UCL

The aggregation of variables in distance design -

How to get more out of distance-based methods

Christian Hennig

August 30, 2011

Using distances for high-dimensional data

Distance-based methods:

- k-nearest neighbours,
- most hierarchical clustering,
- "partitioning around medoids",
- multidimensional scaling.

Consider classification problems, $i = 1, \dots, n$,

$$\mathbf{x}_i \in \mathbb{R}^p,\ Y_i \in \{1, \dots, s\},\ d: \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}_0^+.$$

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p
- *Definition of distance measure allows some flexibility that is often not exploited*

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p
- *Definition of distance measure allows some flexibility that is often not exploited*

What's wrong with distance-based methods?

- "Curse of dimensionality"
  - isn't every point far from every other point?

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p
- *Definition of distance measure allows some flexibility that is often not exploited*

What's wrong with distance-based methods?

- "Curse of dimensionality"
  - isn't every point far from every other point?
- Model-based theory tedious

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p
- *Definition of distance measure allows some flexibility that is often not exploited*

What's wrong with distance-based methods?

- "Curse of dimensionality"
  - isn't every point far from every other point?
- Model-based theory tedious
- Ignore distributional shapes

What makes distance-based methods attractive?

- No variable selection necessary $\Rightarrow$ no loss of information
- Supposedly model-free
- Direct interpretation of distance measure
- No computational problem for large p
- *Definition of distance measure allows some flexibility that is often not exploited*

What's wrong with distance-based methods?

- "Curse of dimensionality"
  - isn't every point far from every other point?
- Model-based theory tedious
- Ignore distributional shapes
- Computationally bad for large n

Distances vs. dimension reduction

Core assumption for dimension reduction is that

1. *relevant information* is of much lower dimensionality than the data,
2. it is possible to separate *relevant* from *irrelevant* information.

Distances vs. dimension reduction

Core assumption for dimension reduction is that

1. *relevant information* is of much lower dimensionality than the data,
2. it is possible to separate *relevant* from *irrelevant* information.

PCA (and the like) identify variance (or robust variance) with *relevant information*.

Variable selection methods assume that some variables are relevant and most are not.

Distances vs. dimension reduction

Core assumption for dimension reduction is that

1. *relevant information* is of much lower dimensionality than the data,
2. it is possible to separate *relevant* from *irrelevant* information.

PCA (and the like) identify variance (or robust variance) with *relevant information*.

Variable selection methods assume that some variables are relevant and most are not.

Both approaches tend to identify *statistical redundance* with *irrelevance*.

Clustering vs. supervised classification

Major difference between clustering and supervised classification for distance design:

- In supervised classification the aim is to keep the misclassification rate down.
  Distances are a tool to achieve this.

Clustering vs. supervised classification

Major difference between clustering and supervised classification for distance design:

- In supervised classification the aim is to keep the misclassification rate down.
  Distances are a tool to achieve this.
- In clustering, defining distances is part of *defining the clustering problem*.
  Distances are not only a tool to find a "good" clustering, but also part of the quality assessment.

Clustering vs. supervised classification

Major difference between clustering and supervised classification for distance design:

- In supervised classification the aim is to keep the misclassification rate down.
  Distances are a tool to achieve this.
- In clustering, defining distances is part of *defining the clustering problem*.
  Distances are not only a tool to find a “good” clustering, but also part of the quality assessment.

In supervised classification it can be assessed whether a certain distance “does a good job”.

In clustering, need distance to define what good job is.

Aspects of distance design

- Variable transformation
- Variable standardisation
- Variable aggregation

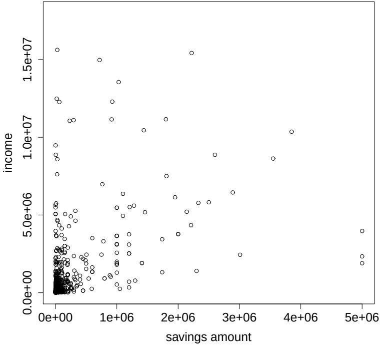
Clustering with mixed type data: social stratification

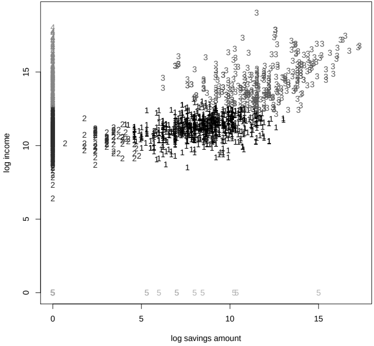
Data from US Survey of Consumer Finances 2007,
provided by Tim Liao (University of Illinois).

