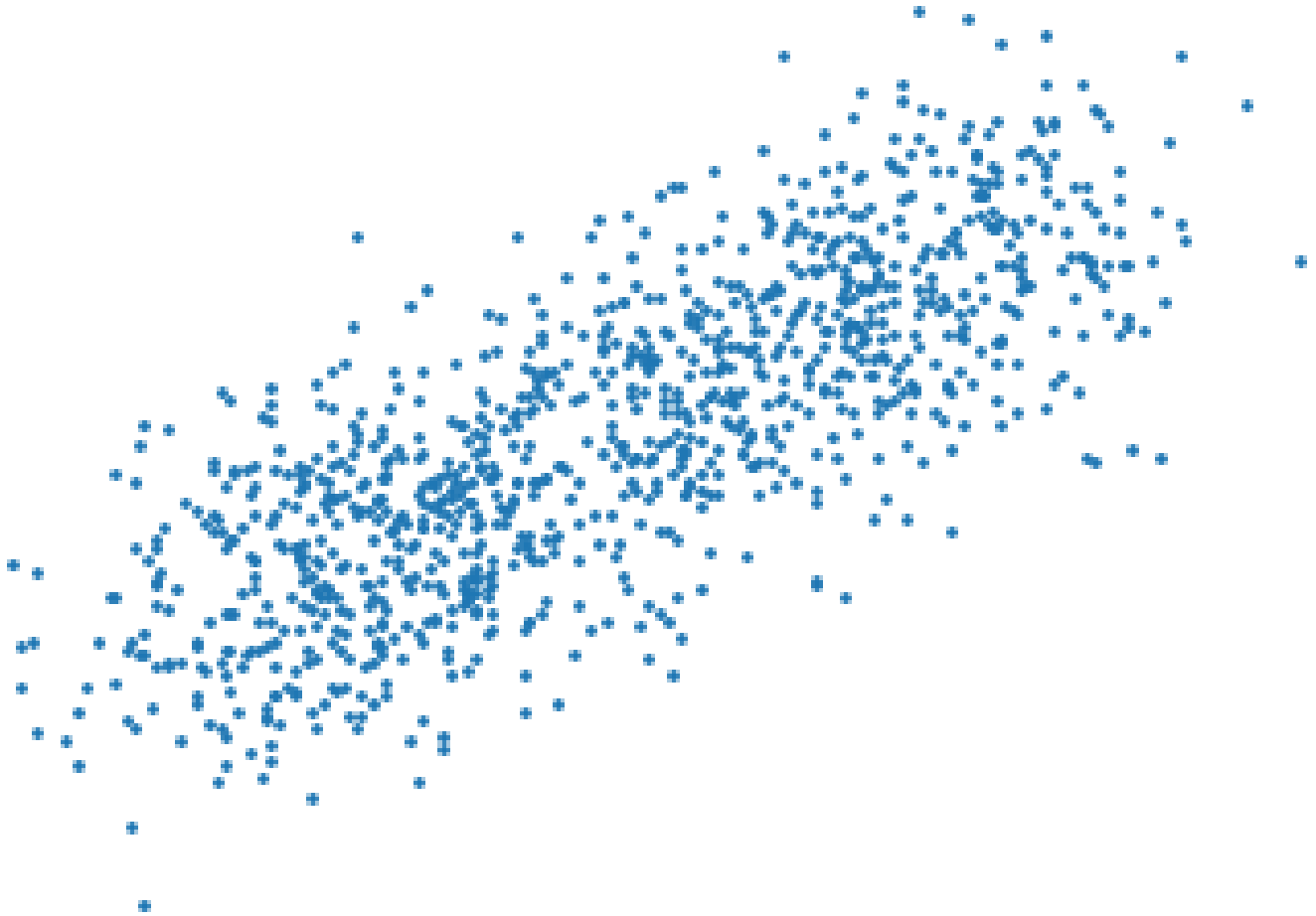
PRINCIPAL COMPONENT ANALYSIS

SERGIO PERNICE

UCEMA

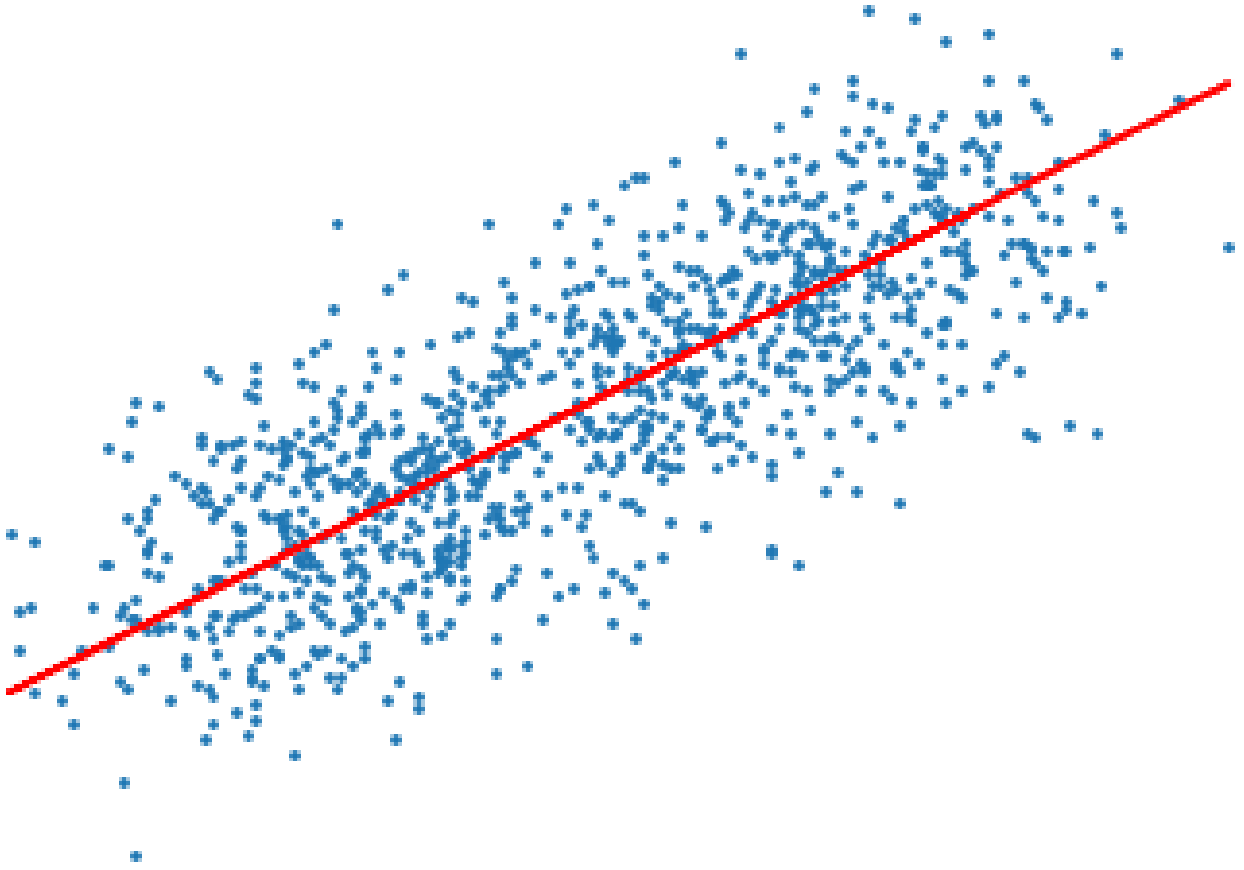
EL PROBLEMA DE REDUCCIÓN DIMENSIONAL: INTUICIÓN

- Cual la “mejor recta” que represente a estos puntos?
- Trate de dibujarla a mano antes de continuar.



