Carga de Trabajo
Caracterización de la carga

- Tendencia central, dispersión
- Histogramas
- Análisis de componentes principales
- Análisis de conglomerados (cluster).

Caracterización: Fase de construcción (I)

- Análisis de los parámetros
  - Una vez determinados los parámetros se realizan las medidas sobre el sistema.
  - Resultado= Conjunto de datos multidimensional
  - Analizar los datos para identificar grupos de componentes con características similares
- Extracción de los valores representativos
  - Utilización de técnicas estadísticas para extraer valores significativos.
  - Determinación de las características de las clases componentes del modelo.

Caracterización: Fase de construcción (II)

- Asignación de valores a los componentes del modelo
  - Transformación de valores significativos en componentes ejecutables.
  - Alternativas:
    - Carga real (natural): Buscar los componentes cuyos parámetros estén próximos a los valores de los componentes representativos
    - Carga sintética: Asignar a los parámetros de control los valores correspondientes a los de los componentes representativos. Se debe calibrar el modelo
  - Nº de componentes influye en su representatividad y en su compacidad.
- Reconstrucción de mezclas de componentes significativos
  - Reproducir en el modelo situaciones similares a las que se producen en la carga real.

Caracterización: Técnicas más comúnes

- Conjunto de técnicas estadísticas para identificar grupos de componentes con características similares.
- Datos multidimensionales.
- Técnicas utilizadas habitualmente
  - Promedios
  - Dispersiones
  - Histogramas de parámetro sencillo
  - Histogramas multiparamétricos
  - Análisis de componentes principales
  - Clustering

Índices de tendencia central

- Resumir el rendimiento de un sistema de computación con un número puede ser engañoso y puede llevar a conclusiones erróneas.
- Reducir el rendimiento de un sistema de computación a un número es una práctica habitual.
- Los valores medios pueden ser útiles para comparaciones elementales de los sistemas.
- La representatividad del valor medio dependerá de la variabilidad presente en los datos de partida.

Especificación de la tendencia central

- Media aritmética: número sencillo que caracteriza un conjunto de datos.
  - Ej: Tiempo entre llegadas.
  - Se utilizan los datos medidos por lo tanto es una media muestral.
- Mediana. Valor medio de la serie. Utilizar con distribuciones sesgadas.
  - Ej: Carga de un ordenador.
  - Reduce el efecto de sesgo que provocan los *outliers* en el valor de la media.
- Moda: Valor más frecuente.
  - Ej: El recurso más utilizado en un sistema
  - Puede no ser único o no existir.

Especificación de la tendencia central: Consideraciones

- La media incorpora la información de todos los datos en la medida final. Es sensible a los *outliers*.
  - La mediana y la moda son menos sensibles a los *outliers*.
- Datos categóricos
  - La moda puede constituir el índice apropiado.
- Si la suma de todas la medidas es un valor útil y significativo entonces la media es un índice apropiado.
- Si los datos de la muestra contienen un conjunto de valores que no están agrupados , la mediana es más significativa e intuitiva para proporcionar información acerca de la tendencia central.

Tipos de medias utilizadas en rendimiento

- Media aritmética

$$\bar{x}_A = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Media armónica

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- Media geométrica

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

Utilización de medias

- El uso de cada tipo de media depende del significado físico que represente cada medida de rendimiento x_i.
- La media aritmética adecuada para tiempos de ejecución.
  - El valor medio es proporcional a la suma de los tiempos de ejecución
- La media armónica: adecuada para ratios.
  - Ej: Ratio entre el número de operaciones y el tiempo total de ejecución.
  - El valor medio que se computa tiene que ser inversamente proporcional a la suma de los tiempos de ejecución.
- La media geométrica no es adecuada ni para tiempos ni para ratios.
  - Es adecuada para números normalizados
  - Apropiada para resumir las medidas con un rango amplio de valores ya que los valores individuales tienen poca influencia en la media.
  - Mantiene un orden consistente cuando se comparan las prestaciones de varios sistemas de computación. El orden es el mismo independientemente de la máquina que se tome como referencia para hacer la comparación. **El orden puede ser erróneo.**

Media aritmética y media armónica

- Un coche realiza un viaje de A a B, que distan 60 km, a una velocidad 60 km/h. Luego regresa a una velocidad de 40 km/h ¿Cuál es la velocidad media?.
  - Ida: 60 km/h Tiempo de viaje: 1h
  - Vuelta: 40 km/h Tiempo de viaje: 1.5h
  - Tiempo total 2.5h
  - Media aritmética (60+40) / 2 = 50 km/h
  - CONTRADICCIÓN: ¡¡ 50 km/h*2.5h = 125 km !!
- Media armónica
  - 2 / (1/60 km/h+1/40 km/h) = 48 km/h
  - CORRECTO: 48 km/h*2.5h = 120 km

Porcentajes y medias

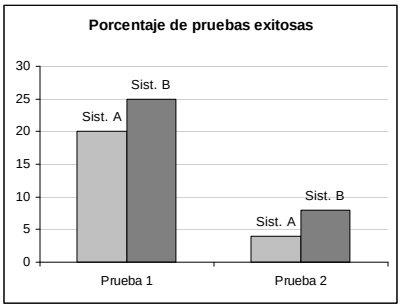
- Dos pruebas realizadas en dos sistemas

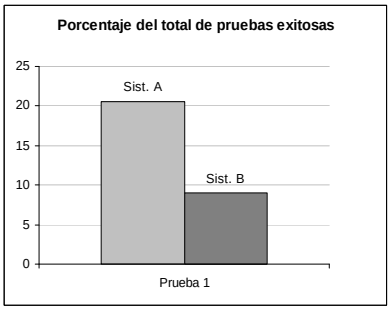
Sistema A

Prueba	Total	Exitos	% éxito
Prueba 1	300	60	20
Prueba 2	50	2	4
	350	62	20,6

Sistema B

Prueba	Total	Exitos	% éxito
Prueba 1	32	8	25
Prueba 2	500	40	8
	532	48	9





Utilización de medias

Tiempos de ejecución

Programa (medida)	Tiempo (s)	F (10^9 FLOP)	Mi (MFLOPS)
1	321	130	405
2	436	160	367
3	284	115	405
4	601	252	419
5	482	187	388
Media aritmética	**425**		**397**
Media armónica	**395**		**396**

Tiempos de ejecución reducidos a la mitad

Programa (medida)	Tiempo (s)	F (10^9 FLOP)	Mi (MFLOPS)
1	161	130	810
2	218	160	734
3	142	115	810
4	301	252	839
5	241	187	776
Media aritmética	**212**		**794**
Media armónica	**197**		**792**

Utilización de medias

Tiempos de Ejecución			
Programa	**Sistema 1**	**Sistema 2**	**Sistema 3**
1	417	244	134
2	83	70	70
3	66	153	135
4	39449	33527	66000
5	772	368	369
Tiempo Total	**40787**	**34362**	**66708**
Media Aritmética	**8157,4**	**6872,4**	**13341,6**
Rango	**2**	**1**	**3**
Media Geométrica	**586,79**	**503,13**	**498,68**
Rango	**3**	**2**	**1**

Normalizado respecto a Sistema 1			
1	1,00	0,59	0,32
2	1,00	0,84	0,84
3	1,00	2,32	2,05
4	1,00	0,85	1,67
5	1,00	0,48	0,48
Media Aritmética	**1**	**1,01465065**	**1,07223929**
Rango	**1**	**2**	**3**
Media Geométrica	**1,00**	**0,86**	**0,85**
Rango	**3**	**2**	**1**

Normalizado respecto a Sistema 2			
1	1,71	1,00	0,55
2	1,19	1,00	1,00
3	0,43	1,00	0,88
4	1,18	1,00	1,97
5	2,10	1,00	1,00
Media Aritmética	**1,32011261**	**1**	**1,08056266**
Rango	**3**	**1**	**2**
Media Geométrica	**1,17**	**1,00**	**0,99**
Rango	**3**	**2**	**1**

Normalizado respecto a Sistema 3			
1	3,11	1,82	1,00
2	1,19	1,00	1,00
3	0,49	1,13	1,00
4	0,60	0,51	1,00
5	2,09	1,00	1,00
Media Aritmética	**1,4952793**	**1,09190074**	**1**
Rango	**3**	**2**	**1**
Media Geométrica	**1,18**	**1,01**	**1,00**
Rango	**3**	**2**	**1**

Especificación de la dispersión: Índices de dispersión

- Rango
  - Diferencia entre el valor máximo y mínimo.
  - Si existen "outliers" no es un buen índice de la variabilidad
  - Depende del número de observaciones.
  - Es útil si el parámetro está acotado.
- Varianza muestral (cuasi-varianza) $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Desviación típica o desviación estándar: s_x
  - Raíz cuadrada de la cuasi-varianza
- Coeficiente de variación $CV \equiv \frac{s_x}{\bar{x}}$
  - CV = 0 indica un parámetro constante. La media es suficiente información
  - CV alto indica una varianza alta. Puede indicar la conveniencia de dividir la población en grupos con parámetros similares

Histogramas

- Muestran las frecuencias relativas de valores del parámetro.
- Parámetros de valores continuos dividir el rango del parámetro en subrangos.
- Los histogramas de un parámetro ignoran la correlación entre los diferentes parámetros
- Los histogramas multi-parámetro muestran la correlación entre los parámetros de la carga de trabajo.
- Representación gráfica: 2-parámetros

Histogramas: Ejemplo

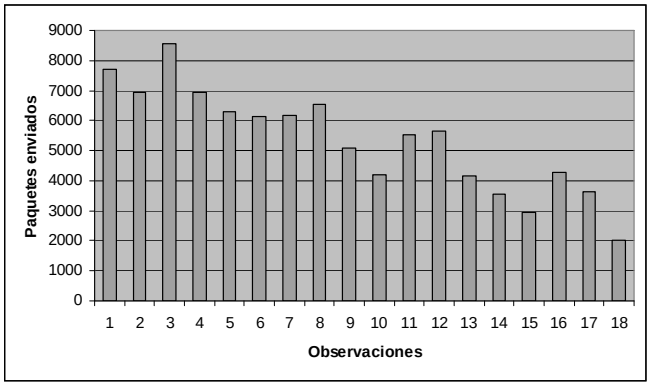
- Se ha medido el número de paquetes enviados (x_s) y recibidos (x_r) por varias estaciones de trabajo en una red de área local.
- Los datos obtenidos se muestran en la siguiente tabla:

Nro. Obs.	Xs	Xr
1	7718	7258
2	6958	7232
3	8551	7062
4	6924	6526
5	6298	5251
6	6120	5158
7	6184	5051
8	6527	4850
9	5081	4825
10	4216	4762
11	5532	4750
12	5638	4620
13	4147	4229
14	3562	3497
15	2955	3480
16	4261	3392
17	3644	3120
18	2020	2946

[Jain91, pp.77]

Histogramas: Ejemplo 1

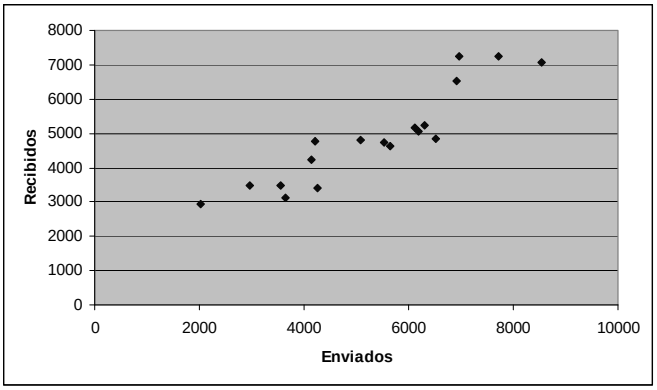
- Enviados



[Jain92, cap.6]

Histogramas: Ejemplo 2

- Recibidos vs. Enviados



[Jain91, pp. 78]

Análisis de componentes principales

- La determinación inicial de las variables se basa en conocimiento *apriorístico*.
- Determinar si se han elegido variables redundantes
  - La información que aportan ya está contenida en la información de otras variables.
  - Su información es independiente del fenómeno en estudio.
- El análisis factorial opera sobre las varianzas y correlaciones entre variables.
  - Identifica o revela conjuntos de relaciones.
- El análisis de componentes principales es una técnica de análisis factorial
  - Técnica para simplificar la estructura de los datos sin modelo prefijado
  - Explicación en pocas componentes de la mayor parte de la información que contienen las variables. Explicar las correlaciones entre un conjunto de variables.
  - Pretende estudiar relaciones de dependencia entre las variables, expresadas a través de un modelo de factores comunes y únicos.
  - Los componentes reproducen o explican esa varianza de la forma más compacta y simple posible.
  - **ATENCION**: Componentes en este contexto no es igual a componente de la carga de trabajo.
  - Tiene que ver con la descripción de la variación de la varianza que es compartida por los datos en tres o más variables.
  - El análisis de datos se realiza sobre las variables derivadas, componentes

Análisis de componentes principales

- Tiene que existir un correlación significativa entre las variables originales.
- Modelo
  - Dado un conjunto de variables $\{x_1, x_2, ..., x_n\}$
    Determinar un conjunto de factores $\{y_1, y_2, ..., y_n\}$ tal que:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j$$

  - a_{ij} expresa la correlación entre la variable x_j y la componente y_i.
  - Las variables y_i son variables aleatorias que no están correlacionadas, forman un conjunto ortogonal, son linealmente dependientes de los parámetros y contienen las proporciones máximas de variabilidad.

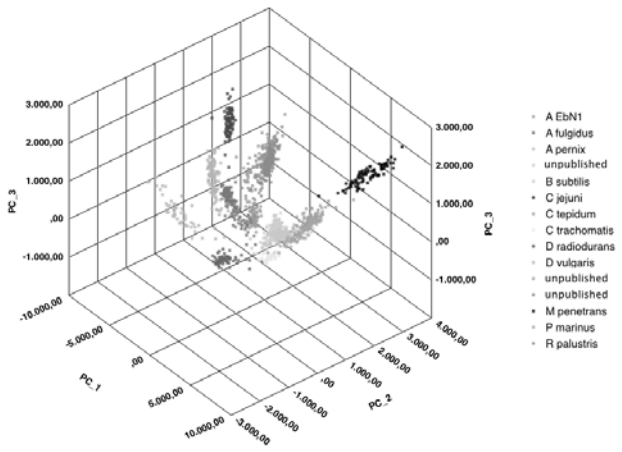
$$var(y_1)=\lambda_1 \geq var(y_2)=\lambda_2 \geq ... \; var(y_n)=\lambda_n$$

