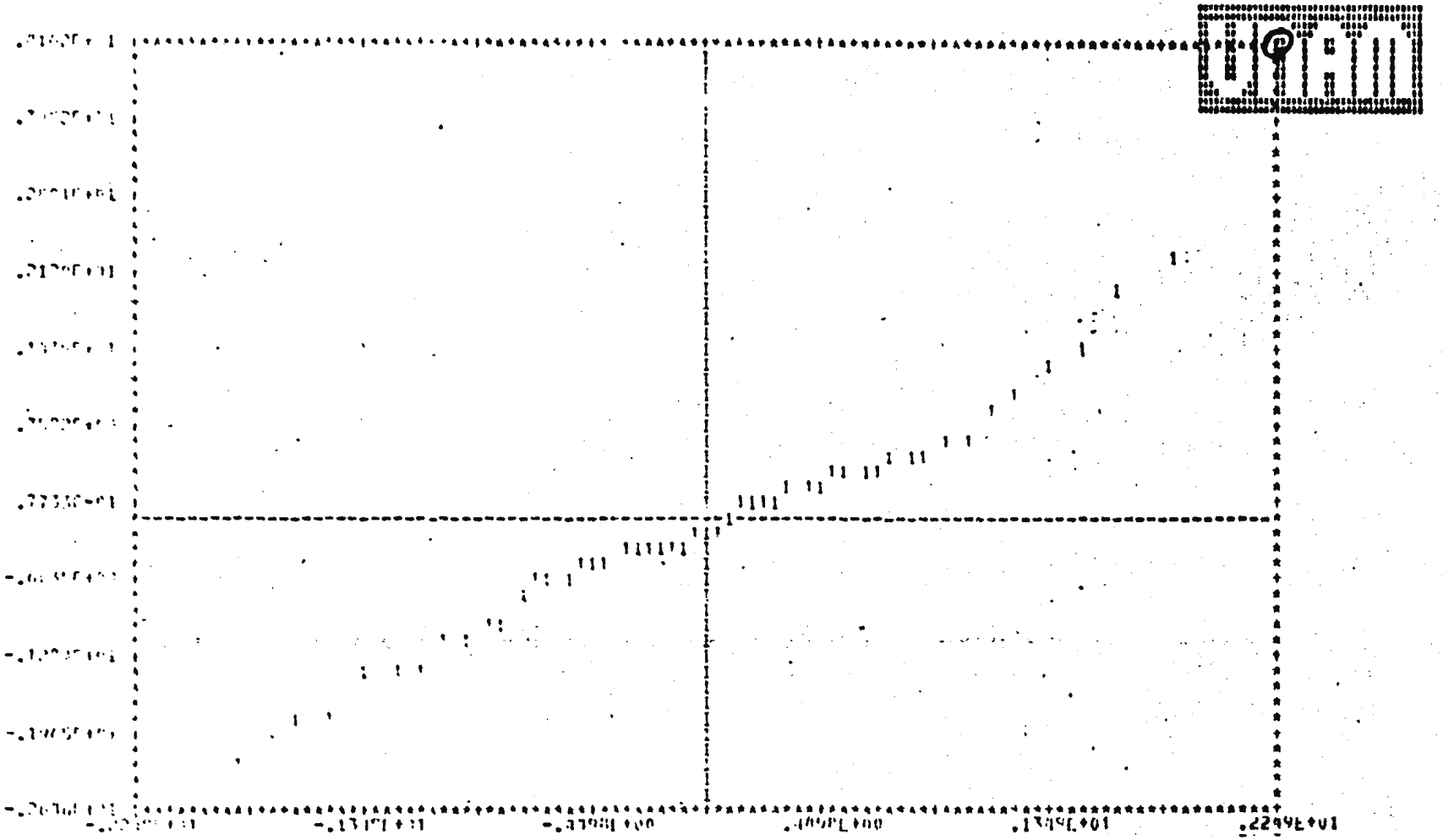
Quinta componente principal contra cuarta componente principal

N= 100, 10 aberrantes, bajo d_2

Los 10 aberrantes dentro del conjunto de datos

A N E X O 2

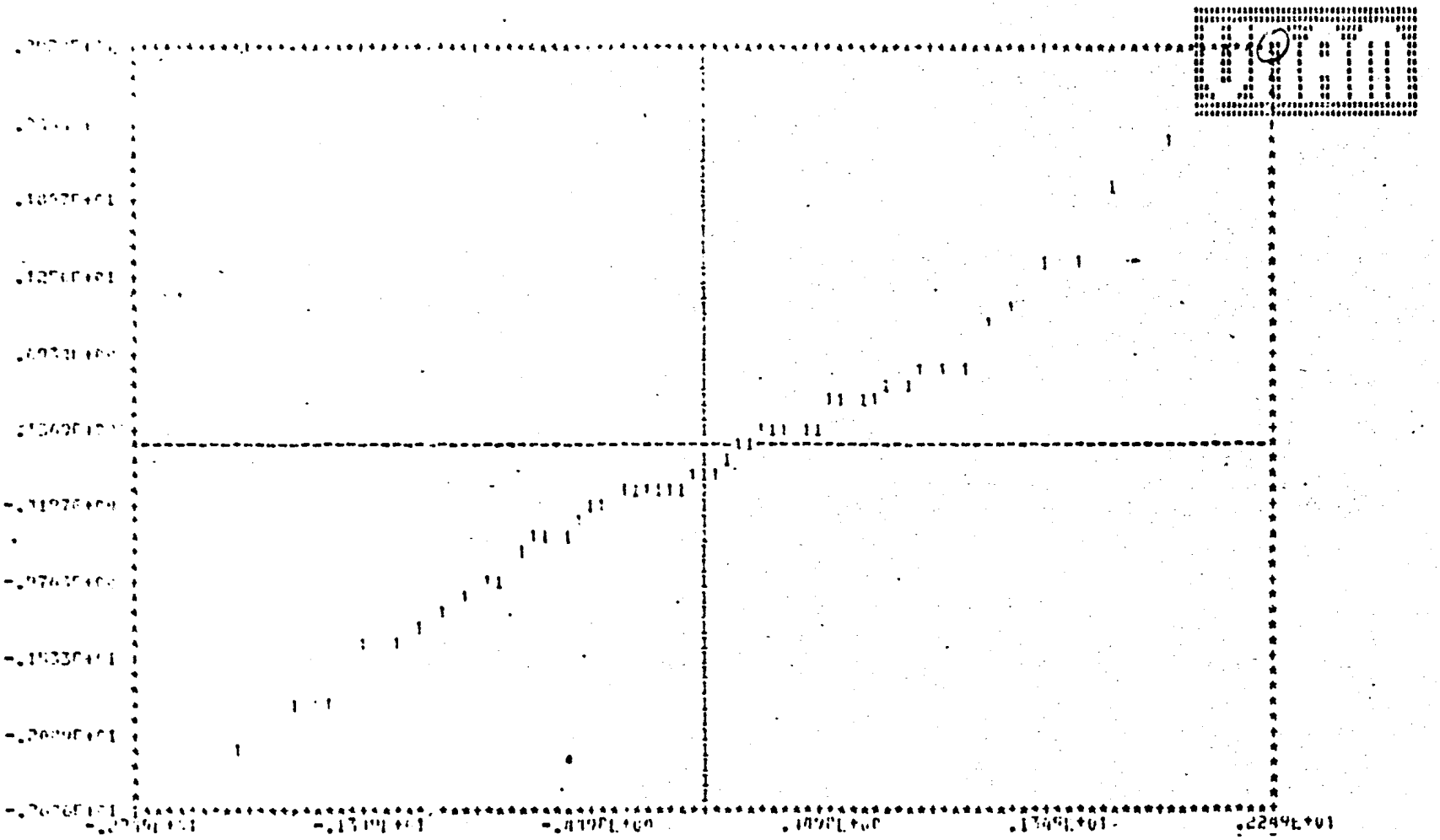
Los resultados de las gráficas de las muestras marginales, se exponen en este anexo, graficando cada una de las variables ordenadas, contra el valor esperado de las estadísticas de orden. Se presentan primero las muestras de tamaño 50 y luego las de 100, habiéndose seleccionado los casos más representativos. Las observaciones aberrantes se encierran también en un círculo.



$\overset{X}{=}_1$ contra. $\hat{\alpha}_i$

N= 50, 1 aberrante, bajo d_1

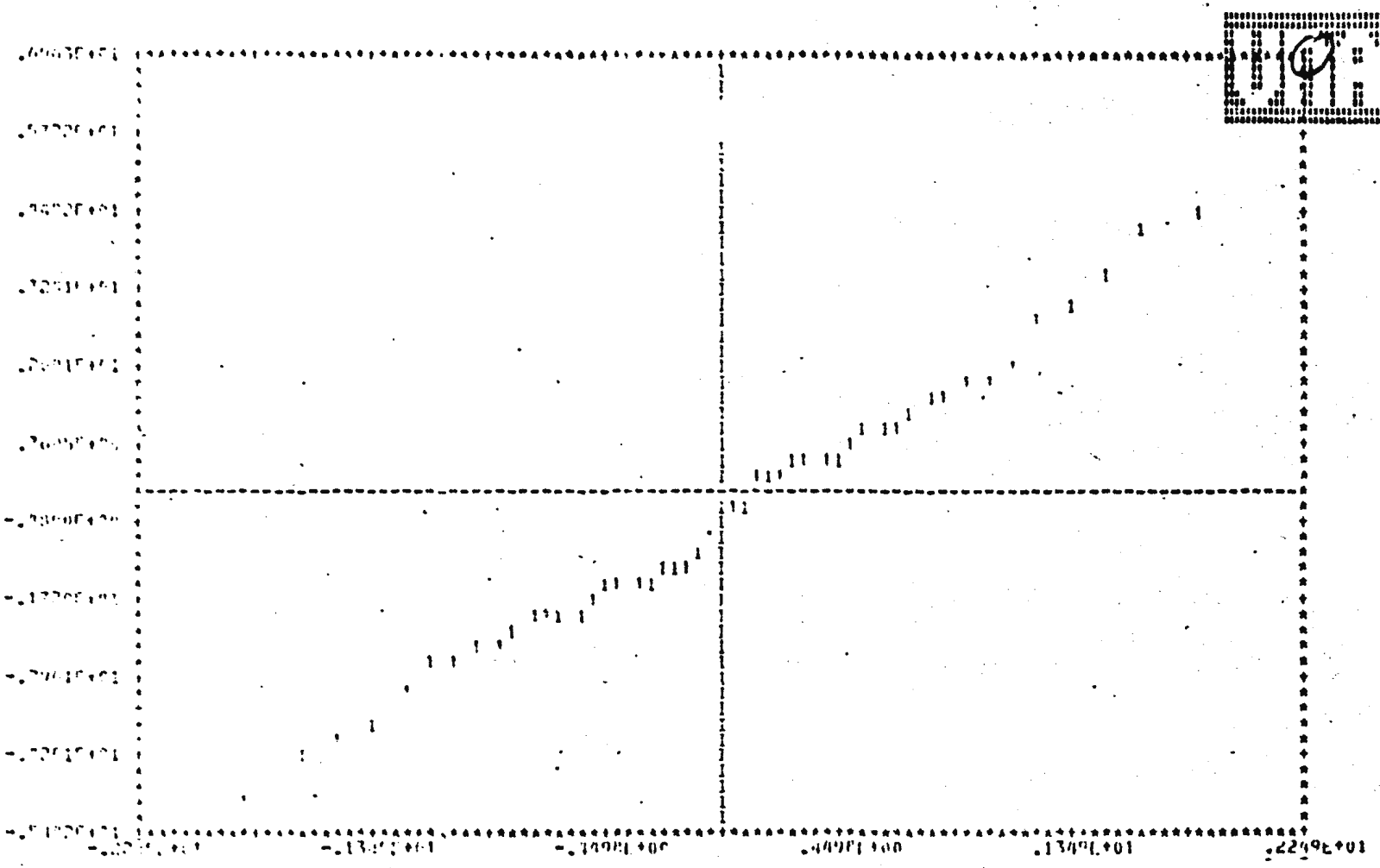
El aberrante fuera de la tendencia lineal y como extremo. Las otras marginales exhiben un comportamiento similar