“Continuous” variables: save.amount, income.
Ordinal categorical variables: check.account, save.account.
Nominal variable: housing.
Binary variables: life.insurance, add.assets.

Transformation

Rationale: model “interpretative distance”





Problem: how to make (mixed type) variables comparable?

- Replace nominal variables by dummies.

Problem: how to make (mixed type) variables comparable?

- Replace nominal variables by dummies.
- Use scores for ordinal variables.
  - Decide “interpretative distance”
  - Standard (Likert) scores
  - Data-dependent scores, e.g., mean ranks
    (makes distances between dense categories larger)

Standardisation

Possible standardisation methods:

- Range
- Standard deviation
- MAD/IQR

Standardisation

Possible standardisation methods:

- Range
- Standard deviation
- MAD/IQR

MAD/IQR is bad for dummies.

No problem here with standard deviation (robustness discussed later).

Dummy variables

Assuming Euclidean aggregation, for I categories:

$$\sum_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 \stackrel{!}{=} qE(X_1 - X_2)^2$$

Assume $P_Y\{c_i\} = \frac{1}{I}$ (could estimate this).

Dummy variables

Assuming Euclidean aggregation, for I categories:

$$\sum_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 \stackrel{!}{=} qE(X_1 - X_2)^2$$

Assume $P_Y\{c_i\} = \frac{1}{I}$ (could estimate this).

Dummy variables

Assuming Euclidean aggregation, for I categories:

$$\sum_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 \stackrel{!}{=} qE(X_1 - X_2)^2$$

Assume $P_Y\{c_i\} = \frac{1}{I}$ (could estimate this).

Need $q < 1$ to prevent gaps from dominating the clustering.
(This depends on clustering method.)

Ordinal variables

$$E(Y_1 - Y_2)^2 \stackrel{!}{=} qE(X_1 - X_2)^2, \; q = \frac{1}{1 + 1/(I-1)}.$$

Ordinal variables

$$E(Y_1 - Y_2)^2 \stackrel{!}{=} qE(X_1 - X_2)^2, \; q = \frac{1}{1 + 1/(I-1)}.$$

May weight variables according to importance.
Weight “account number” variables by $\frac{1}{2}$.

Ordinal variables

$$E(Y_1 - Y_2)^2 \stackrel{!}{=} qE(X_1 - X_2)^2,\ q = \frac{1}{1 + 1/(I-1)}.$$

May weight variables according to importance.
Weight “account number” variables by $\frac{1}{2}$.

Double weight of housing dummies “rented”, “owns” which locates the other ones “in between”.

Aggregation

- Manhattan (L1) $\sum_{l=1}^{p} d_l(x_{il}, x_{jl})$
- Euclidean (L2) $\sqrt{\sum_{l=1}^{p} d_l(x_{il}, x_{jl})^2}$
- Minkowski (Lr) $\left(\sum_{l=1}^{p} d_l(x_{il}, x_{jl})^r\right)^{\frac{1}{r}}$
- Mahalanobis $(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$

