CEU
Universidad
San Pablo

Dimensionality reduction

Alberto Pascual-Montano
Complutense University of Madrid
http://www.dacya.ucm.es/apascual

Madrid, July 2007

Summary of the course

- **Introduction:**
  - Why dimensionality reduction
  - Curse of dimensionality
  - Feature selection vs. feature extraction
  - Linear vs. no linear
  - Accuracy vs. Interpretation
- **Matrix factorization methods**
  - Principal Component Analysis
  - Singular Value Decomposition
  - Factor analysis
  - Non-negative matrix factorization
  - Independent Component Analysis
- **Projection methods**
  - Multidimensional scaling
  - Sammon mapping
  - Self-organizing maps
  - Other clustering techniques
  - Isomap
  - Locally linear embedding (LLE)
- **Applications**
  - Pattern recognition
  - Image classification
  - Gene expression analysis
  - Text mining
- **Practical exercises**
  - Image classification
  - Gene expression analysis
  - Scientific text analysis

Practical guide for this course

MATLAB: http://www.mathworks.com/

TOOLBOXES:

- Statistics
- Bioinformatics
- Neural Networks
- Specific code for this course (available in the web page)

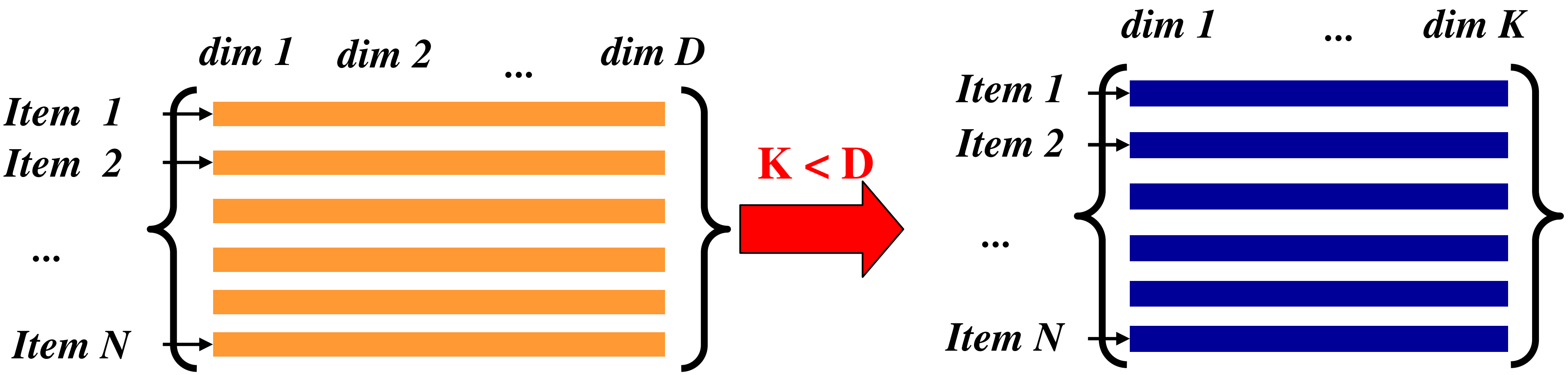
Course web page:

MATLAB: http://www.dacya.ucm.es/apascual/dimred2007.html

Dimensionality Reduction: main motivation

- More features implies more information and potentially higher accuracy
- Important paradox: the more features we have, the more difficult information extraction is.
- Unfortunately, more features means harder to train a classifier:
  - The curse of dimensionality
- Solution: start with as many potentially useful features as possible, and then reduce the number of features

Dimensionality reduction: what is?



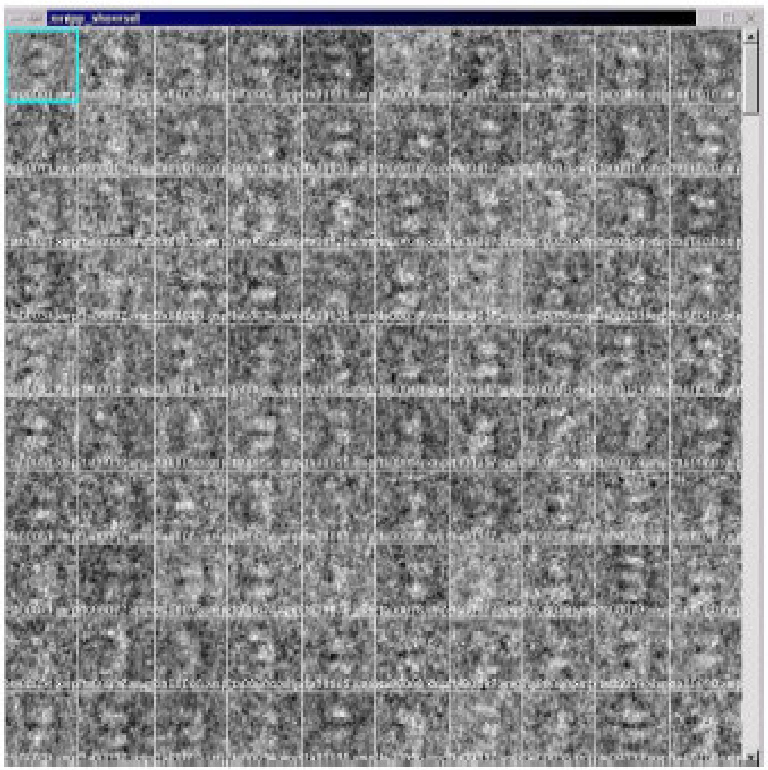
reduce dimensionality of data (number of columns).

Why Dimensionality Reduction

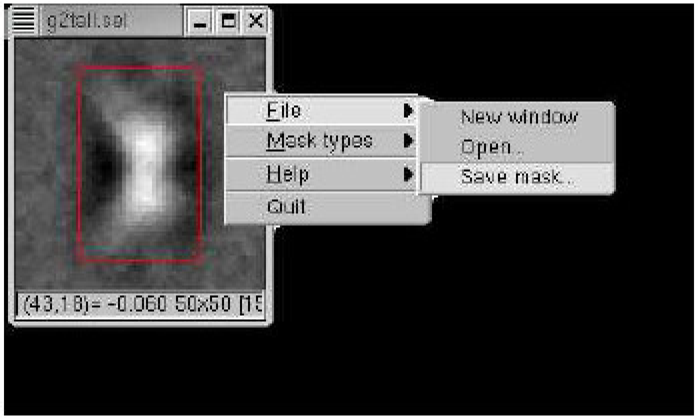
- Number of potential features can be huge
  - Image data: each pixel of an image
    - A 64x64 image = 4096 features
  - Genomic data: expression levels of the genes
    - Several thousand features
  - Text categorization: frequencies of terms in a corpus of documents:
    - More than ten thousand features

Why Dimensionality Reduction: real case scenarios

Electron microscopy images:



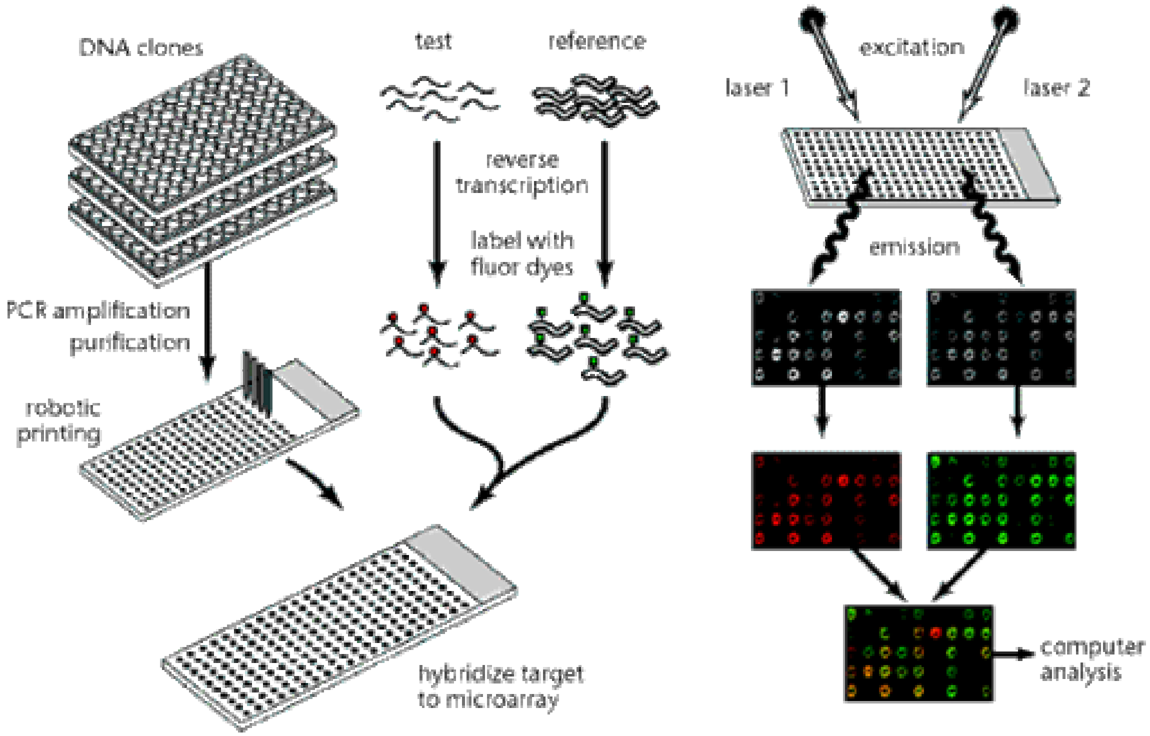
8000 64x64 images = 8000x4096 data matrix

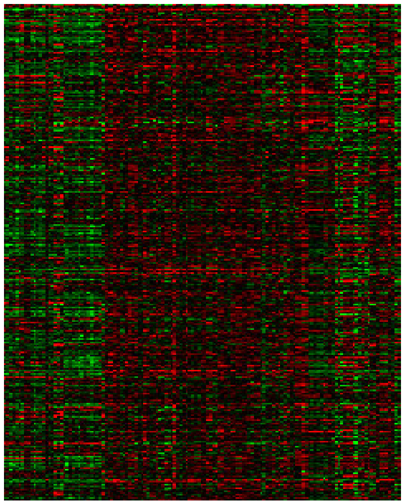


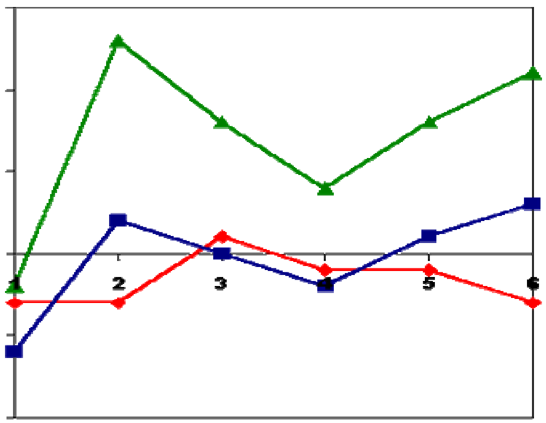
Some feature selection is usually carried out (e.g. a mask)
New reduced matrix: 8000x1000 data matrix

Why Dimensionality Reduction: real case scenarios

Gene expression data:







Why Dimensionality Reduction: real case scenarios

Text analysis:

ID	Free-form text	ID	Free-form text
0	Other	58	General System
2	General System	59	Connecting To & Using the Internet
7	General System	60	Operating Systems
8	Software Applications	61	Operating Systems
21	Games, Sound & Video	62	Connecting To & Using the Internet
22	General System	63	Hard Disk & Other Storage Devices
23	Operating Systems	64	Software Applications
24	Home Networking	66	Games, Sound & Video
25	Connecting To & Using the Internet	67	Keyboard, Mouse & Other Devices
26	Connecting To & Using the Internet	68	Software Applications
27	Printing, Scanning & Photos	70	Connecting To & Using the Internet
33	Operating Systems	71	General System
34	Operating Systems	72	Keyboard, Mouse & Other Devices
35	General System	73	General System
36	Operating Systems	74	Keyboard, Mouse & Other Devices
44	Hard Disk & Other Storage Devices	75	Operating Systems
53	Connecting To & Using the Internet	76	General System
54	Home Networking	77	Operating Systems
55	Connecting To & Using the Internet	78	Connecting To & Using the Internet
57	General System	79	General System

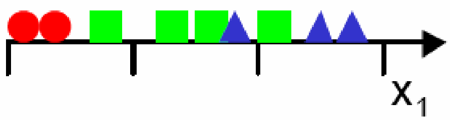
Why Dimensionality Reduction: real case scenarios

Text analysis:

WORDS	applications	connecting	devices	disk	games	general	hard	home	internet	keyboard	mouse	networking	operating	photos	Printing	scanning	software	sound	storage	Systems	using	video
DOCUMENTS																						
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Software Applications	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Games, Sound & Video	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Operating Systems	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
Home Networking	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Connecting To & Using the Internet	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Connecting To & Using the Internet	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Printing, Scanning & Photos	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
Operating Systems	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
Operating Systems	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Operating Systems	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
Hard Disk & Other Storage Devices	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Connecting To & Using the Internet	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Home Networking	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Connecting To & Using the Internet	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
General System	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Curse of dimensionality

- **The curse of dimensionality**
  - A term coined by Bellman in 1961
  - Refers to the problems associated with multivariate data analysis as the dimensionality increases
  - We will illustrate these problems with a simple example
- **Consider a 3-class pattern recognition problem**
  - A simple approach would be to
    - Divide the feature space into uniform bins
    - Compute the ratio of examples for each class at each bin and,
    - For a new example, find its bin and choose the predominant class in that bin
  - In our toy problem we decide to start with one single feature and divide the real line into 3 segments



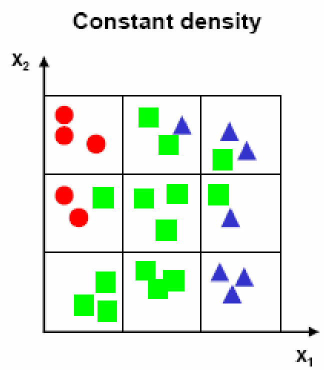
  - After doing this, we notice that there exists too much overlap among the classes, so we decide to incorporate a second feature to try and improve separability

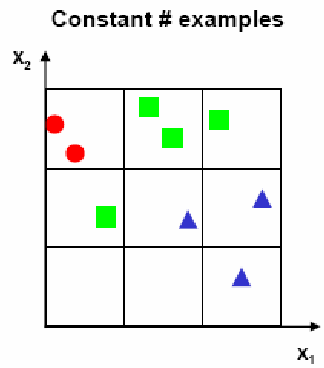
CEU Universidad San Pablo

Taken from Ricardo Gutierrez-Osuna. Texas A&M University

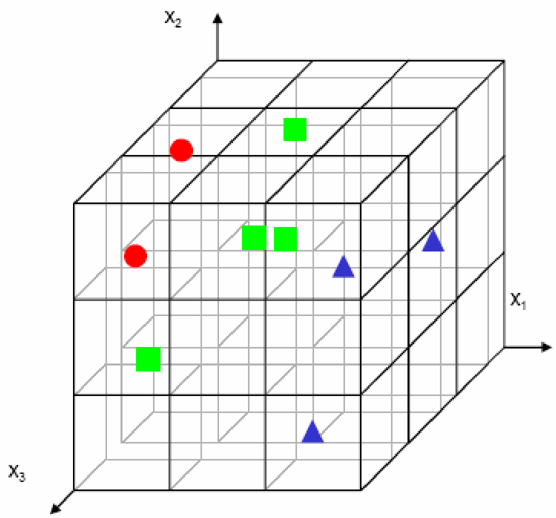
Curse of dimensionality

- **We decide to preserve the granularity of each axis, which raises the number of bins from 3 (in 1D) to 3^2=9 (in 2D)**
  - At this point we need to make a decision: do we maintain the density of examples per bin or do we keep the number of examples had for the one-dimensional case?
    - Choosing to maintain the density increases the number of examples from 9 (in 1D) to 27 (in 2D)
    - Choosing to maintain the number of examples results in a 2D scatter plot that is very sparse