$\underline{X}_1$ contra α_j

N= 50, 1 aberrante, bajo d_2

No se detecta fácilmente al aberrante

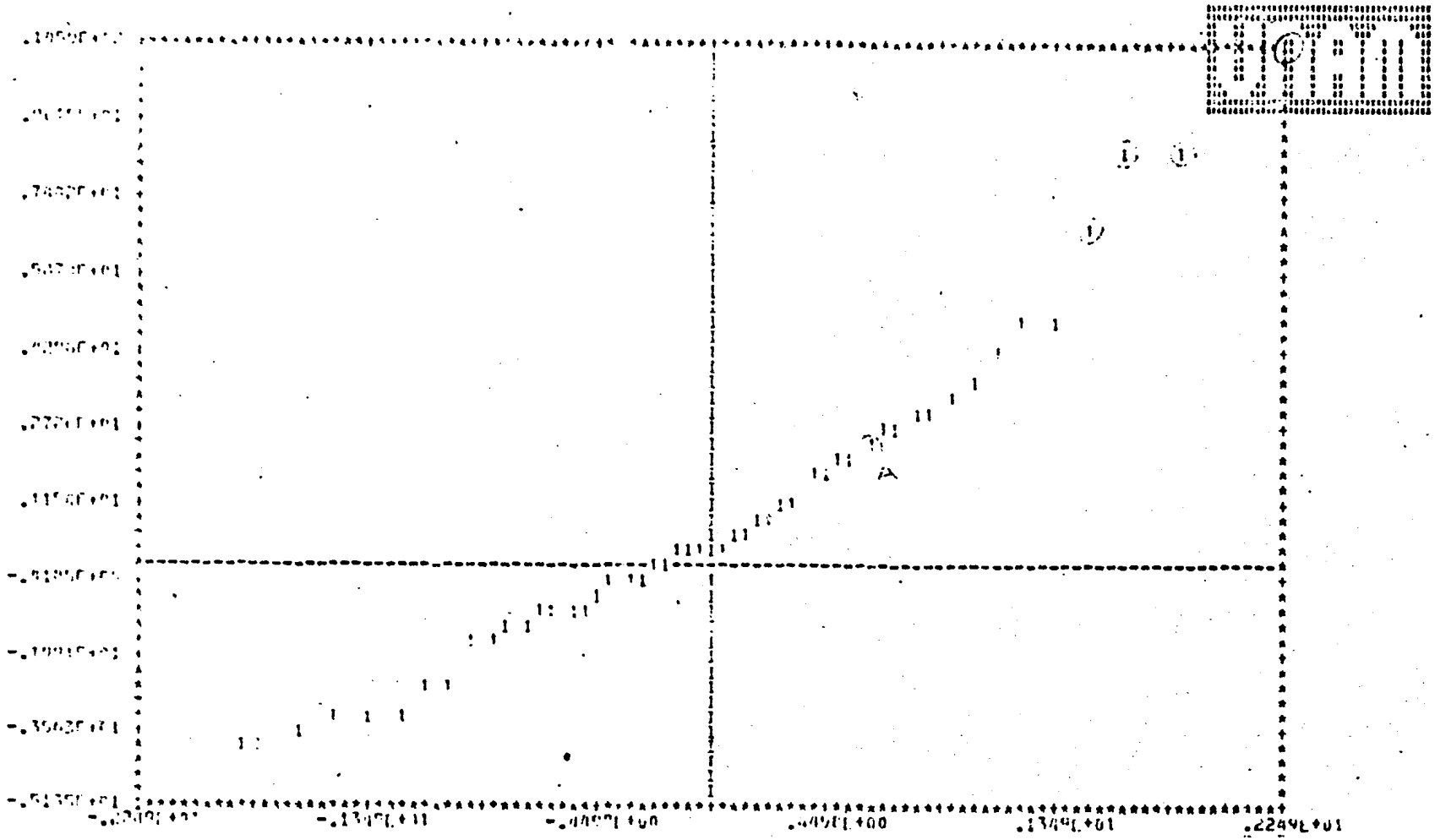


$\underline{X}_2$ contra α_j

N= 50, 1 aberrante, bajo d_2

El aberrante como extremo y fuera de la tendencia lineal

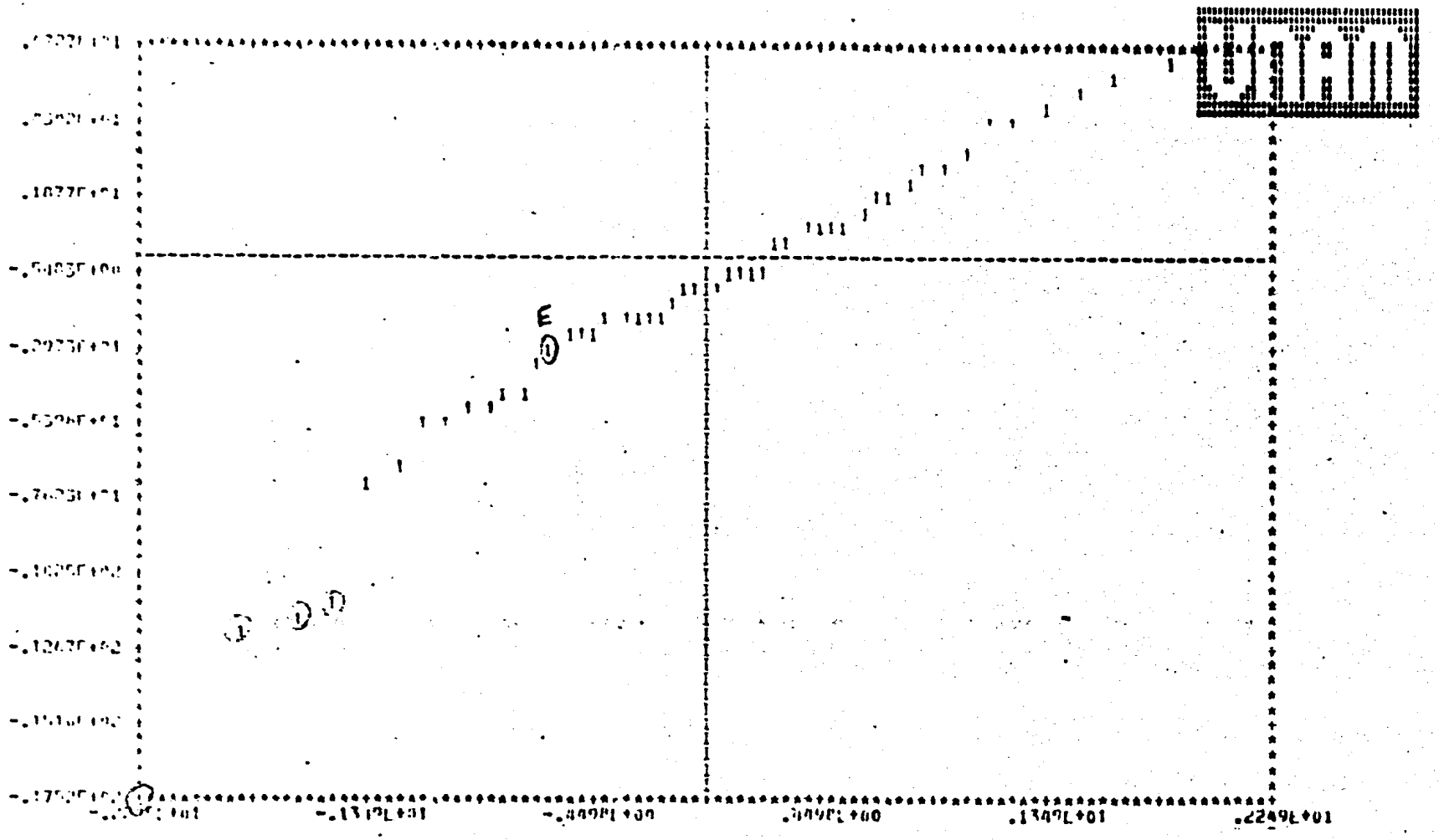
Los marginales 3, 4 y 5,.exhiben un comportamiento semejante



X_2 contra α_j

N= 50, 5 aberrantes, bajo d_1

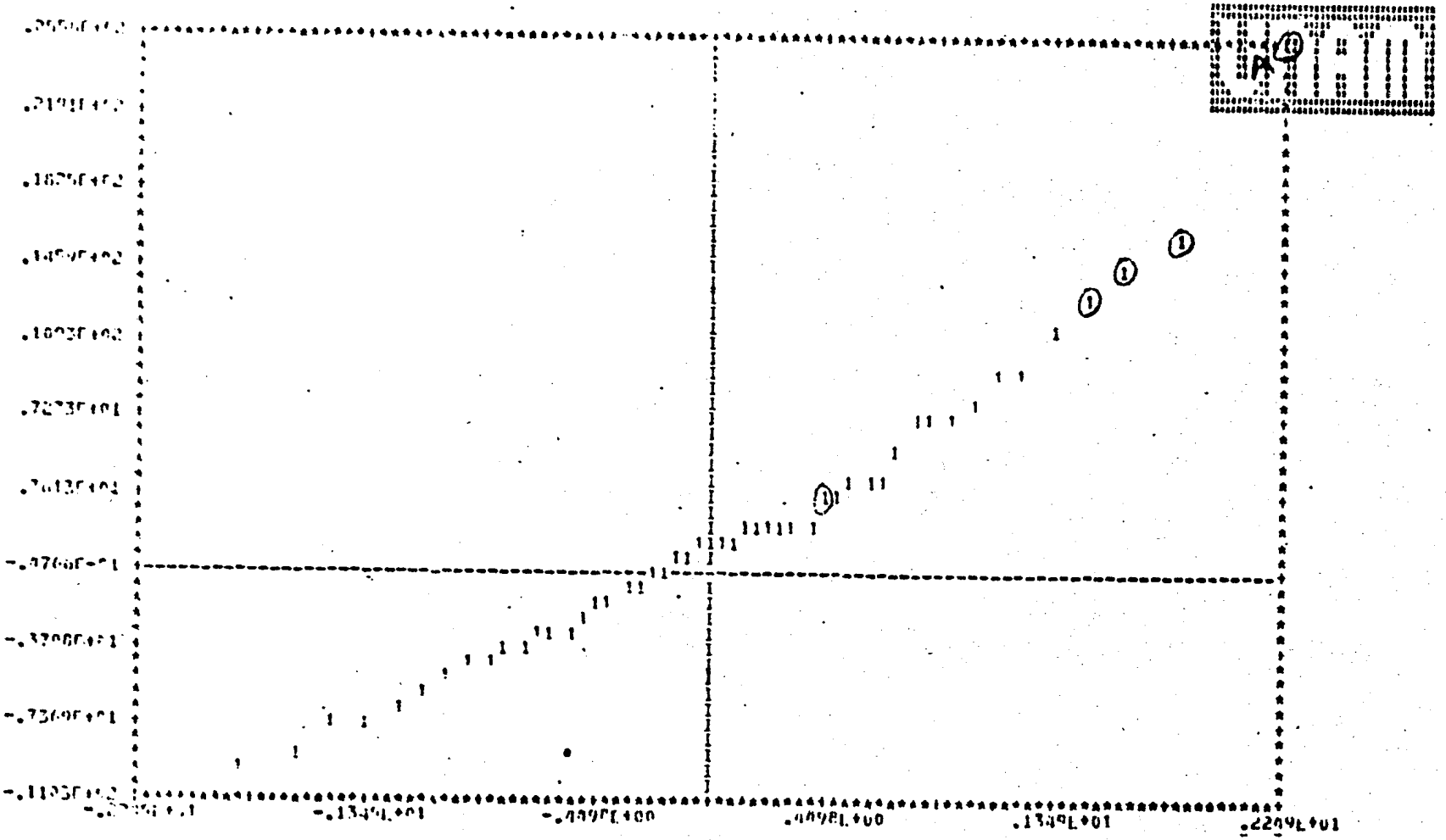
Los cuatro extremos son sospechosos a ser aberrantes,
pero no fácilmente los declaro como tal, debido
a que parece que están dentro del patrón lineal



$\underline{X}_3$ contra σ_j

N= 50, 5 aberrantes, bajo d_1

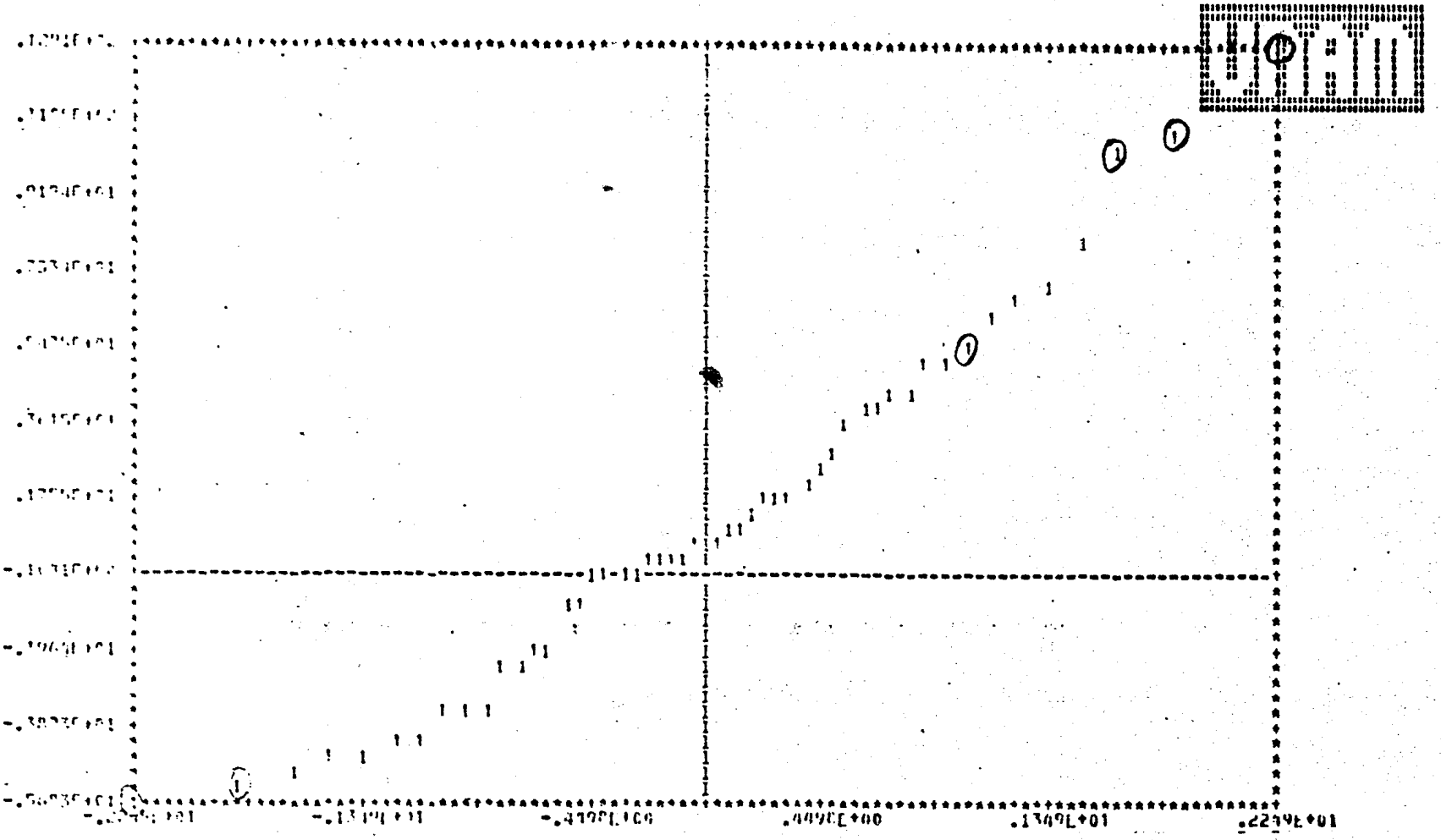
A excepción de la observación E, el resto de los aberrantes son muy obvios



$\underline{X}_4$ contra d_j

N= 50, 5 aberrantes, bajo d_1

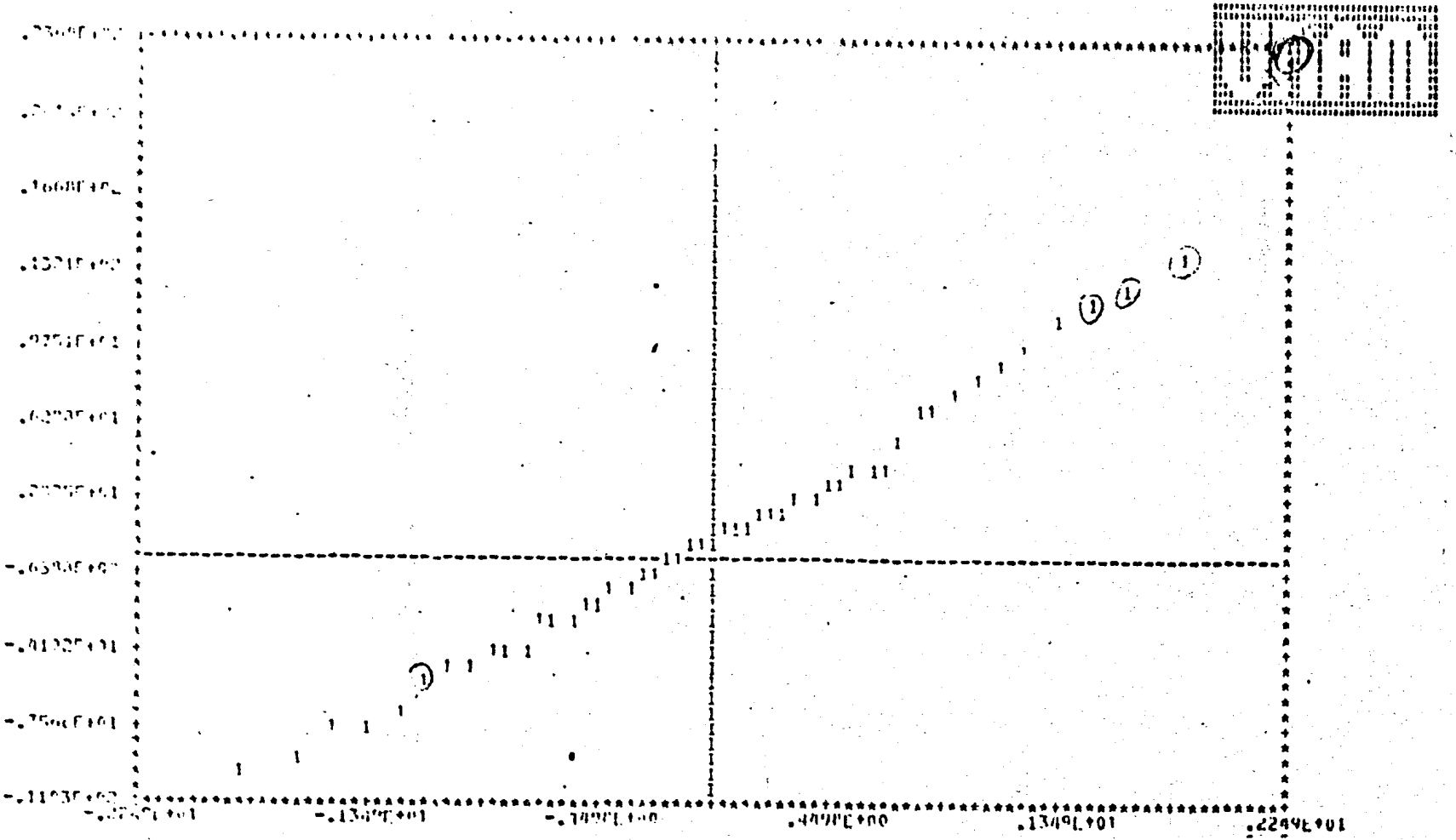
Solo el aberrante A es muy obvio



$\bar{X}_5$ contra α_j

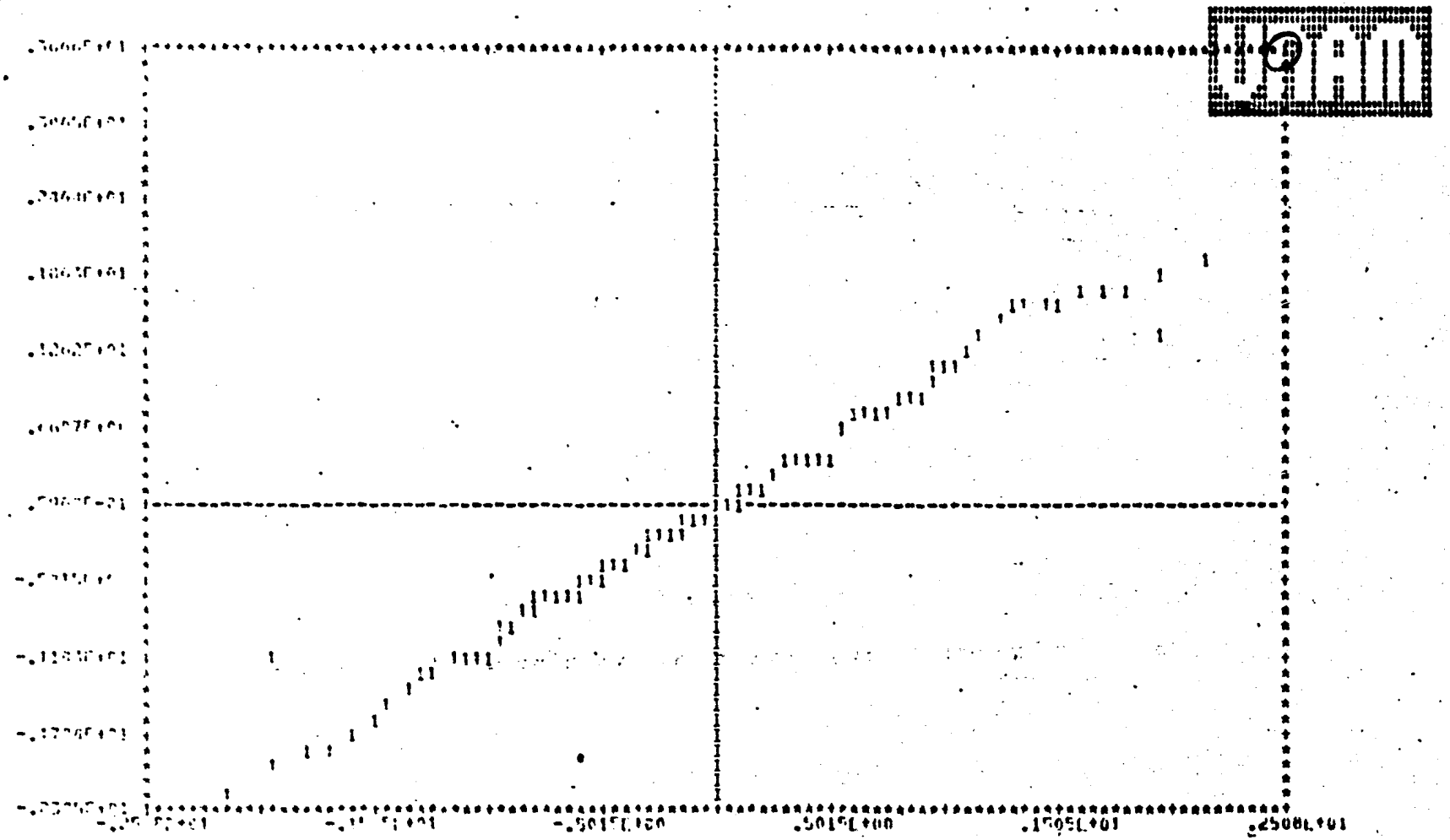
N= 50, 5 aberrantes, bajo d_1

Las observaciones mas sospechosas a ser aberrantes, serían C,D y E, pero no fácilmente las declararía como tal



$\underline{X}_4$ contra α_j

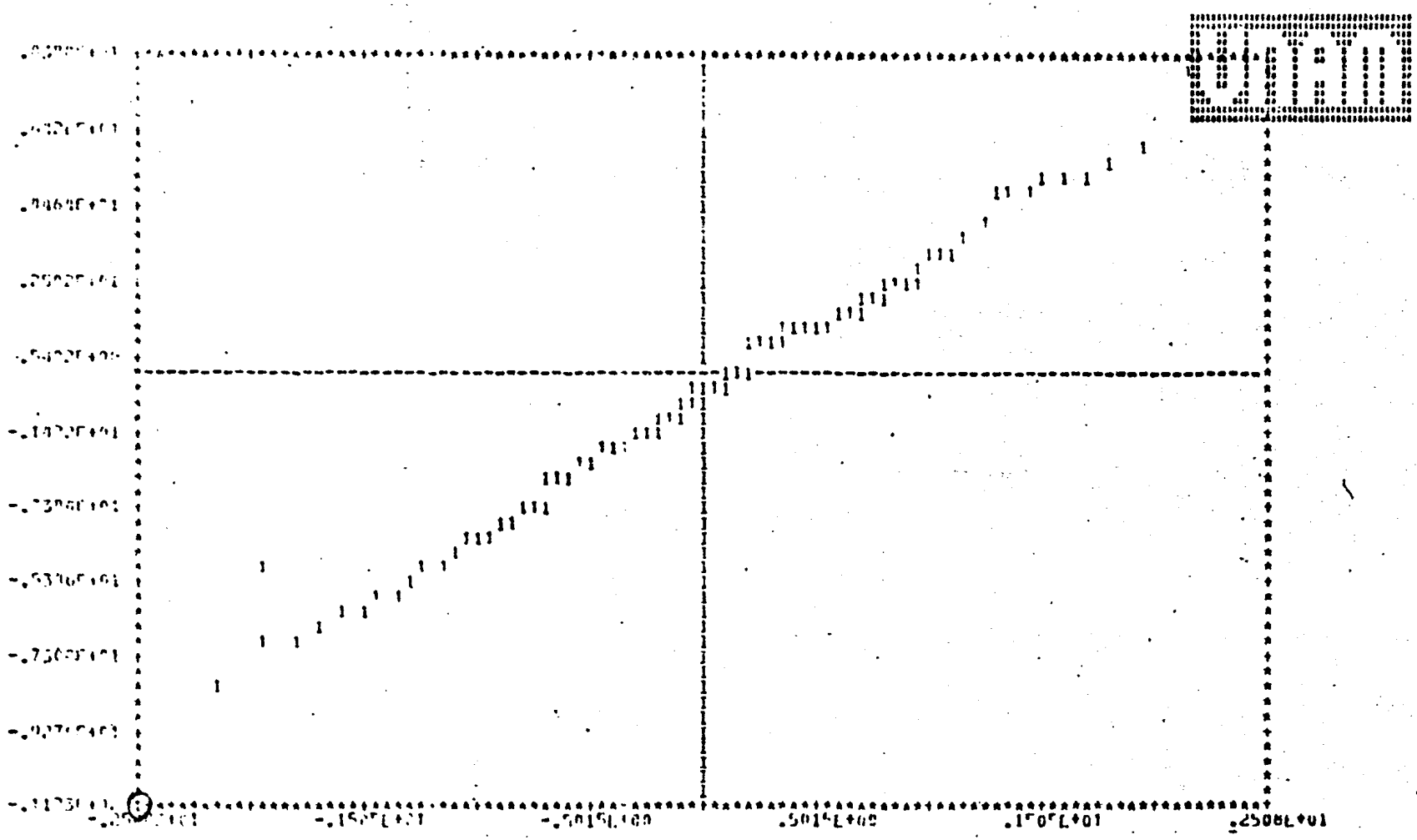
La observación A, es el único aberrante fuera de la tendencia lineal y como extremo



$\underline{X}_1$ contra α_j

N= 100, 1 aberrante, bajo d_1

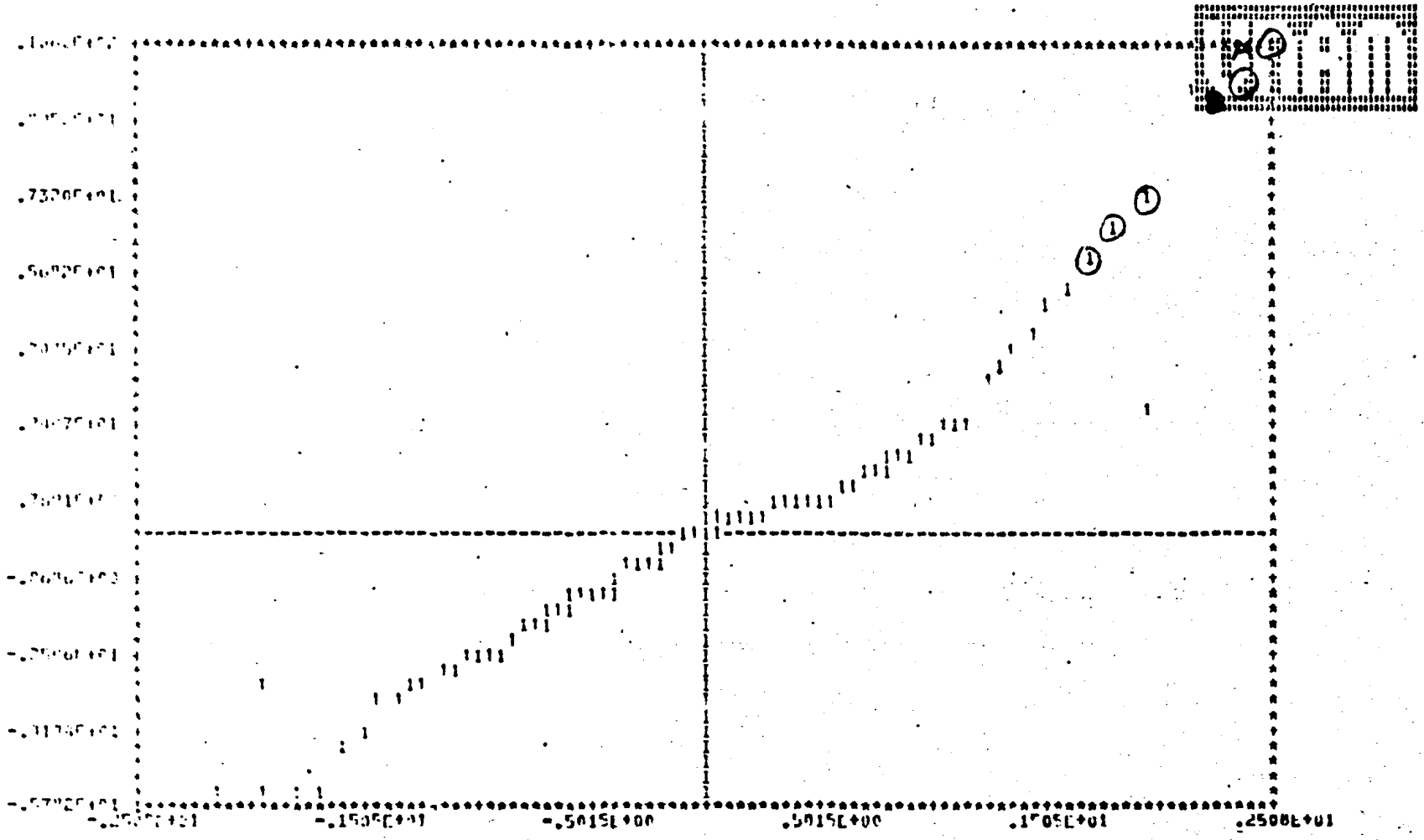
Muy obvio el aberrante



$\underline{X}_3$ contra α_j

N= 100, 1 aberrante, bajo d_1

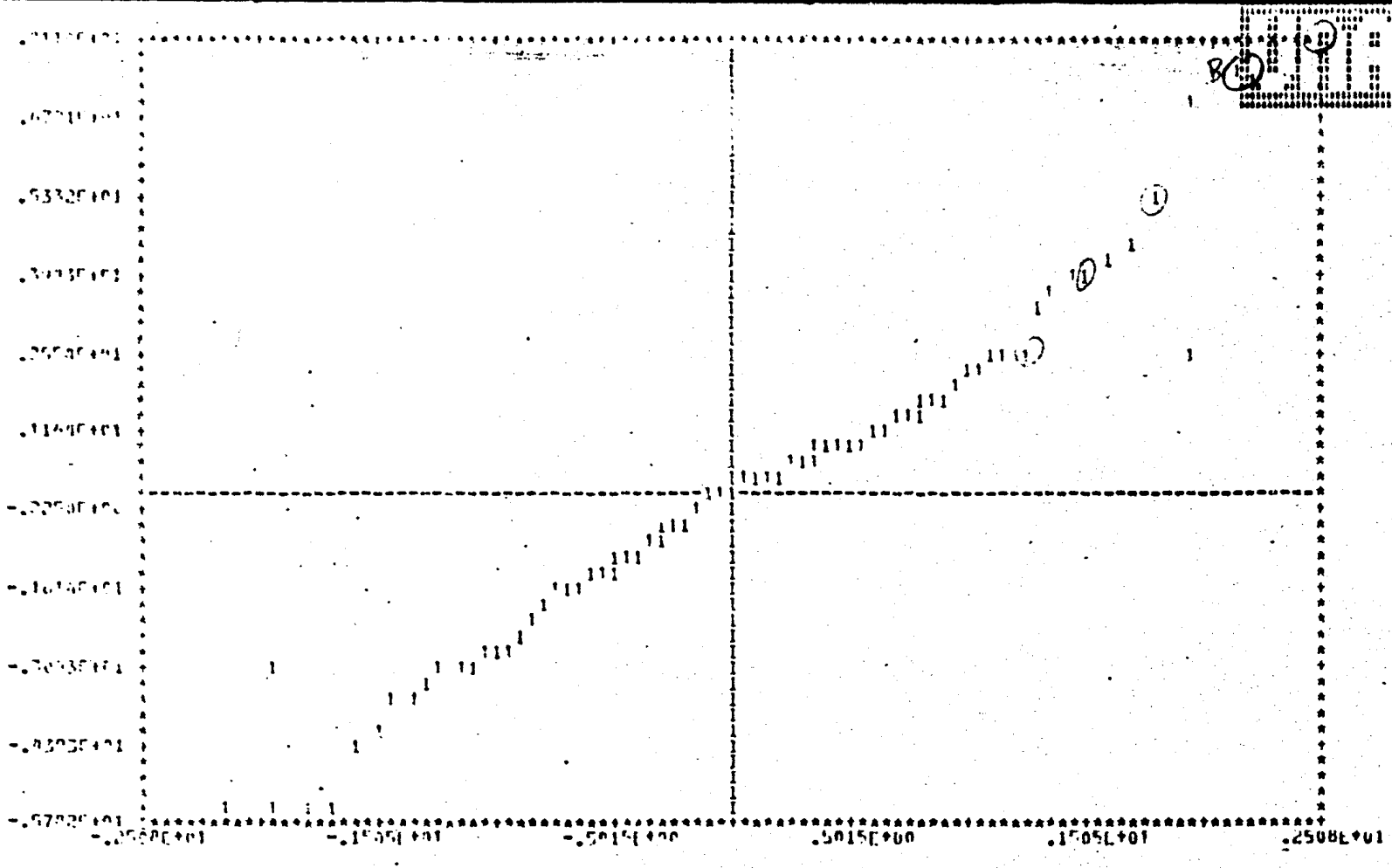
Sospecharía de la observación mas alejada, pero no fácilmente la declararía como aberrante



$\underline{X}_2$ contra α_j

N= 100, 5 aberrantes, bajo d_1

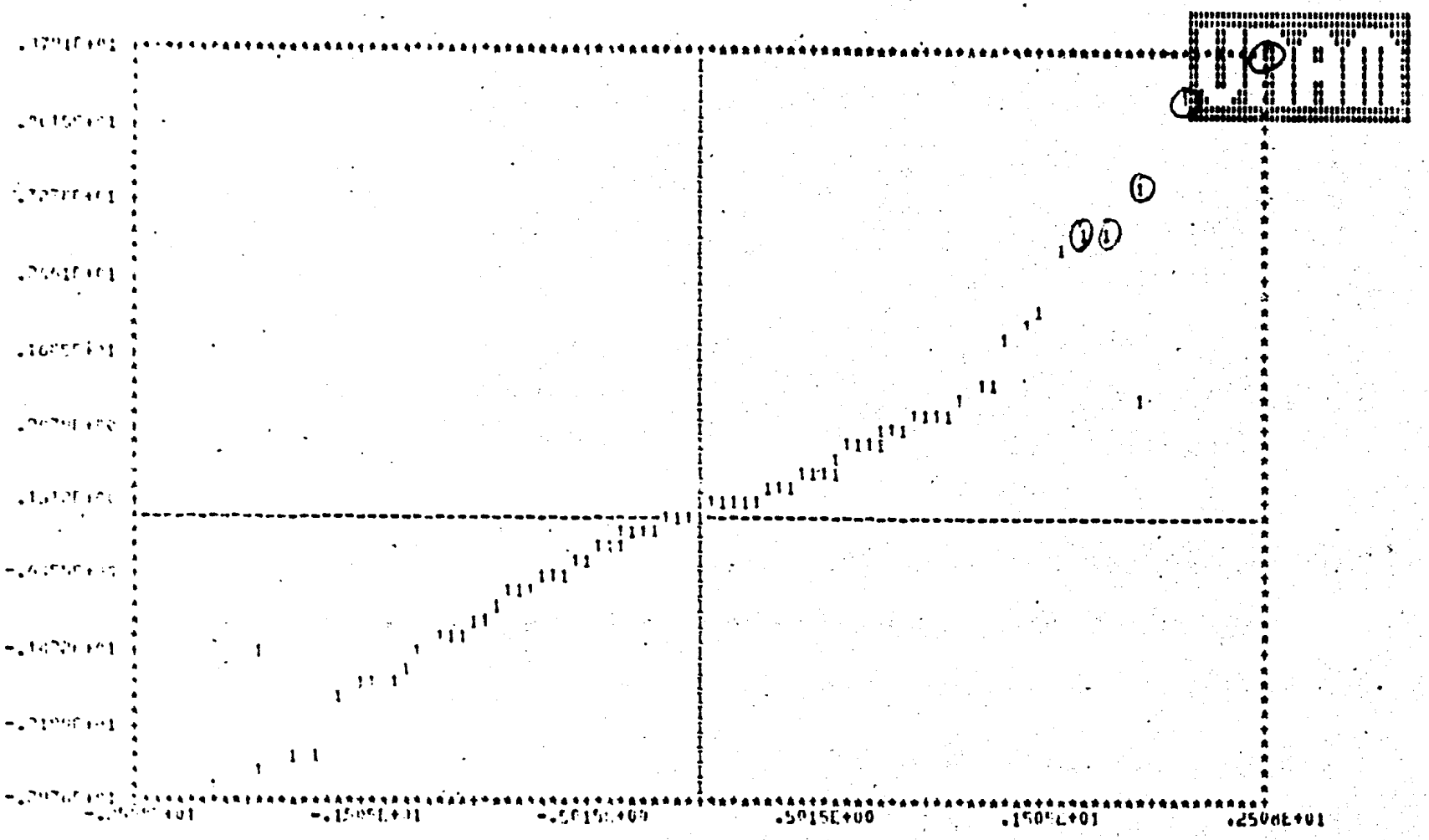
Las observaciones mas sospechosas a ser aberrantes son A y B; pero no fácilmente las declararía como tal



$\underline{X}_2$ contra α_j

N= 100, 5 aberrantes, bajo d_2

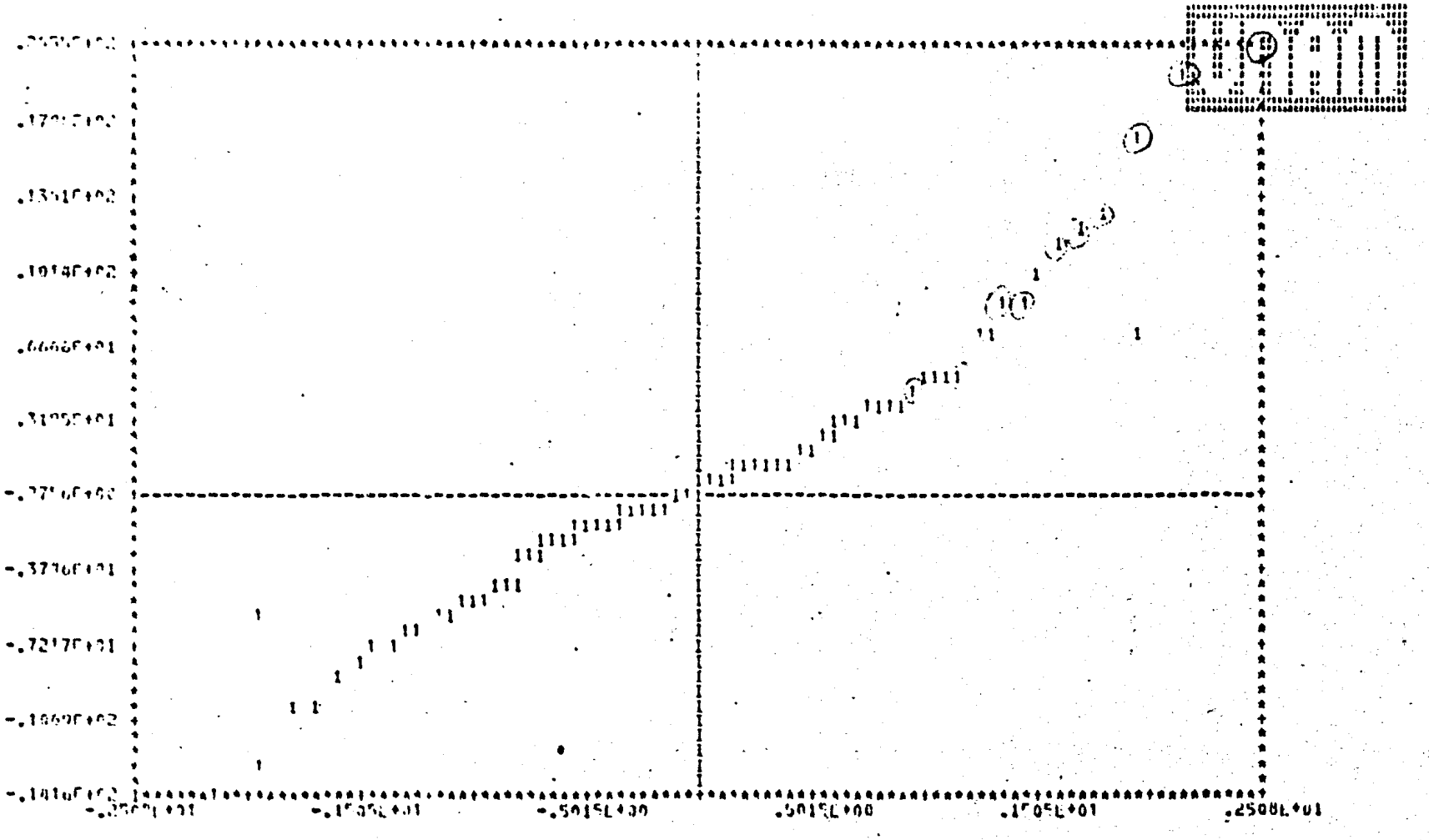
Las observaciones mas sospechosas son A y B



$\underline{X}_1$ contra α_j

N= 100, 1 aberrante, bajo d_2

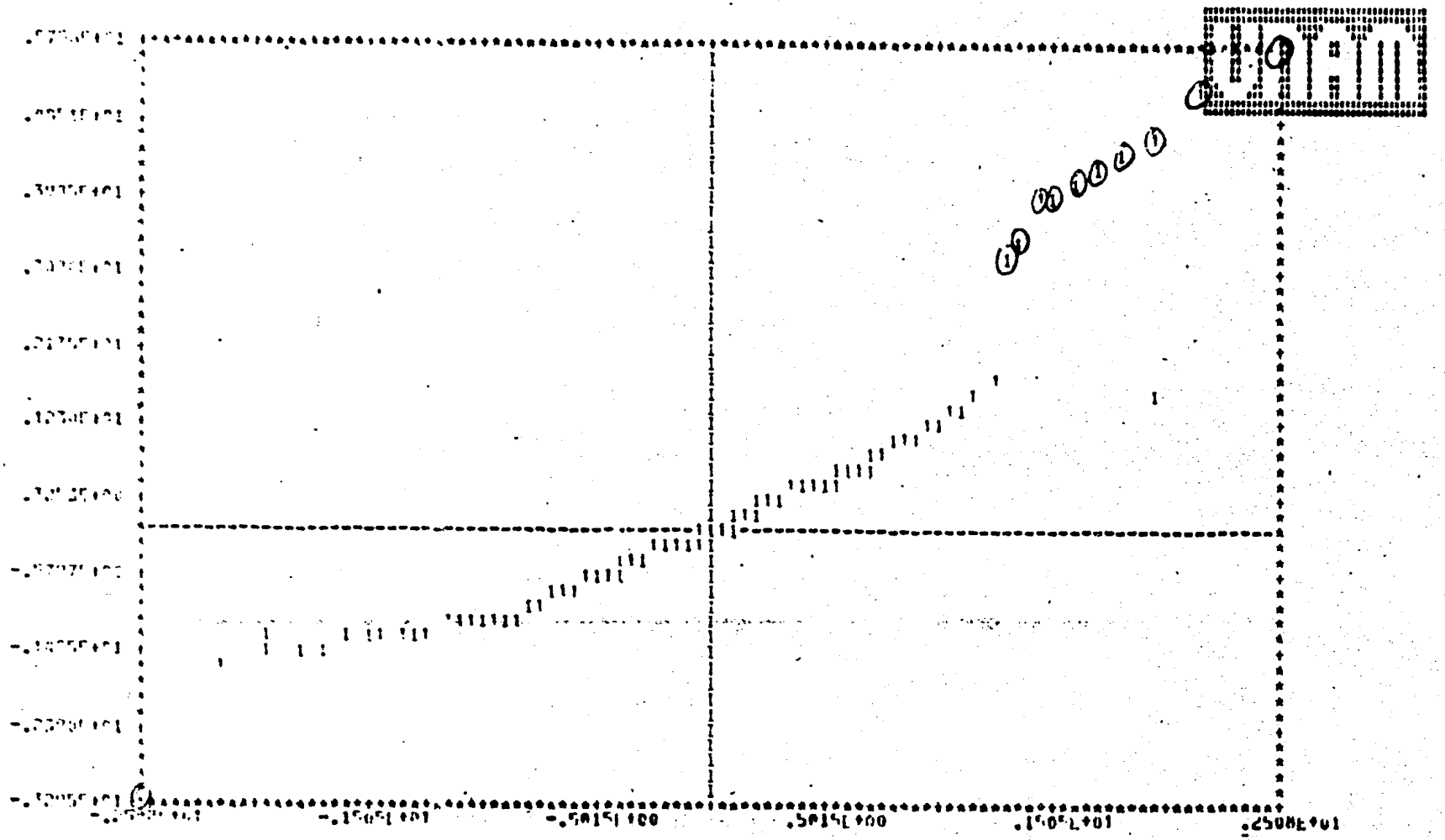
Dudaría en declarar a alguna observación como aberrante



$\underline{X_4}$ contra α_j

N= 100, 5 aberrantes, bajo d_2

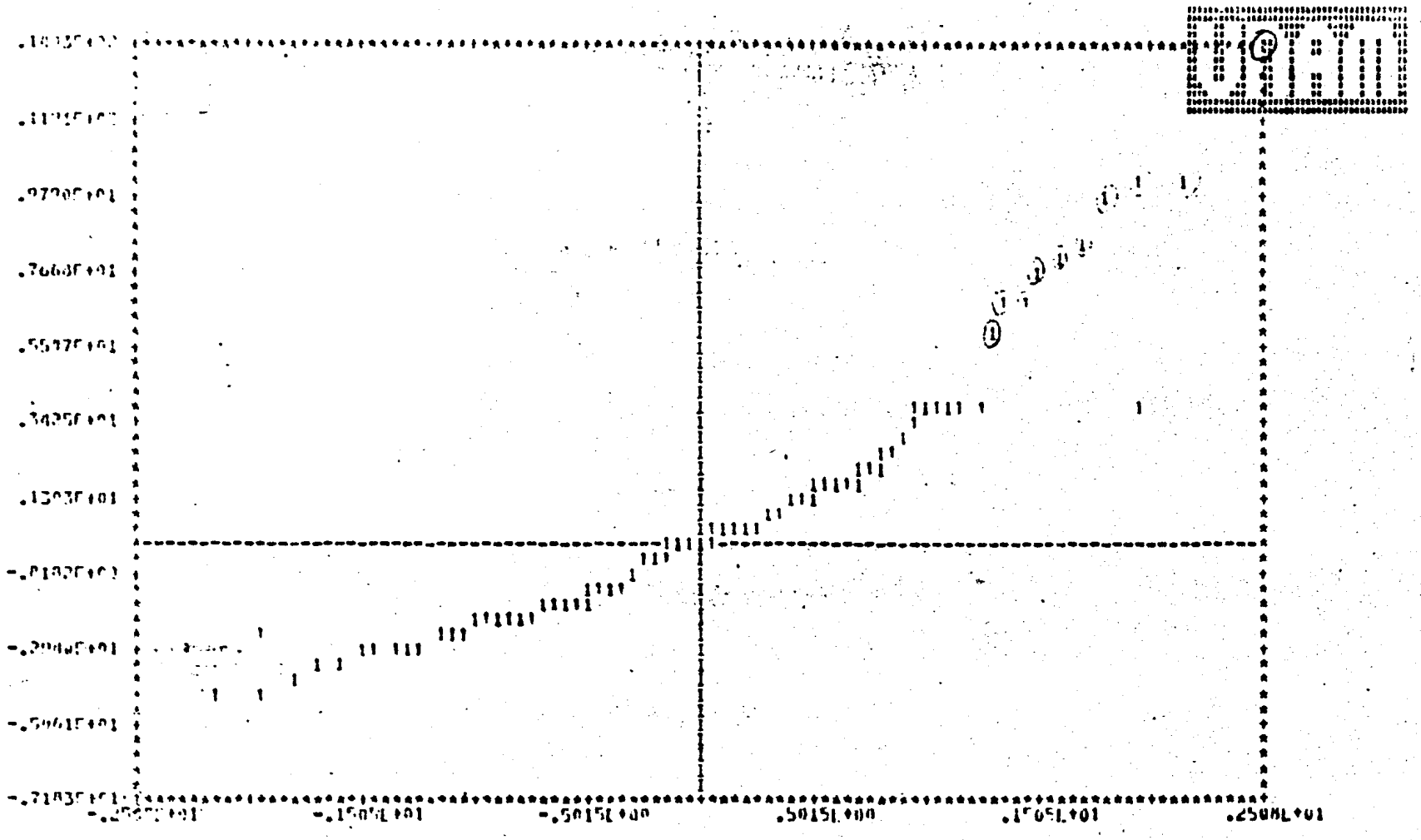
Dudaría en declarar a alguna observación como aberrante



$\underline{X}_1$ contra α_j

N= 100, 10 aberrantes, bajo d_1

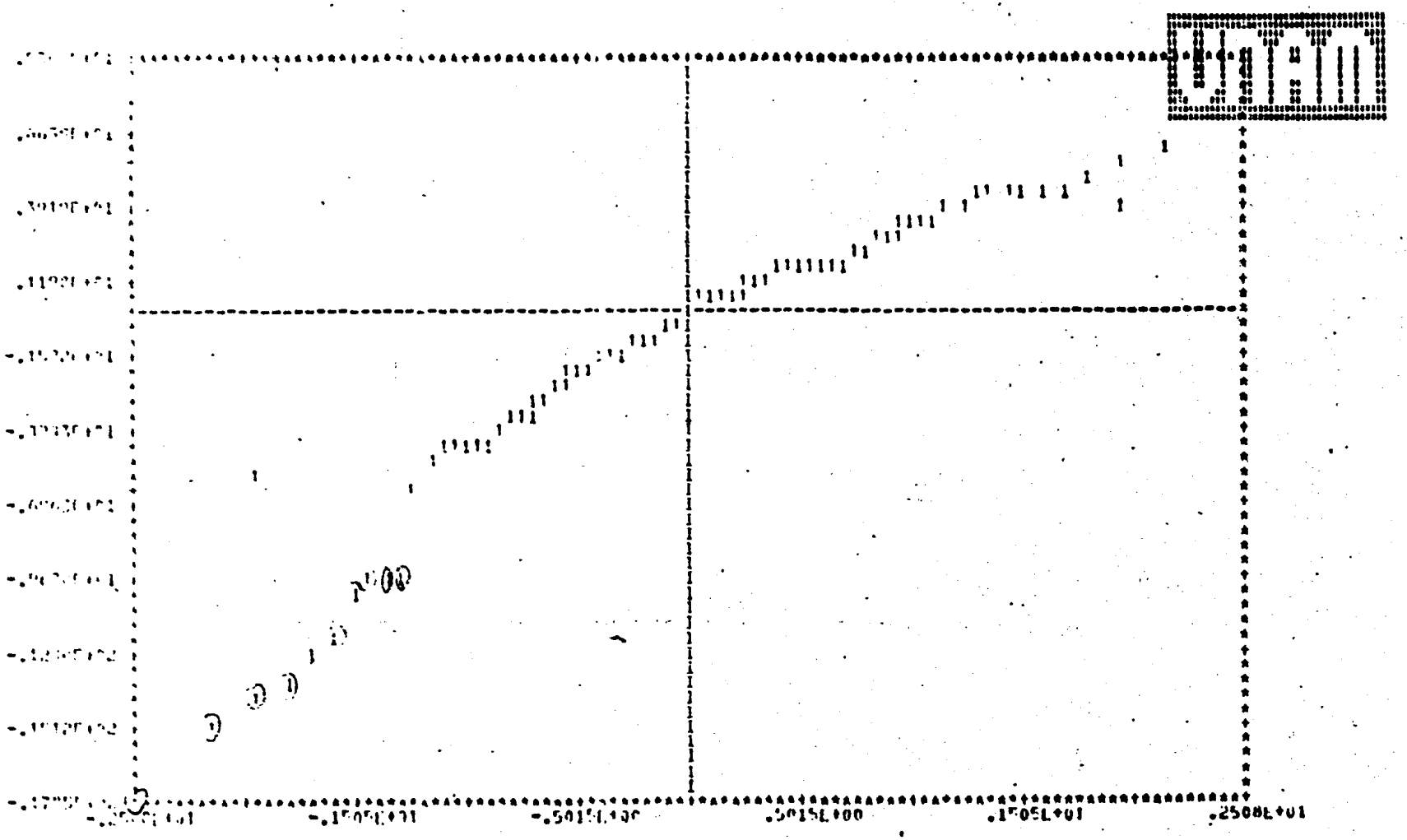
Los 10 aberrantes se detectan fácilmente



$\underline{X}_2$ contra α_j

N= 100, 10 aberrantes, bajo d_1

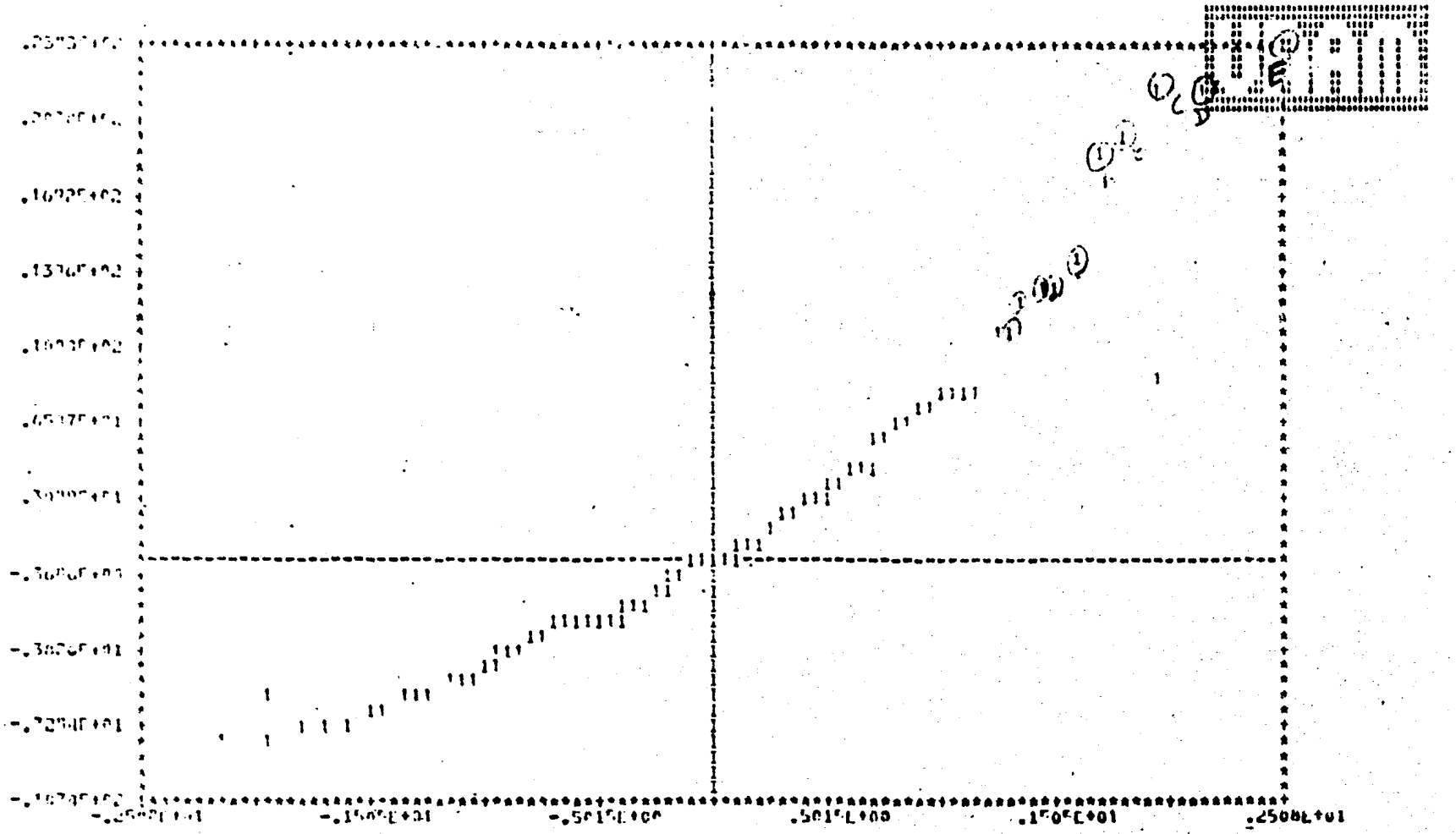
Todos los aberrantes se detectan fácilmente



$\underline{X}_3$ contra α_j

N= 100, 10 aberrantes, bajo d_1

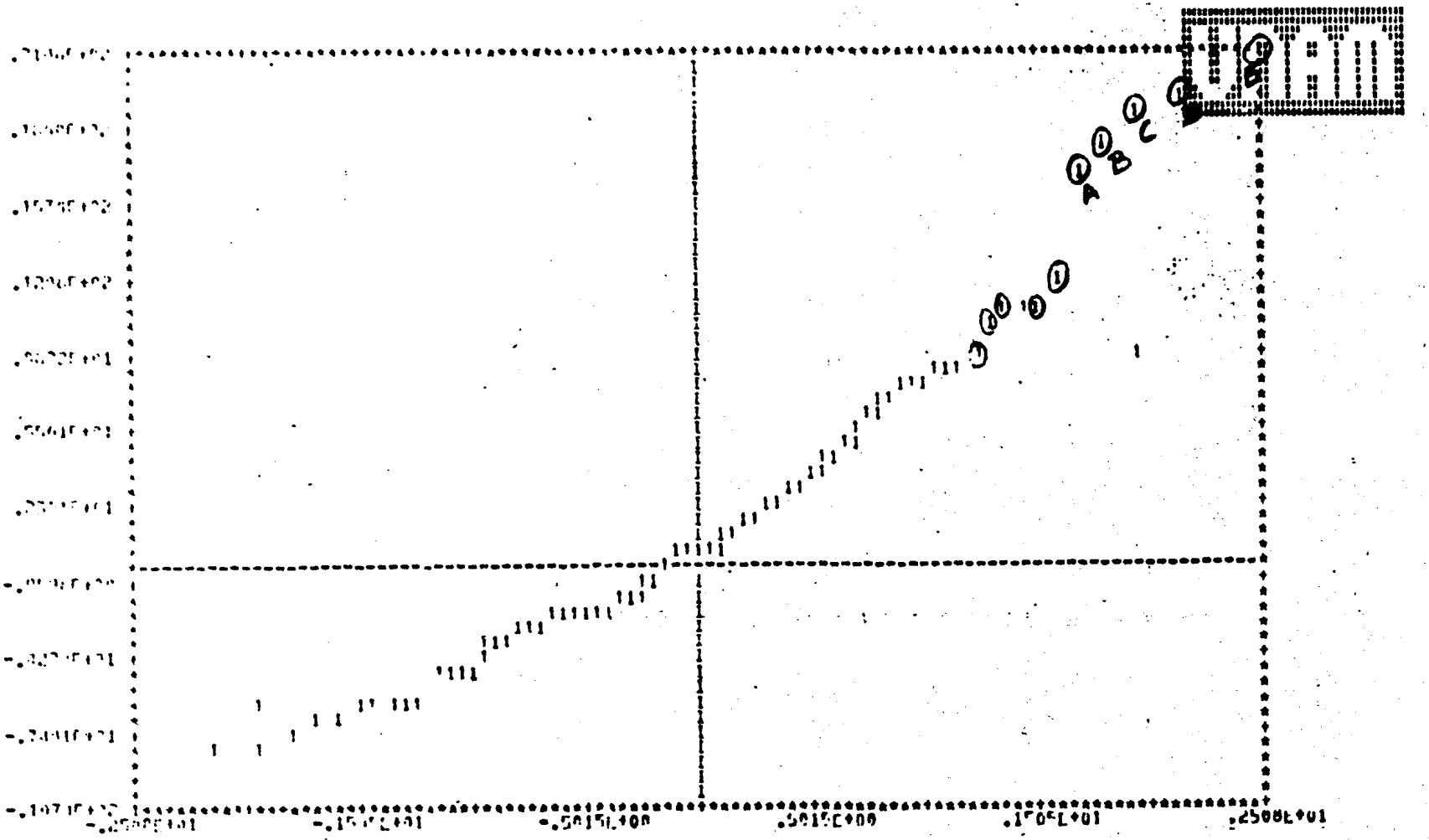
Todos los aberrantes se detectan fácilmente



$\underline{X}_q$ contra α_j

N= 100, 10 aberrantes, bajo d_1

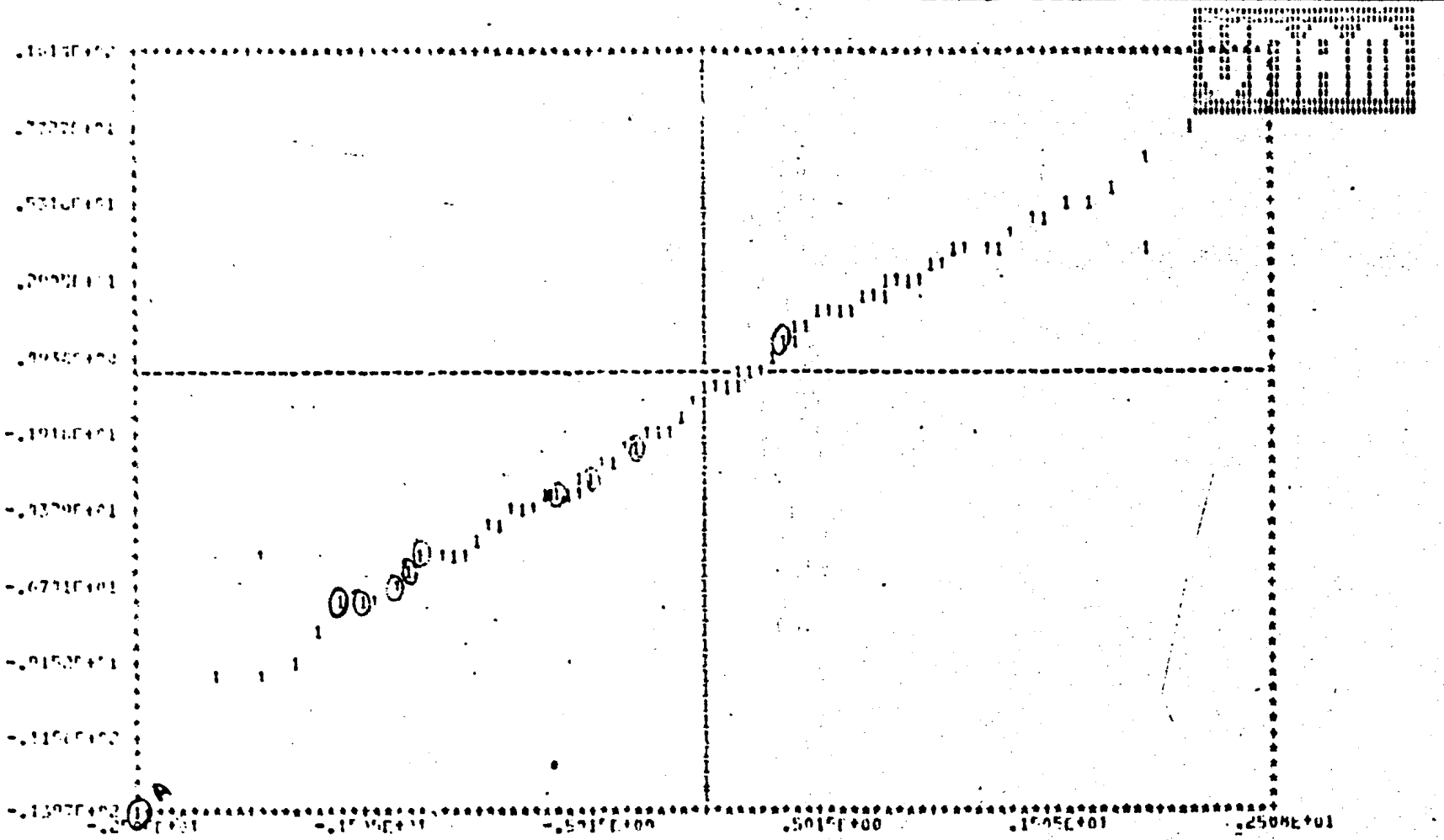
Me parecen mas obvios los aberrantes A,B, C y D, y el resto,
no muy fácilmente los detectaría



$\underline{X}_4$ contra α_j

N= 100, 10 aberrantes, bajo d_2

Las observaciones A, B, C, y D fácilmente las declararía como aberrantes



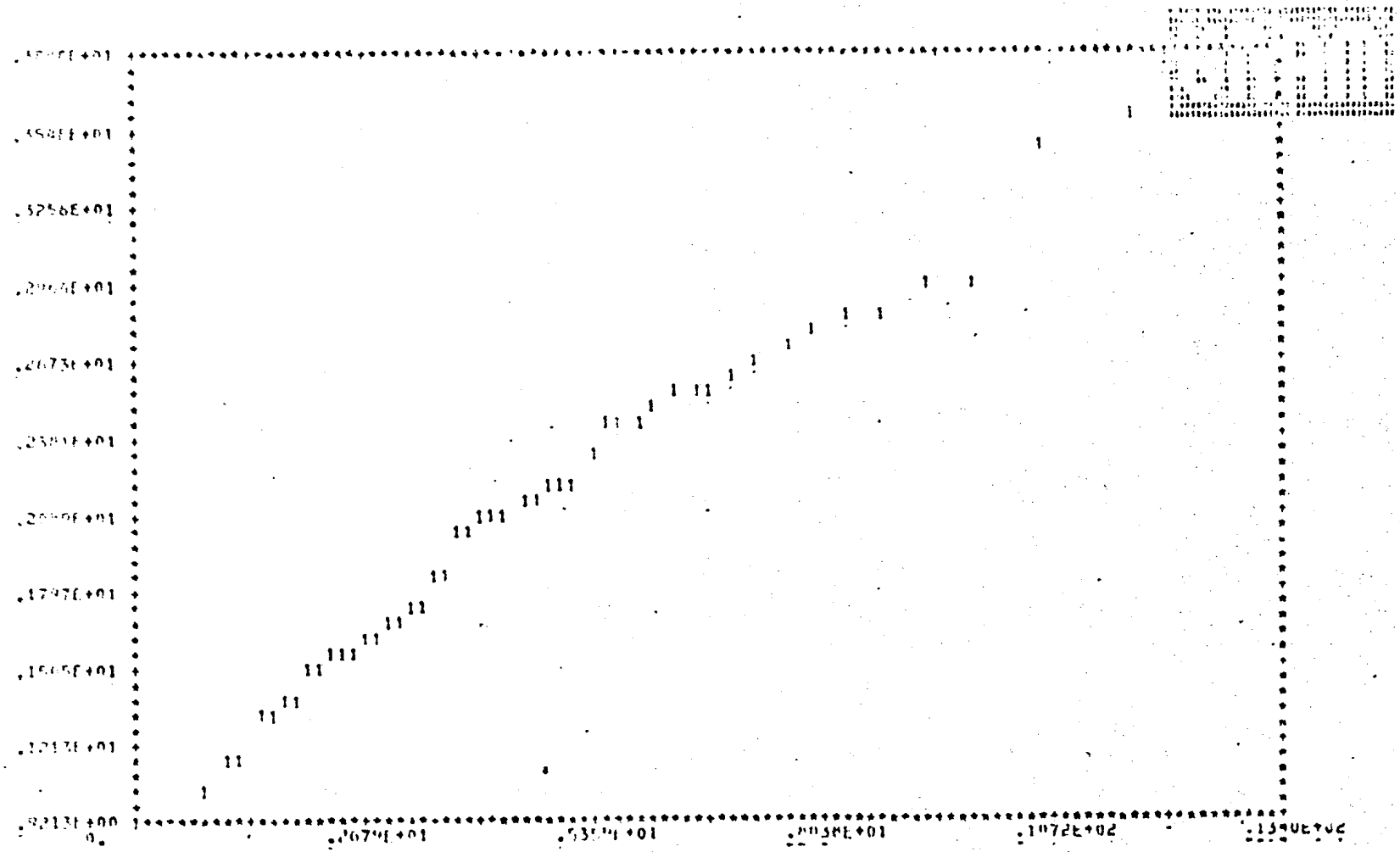
$\underline{X}_5$ contra α_j

N= 100, 10 aberrantes, bajo d_2

La observación A es la única sospechosa a ser aberrante

A N E X O 3

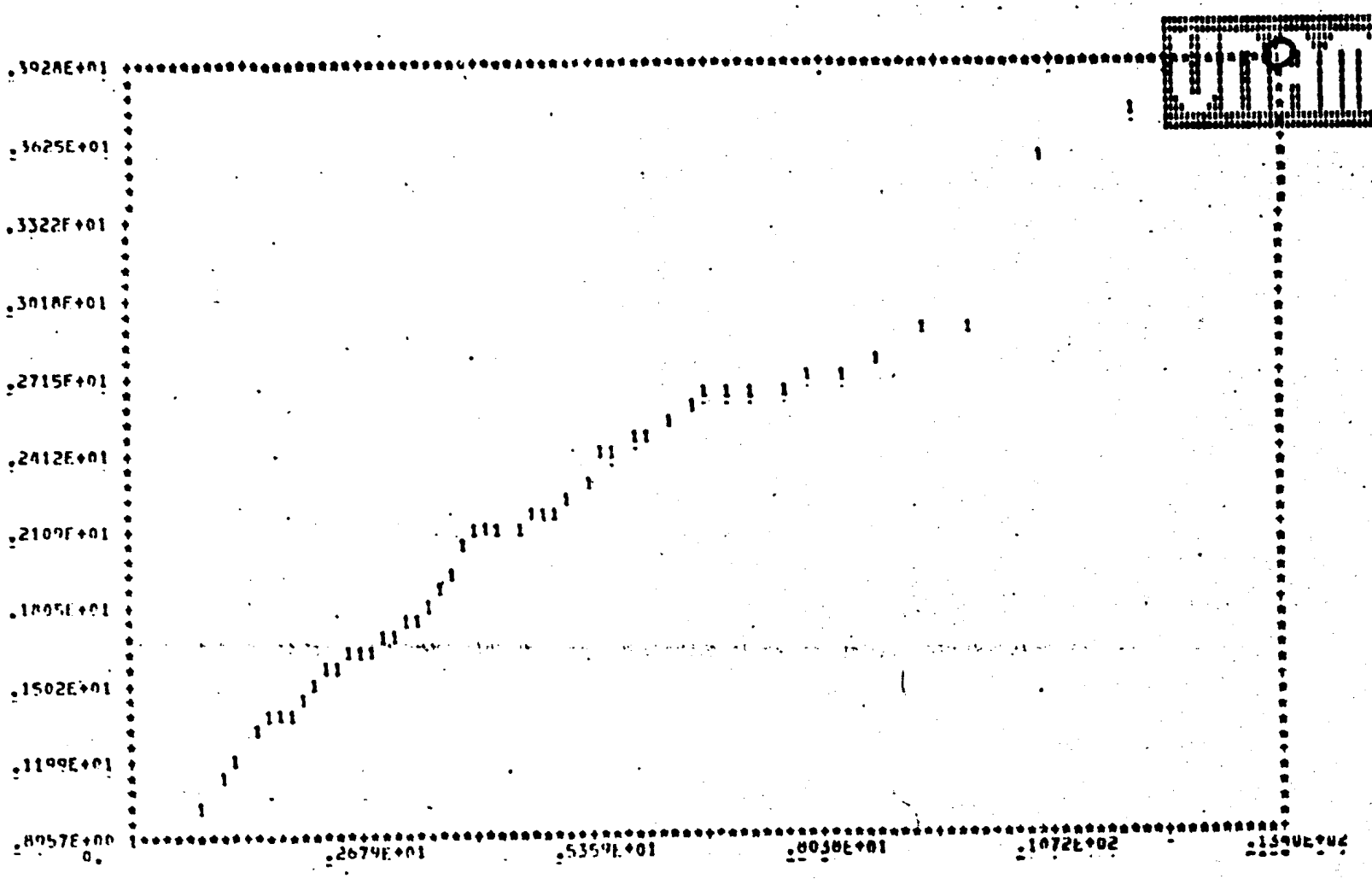
Los resultados que se obtuvieron al graficar la raíz cuadrada de las distancias ordenadas $(X_i - \bar{X})' S^{-1}(X_i - \bar{X})$ contra χ^2_5, son presentados en este anexo, exhibiéndose a los aberrantes encerrados en un círculo.



Raiz cuadrada de las distancias contra d_1

N= 50, 1 aberrante, bajo d_1

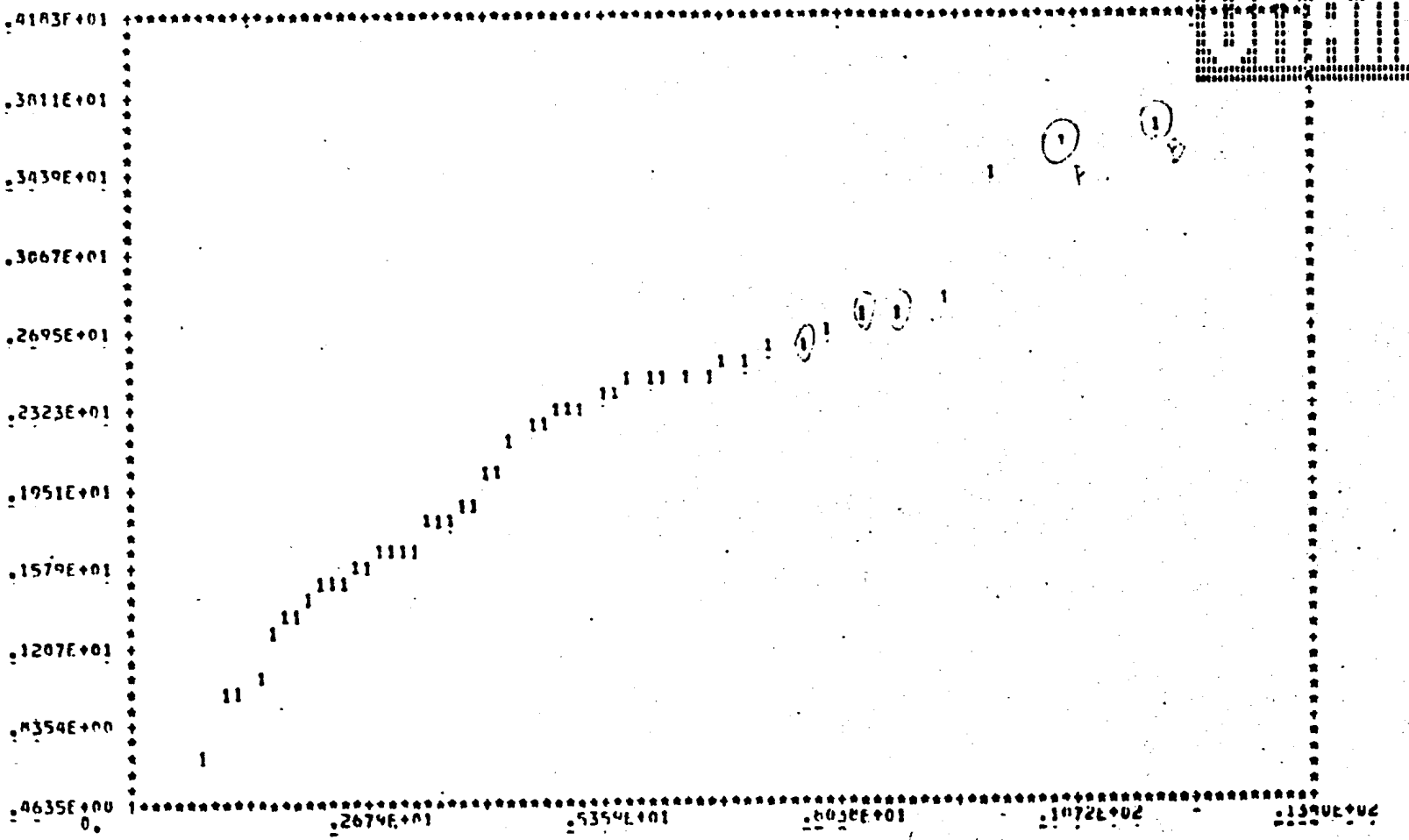
Se detecta fácilmente el aberrante



Raíz cuadrada de las distancias contra χ^2_5

N= 50, 1 aberrante, bajo d_2

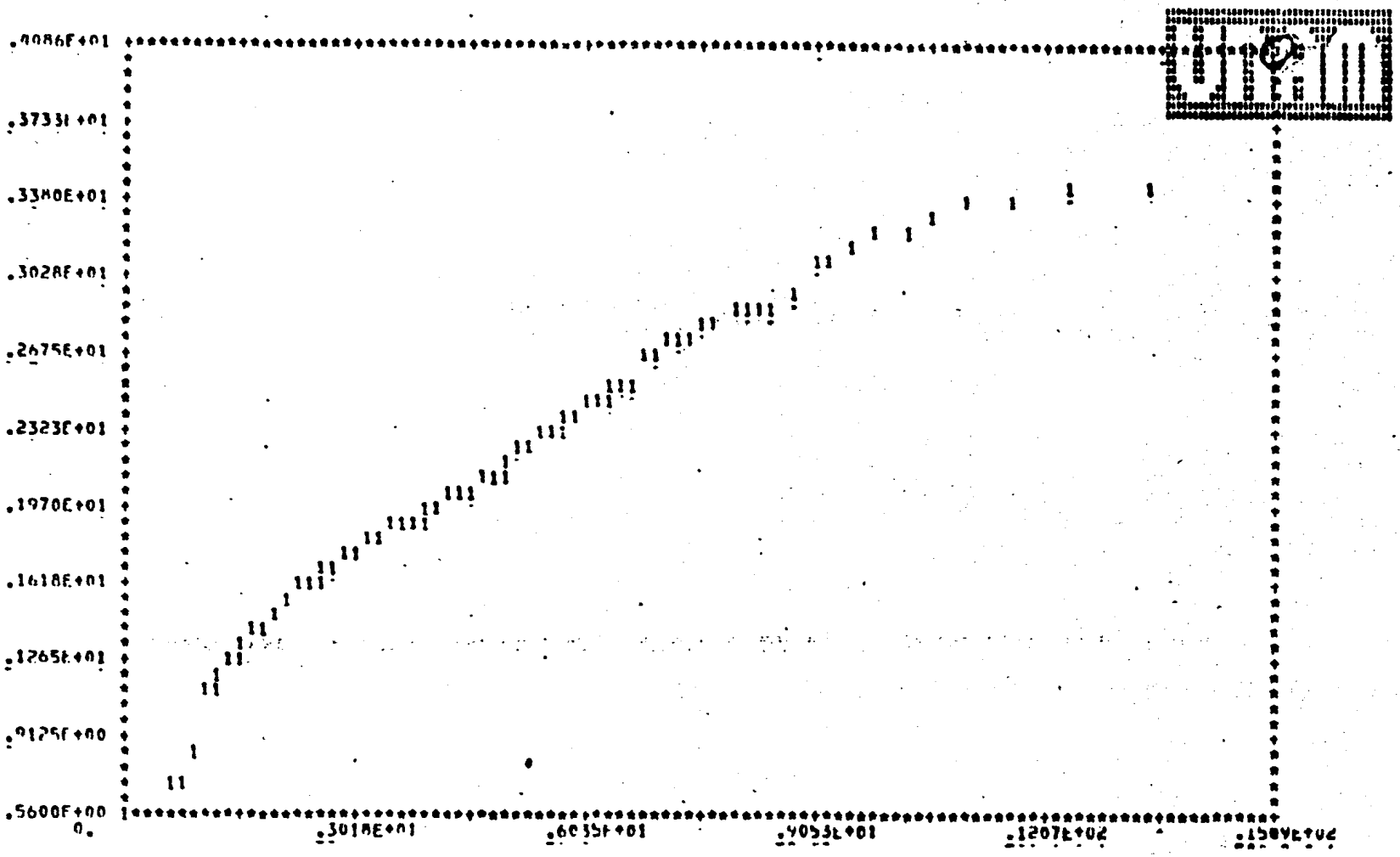
El aberrante bastante alejado del conjunto de datos



Raíz cuadrada de las distancias contra χ^2_5

N= 50, 5 aberrantes, bajo d_2

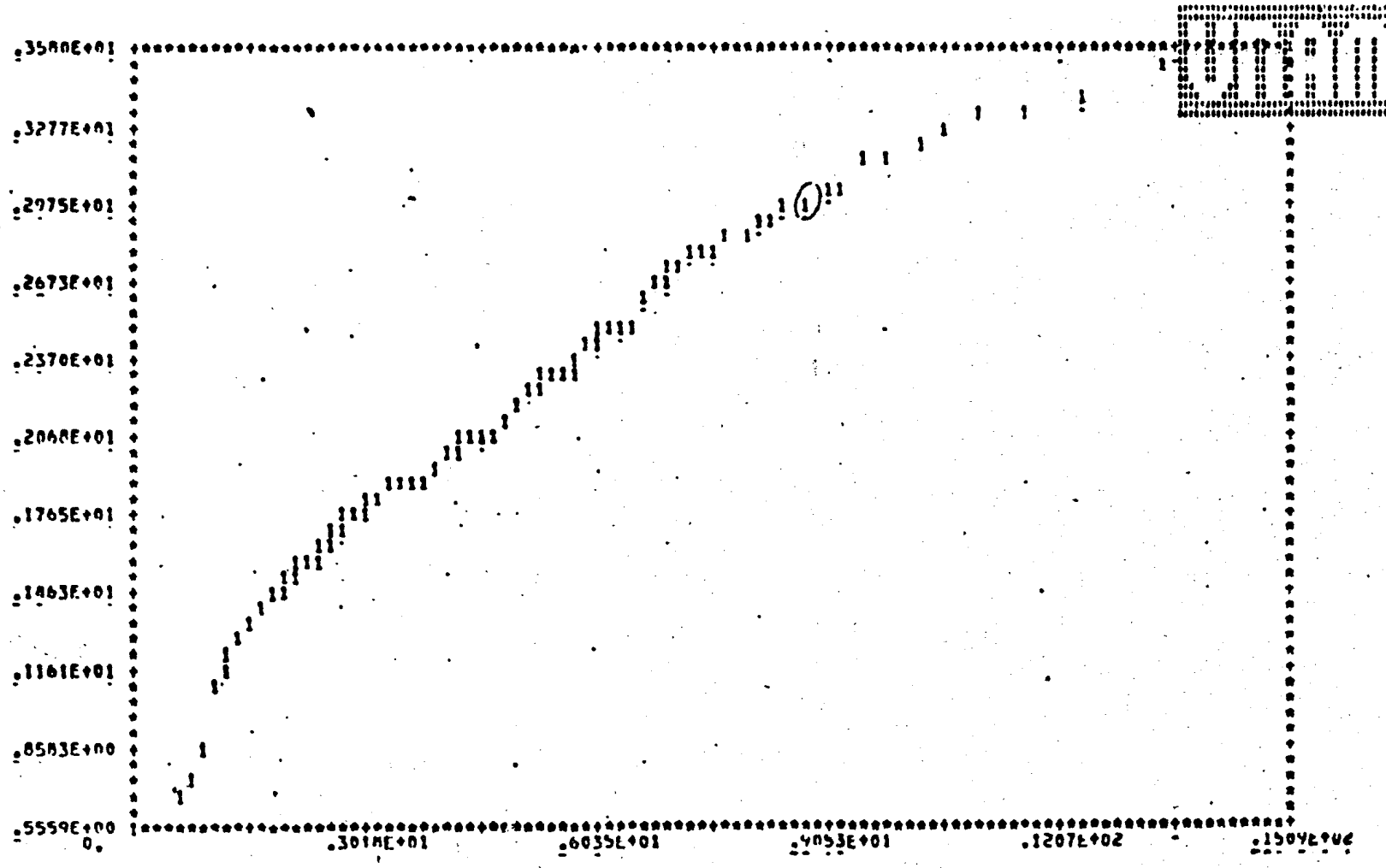
Las mas sospechosas a ser aberrantes, son las observaciones A y B



Raíz cuadrada de las distancias contra χ^2_2

N= 100, 1 aberrante, bajo d_1

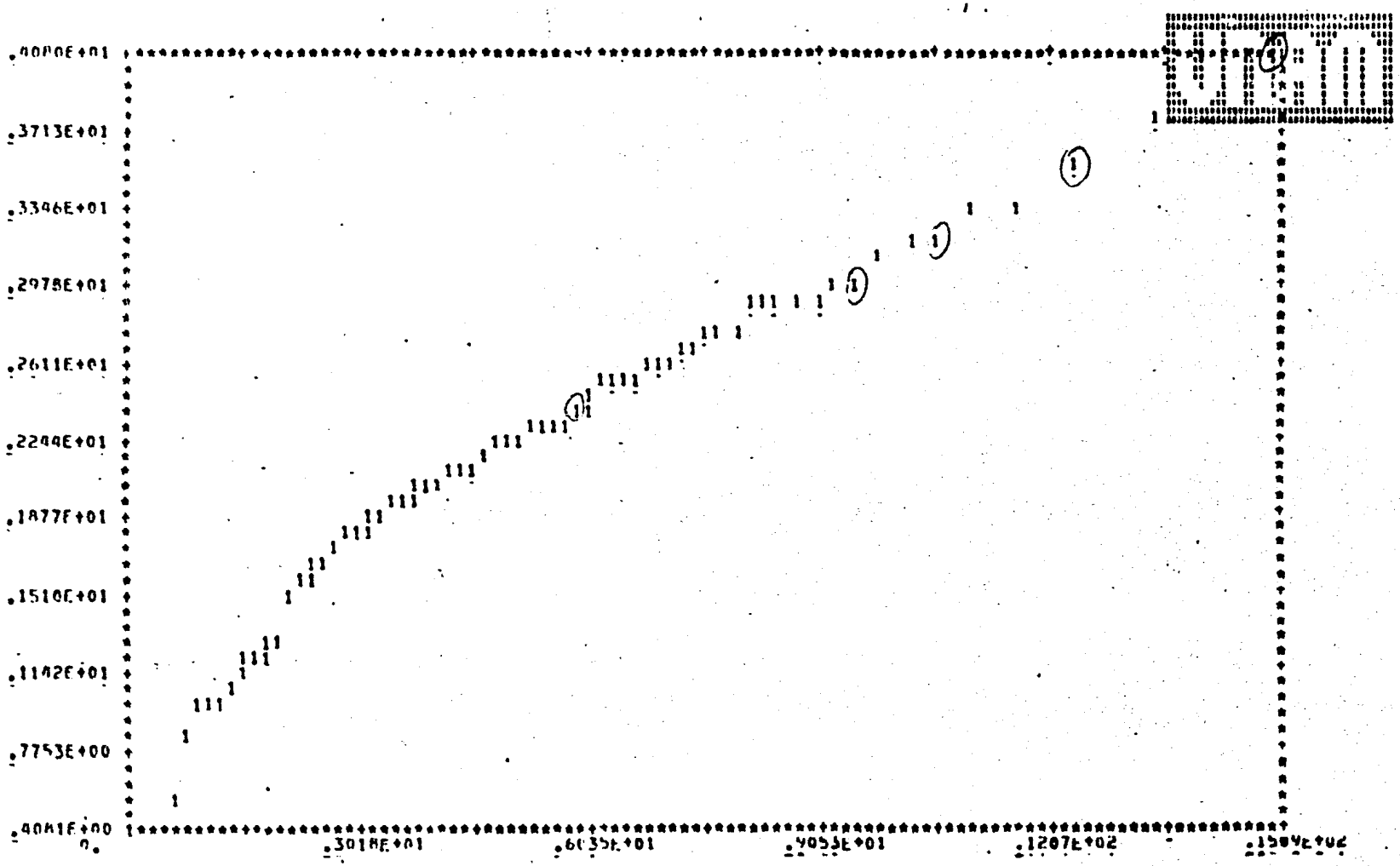
El aberrante se detecta fácilmente



Raíz cuadrada de las distancias contra χ^2

N= 100, 1 aberrante, bajo d_2

El aberrante cae dentro del conjunto de datos

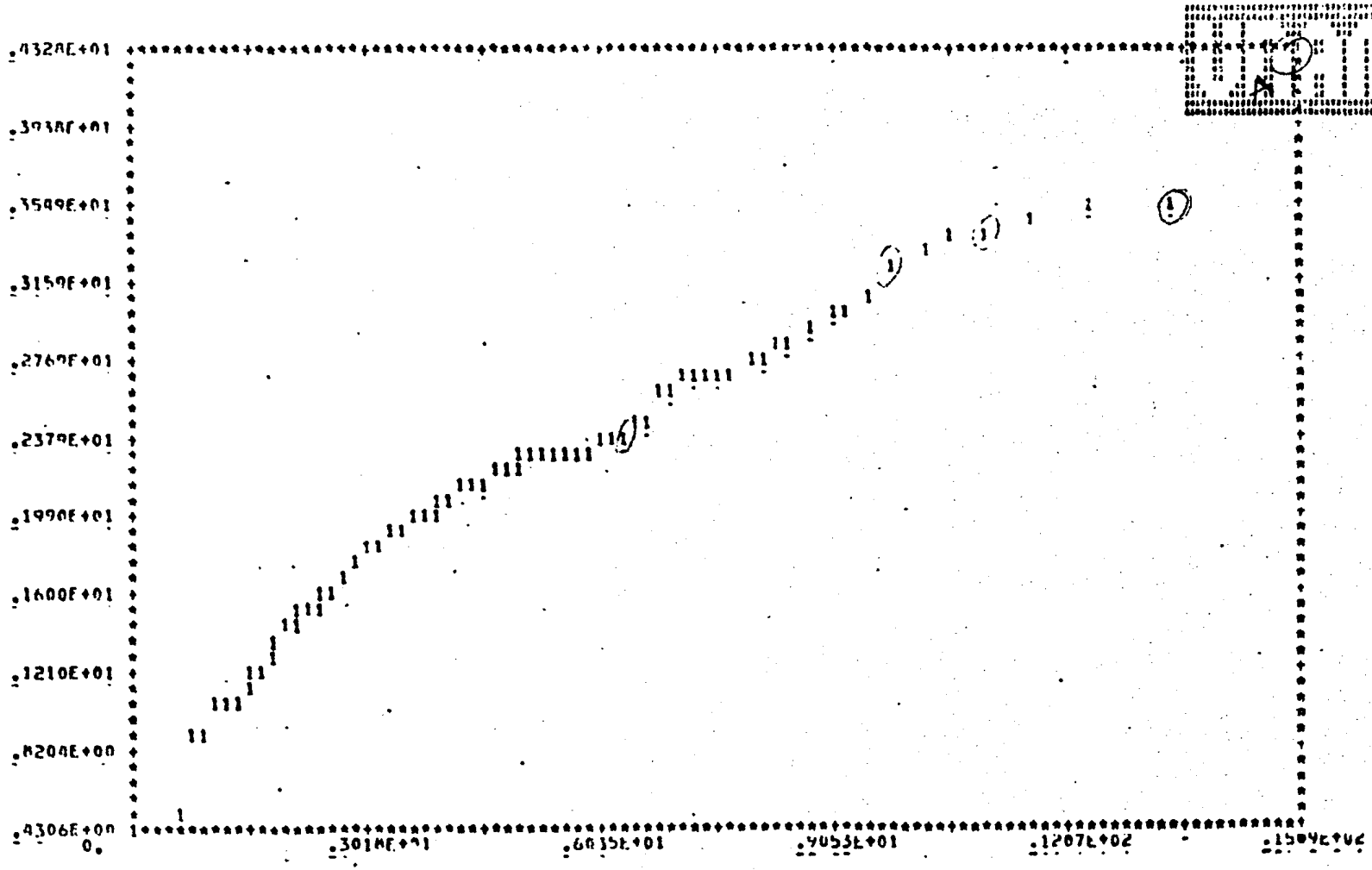


Raíz cuadrada de las distancias contra χ^2_5

N= 100, 5 aberrantes, bajo d_1

Dudaría en declarar alguna observación

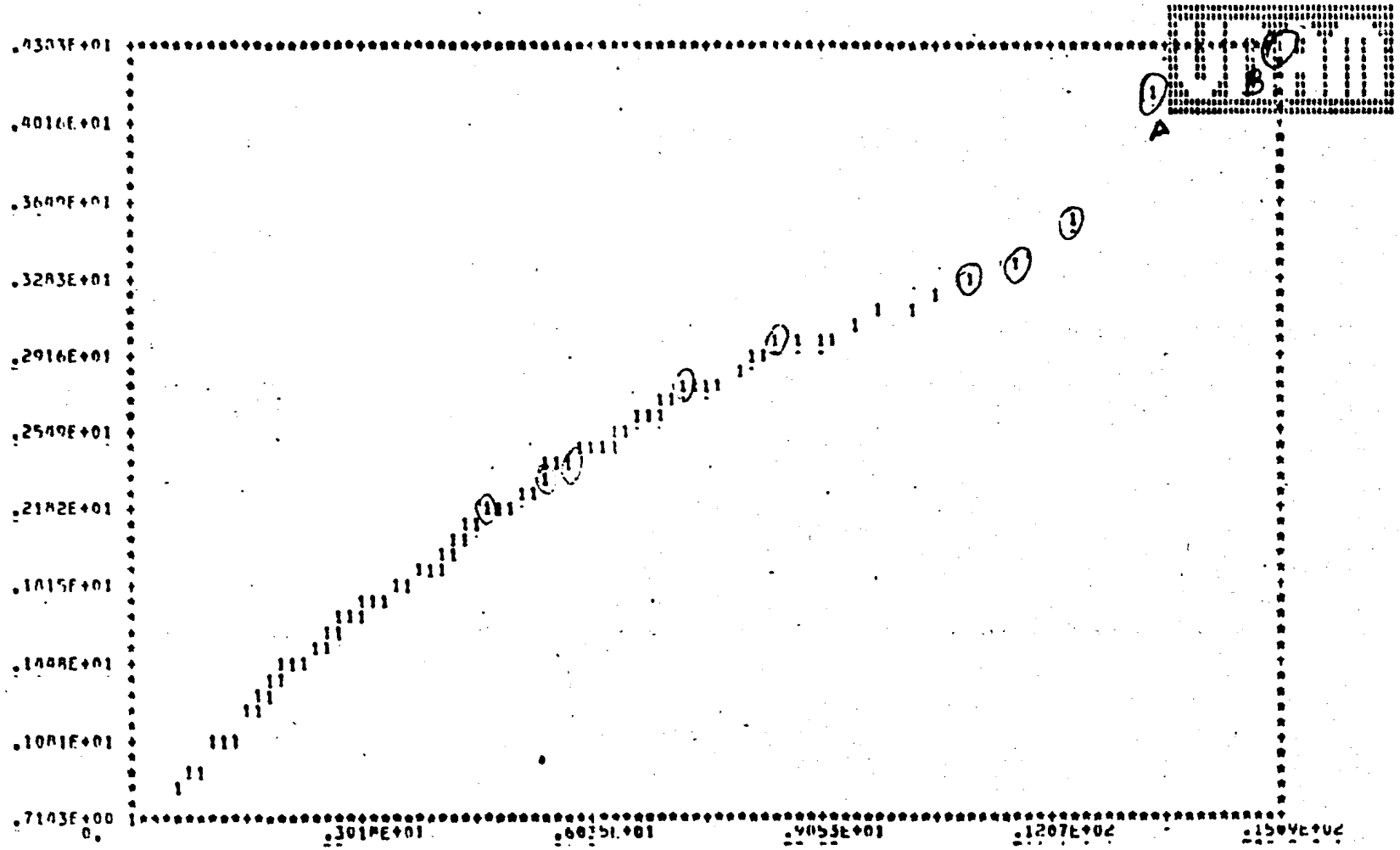
como aberrante



Raíz cudrada de las distancias contra χ^2_5

N= 100, 5 aberrantes, bajo d_2

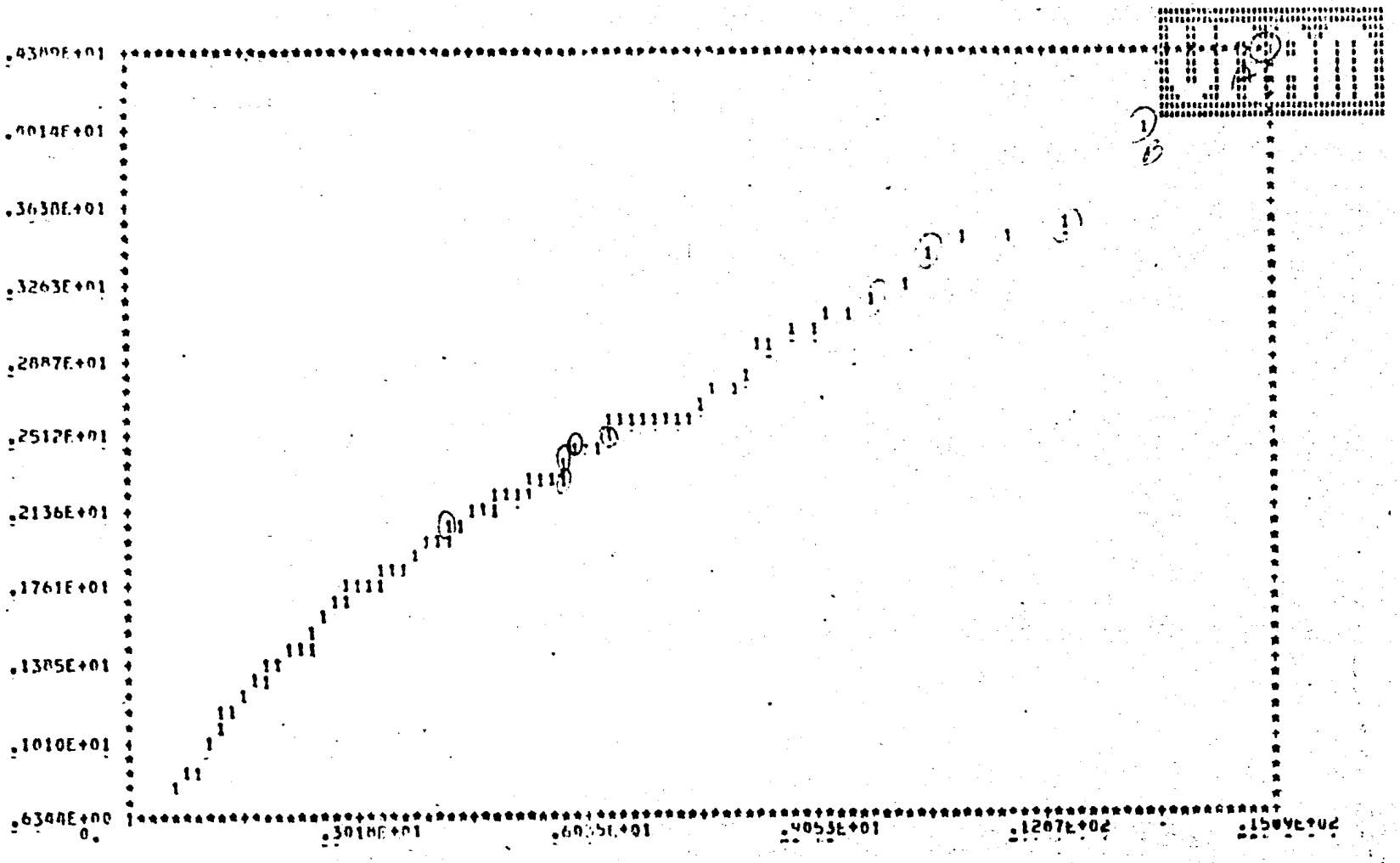
Se detecta fáciomente el aberrante A



Raíz cuadrada de las distancias contra χ^2_5

N= 100, 10 aberrantes, bajo d_1

Los aberrantes A y B, se detectan fácilmente



Raiz cuadrada de las distancias contra χ^2

N= 100, 10 aberrantes, bajo d_2

Los aberrantes mas sospechosos son las observaciones

A y B

APENDICE

N	p = 2		p = 3		p = 4		p = 5	
	5%	1%	5%	1.%	5%	1%	5%	1%
5	3.17	3.19						
6	4.00	4.11	4.14	4.16				
7	4.71	4.95	5.01	5.10	5.12	5.14		
8	5.32	5.70	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.80	6.97	7.01	7.08
10	6.32	6.97	7.01	7.47	7.50	7.79	7.82	7.98
12	7.10	8.00	7.99	8.70	8.67	9.20	9.19	9.57
14	7.74	8.84	8.78	9.71	9.61	10.37	10.29	10.90
16	8.27	9.54	9.44	10.56	10.39	11.36	11.20	12.02
18	8.73	10.15	10.00	11.28	11.06	12.20	11.96	12.98
20	9.13	10.67	10.49	11.91	11.63	12.93	12.62	13.81
25	9.94	11.73	11.48	13.18	12.78	14.40	13.94	15.47
30	10.58	12.54	12.24	14.14	13.67	15.51	14.95	16.73
35	11.10	13.20	12.85	14.92	14.37	16.40	15.75	17.73
40	11.53	13.74	13.36	15.56	14.96	17.13	16.41	18.55
45	11.90	14.20	13.80	16.10	15.46	17.74	16.97	19.24
50	12.23	14.60	14.18	16.56	15.89	18.27	17.45	19.83
100	14.22	16.95	16.45	19.26	18.43	21.30	20.26	23.17
200	15.99	18.94	18.42	21.47	20.59	23.72	22.59	25.82
500	18.12	21.22	20.75	23.95	23.06	26.37	25.21	28.62

TABLA DE VALORES CRITICOS AL 1% y 5% PARA PRUEBAS DE UN ABERRANTE EN UNA MUESTRA NORMAL MULTIVARIADA CUANDO $\underline{\mu}, V$ SON DESCONOCIDAS Y LA PRUEBA ESTADISTICA ES

$$R_N(\underline{\bar{X}}, S) = \max (\underline{X}_j - \underline{\bar{X}})' S^{-1} (\underline{X}_j - \underline{\bar{X}})$$

$$j = 1, 2, \ldots N$$

N	p = 2		p = 3		p = 4		p = 5	
	5%	1%	5%	1%	5%	1%	5%	1%
5	0.0025	0.0005	0.0000					
6	0.0337	0.0150	0.0011	0.0002				
7	0.0860	0.0498	0.0202	0.0090	0.0006	0.0001		
8	0.1417	0.0937	0.0580	0.0335	0.0136	0.0060	0.0004	0.0001
9	0.1942	0.1393	0.1024	0.0674	0.0425	0.0245	0.0098	0.0043
10	0.2419	0.1831	0.1470	0.1049	0.0788	0.0518	0.0327	0.0189
12	0.3229	0.2615	0.2288	0.1791	0.1549	0.1163	0.0966	0.0686
14	0.3879	0.3275	0.2982	0.2460	0.2246	0.1804	0.1631	0.1270
16	0.4410	0.3828	0.3563	0.3040	0.2853	0.2389	0.2242	0.1838
18	0.4850	0.4295	0.4054	0.3542	0.3376	0.2908	0.2782	0.2360
20	0.5221	0.4694	0.4472	0.3976	0.3828	0.3366	0.3257	0.2830
25	0.5935	0.5472	0.5288	0.4839	0.4722	0.4290	0.4211	0.3798
30	0.6451	0.6041	0.5882	0.5478	0.5380	0.4984	0.4923	0.4537
35	0.6842	0.6475	0.6335	0.5969	0.5885	0.5523	0.5473	0.5116
40	0.7150	0.6818	0.6693	0.6360	0.6285	0.5993	0.5911	0.5580
45	0.7399	0.7097	0.6982	0.6677	0.6610	0.6304	0.6267	0.5961
50	0.7605	0.7328	0.7222	0.6941	0.6880	0.6596	0.6564	0.6279
100	0.8629	0.8477	0.8417	0.8269	0.8225	0.8065	0.8047	0.7889
200	0.9232	0.9152	0.9118	0.9035	0.9015	0.8929	0.8918	0.8830
500	0.9650	0.9618	0.9602	0.9568	0.9558	0.9523	0.9517	0.9486

TABLA DE VALORES CRITICOS DE $\sqrt{\ }$ AL 1% Y 5% PARA PRUEBAS DE 2 ABERRANTES EN UNA MUESTRA NORMAL MULTIVARIADA CUANDO μ Y V SON DESCONOCIDAS, Y LA PRUEBA ESTADISTICA ES

$$r_2 = \min$$

B I B L I O G R A F I A

Anderson, D. F., Gnanadesikan, R. and Warner, J.L. " Methods for Assessing Multivariate normality". In multivariate analysis III. (P. R. Krishnaiah, ed.) Academic Press, New York, p.p. 95- 116. 1973.

Barnett, V. and Lewis, T. "Outliers in Statistical Data". John Wiley and Sons. 1978.

Blackith, and Reyment. "Multivariate Morphometrics". Academic Press. London and New York. 1971.

Gnanadesikan, R. " Methods for Statistical Data . Analysis of Multiple Observations". John Wiley and Sons. 1977.

Healy, M.J.R. " Multivariate Normal Plotting". Appl. Stat. 17, p.p. 157-161. 1968.

Marriot, F.H.C. "The Interpretation of Multiple Observations". Academic Press.1974.

Morrison, D.F. "Multivariate Statistical Methods". Mc Graw Hill. 1976.

Overall, J.E., and Klett, C. J. "Applied Multivariate Analysis" Mc Graw Hill. 1972.

Seal, H.L. " Multivariate Statistical Analysis for Biologists". Methuen and Co. LTD. 1964.

Bryant,E. H. and Atchley, W.R. "Multivariate Statistical Methods Within Groups Covariation ". Ed. Halsted Press. 1975.