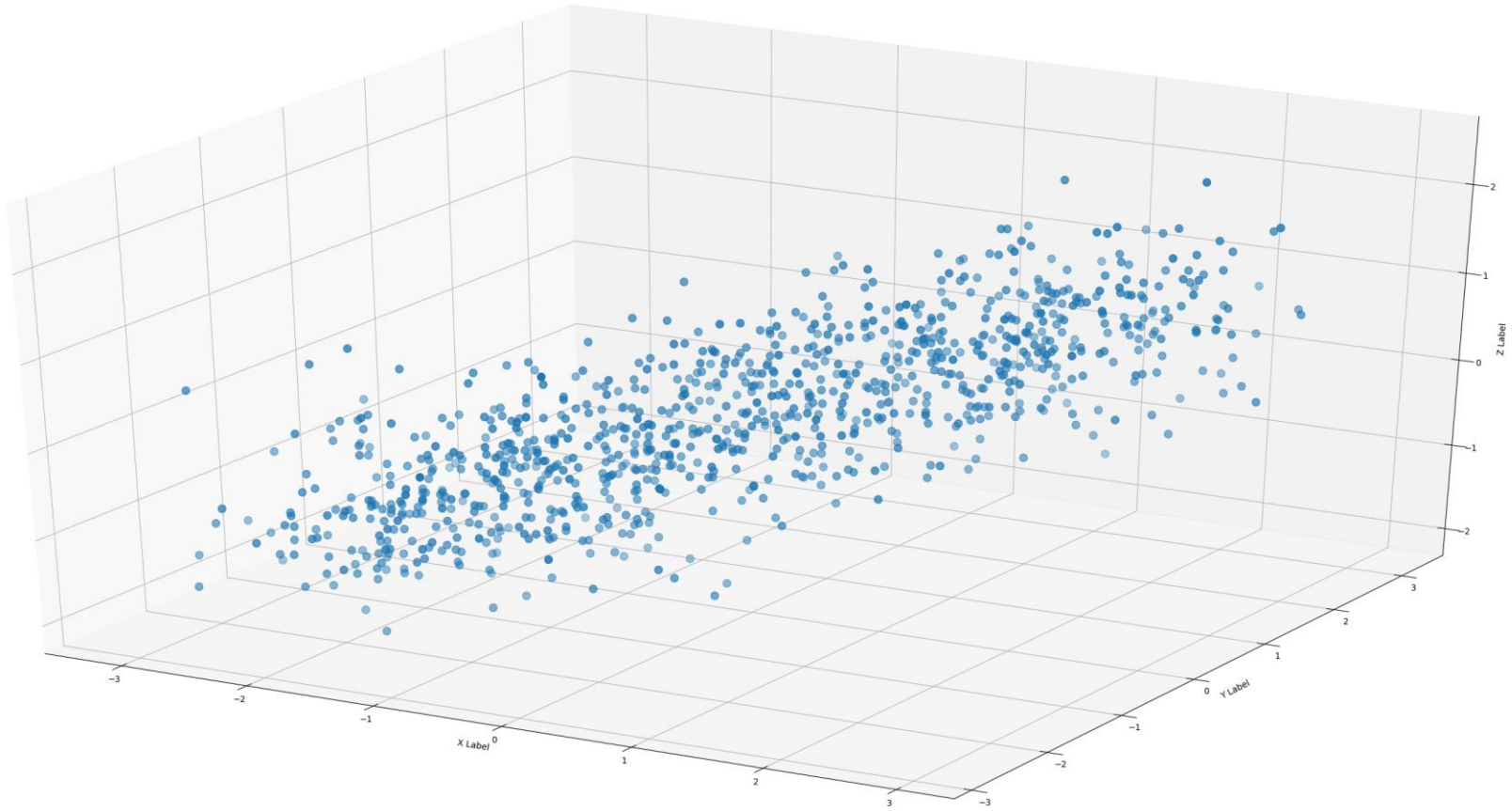
EL PROBLEMA DE REDUCCIÓN DIMENSIONAL: INTUICIÓN

- Por supuesto que frente a una pregunta como esa la respuesta natural es “mejor con respecto a qué criterio? Ya vamos a ir a eso, pero apuesto a que su respuesta no fue muy diferente a esta.
- Esta recta es la “mejor” representación 1-D de los puntos en 2-D.



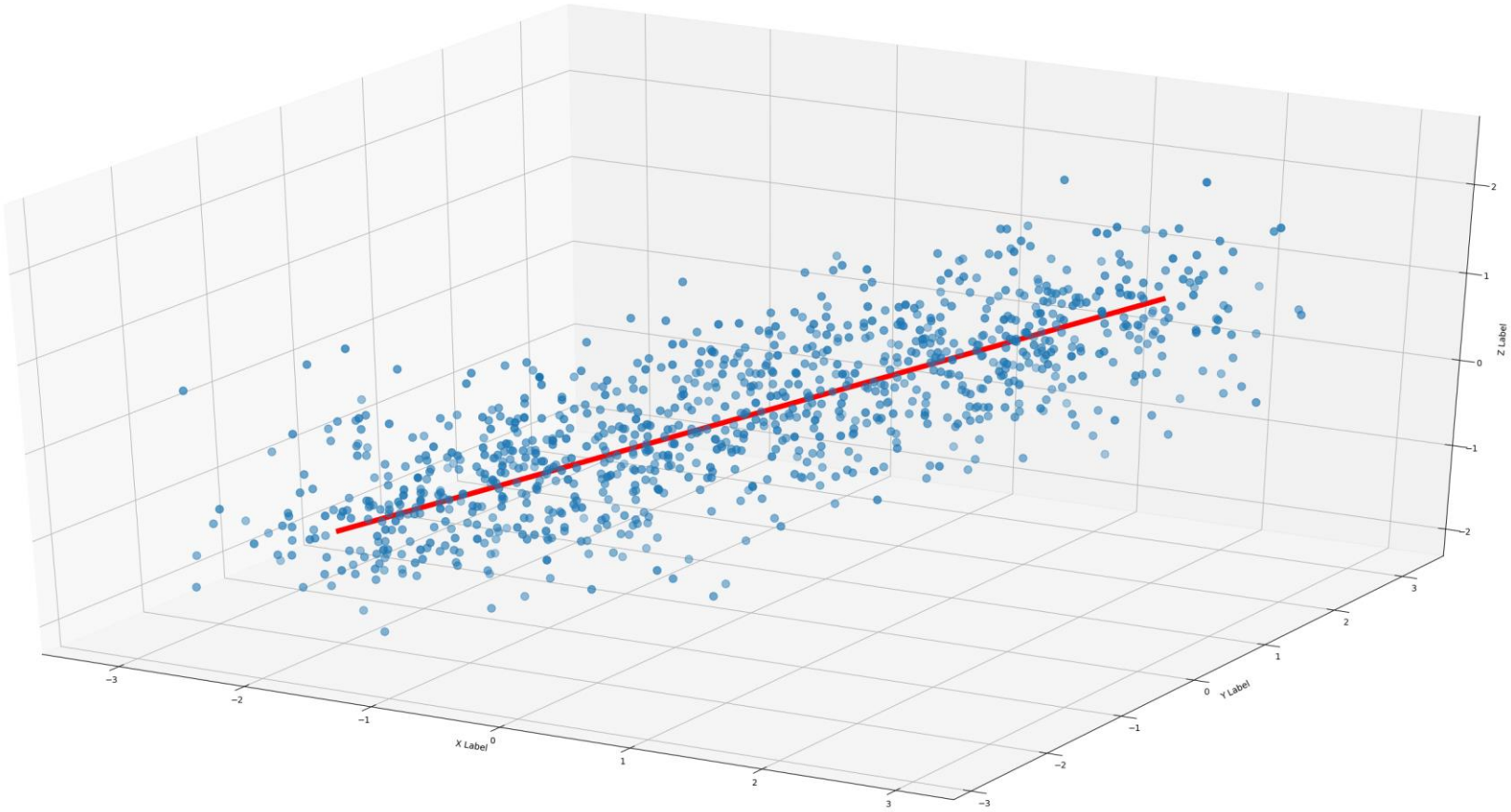
EL PROBLEMA DE REDUCCIÓN DIMENSIONAL : INTUICIÓN

- Cual la “mejor recta” que represente a estos puntos en 3-D?