  - λ_i son los autovalores o valores propios de la matriz de covarianzas (o la de correlación si se utilizan las variables reducidas, media cero y varianza 1)
  - a_i es el autovector de norma 1 asociado al autovalor de λ

Cálculo de componentes principales

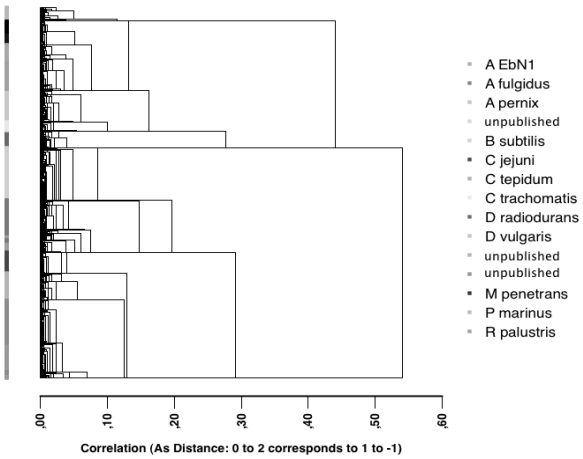
- Examinar la matriz de correlaciones. Si son distintas de cero existe una estructura en los datos indicando la existencia de algún factor subyacente.
- Algún autor estima que si las correlaciones no exceden de 0.3 no tiene sentido el análisis.
- Nº de variables mayor que 3. Nº de muestras varía, algún autor establece un mínimo de 10 observaciones por variable.
- ¿Selección del nº de factores?
  - Varianza explicada: > 80%
  - Elegir factores cuyo valor propio sea >1 con variables reducidas (var=1)
- El método funciona bien con muestras "grandes" (tamaño de muestra mayor a 300) y pocas variables (número de parámetros menor a 40).
- Si las unidades de medida son distintas utilizar la matriz de correlaciones (implica reducir las variables). Si son las mismas o parecidas es mejor hacerlo sobre la matriz de covarianzas, es menos artificial.
- Se debe tener cuidado con las interpretaciones de las resultados.

Análisis de Componentes Principales: Ejemplo

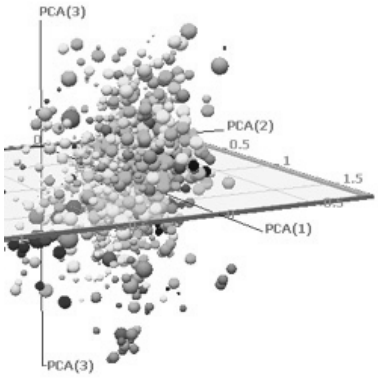


Análisis de Componentes Principales: Ejemplo

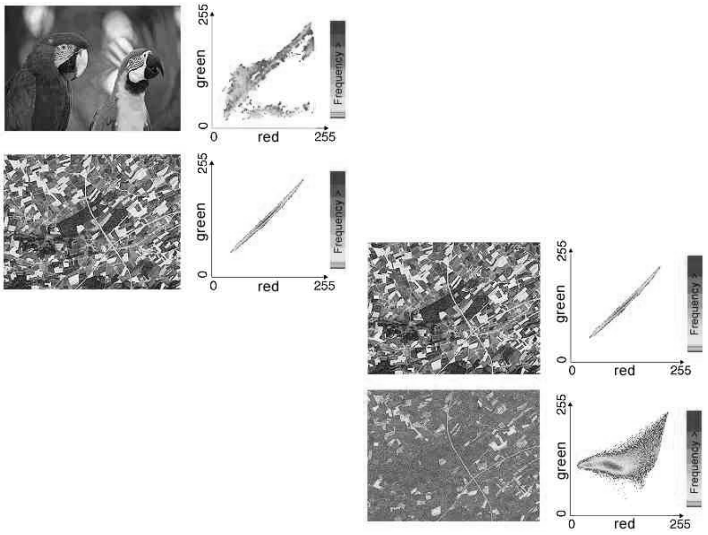
www.megx.net/tetra_new/html/application.html



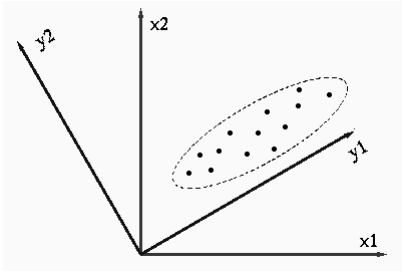
Análisis de Componentes Principales: Ejemplo



Análisis de Componentes Principales: Ejemplo



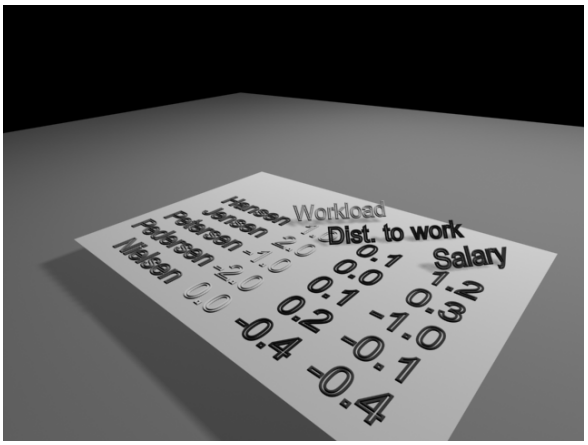
Análisis de Componentes Principales: Ejemplo



The figure above illustrates an orthogonal transform in a $n = 2$ dimensional space. The signal components x_1 and x_2 are highly correlated, and they carry about the same amount of information (dynamic energy). But after the rotation corresponding to an orthogonal transform $Y = A^T X$, the components y_1 and y_2 are decorrelated, and most of the information is concentrated in component y_1. If we only keep y_1 but ignore y_2, a 50% compression rate can be achieved without losing much information in the signal.

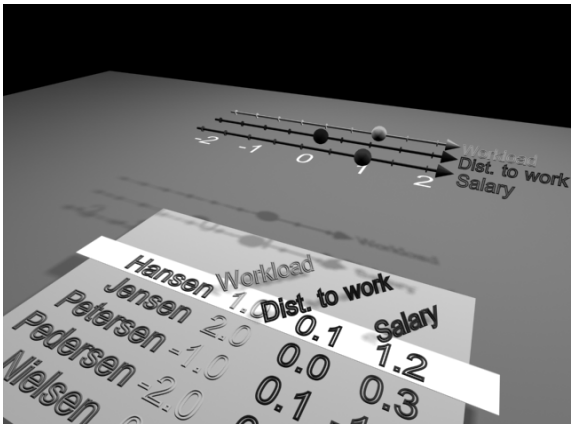
Principal Component Analysis

- http://www.models.kvl.dk/research/Factor_Model_Movies/PCA_story/
- To show the basic idea behind PCA a small data set is generated. It consists of **five persons/samples** (Hansen, Jensen, Petersen, Pedersen and Nielsen) and for each person **three variables** exists describing the added workload, the added salary and the "distance to work".



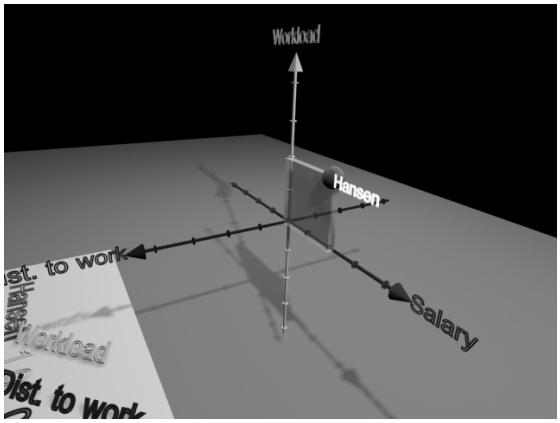
Principal Component Analysis

- When plotting these five samples a three-dimensional space is required where the axes would be the three variables.
- In the figure to the right the three variables for the first sample - Hansen - are marked on the three separate axes. The specific location of the three black spheres represent the variable values of Hansen: 1.0, 0.1 and 1.2 for the added workload, the "distance to work", and the added salary, respectively.



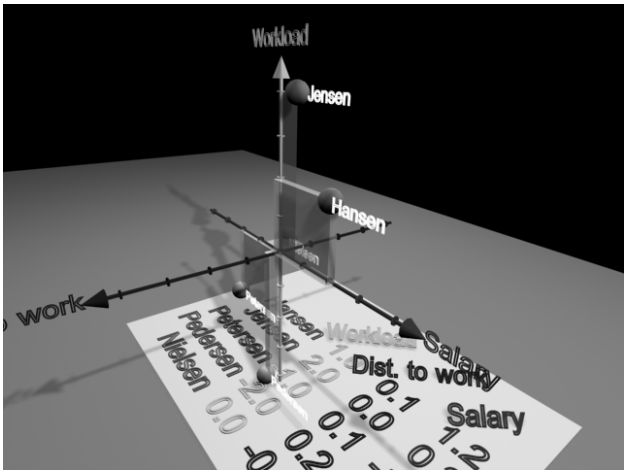
Principal Component Analysis

- The three axes are then turned into orthogonal positions and thereby form the three-dimensional space.
- The location of Hansen is noted to be:
  - 1.0 up of the green workload axis
  - 0.1 out of the blue "distance to work" axis
  - and 1.2 out of the red salary axis



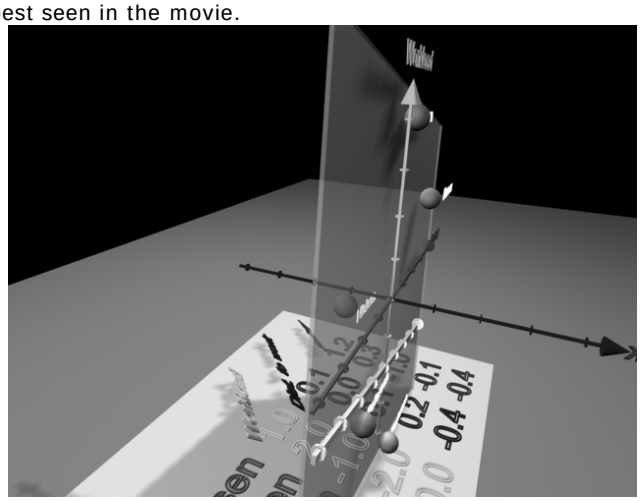
Principal Component Analysis

- Now all five samples are in place, all in direct relation to their values of added workload, "distance to work" and added salary from the original data-table.



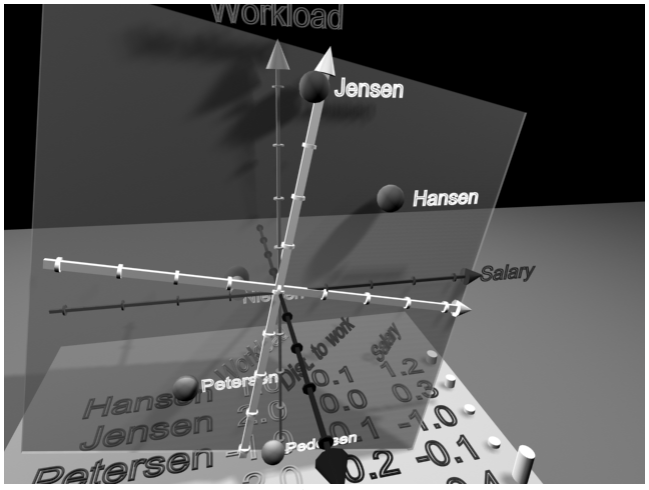
Principal Component Analysis

- It is observed that all samples seem to be located close to a flat two-dimensional plane in the three-dimensional box. This is an important fact - best seen in the movie.



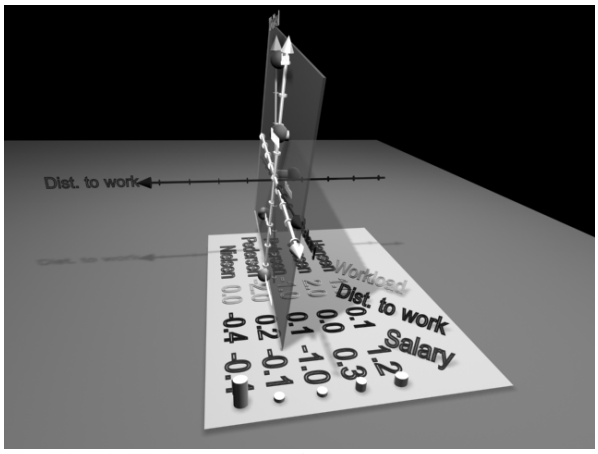
Principal Component Analysis

- This plane - seen as the transparent square - may be described using two new axes (white). The exact spatial location is determined by a minimization of a least squares residual.



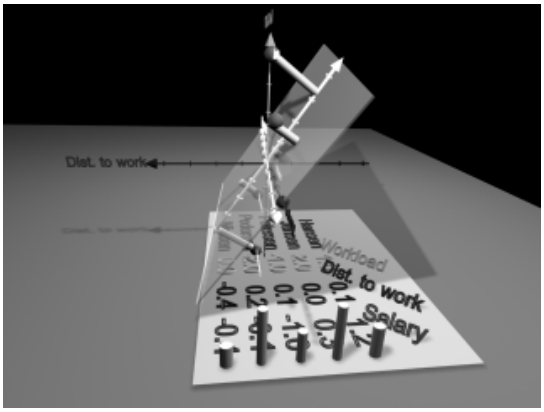
Principal Component Analysis

- The frame to the right is a zoom from the backside of the system of coordinates. The distance from each sample into the transparent plane is the residual of that sample (the cylinder) - the length is equal to the size of the residual.
- The sum of the (squared) residuals from all the samples are made as small as possible when placing the white axes.



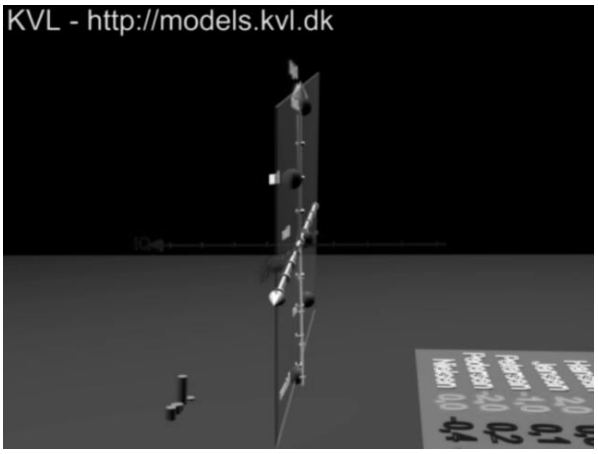
Principal Component Analysis

- The residuals are removed and the samples projected orthogonal onto the new plane described by the two white axes. These are also known as **principal components** or **latent factors**.



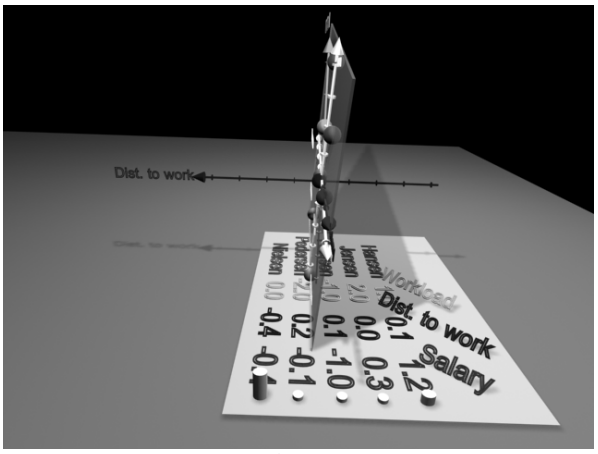
Principal Component Analysis

- The new coordinates of the samples in the new flat two-dimensional space described by the principal components (white axes) are known as **scores** and together with their definition (the corresponding loading), they are called **principal components.**



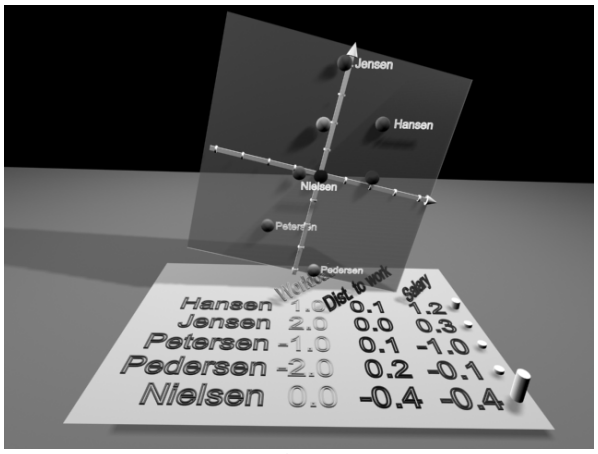
Principal Component Analysis

- Now three unit vectors are placed in origin of the original three-dimensional space along the axes green, red and blue axes and in their terminal point a sphere is placed. These are also projected onto the transparent flat plane described by the principal components.
- This leads to the definition of **loadings**, which describes the relationship between the original variables.



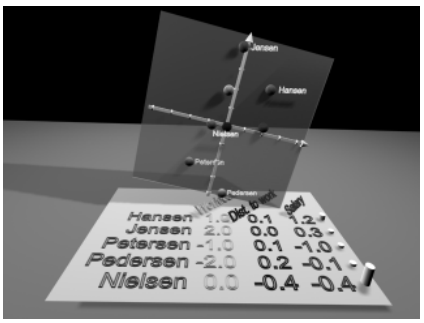
Principal Component Analysis

- When removing the original system of coordinates (the red, blue and green axes) and the unit vectors it leaves the new (reduced) system of coordinates represented by the principal components (white axes). In this system of coordinates both samples and variables are located - the plot is also called bi-plot because it describes the relationship between both the samples/persons and the variables.



Principal Component Analysis

- One may now conclude:
  - Jensen and Hansen are approximately similar because they are located relatively close to each other. Additionally it can be seen that Hansen and Jensen have high values of workload because the green workload sphere is located closely to these samples/persons.
  - Jensen and Hansen posses reverse properties compared to Petersen and Pedersen because they are located opposite from each other.
  - The variable "distance to work" does not describe the samples in any way because the blue "distance to work" sphere is located in origin.



Principal Component Analysis

- The above conclusions are drawn from visual exploration of the bi-plot only. One may correctly argue that similar conclusion could have been drawn from just inspecting the raw data - but when analyzing much larger data sets this is no longer an option. Similar conclusions may also be drawn by using traditional statistics - but it does not e.g. automatically point to important variables. Using traditional statistics one must manually test all possible correlations - a cumbersome job when having thousands of samples each consisting of thousands of variables.

- In conclusion, four types of information may be retrieved using PCA:
  - The relationship between the samples are described by the score values - closely located samples are correlated;
  - The relationship between the variables is described by the loading plot - closely related variables are correlated;
  - The distances from the samples to the principal components are described by the residuals;
  - The relationship between samples and variables are depicted in a bi-plot.

Método de Cálculo [Jain92]

- Calcular la media y la desviación estándar de las variables.
- Normalizar las variables a media cero y desviación estándar 1. (obtener las variables reducidas) $x_i = \frac{x_i - \bar{x}_i}{s_{x_i}}$
- Calcular la correlación entre variables.
- Preparar la matriz de correlaciones C $R_{ij} = \left(\frac{1}{n}\right) \frac{\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right)}{s_i s_j}$
- Resolver $|\lambda I - C|$
- Calcular los autovectores de la matriz de correlación con norma 1.
  $C\, q_l = \lambda_l q_l$
- Obtener los factores principales multiplicando los vectores propios por los vectores normalizados.
  $Y = QX$
- Calcular los valores de los factores principales.
- Establecer la gráfica de los valores de los factores principales.

Ejemplo

Nro. Obs.	Xs	Xr
1	7718	7258
2	6958	7232
3	8551	7062
4	6924	6526
5	6298	5251
6	6120	5158
7	6184	5051
8	6527	4850
9	5081	4825
10	4216	4762
11	5532	4750
12	5638	4620
13	4147	4229
14	3562	3497
15	2955	3480
16	4261	3392
17	3644	3120
18	2020	2946

```
correlations
         enviados recibidos
---------------------------
enviados           0,9156
recibidos  0,9156
---------------------------
```

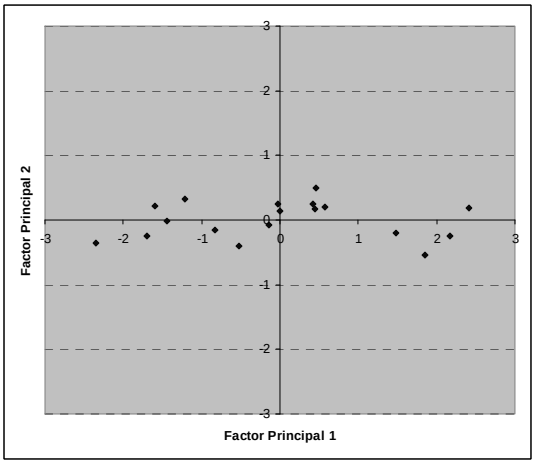

Ejemplo

```
Analysis Summary
Data variables:
recibidos
enviados
Data input: observations
Number of complete cases: 18
Missing value treatment: listwise
Standardized: yes
Number of components extracted: 2
Principal Components Analysis
-------------------------------------------
Component   Percent of Cumulative
  Number    Eigenvalue Variance Percentage
     1     1,91561     95,781   95,781
     2     0,0843894    4,219  100,000
-------------------------------------------
Table of Component Weights
          Component 1 Component 2
          ----------- -----------
recibidos  0,707107      0,707107
enviados   0,707107     -0,707107
```

```
Table of Principal Components
          Component   Component
Row         1             2
------ ------------ ------------
 1        2,17506      0,253176
 2        1,85306      0,54852
 3        2,41291     -0,185611
 4        1,47737      0,200443
 5        0,569572    -0,198858
 6        0,449607    -0,174234
 7        0,420754    -0,255074
 8        0,457032    -0,497412
 9       -0,143071     0,0770608
10       -0,52668      0,396085
11        0,00165749  -0,144555
12       -0,0219272   -0,254243
13       -0,827913     0,1509
14       -1,44072      0,0132823
15       -1,69597      0,251099
16       -1,21065     -0,324436
17       -1,60066     -0,213267
18       -2,34944      0,357125
```

Ejemplo

Componente ppal 1 vs Componente ppal 2



Partición de la carga de trabajo.

- Partición de la carga de trabajo.
- Las cargas reales pueden verse como una colección de componentes heterogéneos.
- Se mejora la representatividad de la caracterización y se incrementa el poder predictivo del modelo.
- Las técnicas de particionamiento dividen la carga de trabajo en una serie de clases de forma que sus poblaciones estén formadas por componentes bastante homogéneos.
- Objetivo: Agrupar componentes que sean similares.
  - ¿ Qué atributo utilizar para establecer la medida de similaridad?
- Se abordará el problema en dos fases.
  - Analizando algunos atributos utilizados para particionar la carga en clases de componentes similares.
  - Análisis cluster que permiten calcular los valores de los parámetros que representan a la clase de componentes.

Partición de la carga de trabajo: Atributos

- Utilización de recursos.
  - El consumo de recursos por cada componente puede utilizarse para dividir la carga de trabajo en clases.
  - Ejemplo: Clases de transacciones en un entorno interactivo.
    - Se considera que el procesador y las E/S son los elementos críticos del sistema.
    - Se establece la siguiente división

Transacción	Frecuencia	Tiempo Máximo de CPU (mseg)	Tiempo Máximo de E/S (ms)
Trivial	40%	8	120
Ligera	30%	20	300
Media	20%	100	700
Pesada	10%	900	1200

Partición de la carga de trabajo: Atributos (cont.)

- Aplicaciones
  - Se pueden agrupar los componentes de la carga basándose en la aplicación a la que pertenecen.
  - Ej: Contabilidad, inventario, servicios de cliente, etc.
  - Ej: Aplicaciones de Internet.
    - Transacciones: http, ftp, email (smtp, pop3)
    - Aplicaciones multimedia (streaming media)
    - Otros

Aplicación	% tráfico total
Http	29
Ftp	20
Smtp y pop3	9
Streaming	11
P2P	14
Otros	17

Partición de la carga de trabajo: Atributos (cont.)

- Objetos
  - Se puede dividir la carga siguiendo el criterio de tipo de objeto manejado por las aplicaciones.
  - Ej: En la WWW, la carga puede ser dividida por los tipos de documentos accedidos en el servidor.

Clase de documento	% de accesos
HTML (ej .html, .htm)	30.0
Imágenes (ej .gif, .jpeg, ...)	40.0
Sonido (ej .au, .mp3, .wav,...)	4.5
Video (ej .mpeg, .avi, .mov)	7.3
Dinámicas (ej cgi o perl)	12.0
Formateadas (ej .ps, .pdf, .doc)	5.4
Otros	0.8

Partición de la carga de trabajo: Atributos (cont.)

- Distribución geográfica
  - Se distinguen entre peticiones y transacciones locales y remotas.
  - Tiene sentido cuando hay que considerar demoras provocadas por las WAN.
    - Ej: local y remota. Por dirección IP agrupando regiones geográficas.
- Unidades organizativas
  - Se tiene en cuenta la organización de la compañía. Ej: Finanzas, Marketing, Produccción,etc.
- Modo
  - El modo de procesamiento o el tipo de interacción con el sistema puede ser utilizado para categorizar los componentes de la carga de trabajo.
    - Interactivo: componentes generados por un nº determinado de PC, estaciones de trabajo con un tiempo de pensar dado. Ej: sistema de e-learning
    - Transacción: conjunto de componentes que llegan a un sistema de computación con una tasa dada de llegadas.Ej: banco en línea.
    - Continuo: componentes ejecutados de forma continua como trabajos batch, y puede ser descrito como el número de componentes activos en el sistema. Ej: Servidor DNS.

Partición de la carga de trabajo: Atributos (cont.)

- Funcional.
  - Se pueden agrupar siguiendo el criterio de las funciones que se realizan.
  - Se recomienda establecer el mínimo número de clases que son necesarias para el estudio de rendimiento.
  - Establecer las clases relevantes para el objetivo del estudio y la organización, el resto de las clases se agrupan en una única.
  - Ej: Carga de trabajo en un sistema UNIX
    - Se puede caracterizar asociando los nombres de proceso con las funciones realizadas por el sistema.

Clases	Nº de comandos	Tiempo total CPU (sec)
Oracle	1515	19350
SAS	58	18020
cp	950	7500
date	225	26
ls	90	115
find	50	60

Cálculo de los parámetros de las clases

- Clases de componentes lo suficientemente homogéneos.
  - Medias
  - Medidas de variabilidad.
- Método más habitual utilizar clustering.
  - Concepto aplicado a un conjunto de técnicas para seleccionar clases de objetos similares.
- Se necesita que la semejanza entre los objetos esté cuantificada.
- Los objetos (componentes de carga) que han de ser clasificados se presenten como elementos de un espacio cuyas dimensiones son los caracteres (parámetros) en base a los que se establece la clasificación.
- En análisis cluster se conoce muy poco o nada acerca de la estructura de las clases.
- El objetivo operacional consiste en descubrir una estructura de clases que se ajuste a las observaciones.

Principios del análisis de conglomerados

- El principio básico radica en establecer grupos de elementos tales que cada miembro de un grupo sea más parecido o esté más cohesionado con los de su mismo grupo que con los de cualquier otro.
- Los grupos obtenidos son disjuntos, es decir, ningún elemento pertenece a más de un grupo.
- La totalidad de los grupos agota el campo bajo estudio, es decir, no hay individuos sin clasificar.
- El nexo o relación global entre dos individuos es función de su similitud en cada uno de los aspectos considerados como relevantes.

Elementos del proceso de análisis de conglomerados

- Elección de los objetos a clasificar
  - Muestra de los componentes de la carga
- Elección de las características o variables observadas sobre cada objeto.
  - Selección de los parámetros de la carga de trabajo
- Estandarización de las variables
  - Transformación de parámetros. El análisis funciona mejor con unidades de datos relativamente similares.
  - Eliminación de valores anómalos (outliers).
  - Normalización y escalado de las variables
- Elección de la medida de semejanza. Medida de distancia.
- Implementación de los algoritmos para computar los conglomerados.
- Interpretación de los resultados.

Elección de variables

- Elección de variables para el análisis cluster
  - Poca "variabilidad" = poco poder discriminatorio.
    Pueden dar lugar a clases que carecen de sentido.
  - Demasiada "variabilidad" = fuertes discriminadores.
    Pueden enmascarar la busca de los cluster y dar lugar a resultados erróneos.
  - Tienen que ser efectivas para la clasificación.
  - Elegir variables que muestren diferencias consistentes entre subgrupos.
- Criterio en la selección de variables de la carga de trabajo para el análisis cluster.
  - Varianza
  - Impacto en el rendimiento.
  - Ej: si el *nº de líneas impresas* no tiene impacto en el rendimiento: eliminarla.

Tipos de variables

- Variables
  - Pueden ser: Continua, Discreta, Binaria o dicotómica
- Escalas de medidas. Cada escala cumple todas las propiedades de las anteriores.
  - Nominal: Distingue entre clases. $X_A = X_B \; o \; X_A \neq X_B$
    - Discretas: Estado de la CPU
    - Binarias: Conectado-Desconectado
  - Ordinal: Introduce un orden entre los objetos. $X_A > X_B \; o \; X_A < X_B$
    - Continuas: Juicios humanos sobre intensidades.
    - Discretas: Tamaños de documento web (pequeño, medio, grande)
  - Intervalo: asigna una medida significativa a la diferencia entre dos objetos. Se puede decir que A es $X_A - X_B$ unidades mayores que B.
    - Continua: Grados centígrados.
    - Binarias: Respuestas 1 para terminal activo 0 en otro caso.
  - Razón: es una escala de intervalo en la cual existe un origen inferior de medidas denominado punto 0. Si $X_A > X_B$ entonces se dice que A es X_A / X_B veces superior a B.
    - Continua: Tiempo de respuesta.
    - Discretas: Nº de procesos.

Estandarización de variables

- Homogeneización.
  - Pueden existir variables pertenecientes a distintos tipos.
  - Se necesita preparar las variables para que las unidades de datos sean similares.
  - Conversiones de escala.
    - Elegir un tipo particular de escala y transformar las variables hasta alcanzar la homogeneidad.
    - Orden de escala (información creciente): nominal, ordinal, intervalo, razón.
    - Ascender una variable en la escala implica información adicional o la aceptación de una nueva hipótesis.
    - El descenso en el nivel implica la renuncia a parte de la información.
  - Se puede elegir una variable dominante convirtiendo a este tipo tantas variables como sean necesarias.

Estandarización de variables

- Elementos anómalos (*outliers*).
  - Puntos con valores de los parámetros extremos.
  - Tienen efectos significativos en la media y varianza.
  - Aunque se incluyen en la normalización su inclusión o exclusión puede tener efectos significativos en los resultados finales del clustering.
- Escalado
  - Normalización.
    - Reducción de variables a forma estándar (media cero y varianza unidad)
  - Pesos
    - Parámetro k de valores $\{x_{1k}, x_{2k}, ...,x_{mk}\}$
    - $x'_{ik} = w_k x_{ik}$
    - El peso wk puede ser asignado dependiendo de la importancia relativa del parámetro o ser inversamente proporcional a la desviación estándar de los valores del parámetro.
  - Normalización de rangos. Igualdad de amplitud de variables
    - El rango es cambiado de $[x_{min,k}, x_{max,k}]$ a $[0,1]$
    - $x'_{ik} = (x_{ik} - x_{min,k}) / (x_{max,k} - x_{min,k})$

Algunas medidas de distancia

- Variables de tipo cuantitativo.
  - Sean dos componentes $\{x_{i1}, x_{i2}, ...,x_{in}\}$ y $\{x_{j1}, x_{j2}, ...,x_{jn}\}$
- Distancia euclidiana.

$$d = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Distancia euclidiana ponderada.

$$d = \sqrt{\sum_{k=1}^{n} a_k (x_{ik} - x_{jk})^2}$$

- Distancia Chi-cuadrado.

$$d = \sqrt{\sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{x_{ik}}}$$

Obtención de las clases. Algoritmos

- Ninguno de los algoritmos existentes posibilita obtener una agrupación óptima según un criterio dado (y tampoco es posible examinar todas las posibles agrupaciones).
- Minimizar la varianza intragrupo y maximizar la intergrupos.
  - Varianza total = varianza intragrupos+varianza intergrupos
- Métodos jerárquicos
  - La secuencia de formación de grupos ofrece un orden o jerarquía, representable gráficamente bajo la forma de un árbol o dendograma.
- Métodos no jerárquicos
  - Se parte de un conjunto arbitrario de k clases y los miembros de las clases se van moviendo hasta que la varianza intragrupo es mínima.
- El análisis cluster se realiza habitualmente con paquetes software

Métodos Jerárquicos

- Dos modalidades: aglomerativas y las divisivas.
  - Aglomerativas: Se parte de n clusters (n componentes) se procede uniendo grupos vecinos hasta que se obtiene el nº de clusters deseados.
    - Los componentes de un grupo creado en una determinada fase permanecen unidos en las sucesivas.
  - Divisivas: Se parte de un cluster con todos los componentes. Se van dividiendo sucesivamente los clusters hasta que se obtiene el nº de clusters deseado.
- Métodos aglomerativos
  - Parten de la matriz de similaridades (o distancias)
  - Se identifican los dos componentes más relacionados que pasan a formar parte del primer grupo.
  - Variantes para definir (calcular) la relación entre dicho grupo y los componentes o entre dos grupos.

Algoritmos aglomerativos

- Variantes para definir (calcular) la relación entre dicho grupo y los componentes o entre dos grupos.
  - a. Similitud máxima (single linkage) :
    - Índice de relación entre A y un individuo, u otro grupo B, se define como el índice de similitud máxima (distancia mínima) existente entre un miembro de A y otro de B.
    - Opera bien ante grupos bien separados entre sí, pero no cuando hay casos aislados intermedios.
  - b. Similitud mínima (complete linkage):
    - Índice de relación entre A y un individuo, u otro grupo B, se define como el índice de similitud mínima (distancia máxima) existente entre un miembro de A y otro de B.
    - Adecuado para grupos compactos, pero no ante grupos internamente heterogéneos, aunque estén bien separados. Se ve poco afectado por la existencia de casos intermedios.

Algoritmos aglomerativos

- c. Similitud media (average linkage)
  - Índice de relación entre A y un individuo, u otro grupo B, se define como la media de todos los índices de relación entre pares de miembros de A y de B.

$$S_{A,B} = \frac{\sum_{i,j} S_{ij}}{N_A N_B} \; ; \; i \in N_A, j \in N_B,$$

Algoritmos Aglomerativos: Variantes

- d. Centroide
  - (Def.: centroide de un cluster es el punto cuyos valores de las variables son las medias de los valores de las variables de todos los puntos del cluster)
  - Para cada grupo formado se obtiene un perfil (centroide) como la media en cada variable de los valores individuales o de los grupos previos.
  - El índice de relación entre grupos (la distancia) se aplica ahora a los vectores que los representan.
- e. Mínima variación intragrupo (Ward)
  - Principio de minimizar en cada etapa de fusión la heterogeneidad dentro de cada grupo.
  - En cada fase, se evalúan todas las posibles fusiones y se realiza aquella que provoca el menor incremento de la suma total de las variaciones intragrupo.
  - Variaciones intragrupo se evalúan como la suma de las diferencias al cuadrado entre los valores de cada componente y sus medias de grupo.
  - Tiende a favorecer la fusión de grupos pequeños, compacto y está fuertemente condicionado por la estandarización de las variables.

Métodos no jerárquicos

- Elección del conjunto inicial de las k clases
  - Elegir k componentes al azar para conformar "nucleos" y asignar los restantes a aquel núcleo más próximo.
  - Tomar una agrupación al azar de los componentes en k grupos mutuamente excluyentes y computar sus centroides (valores medios de los componentes en cada grupo en las diferentes variables)
  - Usar los grupos resultantes de una clasificación jerárquica previa.

- Trasvase de componentes entre clases.
  - K-medias
  - Determinar los k-centroides y asignar cada uno de los restantes componentes al centroide más próximo [la suma de las diferencias al cuadrado para cada variable respecto al centroide de cada grupo].
  - Después de cada asignación se calcula de nuevo el centroide.
  - Una vez que todos los componentes están agrupados los centroides se toman como fijos y se ejecuta una última revisión para reubicar, si procede, los componentes al centroide más cercano. Variante: seguir calculando hasta que la partición se estabilice

Minimal Spanning Tree (MST) Method [Menascé02, pp.242]

- Técnica jerárquica aglomerativa.
  - Establecer el número de clusters inicial igual al número de componentes de la carga de trabajo ($j=p$)
  - Repetir los siguientes pasos hasta que el número deseado de clusters sea obtenido.
  - Determinar los valores de los parámetros del centroide Cj de cada uno de los j clusters. Estos valores de los parámetros son las medias de los valores de los parámetros de todos los puntos en el cluster.
  - Calcular la matriz de distancias intercluster $j \times j$, donde cada elemento (m,n) representa la distancia entre los centroides de los clusters m y n
  - Determinar el elemento mínimo distinto de cero (q,r) de la matriz de distancias. Indica que los clusters q y r tienen que ser unidos.
  - Decrementar el numéro de clusters ($j \leftarrow j-1$).

Algoritmo de k-medias [Menascé02, pp.243]

- Técnica no jerárquica
  - Establecer el número de clusters en k
  - Elegir k puntos de inicio, para ser utilizados como estimadores iniciales de los centroides del cluster. Por ejemplo, se pueden seleccionar los primeros k puntos de la muestra o los k puntos que estén más alejados entre ellos.
  - Examinar cada punto de la carga de trabajo y colocarlo en el cluster cuyo centroide esté más cercano. La posición del centroide se recalcula cada vez que un punto nuevo se añade al cluster.
  - Repetir el paso 3 hasta que no existan puntos que cambian su asignación durante un paso completo o hasta que se realice un número máximo de pasos.

Interpretación del agrupamiento (cluster)

- Las salidas de los paquetes software suelen ser una descripción estadística de los centroides del cluster con el nº de componentes de cada cluster.
- Un cluster con muy poca población se descarta, sobre todo si sus miembros tienen un impacto no significativo en el rendimiento.
- Seleccionar uno o mas componentes representativos de cada cluster para utilizarlos como carga de prueba en los estudios de rendimiento.
- El número de representantes puede ser proporcional al tamaño del cluster, a la demanda de recursos totales del cluster o a cualquier combinación de los dos criterios anteriores.
- Los parámetros de caracterización de la clase coinciden con los del centroide del cluster.
- El clustering es mejor que la selección aleatoria de programas. Sin embargo los resultados tienen una variabilidad alta. No existen reglas de selección de parámetros, de métrica o escalado.

Interpretación del agrupamiento (cluster)

- No sirve para la comparación de cargas en sitios diferentes.
- Para estudios de planificación de capacidad es deseable que el número de clusters sea pequeño.
- Criterio de parada
  - Linkage distance: medida que representa la distancia más lejana entre un componente en un cluster a un componente en otro cluster. Se incrementa en función de cómo son de diferentes los componentes que están siendo combinados. Si excede un determinado límite el algoritmo se detiene.
  - Examinar la variación de dos métricas:
    - Distancia intracluster: distancia media entre puntos de un cluster y su centroide.
    - Distancia intercluster: distancia media entre centroides.
    - Esta variación se puede caracterizar por el coeficiente de variación CV. Tener en cuenta que el objetivo del agrupamiento es minimizar el CV intracluster y maximizar el CV intercluster.
    - El ratio entre los CV intra e inter cluster se denota por β_{CV}. El valor más pequeño es el mejor.

Ejemplo [Menascé02, pp.243]

- Web server (procesador rápido y un subsistema grande de discos)
- Sirve documentos html bajo demanda.
- Se registran en el Web log el comportamiento de las peticiones
- Analizar el coste-beneficio de poner en caché los documentos más populares en memoria.
- Estudiar la relación entre el tamaño de un documento y su popularidad.
- Parámetros que caracterizan la carga para este estudio
  - Tamaño
  - Nº de accesos
- Sobre los datos originales se efectúa una transformación log_{10} ya que los datos no son homogéneos.
- Objetivo 3 clusters
- Distancia euclídea

Ejemplo
Tabla de datos

Muestra de la carga

Documento	Tamaño (KB)	Nro. de Accesos
1	12	281
2	150	28
3	5	293
4	25	123
5	7	259
6	4	241
7	35	75

Ejemplo

Transformación Logarítmica de Parametros

Documento	Tamaño (KB)	Nro. de Accesos
1	1,08	2,45
2	2,18	1,45
3	0,70	2,47
4	1,40	2,09
5	0,85	2,41
6	0,60	2,38
7	1,54	1,88

Ejemplo

Centroides de los clusters iniciales

Documento	Tamaño (KB)	Nro. de Accesos
C1	1,08	2,45
C2	2,18	1,45
C3	0,70	2,47
C4	1,40	2,09
C5	0,85	2,41
C6	0,60	2,38
C7	1,54	1,88

Matriz de distancia intercluster

Cluster	C1	C2	C3	C4	C5	C6	C7
C1	0	1,49	0,38	0,48	0,24	0,48	0,74
C2		0	1,79	1,01	1,64	1,83	0,76
C3			0	0,79	0,16	**0,13**	1,03
C4				0,00	0,64	0,85	0,85
C5					0	0,25	0,88
C6						0	1,1
C7							0

Ejemplo

Centroides de los clusters iniciales

Documento	Tamaño (KB)	Nro. de Accesos
C1	1,08	2,45
C2	2,18	1,45
C36	0,65	2,42
C4	1,40	2,09
C5	0,85	2,41
C7	1,54	1,88

Matriz de distancia intercluster

Cluster	C1	C2	C36	C4	C5	C7
C1	0	1,49	0,43	0,48	0,24	0,74
C2		0	1,81	1,01	1,64	0,76
C36			0	0,82	**0,19**	1,05
C4				0	0,64	0,26
C5					0	0,88
C7						0

Ejemplo

Centroides de los clusters iniciales

Documento	Tamaño (KB)	Nro. de Accesos
C1	1,08	2,45
C2	2,18	1,45
C365	0,75	2,42
C4	1,40	2,09
C7	1,54	1,88

Matriz de distancia intercluster

Cluster	C1	C2	C365	C4	C7
C1	0	1,49	0,33	0,48	0,74
C2		0	1,73	1,01	0,76
C365			0	0,73	0,96
C4				0	**0,26**
C7					0

Ejemplo

Centroides de los clusters iniciales

Documento	Tamaño (KB)	Nro. de Accesos
C1	1,08	2,45
C2	2,18	1,45
C365	0,75	2,42
C47	1,47	1,98

Matriz de distancia intercluster

Cluster	C1	C2	C365	C47
C1	0	1,49	**0,33**	0,61
C2		0	1,73	0,89
C365			0	0,84
C47				0

Ejemplo

Centroides de los clusters finales

Documento	Tamaño (KB)	Nro. de Accesos
C1356	0,91	2,43
C2	2,18	1,45
C47	1,47	1,98

Matriz de distancia intercluster

Cluster	C1356	C2	C47
C1356	0	1,60	0,72
C2		0	0,89
C365			0

Ejemplo

Salida del proceso de Clustering / Agrupamiento

Tipo	Clase	Tamaño	Nro. Acc.	Nro. Comp.
Pequeño	C1356	8,19	271,51	4
Medio	C47	29,58	96,05	2
Grande	C2	150	28	1