- **Moving to three features makes the problem worse:**
  - The number of bins grows to 3^3=27
  - For the same density of examples the number of needed examples becomes 81
  - For the same number of examples, well, the 3D scatter plot is almost empty



Curse of dimensionality practical problems:

- the number of samples required per variable increases exponentially with the number of variables
- The rapid increase in volume associated with adding extra dimensions.
- The more dimensions you have, the more similar things appear.

Curse of dimensionality

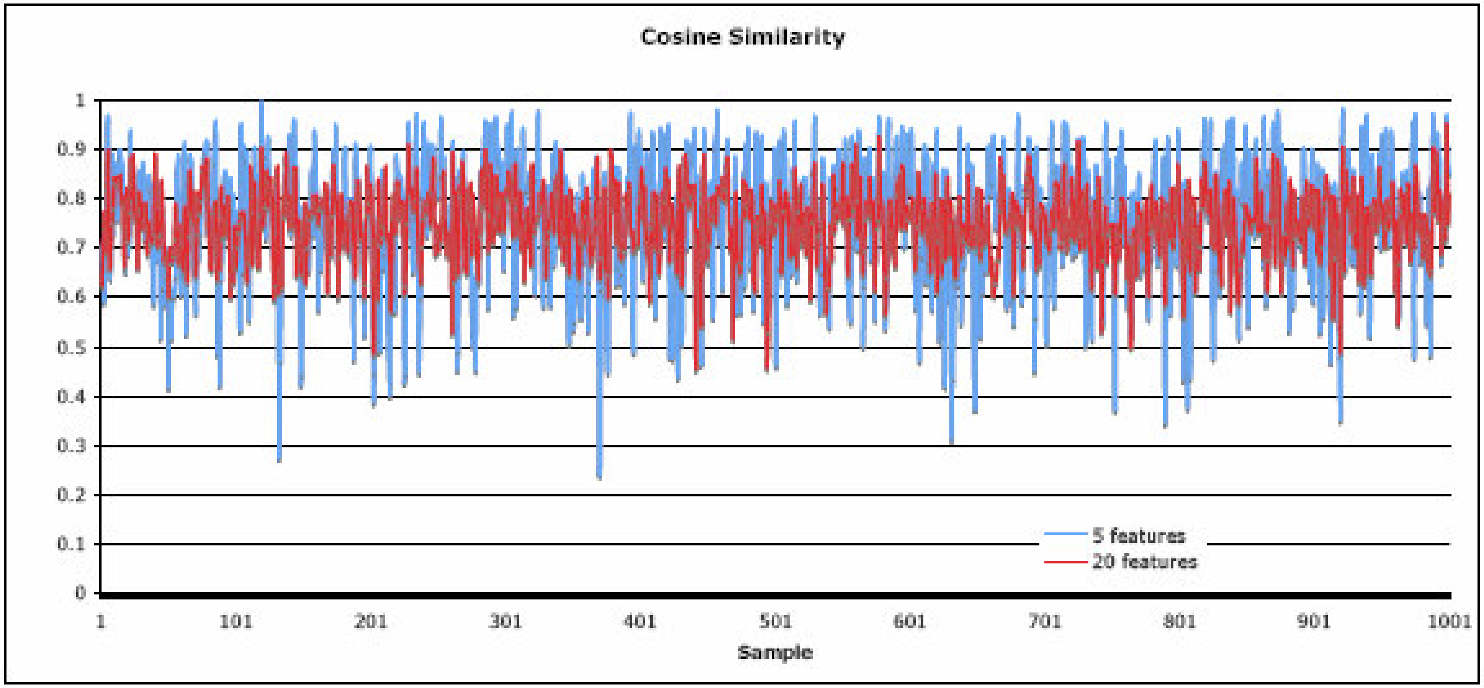
Silverman provides a table illustrating the difficulty of kernel estimation in high dimensions. To estimate the density at 0 with a given accuracy, he reports:

Dimensionality	Required Sample Size
1	4
2	19
5	786
7	10,700
10	842,000

Silverman, *Density Estimation for Statistics and Data Analysis*, 1986, Chapman & Hall.

Curse of dimensionality practical problems:

The more dimensions you have, the more **similar** objects appear:



Curse of dimensionality:

It is NOT a problem of computing effectiveness and hardware requirements!

Feature selection versus Feature extraction

- **Feature selection:** Try to find a subset of the original variables (also called features or attributes). Two strategies are filter (e.g. information gain) and wrapper (e.g. genetic algorithm) approaches.
- **Feature extraction:** A new reduced set of variables is created by applying a mapping of the multidimensional space into a space of fewer dimensions. This means that the original feature space is transformed into a reduced, although informative new space.

Feature selection:

Customer age	Gender	Age group	Number of purchases	Business location	Groceries	Garden	Furniture	Electronics	Toys
68	Female	60	60	New York	5394	5429	5865	4860	18918
41	Male	40	9	Boston	1419	1431	1362	885	0
56	Female	50	12	New York	4286	467	524	216	304
77	Female	70	5	New York	684	0	238	0	0
61	Female	60	42	Los Angeles	5165	6999	3488	10013	11266
45	Female	40	59	Seattle	4449	7156	5774	6396	185
62	Male	60	1	Los Angeles	0	0	0	153	0
44	Female	40	22	Los Angeles	3532	2373	825	1139	0
52	Female	50	20	Boston	649	1582	584	1033	185
18	Female	10	14	Seattle	5061	0	417	0	0
74	Female	70	3	Los Angeles	122	467	0	436	0
55	Female	50	20	Boston	1369	731	1369	5586	354
75	Female	70	20	New York	1478	1626	379	474	298
44	Female	40	88	Los Angeles	1431	464	91	492	0
66	Female	60	33	New York	2223	5535	4377	2593	2216
56	Female	50	41	Boston	1164	4154	219	3846	662
42	Male	40	6	Los Angeles	521	122	282	555	0
58	Female	50	5	Boston	0	0	609	797	756
66	Male	60	30	Seattle	4034	1186	4452	4688	1092
59	Female	50	1	Seattle	0	0	0	527	216
65	Female	60	10	Boston	0	747	0	609	1934
42	Male	40	90	Los Angeles	3388	1434	1761	587	0
42	Female	40	10	Seattle	832	248	1265	379	0
48	Female	40	15	Los Angeles	1651	4606	681	0	0
37	Female	30	34	New York	1855	408	7410	712	0
51	Male	50	24	New York	4980	141	285	916	0
60	Male	60	8	Los Angeles	0	467	2373	81	1004
59	Female	50	45	Seattle	1488	1921	1667	8229	6361
61	Female	60	20	Los Angeles	709	4085	612	5887	2113
27	Female	20	20	Los Angeles	367	1186	3278	50	2150

Feature extraction:

Transformation: e.g. PCA

Customer age	Gender	Age group	Number of purchases	Business location	Groceries	Garden	Furniture	Electronics	Toys
68	Female	60	60	New York	5394	5429	5865	4860	18918
41	Male	40	9	Boston	1419	1431	1362	885	0
56	Female	50	12	New York	4286	467	524	216	304
77	Female	70	5	New York	684	0	238	0	0
61	Female	60	42	Los Angeles	5165	6999	3488	10013	11266
45	Female	40	59	Seattle	4449	7156	5774	6396	185
62	Male	60	1	Los Angeles	0	0	0	153	0
44	Female	40	22	Los Angeles	3532	2373	825	1139	0
52	Female	50	20	Boston	649	1582	584	1033	185
18	Female	10	14	Seattle	5061	0	417	0	0
74	Female	70	3	Los Angeles	122	467	0	436	0
55	Female	50	20	Boston	1369	731	1369	5586	354
75	Female	70	20	New York	1478	1626	379	474	298
44	Female	40	88	Los Angeles	1431	464	91	492	0
66	Female	60	33	New York	2223	5535	4377	2593	2216
56	Female	50	41	Boston	1164	4154	219	3846	662
42	Male	40	6	Los Angeles	521	122	282	555	0
58	Female	50	5	Boston	0	0	609	797	756
66	Male	60	30	Seattle	4034	1186	4452	4688	1092
59	Female	50	1	Seattle	0	0	0	527	216
65	Female	60	10	Boston	0	747	0	609	1934
42	Male	40	90	Los Angeles	3388	1434	1761	587	0
42	Female	40	10	Seattle	832	248	1265	379	0
48	Female	40	15	Los Angeles	1651	4606	681	0	0
37	Female	30	34	New York	1855	408	7410	712	0
51	Male	50	24	New York	4980	141	285	916	0
60	Male	60	8	Los Angeles	0	467	2373	81	1004
59	Female	50	45	Seattle	1488	1921	1667	8229	6361
61	Female	60	20	Los Angeles	709	4085	612	5887	2113
27	Female	20	20	Los Angeles	367	1186	3278	50	2150

Feature extraction:

New variables (features)

Customer age	Gender	Age group	Number of purchases	PCA 1	PCA 2	PCA 3
68	Female	60	60	-11689.13061317	5340.774666753	-3483.138211999
41	Male	40	9	286.4689144358	-339.5693466753	-180.1779096908
56	Female	50	12	507.7173985103	-382.5567038016	229.6156090041
77	Female	70	5	2358.421790129	-253.9924428911	79.67897661748
61	Female	60	42	-11534.88283297	8257.820518043	-2739.450022123
45	Female	40	59	-8862.715962248	959.6640749912	-688.9539078341
62	Male	60	1	2681.406315515	28.90683172963	-58.74969730759
44	Female	40	22	-730.3070502793	-57.83224928944	-880.2865324676
52	Female	50	20	855.5816806707	271.4883592756	-741.9321704666
18	Female	10	14	698.7353441909	-578.7164357212	629.345636479
74	Female	70	3	2274.603583329	198.3003881873	-330.864742706
55	Female	50	20	-1417.734626017	3522.910454314	1731.264314998
75	Female	70	20	879.4957400298	-72.42969068754	-978.6723633218
44	Female	40	88	1732.915428644	125.4499401156	-129.391229542
66	Female	60	33	-5183.699904375	-404.8180005482	-1869.454806482
56	Female	50	41	-1742.185402327	2532.367743932	-2145.795166843
42	Male	40	6	2088.838664385	154.5917778364	154.941852385
58	Female	50	5	1871.75827704	431.9663705002	269.2745291048
66	Male	60	30	-4264.273729785	1205.58595902	2682.573231864
59	Female	50	1	2475.79698357	386.0405919391	10.68622502717
65	Female	60	10	1686.503568202	881.8940711684	-910.4968557223
42	Male	40	90	-530.013445699	-889.7763254622	124.2905005394
42	Female	40	10	1384.93369774	-536.8298520329	497.9854485601
48	Female	40	15	-650.9268310247	-998.1566258202	-3269.252316994
37	Female	30	34	-2974.4810786	-3623.584773952	3479.268857914
51	Male	50	24	344.2079109456	200.9508895213	725.0025872811
60	Male	60	8	798.9073851337	-1044.588044058	461.5159831137
59	Female	50	45	-4668.774003788	7082.070218074	510.2423383574
61	Female	60	20	-2966.461022976	4382.439296706	-1630.872995896
27	Female	20	20	-491.2353453043	-1305.084183826	111.4339578214
61	Male	60	22	-1063.473411997	1764.131519608	-667.4164706311
46	Male	40	8	1488.354725366	-433.3919166003	-474.3377606303