EL PROBLEMA DE REDUCCIÓN DIMENSIONAL : INTUICIÓN

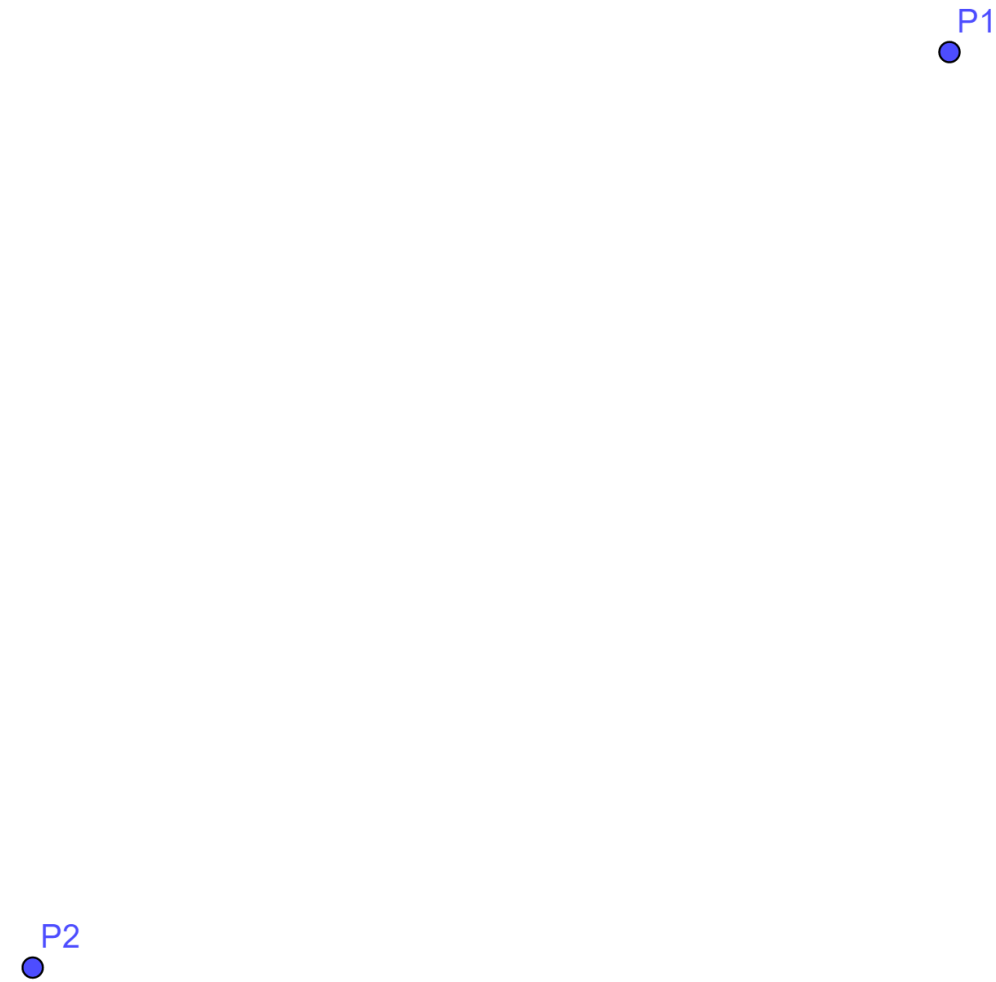
- Esta recta es la “mejor” representación 1-D de los puntos en 3-D.
- También podríamos pensar en el mejor plano (2-D) de los puntos en 3-D.



EL PROBLEMA DE REDUCCIÓN DIMENSIONAL: INTUICIÓN

- En general, nos podemos preguntar cuál es el mejor “k-plano”, o subespacio afine (no necesariamente pasa por el origen) de k dimensiones, de un conjunto de puntos en d dimensiones ($k < d$).
- Ahora tenemos que definir más claramente qué queremos decir por “mejor”.

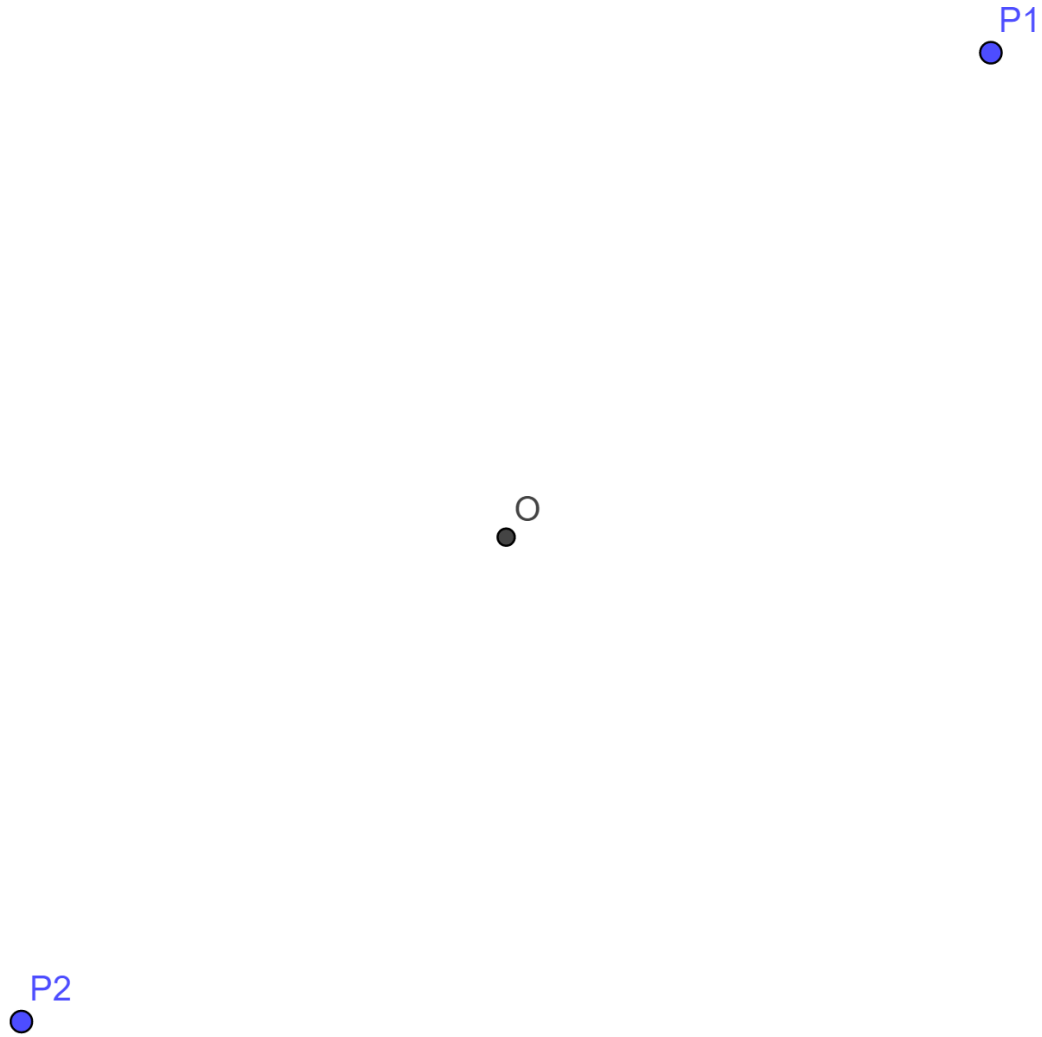
EL PROBLEMA DE LA MEJOR RECTA ENTRE DOS PUNTOS



- Dos puntos en el plano.

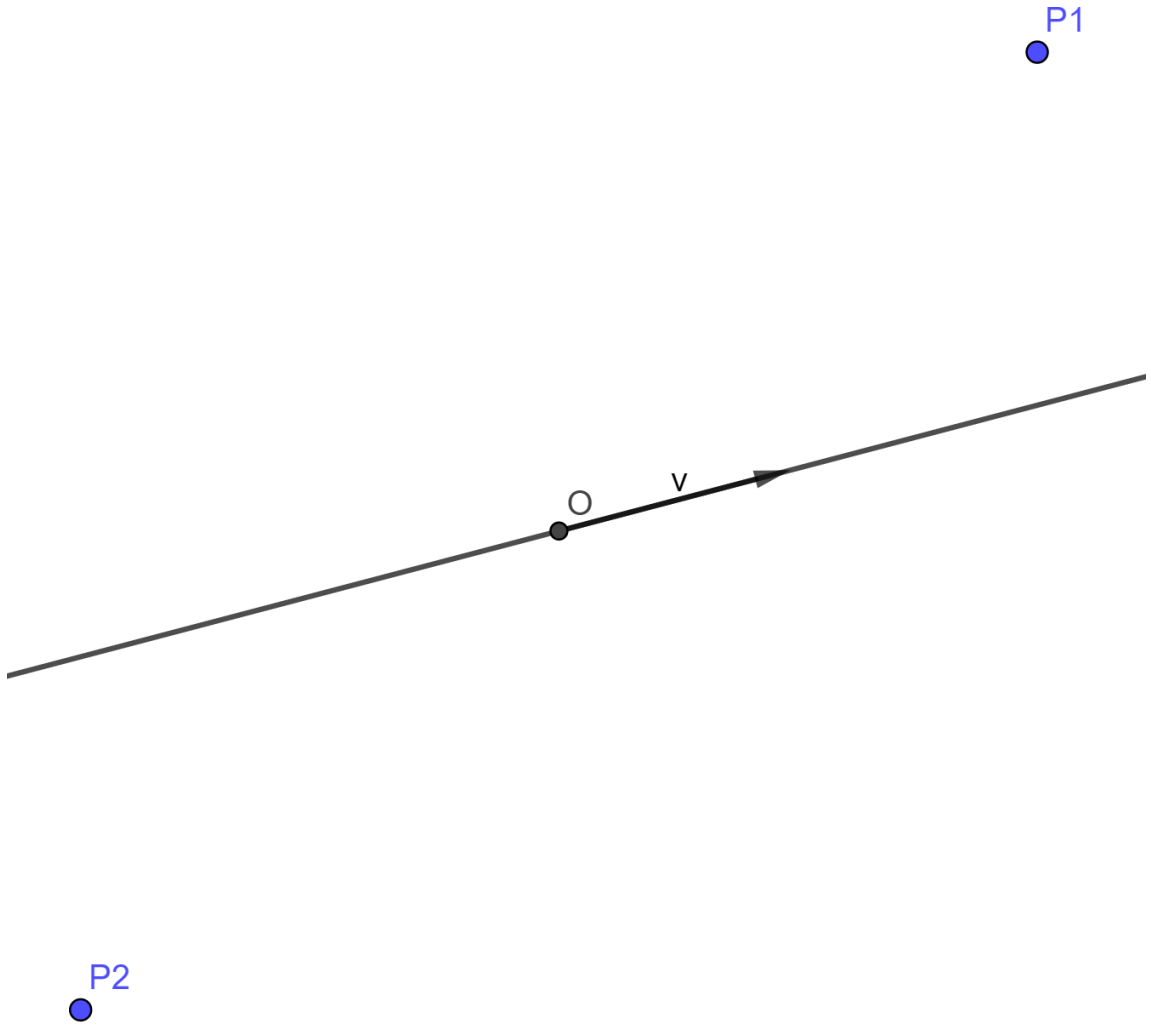
EL PROBLEMA DE LA MEJOR RECTA ENTRE DOS PUNTOS

- Nos paramos en el medio.
- Nos trasladamos desde el origen de coordenadas que tengamos, al centro de los puntos.



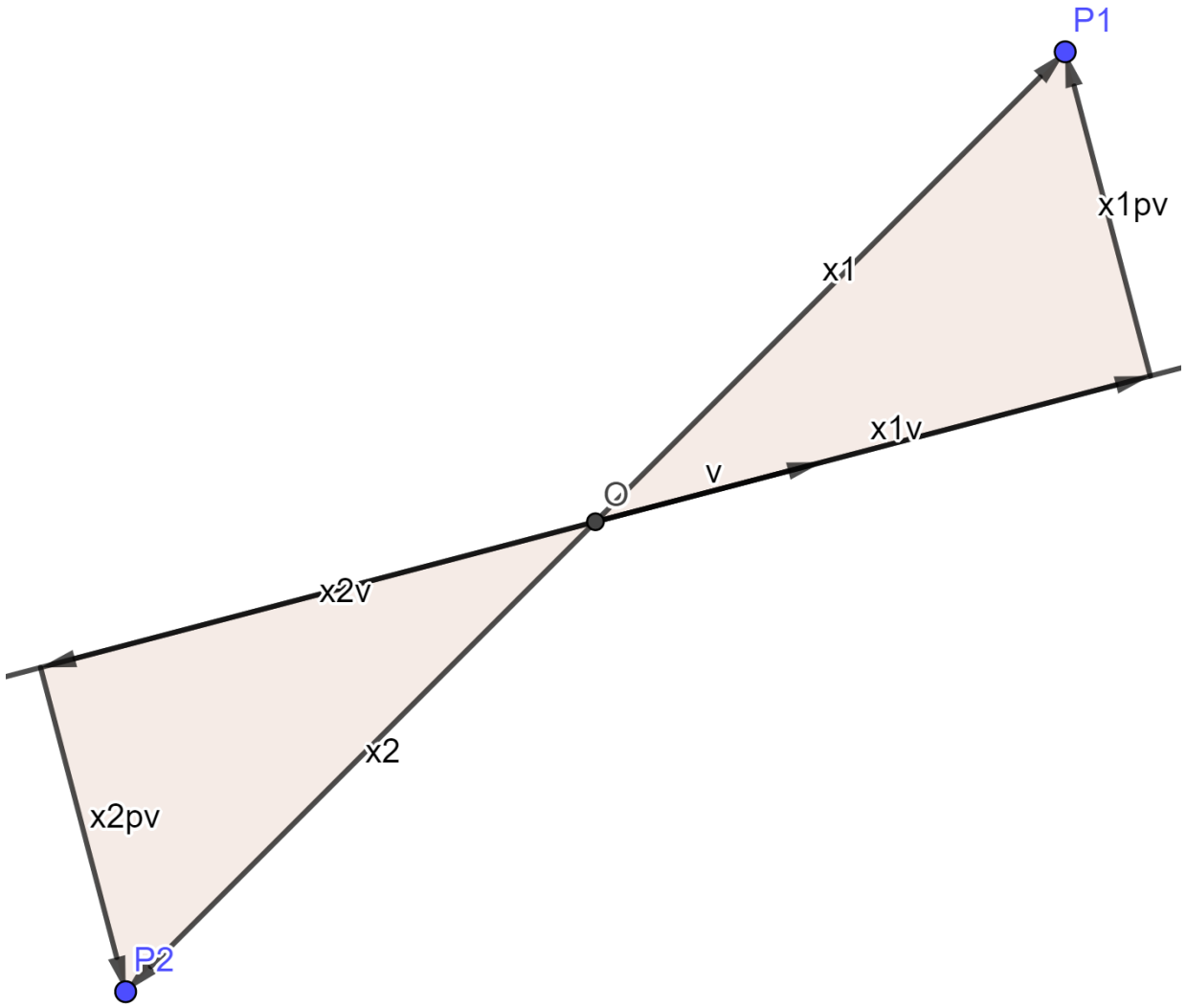
EL PROBLEMA DE LA MEJOR RECTA ENTRE DOS PUNTOS

- Todo vector **v** de norma 1 define una recta.
- La mejor recta es la que minimiza el cuadrado de la distancia de la recta a los puntos.