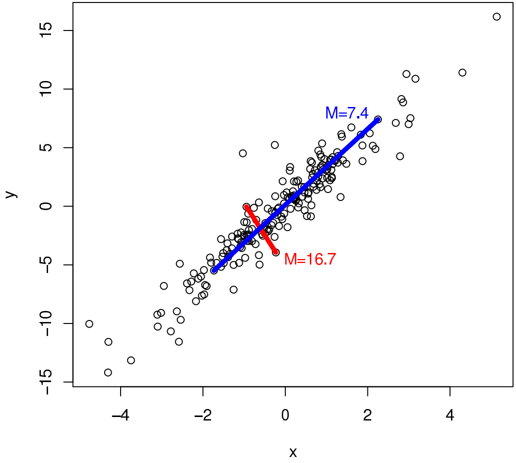
Aggregation

- Manhattan (L1) $\sum_{l=1}^{p} d_l(x_{il}, x_{jl})$
- Euclidean (L2) $\sqrt{\sum_{l=1}^{p} d_l(x_{il}, x_{jl})^2}$
- Minkowski (Lr) $(\sum_{l=1}^{p} d_l(x_{il}, x_{jl})^r)^{\frac{1}{r}}$
- Mahalanobis $(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$

Determines weight of variable-wise distance in aggregation. Higher r Minkowski means that a single large distance dominates overall distance.

Mahalanobis distance

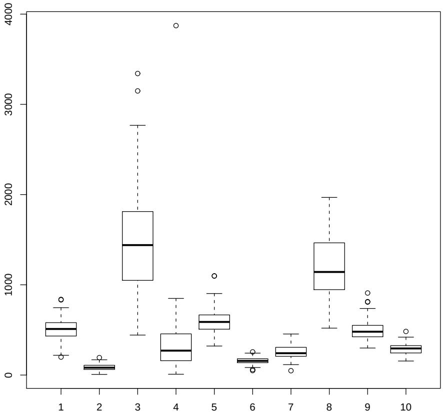
quantifies deviation from general tendency.



Classification: microarrays

79 prostate cancer patients, 39 having disease recurrence, expressions on 22,283 genes (Sun and Goodison 2009).

Try k nearest neighbours with L2-aggregation.



Skew, very different variances, occasional outliers.

LOO-CV: using variables as they are is better than

- doing sd/range/MAD standardisation,
- log-transformation.

Variable variances *are* informative.
Outliers are not.

LOO-CV: using variables as they are is better than

- doing sd/range/MAD standardisation,
- log-transformation.

Variable variances *are* informative.
Outliers are not.

Range standardisation annihilates *variables* with outliers,
MAD-standardisation contaminates *observations* with outliers.

“Boxplot-standardisation”

…keeps distances in centre informative,
but tames outliers.

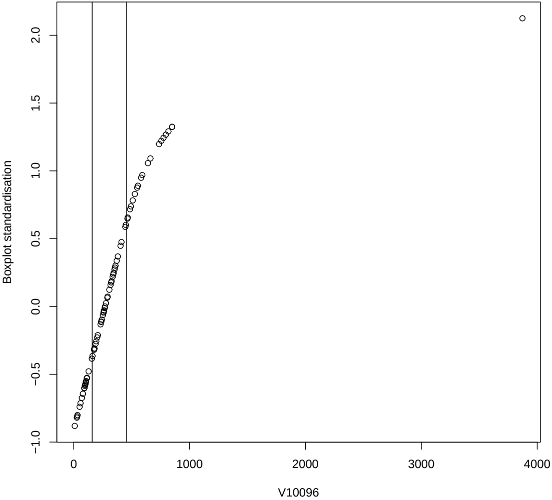
- Compute min, max, all quartiles.
- Center data at median, divide by IQR.
- If all points are now $\in [q_1 - 1.5\mathrm{IQR}, q_3 + 1.5\mathrm{IQR}]$, that’s it.
- Otherwise transform $[q_3, \max]$ to $[q_3, q_3 + 1.5\mathrm{IQR}]$ by $q_3 - \frac{1}{k((x-q_3)+1)^k} + \frac{1}{k}$ with suitable k, and analogously below q_1.

“Boxplot-standardisation”

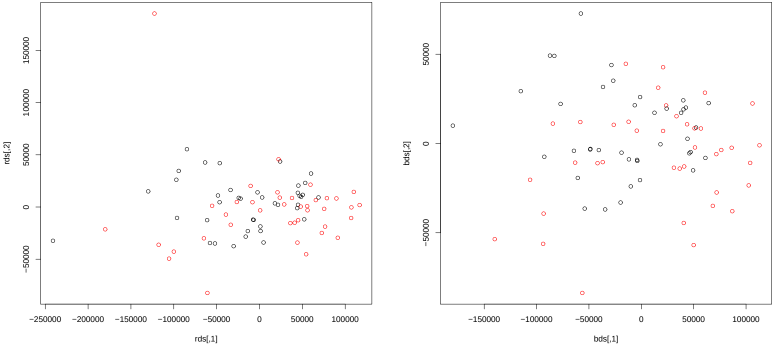
. . . keeps distances in centre informative,
but tames outliers.