EL PROBLEMA DE LA MEJOR RECTA ENTRE DOS PUNTOS

- La distancia al cuadrado del origen a los puntos no depende de **v**, pero el cuadrado de la distancia de los puntos a la recta sí depende de **v.**
- **La mejor recta es la que minimiza la suma del cuadrado de las distancias de la recta a los puntos.**
- Por Pitágoras, eso equivale a maximizar la suma del cuadrado de las proyecciones ortogonales de los puntos sobre la recta.



EL PROBLEMA DE LA MEJOR RECTA ENTRE DOS PUNTOS

- **La mejor recta es la que minimiza la suma del cuadrado de las distancias de la recta a los puntos.**
- Por Pitágoras, eso equivale a maximizar la suma del cuadrado de las proyecciones ortogonales de los puntos sobre la recta.
- Esa recta es la mejor representación 1-D de puntos en 2-D.
- En el caso de dos puntos, conocer la recta y conocer las coordenadas de los dos puntos en la recta, es equivalente a conocer las posiciones de los dos puntos.
- Si son más de dos puntos, conocer la recta y conocer la coordenada de los dos puntos en la recta, no es equivalente a conocer la posición de los puntos, pero es la mejor aproximación 1-D.

UN EJEMPLO DE JUGUETE

- Por qué PCA?
- para visualizar los datos en un espacio de dimensiones inferiores
- para comprender las fuentes de variabilidad de los datos
- para comprender las correlaciones entre las diferentes coordenadas de los datos.

	Tofu	Tacos	Pollo	Papas fritas
Alicia (A)	10	1	2	7
Benjamín (B)	7	2	1	10
Carolina (C)	2	9	7	3
Diego (D)	3	6	10	2

Rating de comidas por diferentes personas

UN EJEMPLO DE JUGUETE

- Número de datos (personas): n = 4
- Dimensionalidad de los datos (tipo de comida): d = 4
- A diferencia de lo que típicamente ocurre con regresiones lineales, no diferenciamos entre una variable a explicar y las otras como explicativas. A priori todas las dimensiones de los datos son idénticas. Sólo a posteriori algunas van a ser más importantes que otras.
- ¿Podemos visualizar este conjunto de datos en menos de 4 dimensiones y de modo que resulte útil para “entenderlos”?
- ¿Podemos entender los razones subyacentes en las diferencias de opinión sobre los diversos alimentos?
- La observación clave es que cada fila es un punto de R^4 que se puede expresar aproximadamente como la suma vectorial

$$\boldsymbol{x}_i \approx \overline{\boldsymbol{x}} + a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2$$

- ver Jupyter Notebook “RankingDeComidas.ipynb”.

UN EJEMPLO DE JUGUETE

$$\bar{\mathbf{x}} + a_1\mathbf{v}_1 + a_2\mathbf{v}_2$$
$$\bar{\mathbf{x}} = (5.5, 4.5, 5, 5.5)^\top$$
$$\mathbf{v}_1 = (0.48, -0.48, -0.56, 0.48)^\top$$
$$\mathbf{v}_2 = (-0.52, 0.52, -0.48, 0.48)^\top,$$

$$\begin{aligned} \mathbf{x} + 6.96\mathbf{v}_1 - 2.27\mathbf{v}_2 &= (9.5, 0.5, 3, 7.5) \\ \text{Alicia} &= (10, 1, 2, 7) \\ \mathbf{x} - 5.41\mathbf{v}_1 + 1.71\mathbf{v}_2 &= (1.5, 8.5, 7, 3.5) \\ \text{Carolina} &= (2, 9, 7, 3) \end{aligned}$$

- Notar que todo vector $\boldsymbol{v} = \bar{\boldsymbol{x}} + a_1\boldsymbol{v}_1 + a_2\boldsymbol{v_2}$ “vive” es un plano 2-D (2-plano) que esta sumergido en el espacio 4-D de los datos.
- Para que sirve esta aproximación?
- Por un lado, proporciona una manera de visualizar los datos como puntos en baja dimensión. Con más datos, podemos imaginar inspeccionar la figura resultante en busca de grupos de puntos de datos similares.
- Por el otro, ayuda a interpretar los datos. Cada punto de datos tiene una coordenada a_1 en la dirección $\mathbf{v}_1$ y una coordenada a_2 en la dirección $\mathbf{v}_2$.

UN EJEMPLO DE JUGUETE

- ¿Qué significa el vector $\mathbf{v}_1$? (recordemos que las dimensiones son los ranking de tofu, tacos, pollo, y papas fritas):

$$\boldsymbol{v}_1 = \begin{bmatrix} 0.48 \\ -0.48 \\ -0.56 \\ 0.48 \end{bmatrix}$$

- Las primera y cuarta coordenadas de $\mathbf{v}_1$ tienen un valor positivo (y por lo tanto están positivamente correlacionadas) mientras que las coordenadas segunda y tercera tienen un valor negativo (y por lo tanto están positivamente correlacionadas entre sí y negativamente con las coordenadas primera y cuarta).
- Teniendo en cuenta que el tofu y las papas fritas forman parte importante de la dieta vegetariana, mientras que los tacos y el pollo no, podemos interpretar que la coordenada correspondiente a $\mathbf{v}_1$ indica el grado en que alguien tiene preferencias vegetarianas.
- Notar además, que los 4 puntos tienen las coordenadas más grandes en la dirección de $\mathbf{v}_1$:

$$\text{Alicia}_1 = 7.0, \qquad \text{Benjamin}_1 = 7.0, \qquad \text{Carolina}_1 = -5.4, \qquad \text{Diego}_1 = -5.6$$

- A la inversa, si supiéramos quiénes son vegetarianos y quiénes no, a partir de sus calificaciones podríamos deducir qué alimentos tienen más probabilidades de ser vegetarianos sin saberlo previamente.

UN EJEMPLO DE JUGUETE

- Mediante un razonamiento similar, podemos interpretar que el segundo vector $\mathbf{v}_2$ indica el grado en que alguien cuida su salud con su dieta.
- El hecho de que $\mathbf{v}_1$ tiene asociados mayores proyecciones que $\mathbf{v}_2$ indica que es el más fuerte de los dos efectos.
- El objetivo de PCA es calcular aproximaciones como las de este ejemplo automáticamente, incluso para grandes conjuntos de datos.

OBJETIVO DEL PCA

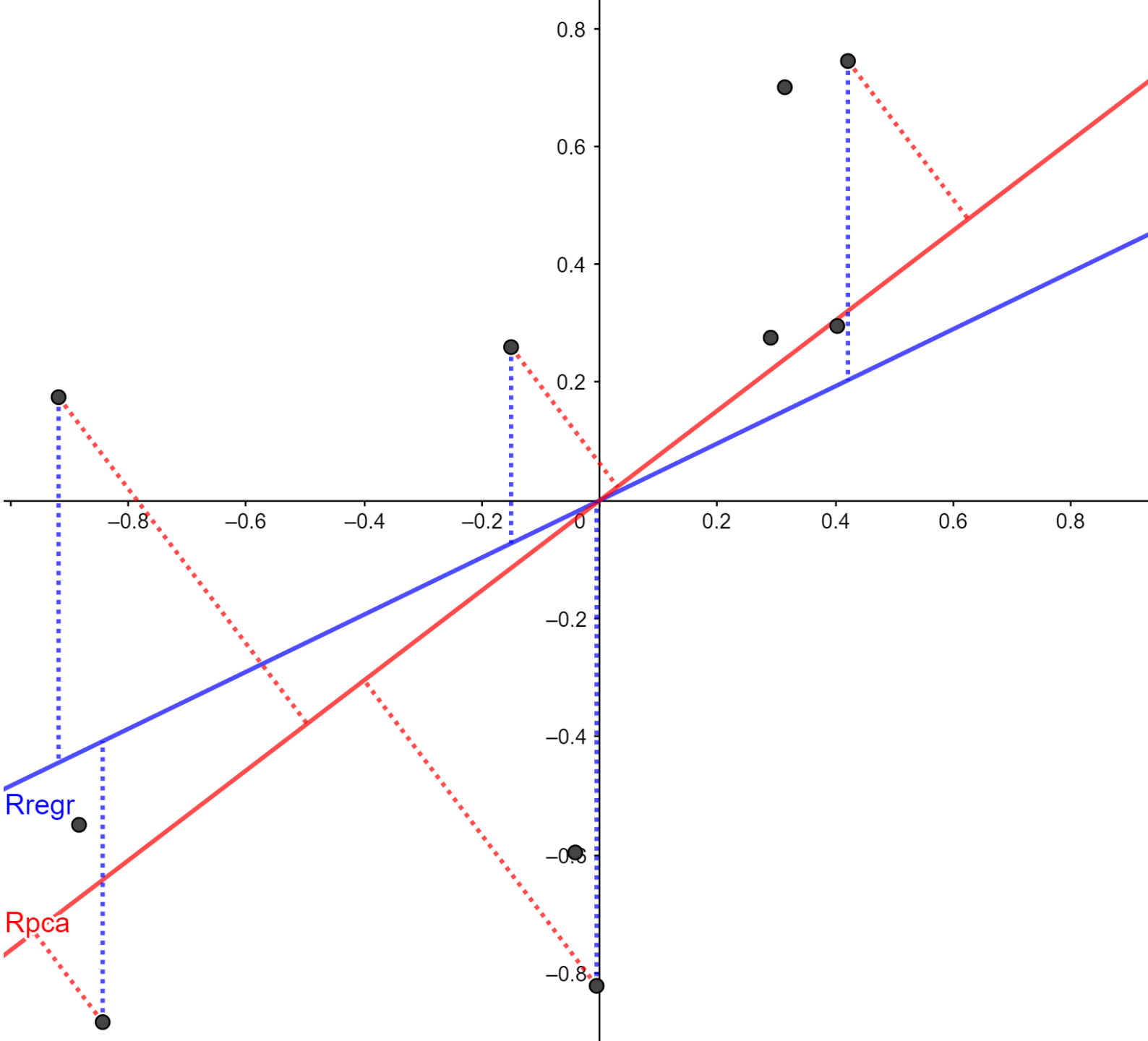
- El objetivo de PCA es encontrar aproximaciones de cada uno de los m vectores $\mathbf{x}_1,\ldots,\mathbf{x}_m \in \mathbb{R}^n$ como combinaciones lineales de d vectores $v_1,\ldots,v_d \in \mathbb{R}^n$, de modo que:

$$\boldsymbol{x}_i \approx \sum_{j=1}^{d} a_{ij}\,\boldsymbol{v}_j, \qquad i = 1,\ldots,m$$

- En el ejemplo m = 4, n = 4, d = 2.
  - m = número de datos
  - n = dimensionalidad de los datos
  - d = dimensionalidad del subespacio afine para aproximar a los datos
- Ejemplos de datos: imágenes (dimensiones = píxeles); mediciones (dimensiones = sensores); documentos (dimensiones = palabras)…
- PCA ofrece una definición formal de qué d vectores son los mejores para este propósito, y algoritmos para calcular estos vectores.

PCA VS. REGRESIONES EN EL CASO 1-D

- Regresiones en azul.
- PCA en rojo.



PRE-PROCESAMIENTO DE DATOS: TRASLACIÓN

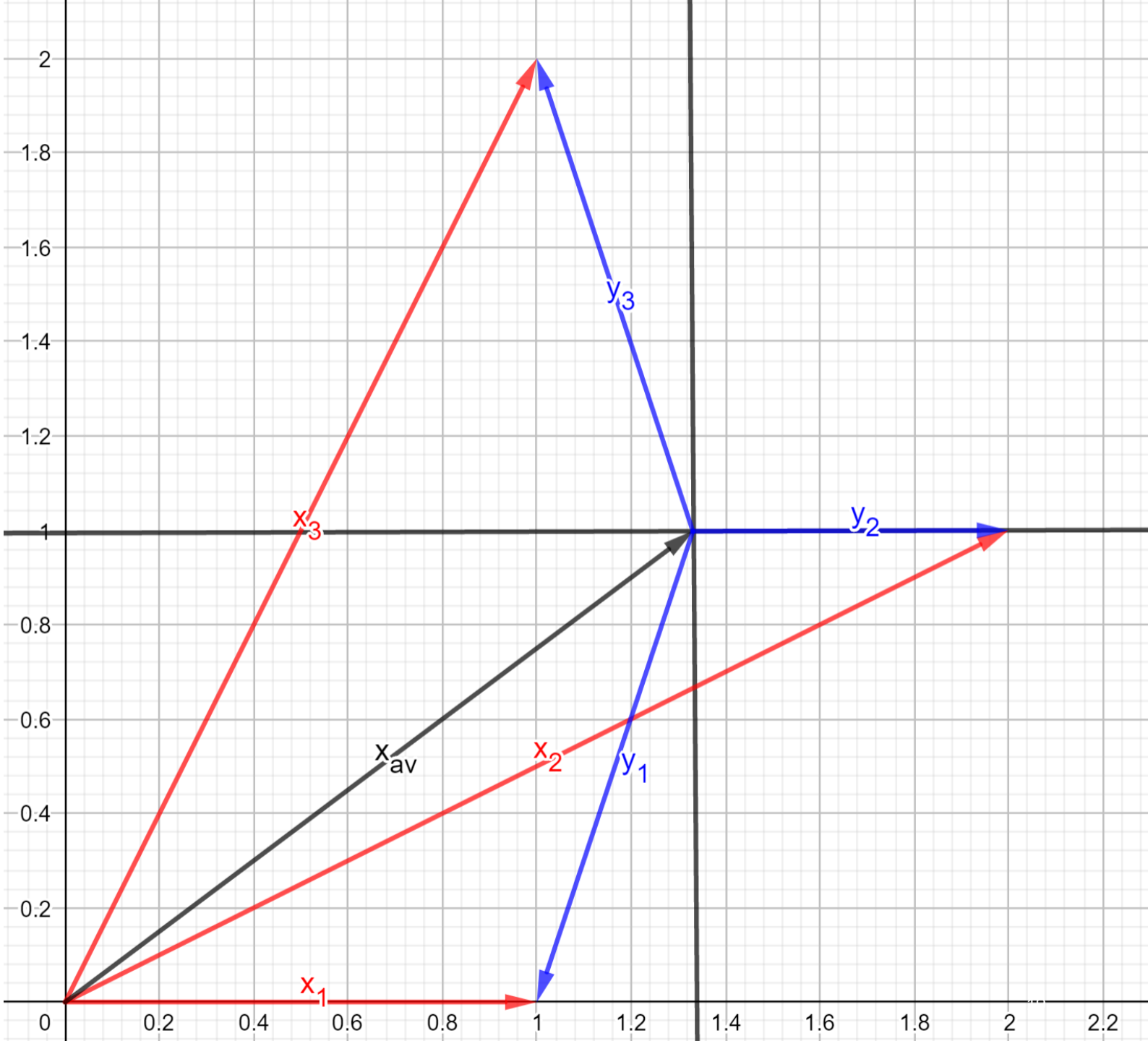
- En el espacio de los datos, nos trasladamos a un sistema de coordenadas con origen en el "centro de los datos":

$$\overline{\boldsymbol{x}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$$

- En el nuevo sistema, al dato $\boldsymbol{x}_i$ le corresponde el vector $\boldsymbol{y}_i$, el $\boldsymbol{y}_i$ "promedio" es $\mathbf{0}$.

$$\boldsymbol{x}_i = \overline{\boldsymbol{x}} + \boldsymbol{y}_i, \rightarrow \overline{\boldsymbol{y}} = \mathbf{0}$$

- Al terminar el análisis volvemos al sistema original.

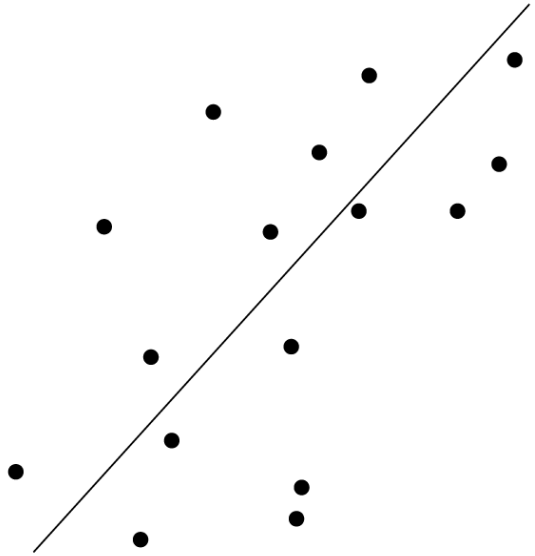


PRE-PROCESAMIENTO DE DATOS: CAMBIO DE ESCALA

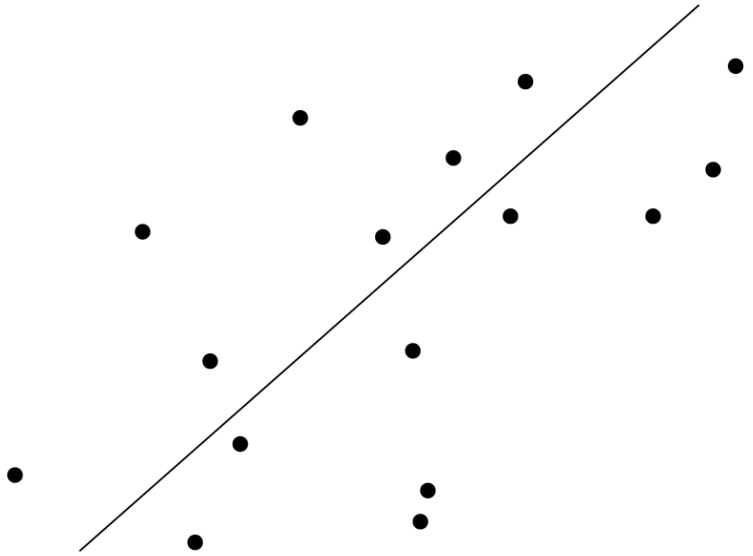
- Suele ocurrir que debemos cambiamos la escala de las diferentes dimensiones de los datos para que todas las dimensiones tengan la misma escala

$$\tilde{y}_{ij} = \frac{y_{ij}}{\sqrt{\sum_{i=1}^{m} y_{ij}^2}}, \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n$$

- Recordemos que m = número de datos y n = dimensionalidad de los datos.
- Es decir que la coordenada j del dato i-ésimo se divide por la raíz cuadrada de la suma de los cuadrados de la coordenada j de todos los datos.



(a)



(b)

Los puntos en (b) son los mismos puntos que en (a) pero cambiando la escala de la coordenada x, y manteniendo la de la coordenada y. Debe resultar obvio que este procedimiento cambia la “mejor” recta.

PRE-PROCESAMIENTO DE DATOS: CAMBIO DE ESCALA

- Por qué se re-escalan? Como los datos vienen dados como vienen dados, no es obvio a priori que las magnitudes de las diferentes coordenadas sean comparables. Este procedimiento las vuelve a todas comparables.

- Notemos que $\sqrt{\sum_{i=1}^{m} y_{ij}^2}$ es la norma ℓ_2 de un vector cuyas coordenadas son las m coordenadas j de nuestros datos. Entonces dividiendo por esta norma, básicamente cambiamos la escala de todas las coordenadas de modo que lo que buscamos no dependa de las longitudes relativas de las diferentes coordenadas.
  - Ejemplo: supongamos que tenemos puntos en un plano y queremos analizar su distribución. Una persona mide las coordenadas x y otra las y. Pero no determinaron a priori las unidades en las que las iban a medir. Una las mide en metros y la otra en pies. El procedimiento propuesto hace que ambas coordenadas sean comparables, independientemente de las unidades en las que se hayan medido.
- En algunas aplicaciones, como en procesamiento de imágenes, donde todas las coordenadas están naturalmente en las mismas unidades (las intensidades de los píxeles), no hay necesidad de realizar dicha cambio de escala de las diferentes coordenadas.
- Lo mismo ocurre con el ejemplo del juguete, donde las calificaciones de las distintas comidas tenían la misma escala de 1 a 10.
- Igualmente podemos desandar esta transformación a las escalas originales una vez que encontramos lo que buscamos.

LA FUNCIÓN OBJETIVO PARA EL CASO D = 1

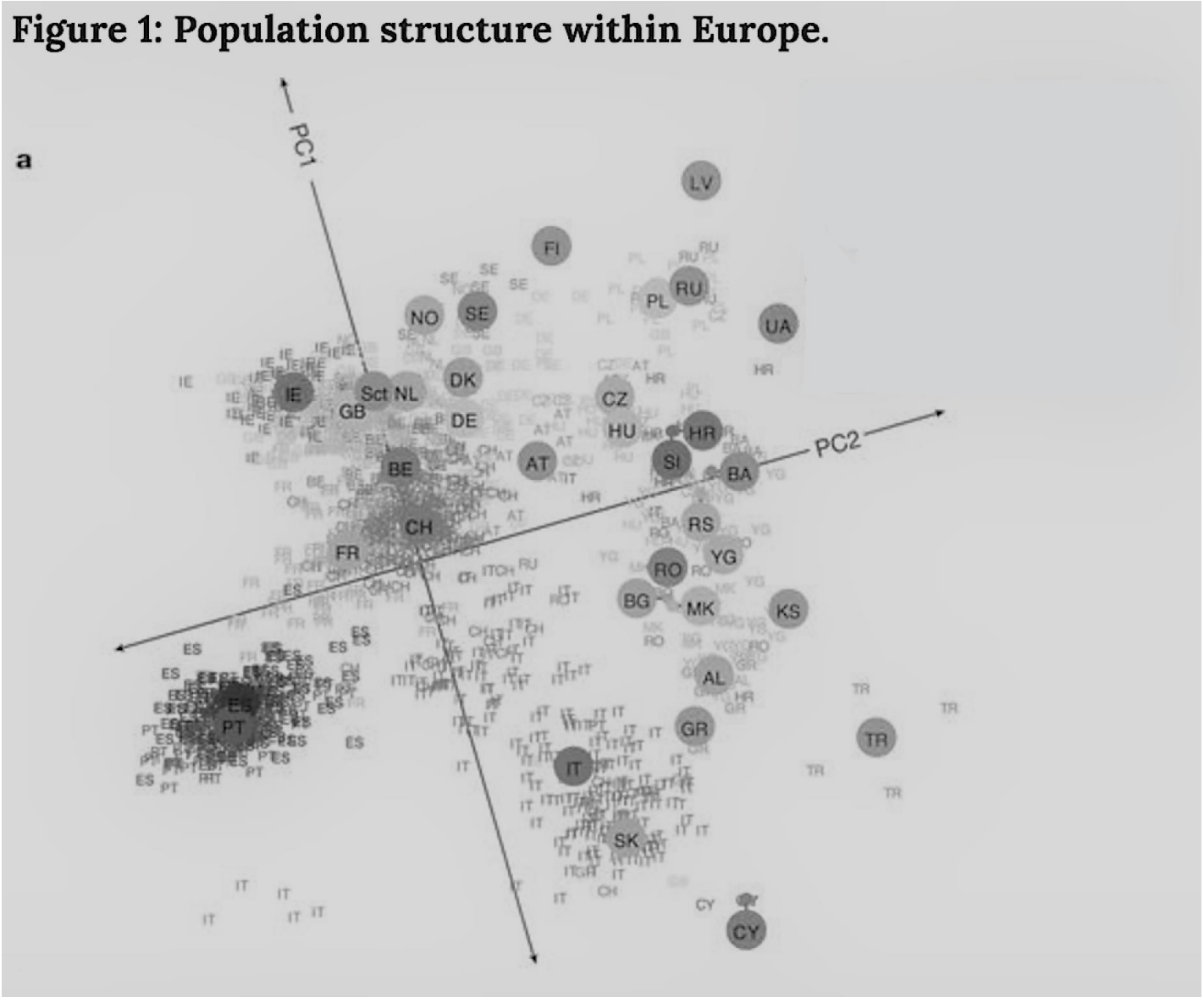
- PCA define a la “mejor línea” como la que minimiza el promedio del cuadrado de la distancia Euclidiana entre la línea y los datos.
- Esta definición está geométricamente motivada.
- Pero además tiene un motivación estadística: es la dirección que maximiza la varianza de los datos.
- Fijado el origen de coordenadas en el centro de los datos, la suma del cuadrado de las distancias desde dicho origen hasta los puntos correspondientes a los datos esta fija. Por el teorema de Pitágoras, minimizar el cuadrado de las distancias ortogonales a la mejor línea es equivalente a maximizar el cuadrado de las distancias a lo largo de dicha línea. Pero esto es proporcional a la varianza de los datos en esta dirección.

$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left((\text{distancia entre } \mathbf{x}_i \text{ y la línea generada por } \mathbf{v})^2 \right)$$

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) DE GENOMAS

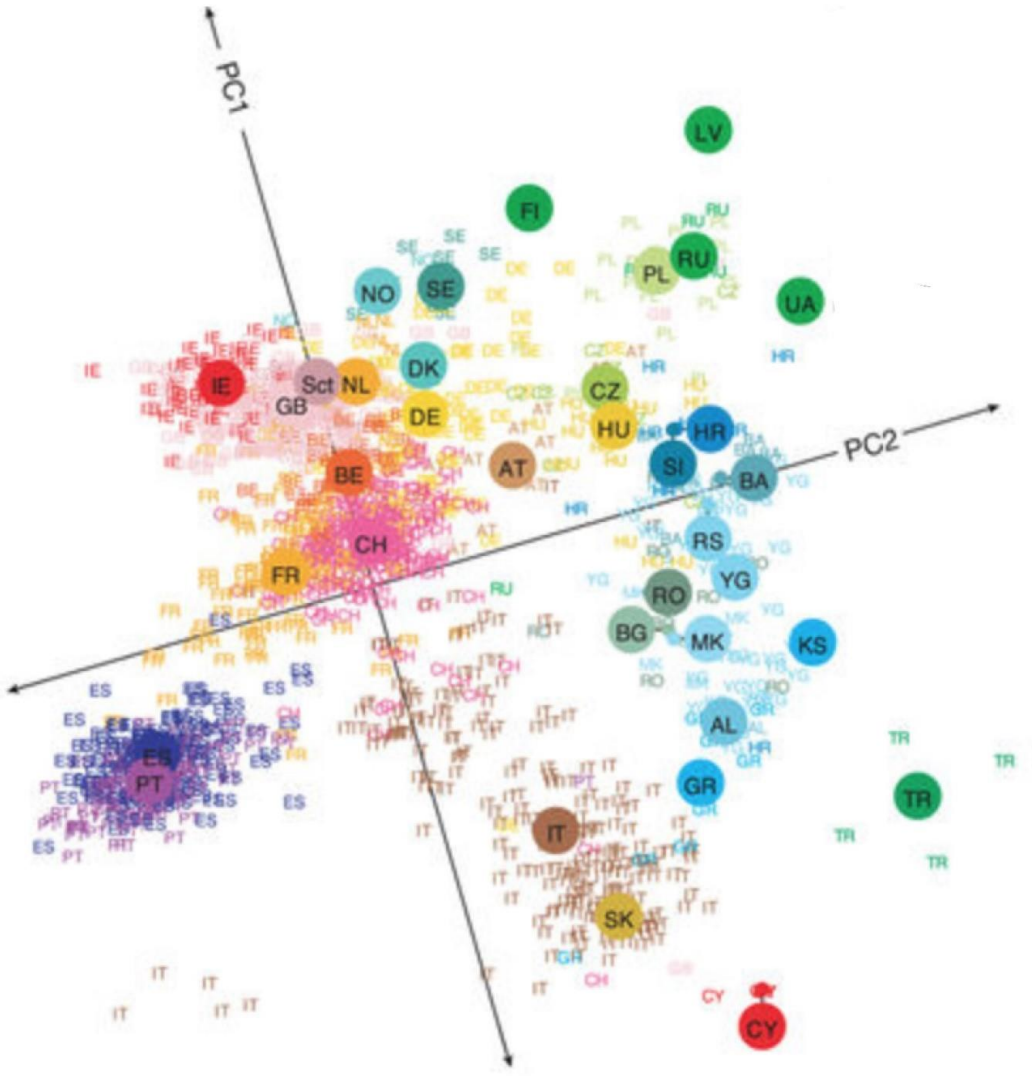
- ¿Podemos inferir dónde nació alguien a partir de su ADN? En un notable estudio, Novembre et al. [1] usaron PCA para mostrar que en algunos casos la respuesta es sí. "Este es un estudio de caso sobre cómo PCA puede revelar que los datos “saben” cosas que no esperarías. Novembre y sus colaboradores consideraron un conjunto de datos compuesto por 1387 europeos (las filas). El genoma de cada persona se examinó en las mismas 200.000 “SNPs” (las columnas) (posiciones del ADN que tienden a exhibir mutaciones genéticas). Entonces, en una de estas posiciones, tal vez el 90% de todas las personas tienen una “C”, mientras que el 10% tiene una “A”. “En pocas palabras, Novembre et al. [1] aplican PCA a este conjunto de datos (con n ~ 1400, d = 200.000 y k = 2) y grafican los datos de acuerdo con los dos componentes principales v1 y v2 (usando exactamente la misma receta que explicamos, con los ejes x e y correspondientes a v1 y v2).
- [1] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltn Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. “Genes mirror geography within Europe”. *Nature*, **456**: 98-101, 2008.

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) DE GENOMAS



ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) DE GENOMAS

Figure 1: Population structure within Europe.



ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) DE GENOMAS

Figure 1: Population structure within Europe.