- Compute min, max, all quartiles.
- Center data at median, divide by IQR.
- If all points are now $\in [q_1 - 1.5\mathrm{IQR}, q_3 + 1.5\mathrm{IQR}]$, that’s it.
- Otherwise transform $[q_3, \max]$ to $[q_3, q_3 + 1.5\mathrm{IQR}]$ by $q_3 - \frac{1}{k((x-q_3)+1)^k} + \frac{1}{k}$ with suitable k, and analogously below q_1.

May use this as IQR-keeping transformation by multiplying by IQR.


Boxplot standardisation
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
V10096

Using this with 3-nearest neighbour, L2 gets 53/79 right.



Simulation: standardisation and aggregation

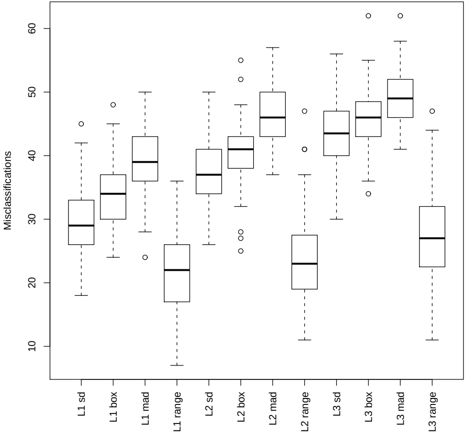
$n = 100,\ p = 500,\ n_1 = 50,\ n_2 = 50.$
Variable 1-5: class 1 t_3, class 2 t_3 centered at 8.
Variable 6-500: t_3.

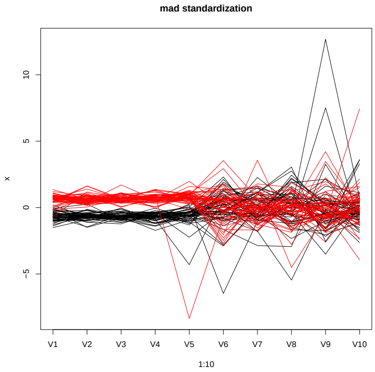
Standardisation: sd, boxplot, mad, range.
Aggregation: L1, L2, L3, L4 (not shown).

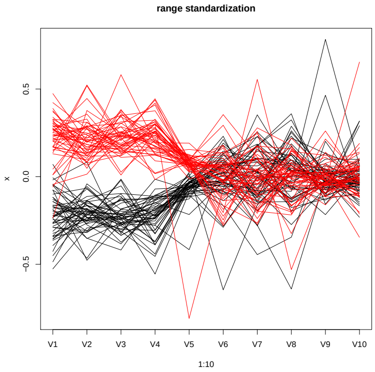
Classify by 1-nn.

Similar results for 3 classes, unequal sizes, normal distribution, clustering.

Being "robust" is apparently bad, but why?




mad standardization
x
10
5
0
−5
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
1:10


range standardization
x
0.5
0.0
−0.5
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
1:10

Conclusion

Distance design gives you flexibility.

Conclusion

Distance design gives you flexibility.
It depends on the situation what is best.
In clustering, the distances defines what is good.

Conclusion

Distance design gives you flexibility.
It depends on the situation what is best.
In clustering, the distances defines what is good.
You learn something from it -
is relevant information lost by
standardisation or Mahalanobis?

Conclusion

Distance design gives you flexibility.
It depends on the situation what is best.
In clustering, the distances defines what is good.
You learn something from it -
is relevant information lost by
standardisation or Mahalanobis?
“Robust” standardisation is not always a good idea.

To do: explore standardisation and aggregation theoretically.
Criteria to enable different within-class variation.

This presentation is supported by

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning