A comparison between Hudap and Correspondence Analysis

Fabrice Buschini

LPS – EHESS (Paris)

The data file:

- Results of a content analysis on 407 articles on SR.
- Meta-analysis conducted by Professor Annamaria de Rosa’s team.
- 30 variables or categories.

The variables:

- The first six (V1 to V6) can be considered as descriptive variables
  - They are related to the form of the articles (language, author's country, publication year, type of publication, etc.)
- The last twenty-four (V7 to V30) are the main variables which can be called active variables
  - They are concerned with the content of the articles (methodology employed, process described, etc.).

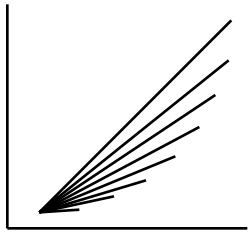
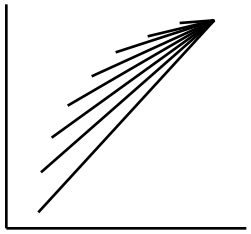
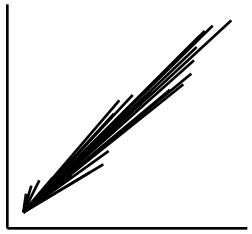
The Hudap's principles
(WSSA procedure)

- The WSSA belongs to the family of MDS (multidimensional scaling)
  - MDS tries to represent in a small space (2 or 3 dimensions) the distances (or proximities) existing between variables.
- In Hudap, the distance measure in an index of proximity : the monotonicity coefficient of Guttman.

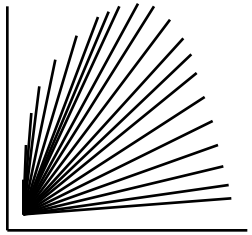
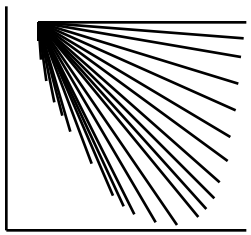
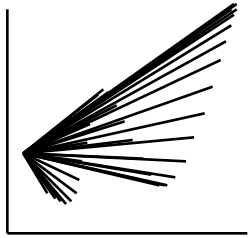
Monotonicity coefficient

- Can be compared to a correlation coefficient, but not necessarily a linear one.
- Measures the way two variables vary broadly in the same sense.
- Then two variables can be considered close in as much as they vary in the same sense.

Examples of monotonous relations between two variables

Examples of non monotonous relations between two variables

Note !

- The Wssa in Hudap can be used only for thoses variables for which the Guttman’s coefficient is meaningful.

Four levels of measurement

- Ratio level: continuous measure with a zero point. It conserves order, deviation, and is proportional (e.g. metric system)
- Interval level: continuous measure with or without zero point. It conserves order and deviation, but it is not proportional (e.g. temperature scale)
- Ordinal level: discontinuous measure. Conserves order, but nothing can be said on deviation (e.g. social classes)
- Nominal level: discontinuous measure. Nothing can be said on the relations between values (e.g. gender, language)

Back to the data file

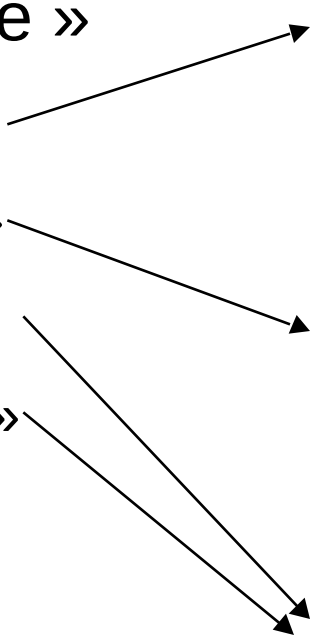
- Most of the variables are nominal ones, some are ordinal
- With this kind of variables, to calculate the monotonicity coefficient is raher meaningless
- If one finds a high positive coefficient between the variables LANGUAGE (1=italian, 2=english, 3=french, 4=spanish) and PROCESS (1=no, 2=anchorage, 3=objectivation, 4=both), it then means thoses two variables are varying in the same sense. But does it really mean anything, especially when one knows that the order of categories was arbitrarily chosen?

Solution

- Transformation of variables with a disjunctive coding
  - The principle is to create for each variable as many new variables as modalities existing for the former one.

Example: variable « language »

- V4 « language »
  - 1 « italian »
  - 2 « english »
  - 3 « french »
  - 4 « spanish »



- V4_1 « italian_lg »
  - 0 « no »
  - 1 « yes »
- V4_2 « english_lg »
  - 0 « no »
  - 1 « yes »
- Etc.

Advantages of disjunctive coding

- Wssa can be run because the monotonicity coefficient makes sense here
- Correspondence analysis can also be conducted on the data
- Therefore, a comparison can be made between the two methods on the same data

Differences between Wssa and Anacor

- Wssa
  - The distance index is the monotonicity coefficient
  - Interpretations are made on proximities and spatiality
- Anacor
  - The distance index is the khi square distance
  - Interpretations are made on factors

Preparing the common data file

- After re-coding, the 30 original variables gave birth to 236 new variables (58 for the descriptive ones and 178 for the others)
- Of those new variables, 16 have a null variance and then should be deleted
  - They correspond to 16 uused modalities in the 30 original variables
- In order to reduce the number of variables and to make the data more homogeneous, the new variables with a frequency lower than 10 (2.5%) were coupled with other close variables
- This procedure is equivalent to come back on the content analysis in order to reduce the number of categories

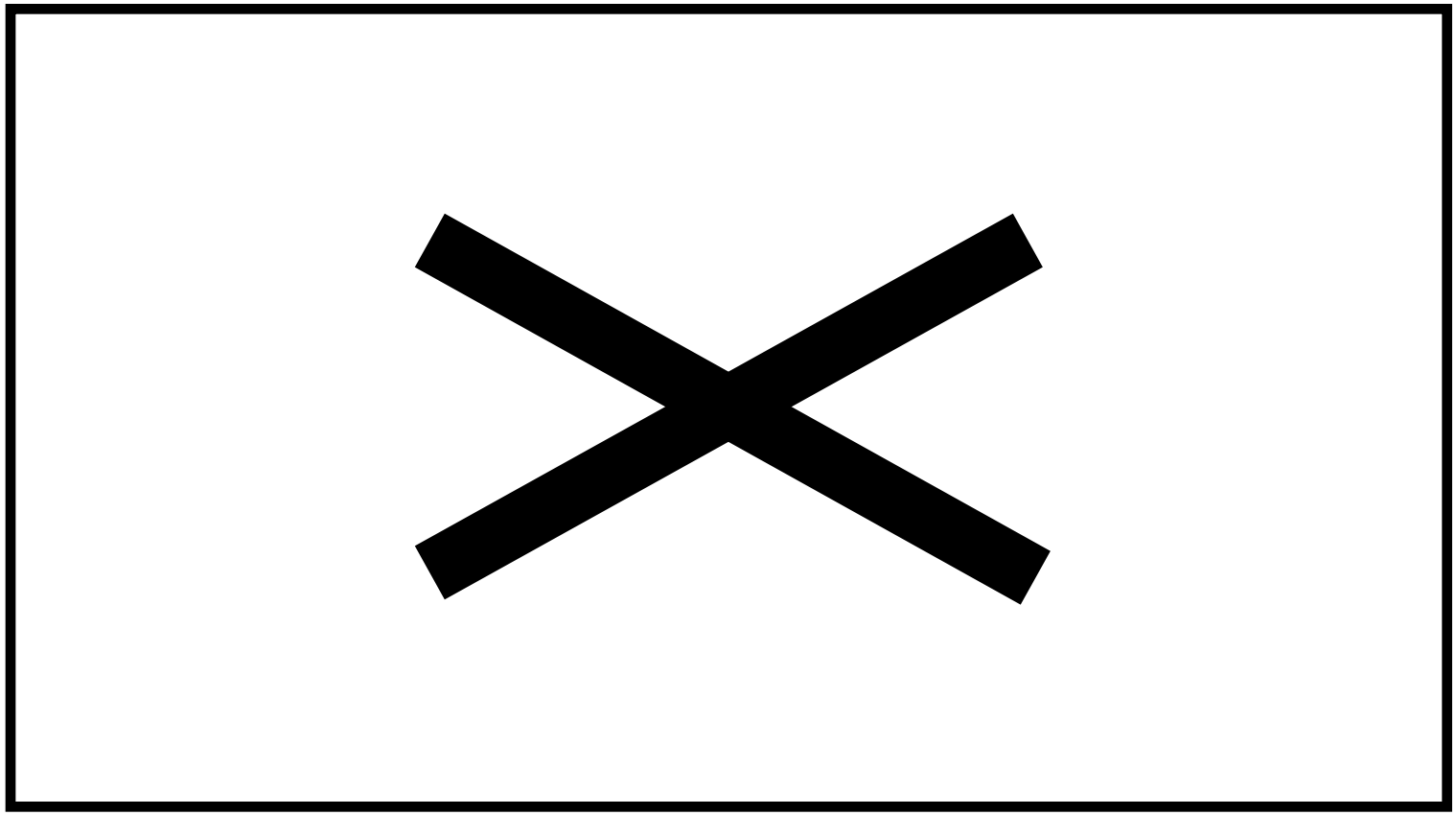
The final data file used for both analyses

- After erasing the problematic variables, 116 remain
  - 27 for the descriptive variables
  - 89 for the active variables
- Some variables could remain problematic
  - One with a frequency lower than 10 (V18_3)
  - Twelve coming from 6 original variables with too unequal categories (>94 % and <6 %)
    - V12_1,2 (383/24) ; V24_1,7 (389/18) ; V25_1,7 (396/11) ; V27_1,5 (388/19) ; V28_1,5 (396/11) ; V29_1,9 (385/22)

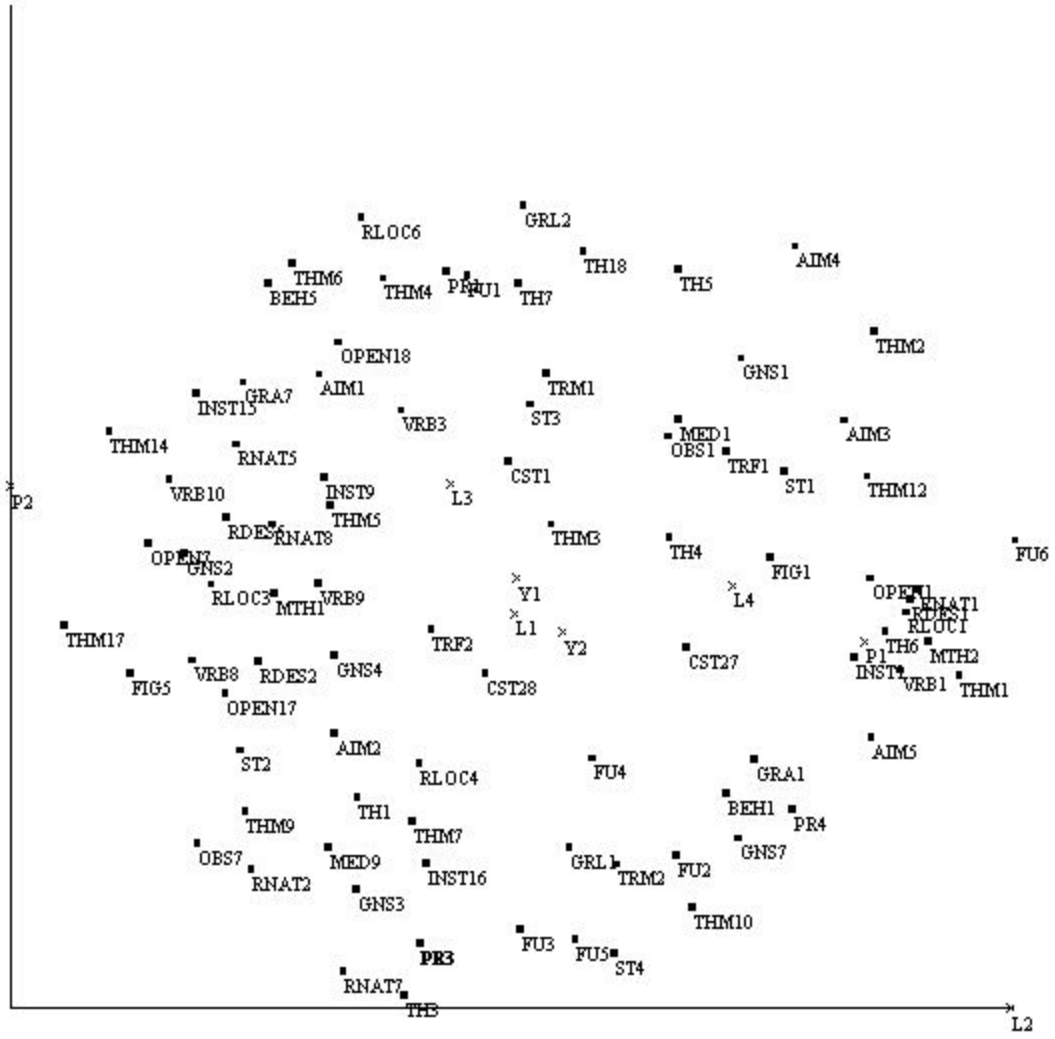
Five analyses were conducted on different numbers of variables

- On all the 89 variables
- After deleting one variable for each of the nine dichotomous ones : 80 variables
- The former minus all the variables coming from the original GNS, MTH, CST and THM : 59 variables
- The former minus all the variables coming from the original GRL, GRA, OBS, VRB, FIG, BEH and MED : 48 variables
- Only on the 16 variables coming from the original ST, AIM, RDES and RLOC

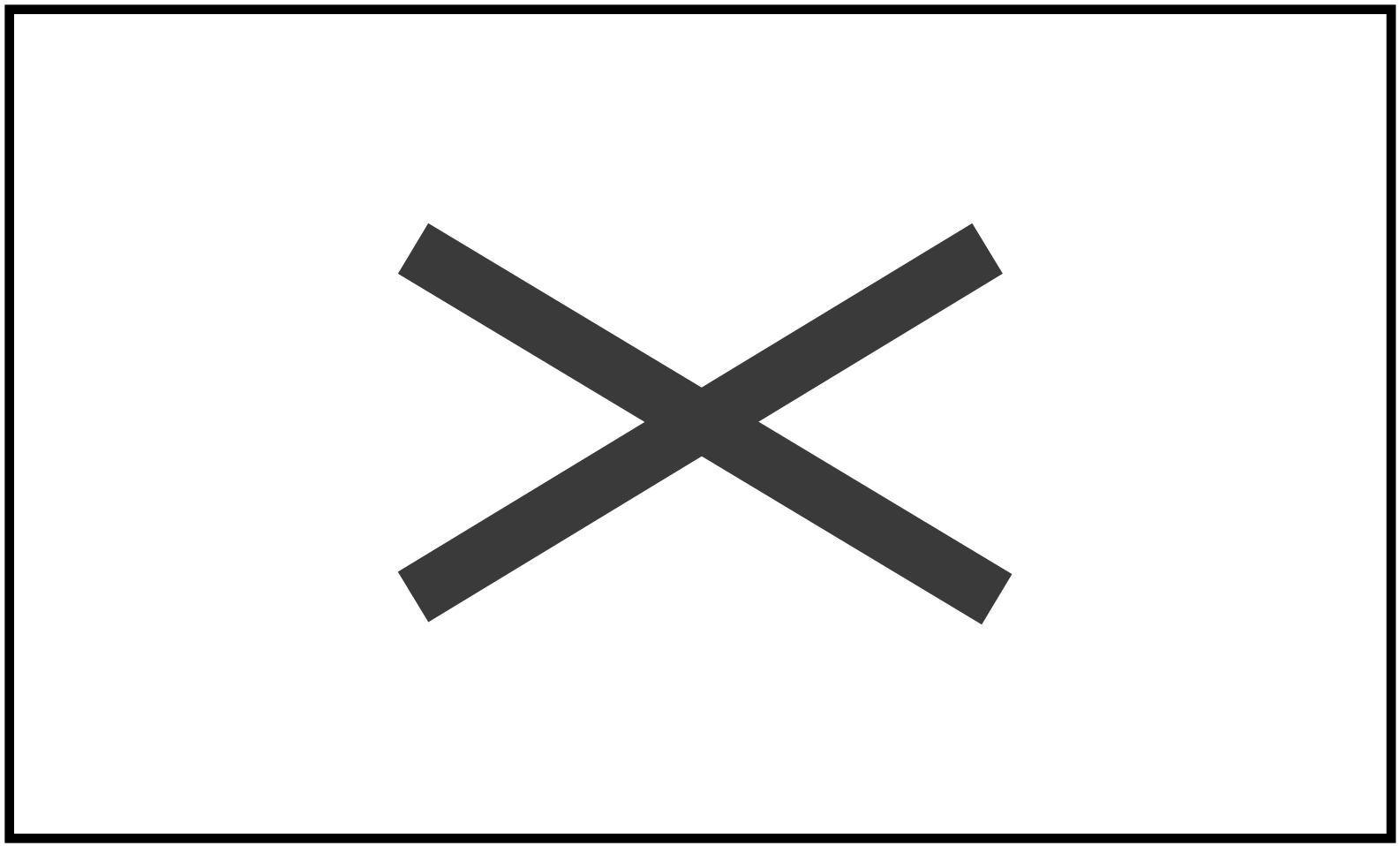
Fit indexes for both methods in function of the number of variables



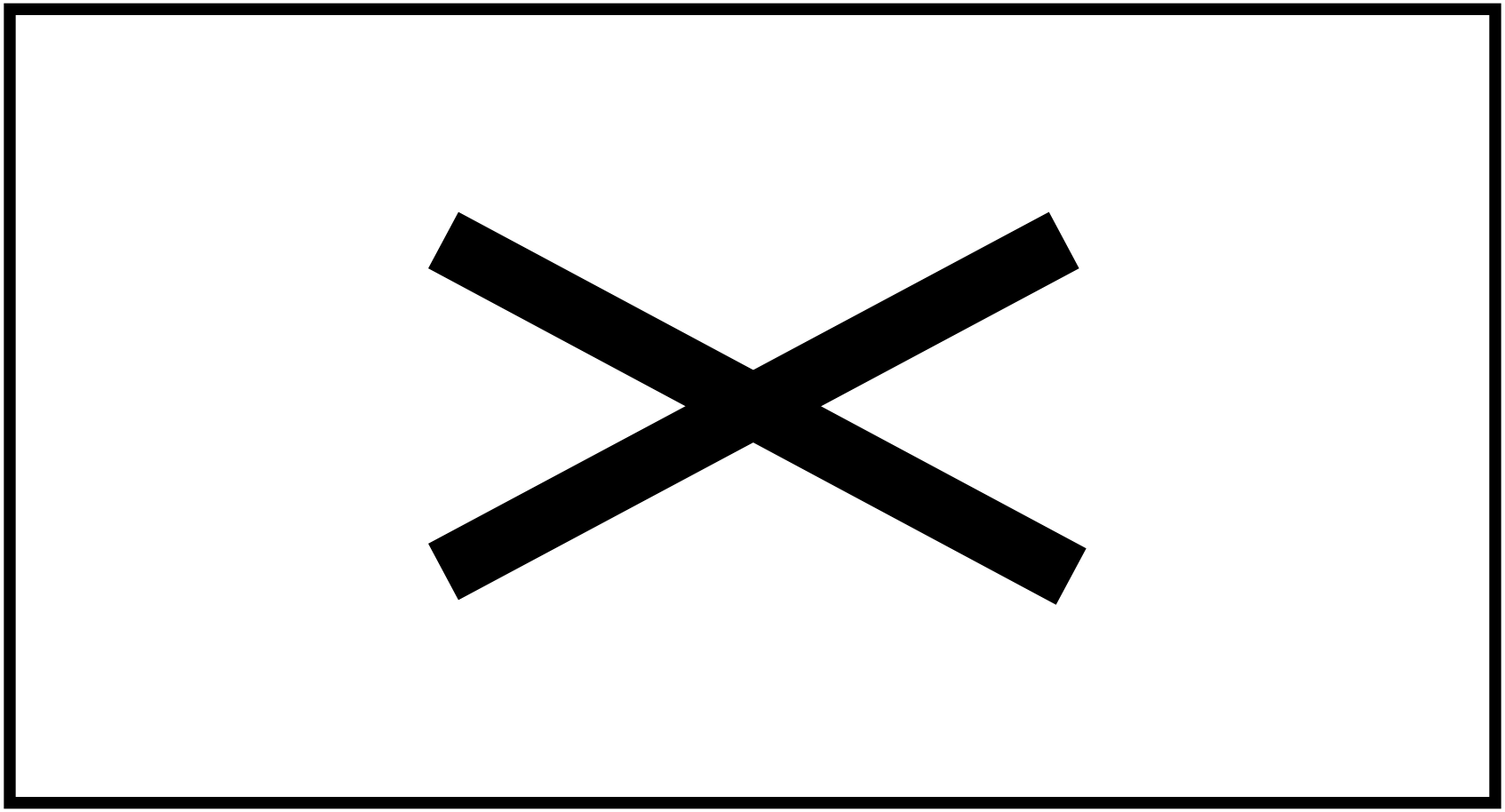
Wssa for the 89 variables



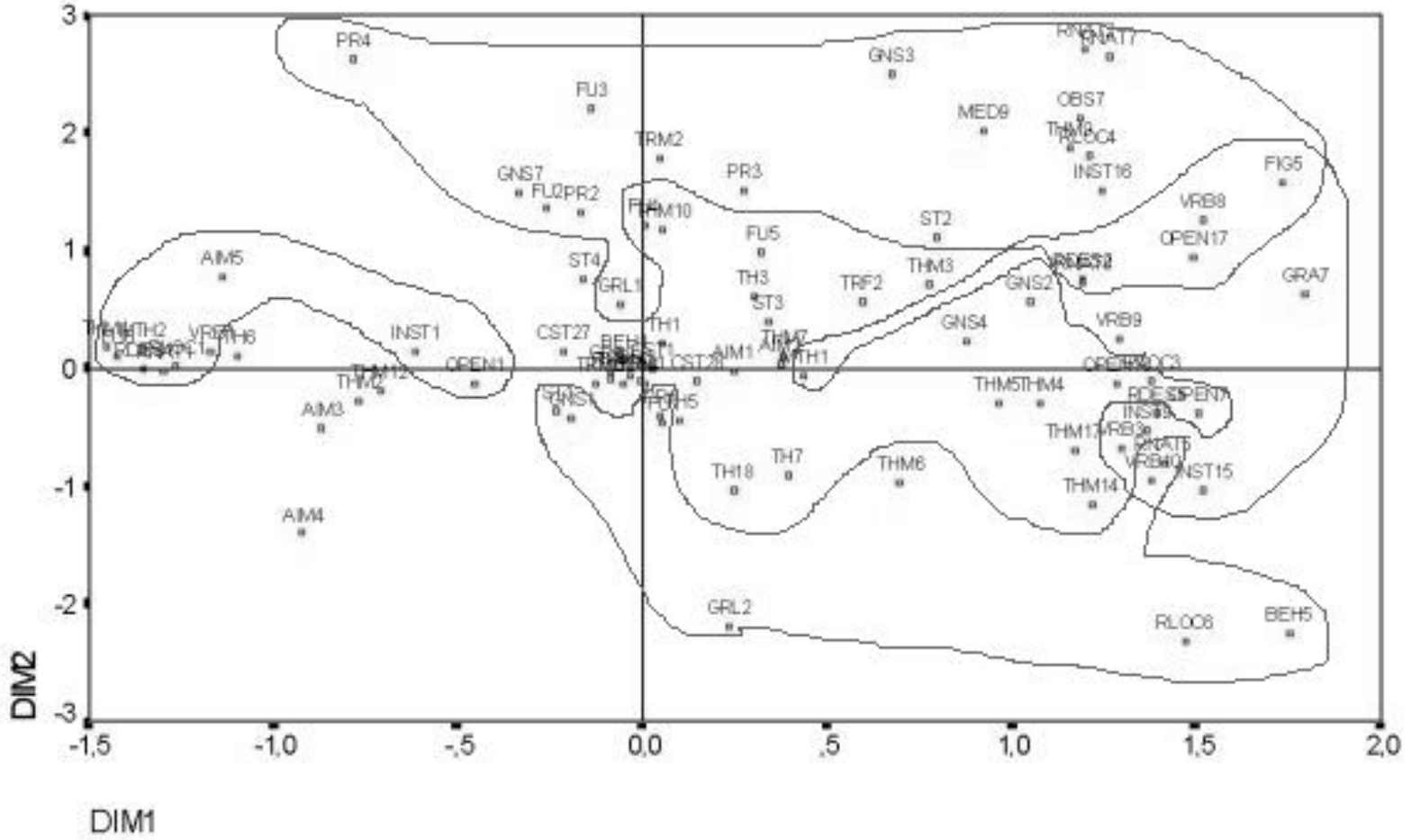
Factorial space 1x2 for the 89 variables



Contributing variables on the two first dimensions (anacor89)

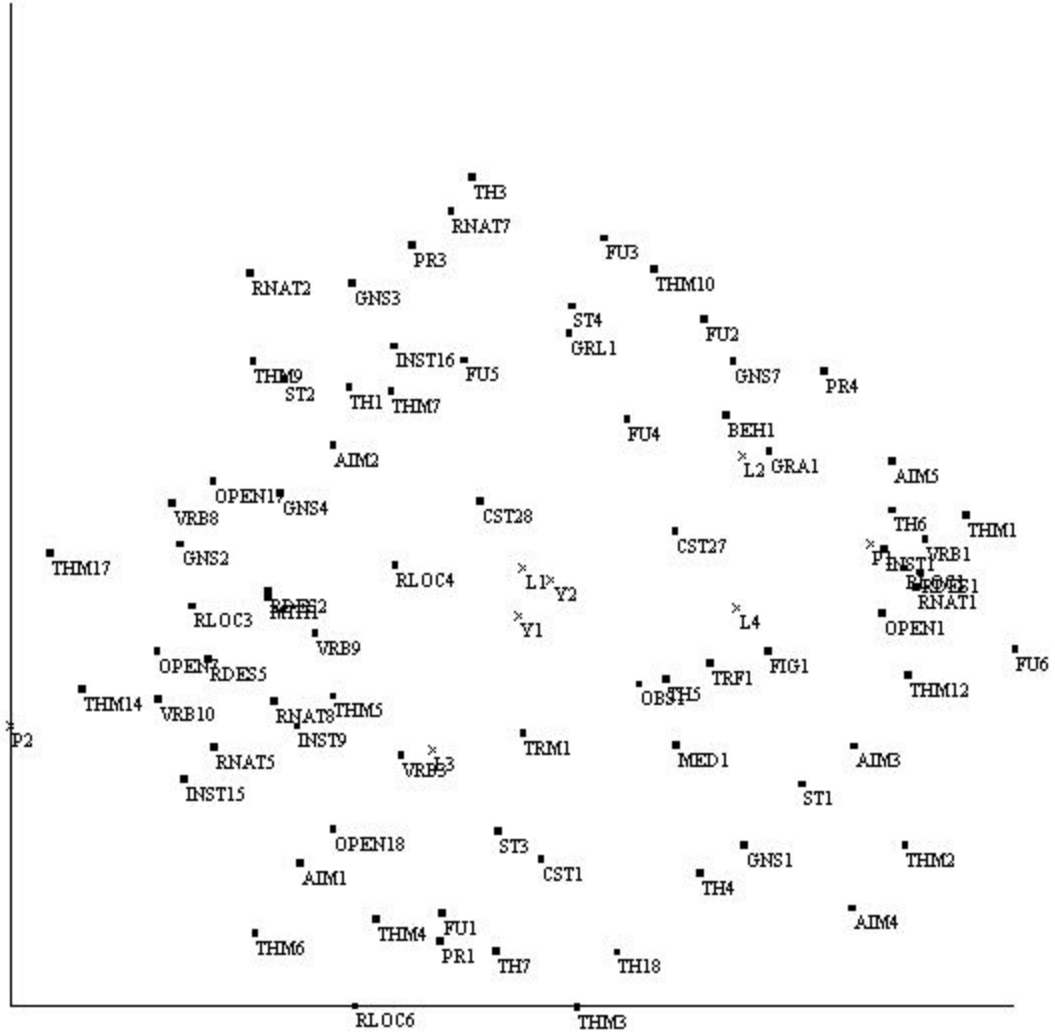


Factorial space 1x2 with contributing points (89)

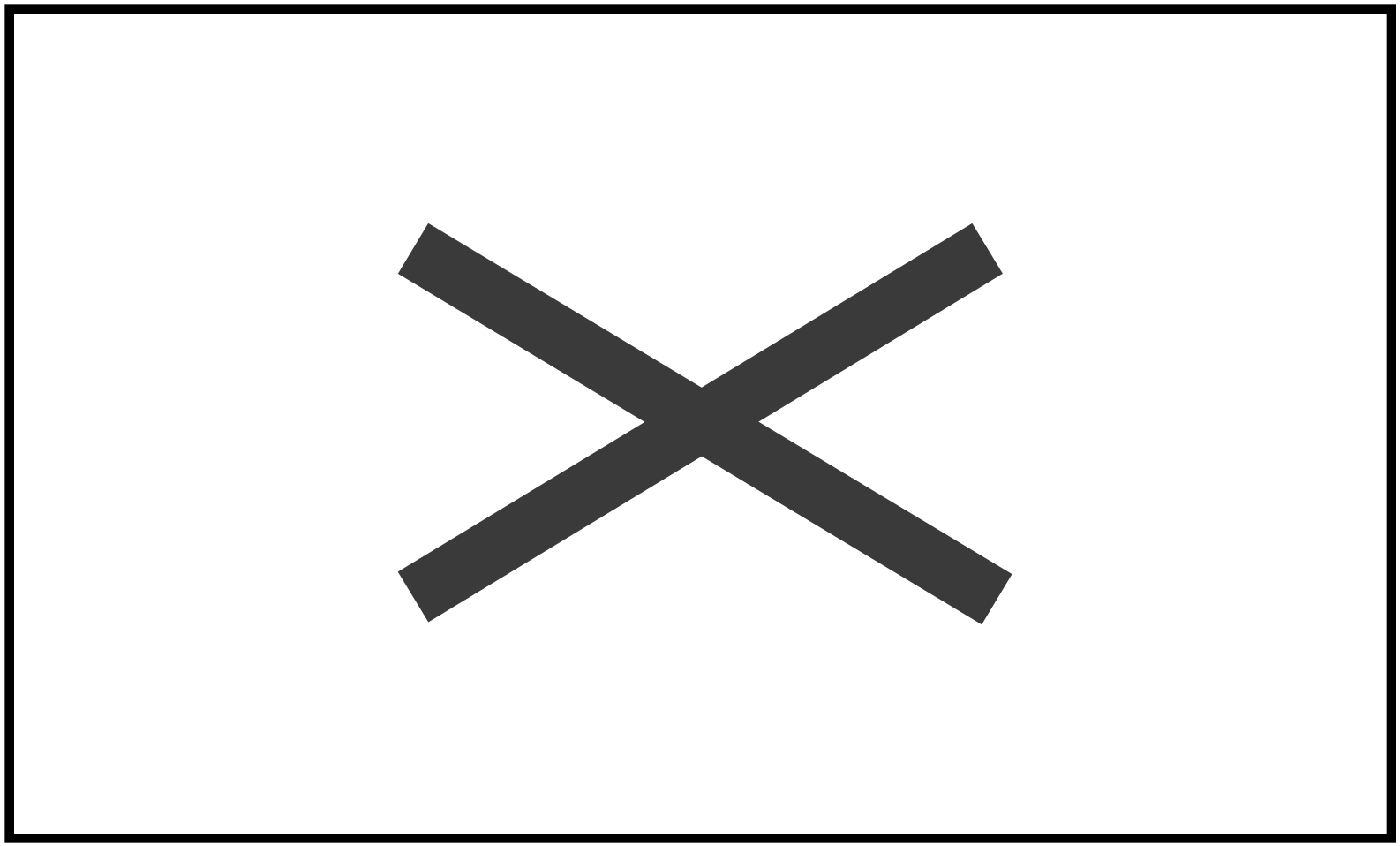


Wssa with contributing point on anacor89

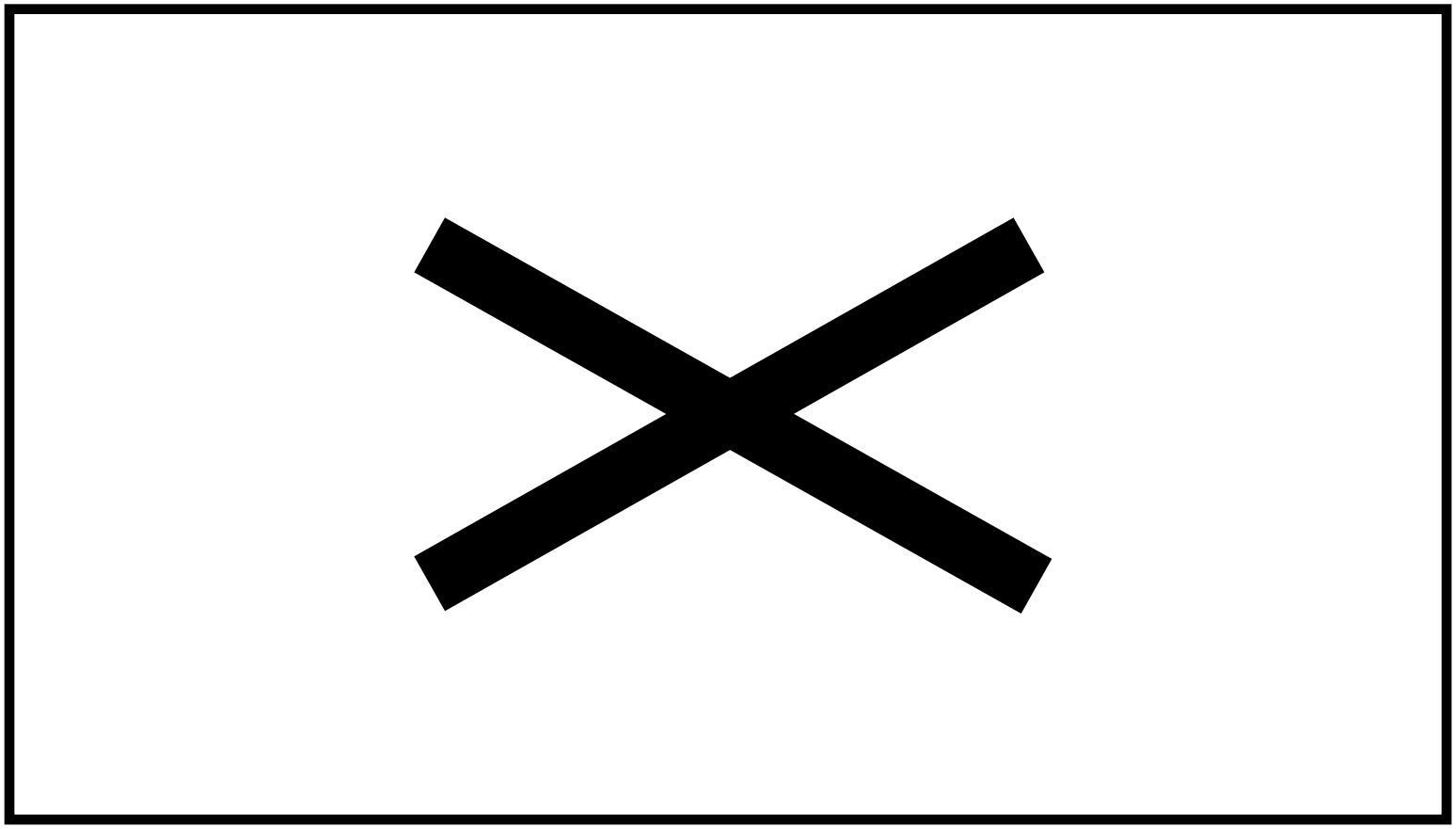
Wssa for the 80 variables



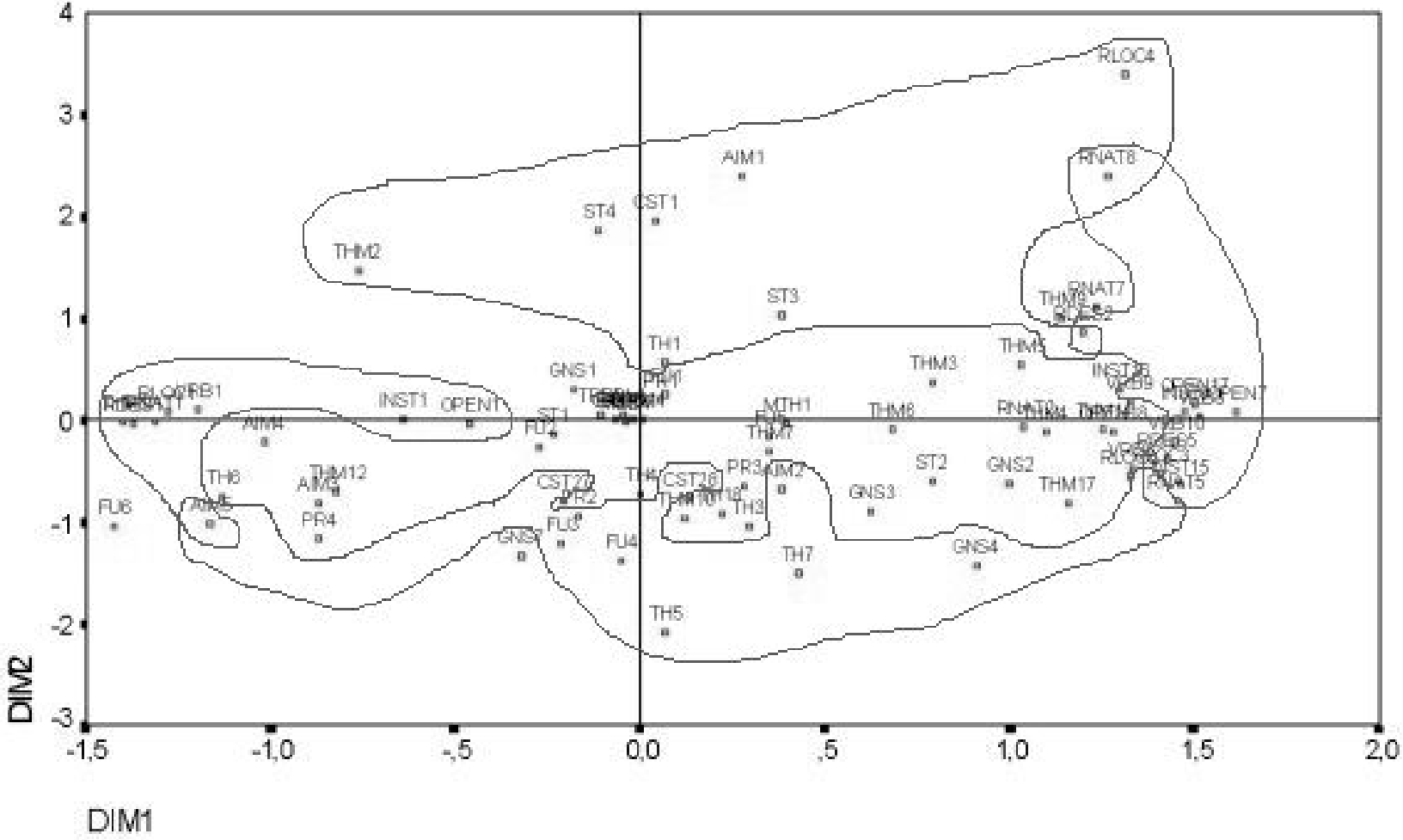
Factorial space 1x2 for the 80 variables



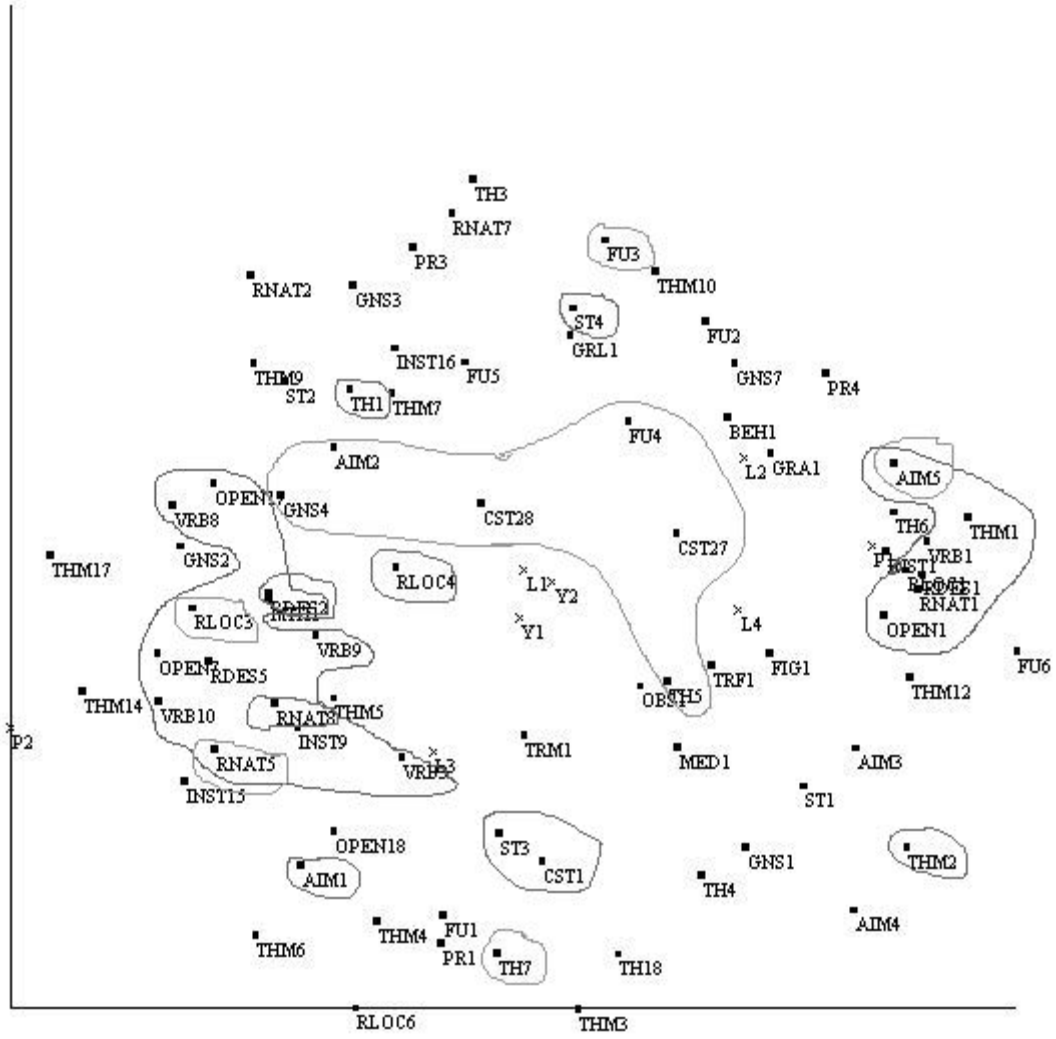
Contributing variables on the two first dimensions (anacor80)



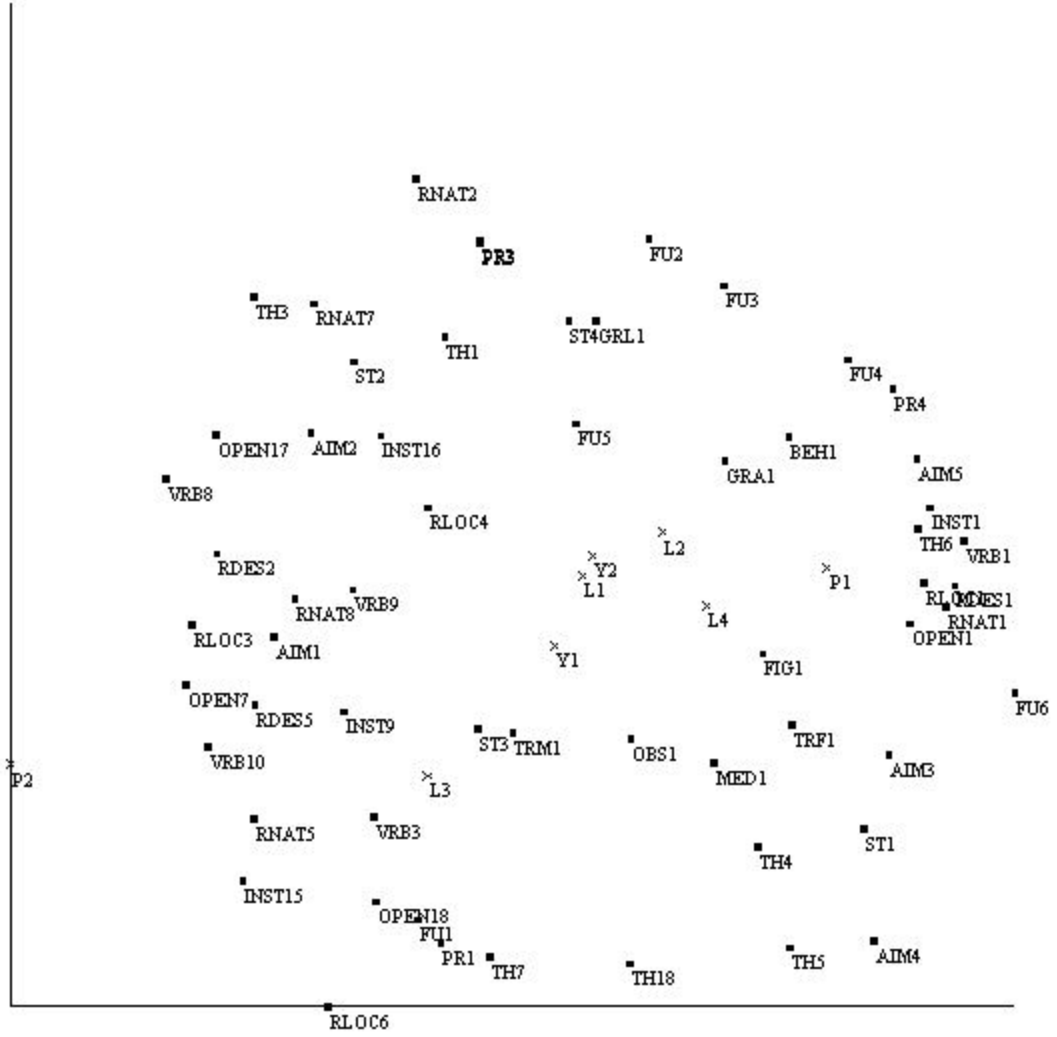
Factorial space 1x2 with contributing points (80)



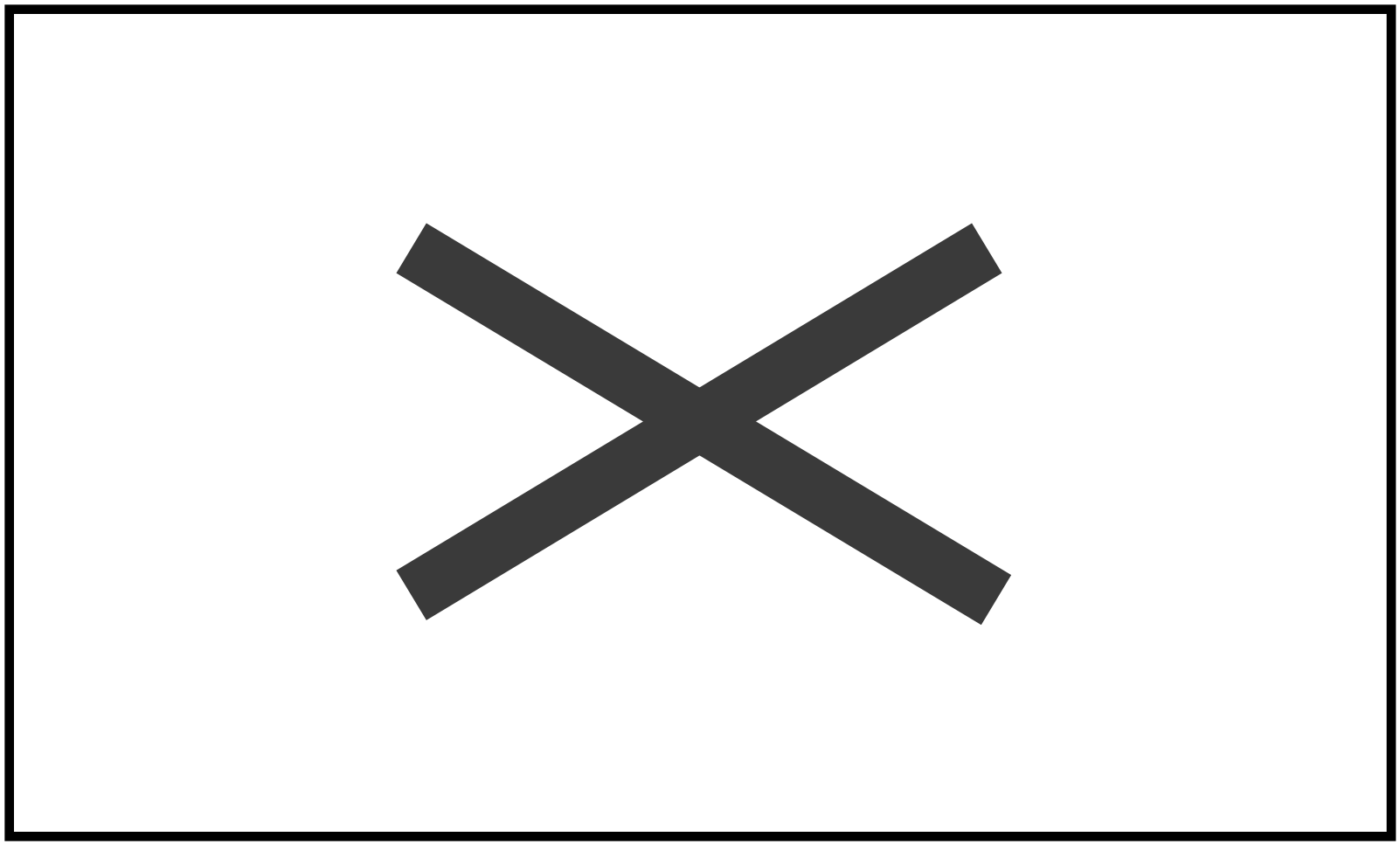
Wssa with contributing point on anacor80



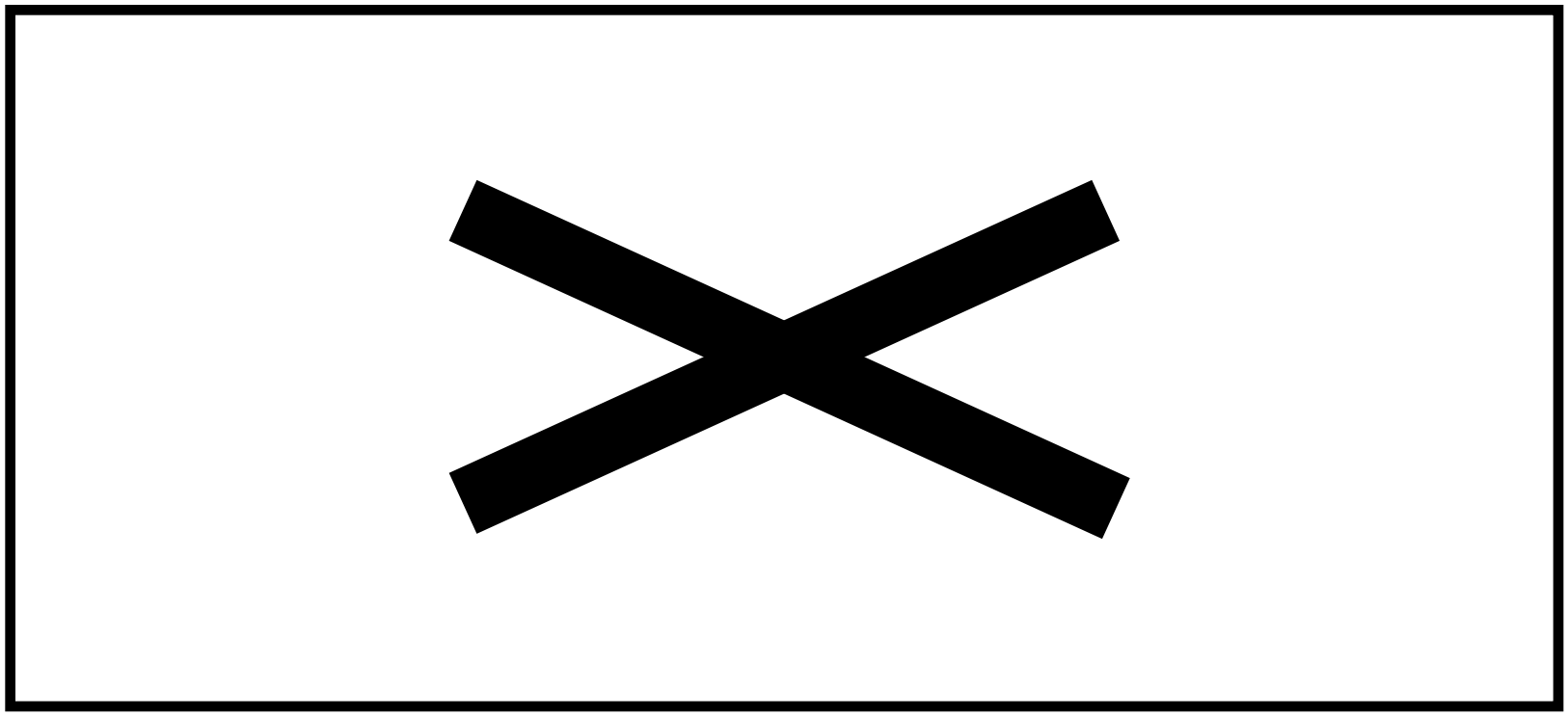
Wssa for the 59 variables



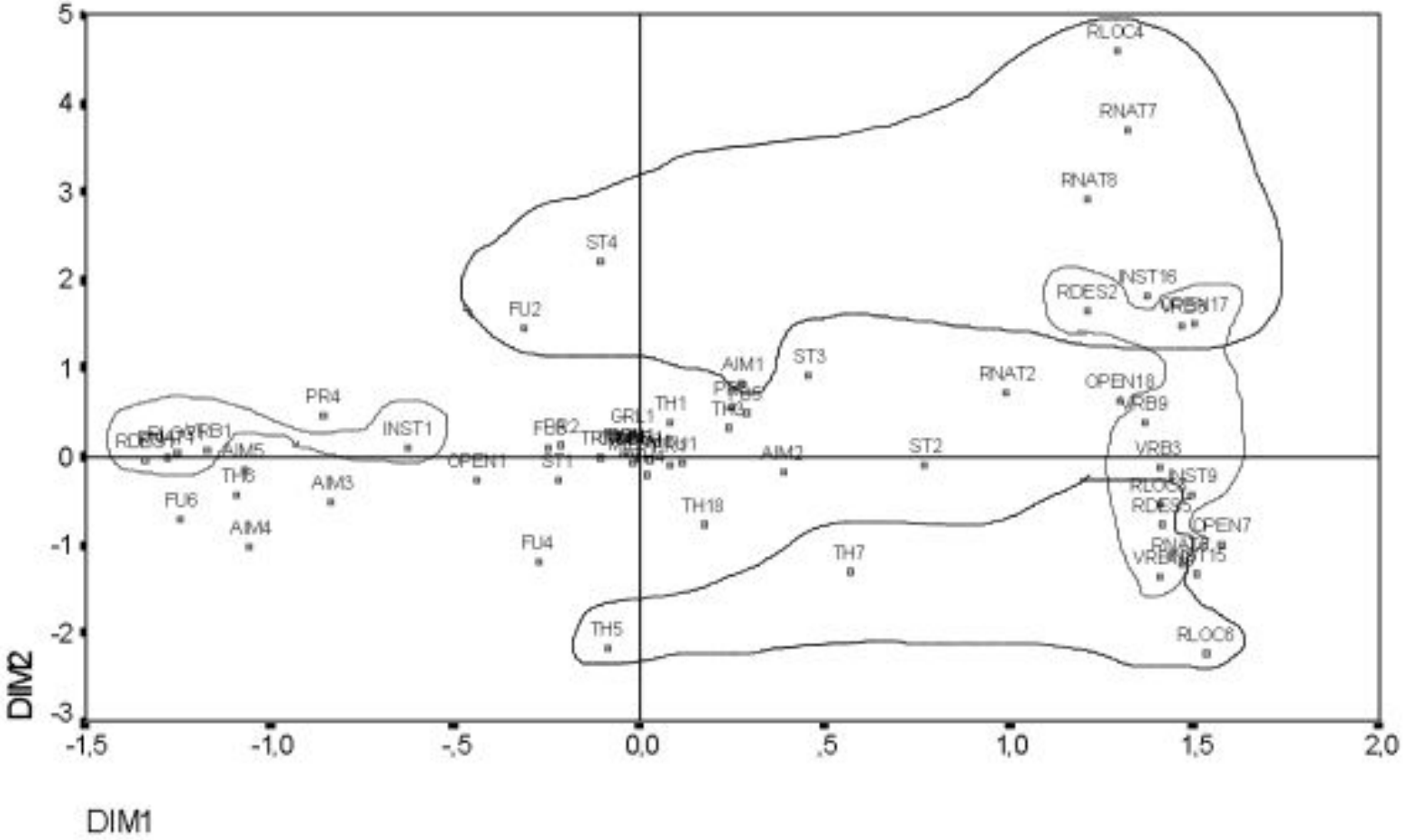
Factorial space 1x2 for the 59 variables



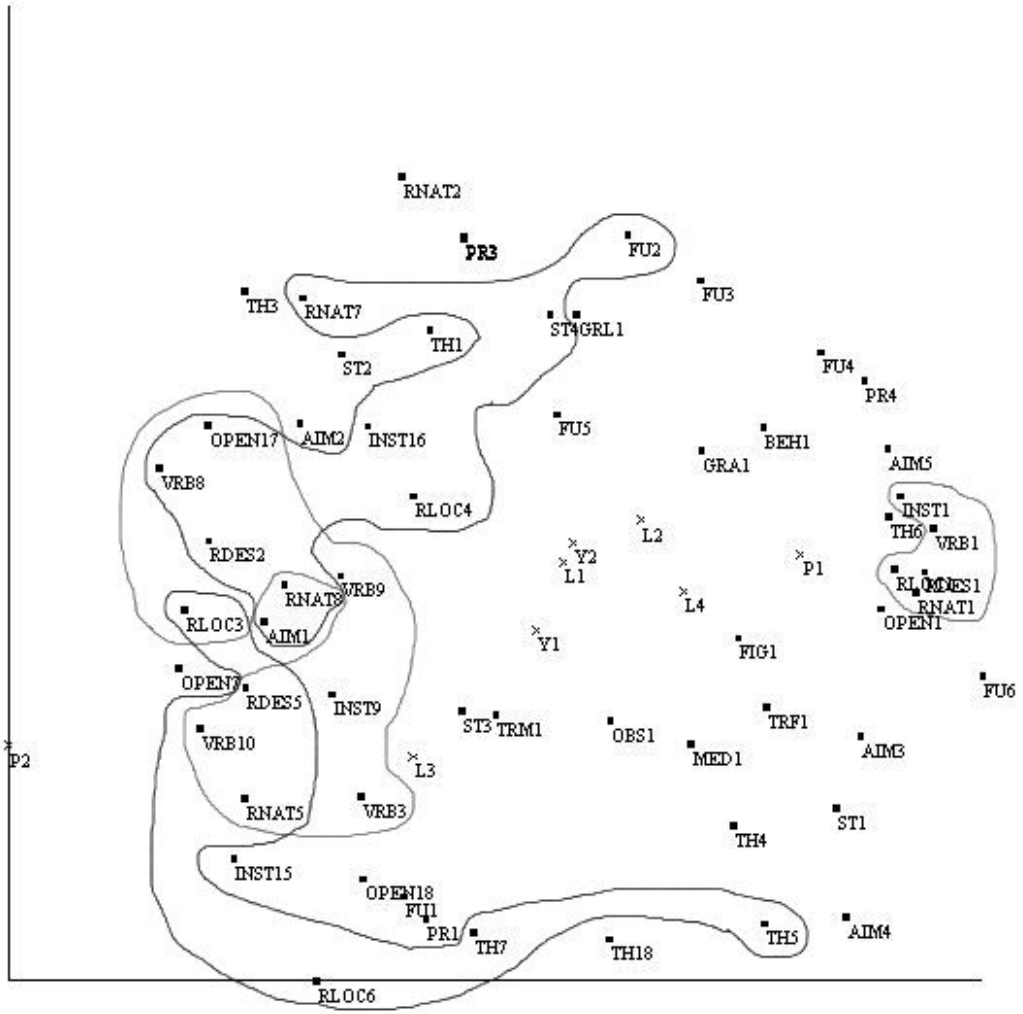
Contributing variables on the two first dimensions (anacor59)



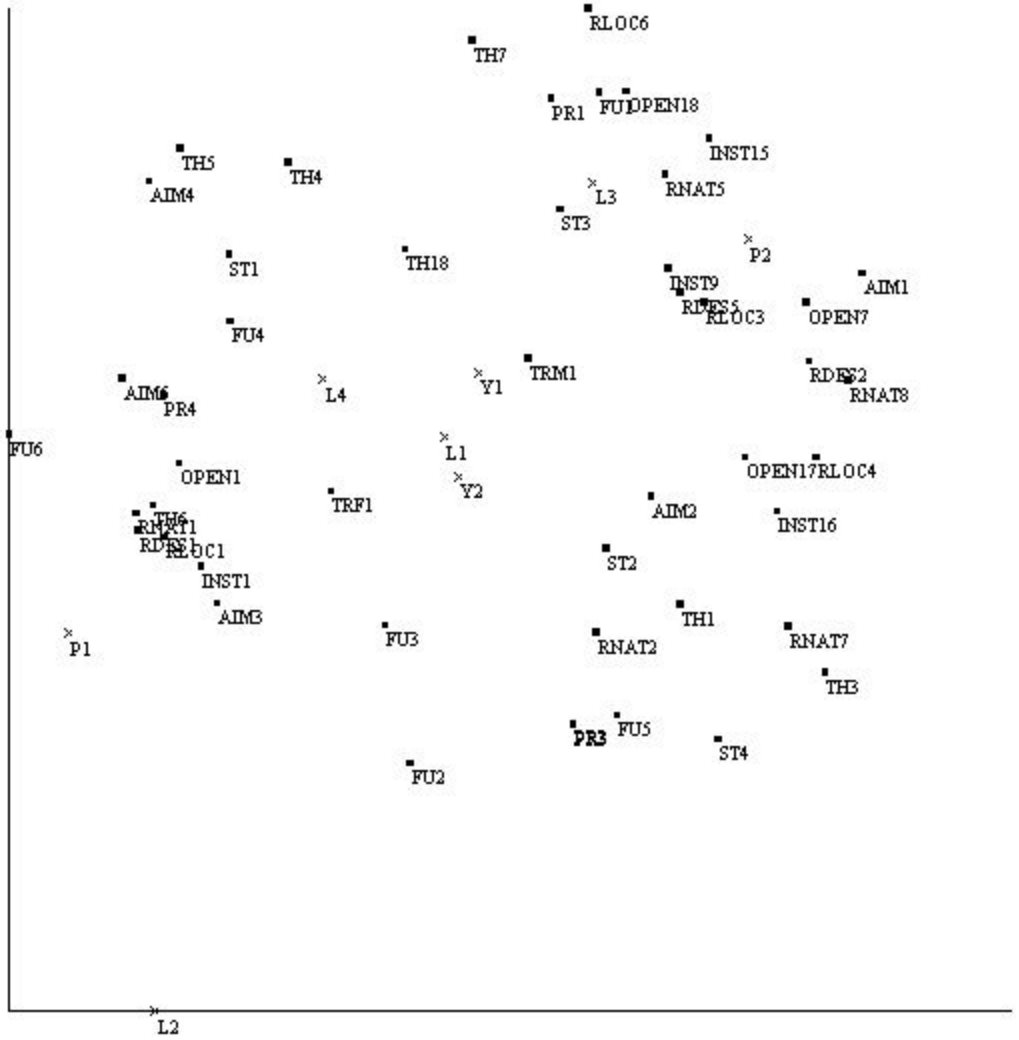
Factorial space 1x2 with contributing points (59)



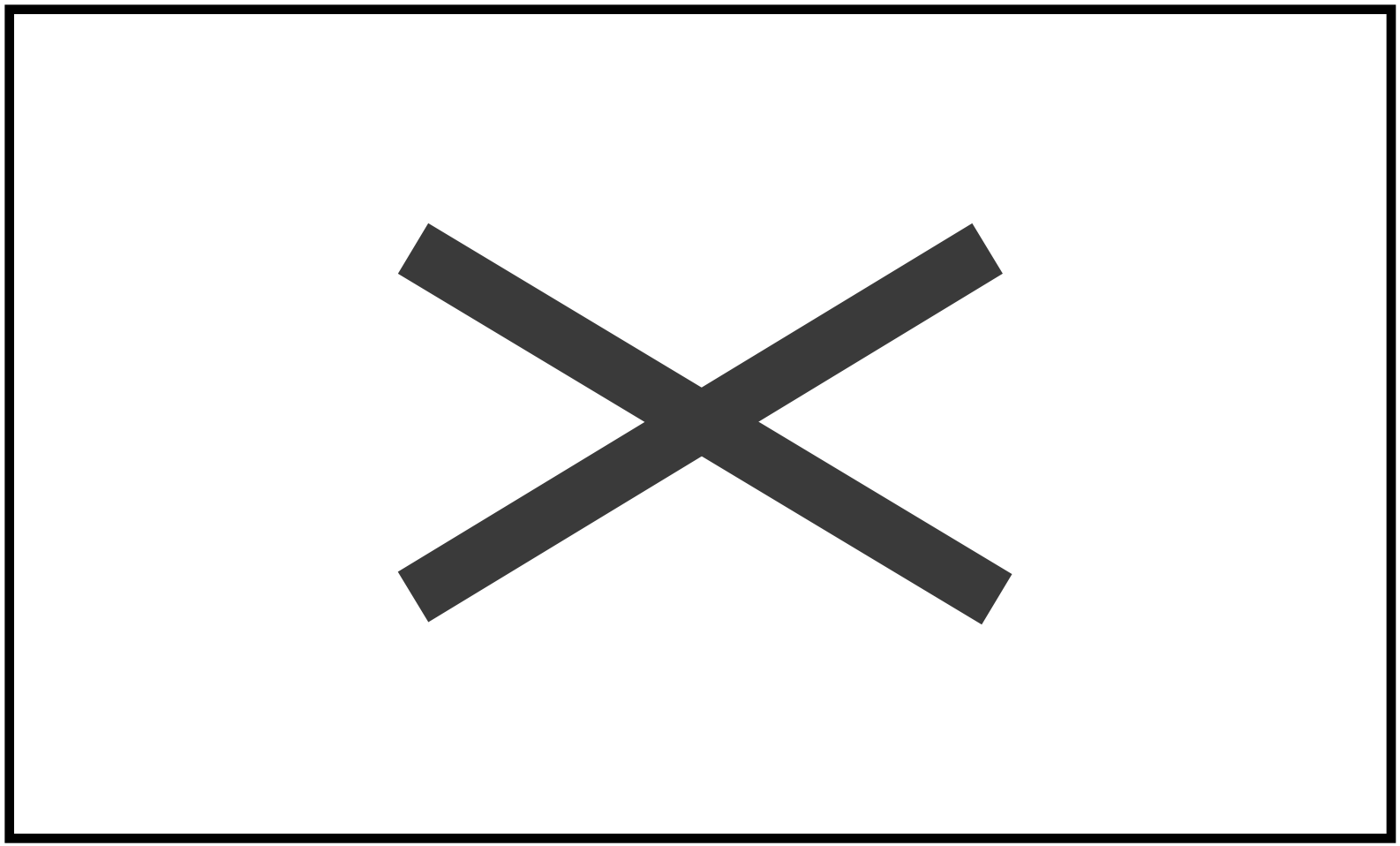
Wssa with contributing point on anacor59



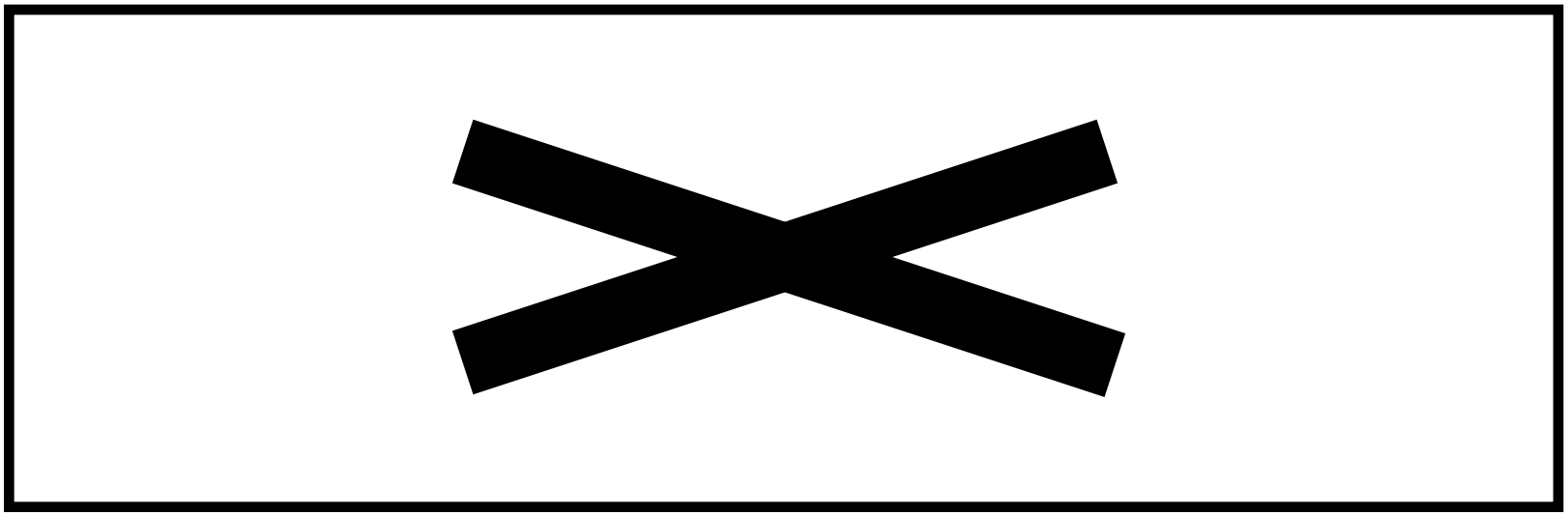
Wssa for the 48 variables



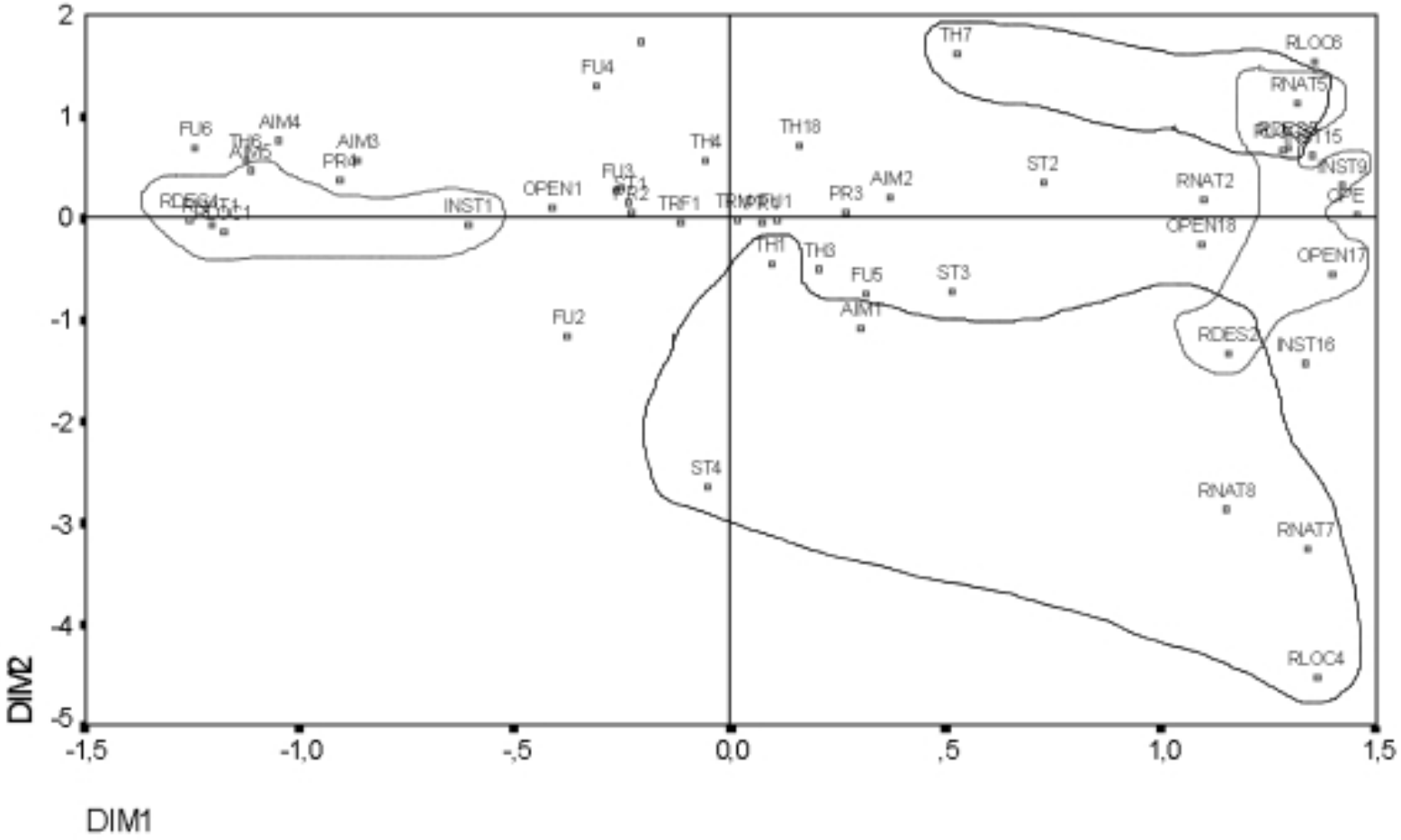
Factorial space 1x2 for the 48 variables



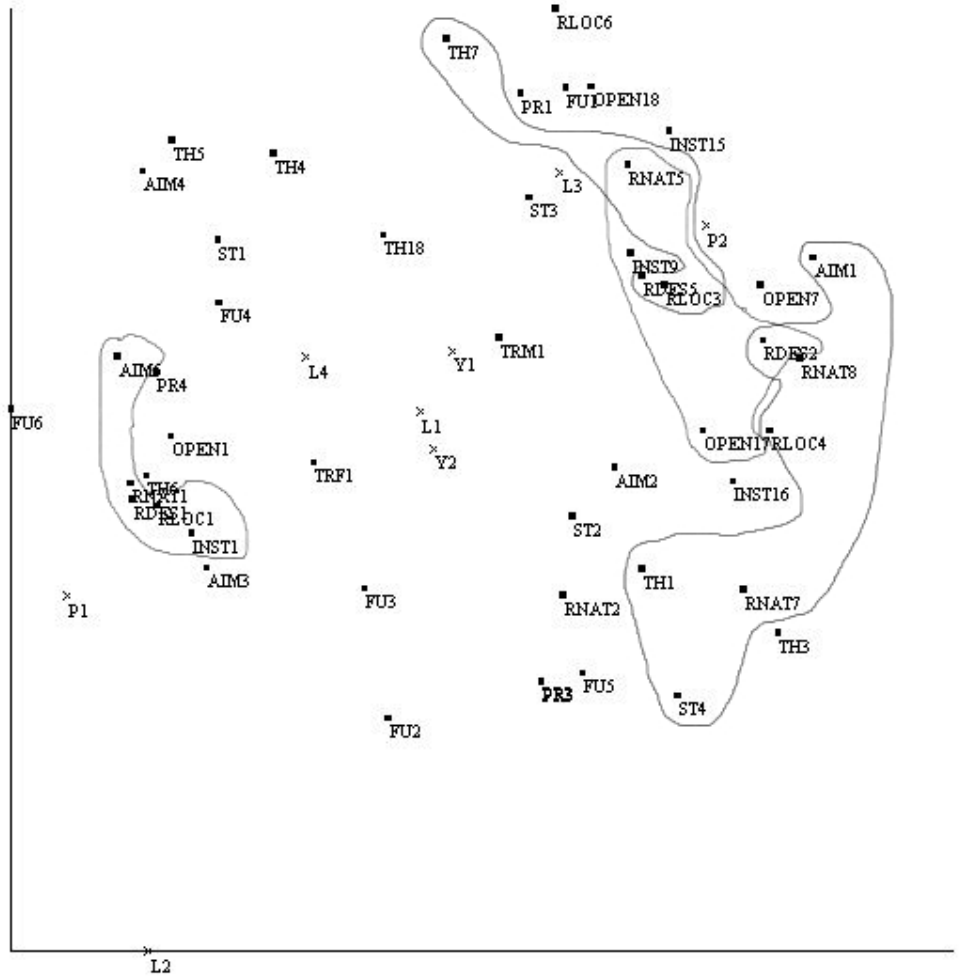
Contributing variables on the two first dimensions (anacor48)



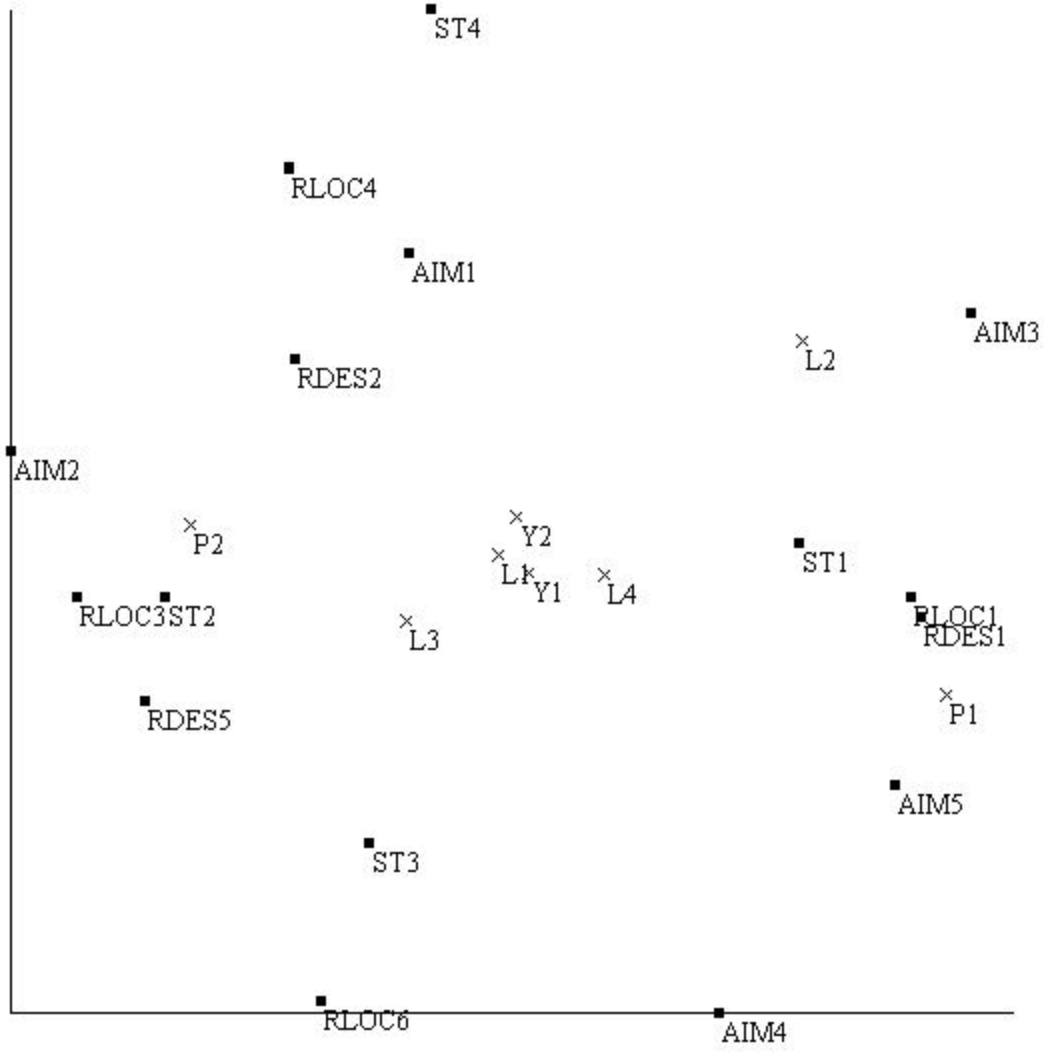
Factorial space 1x2 with contributing points (48)



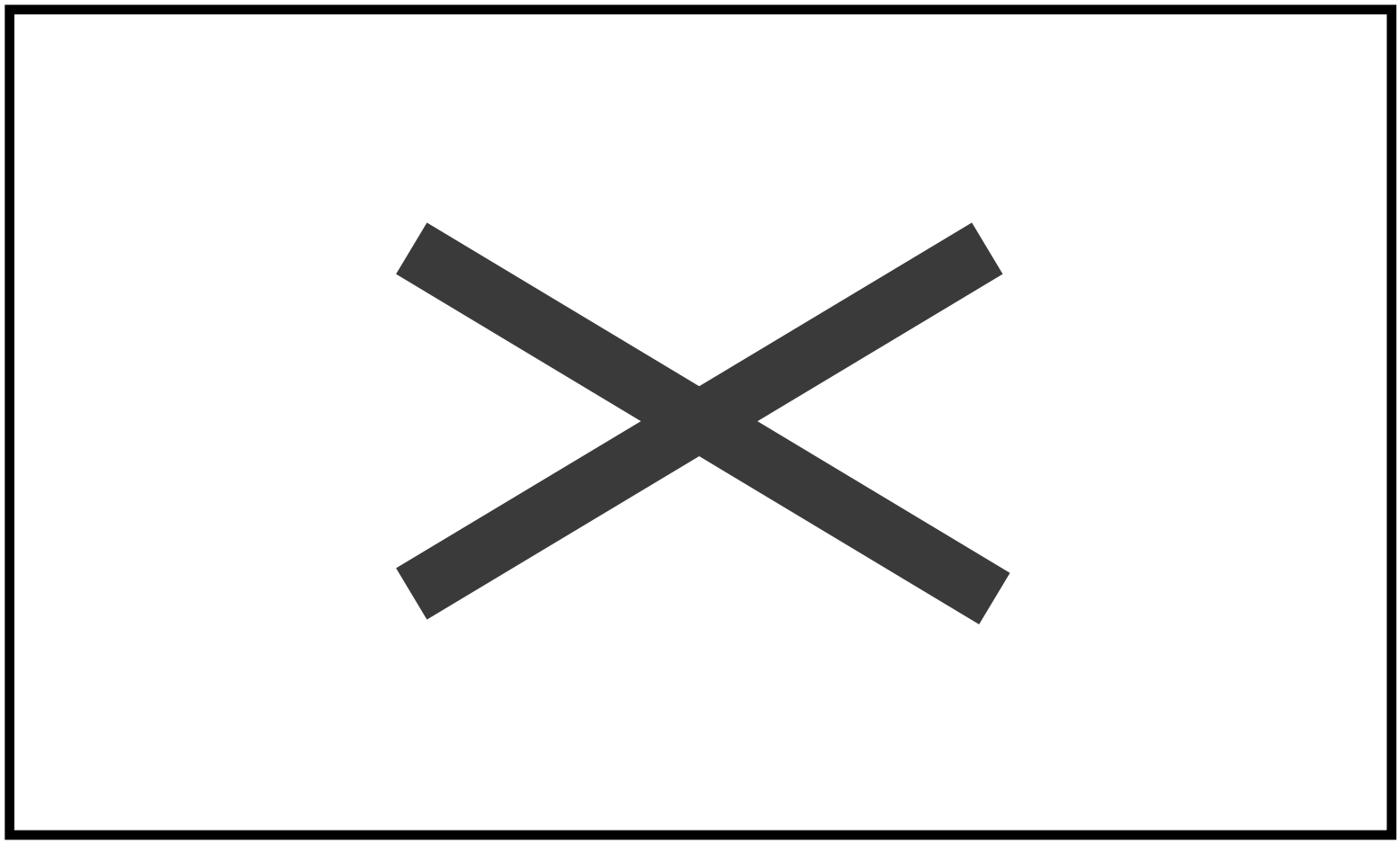
Wssa with contributing point on anacor48



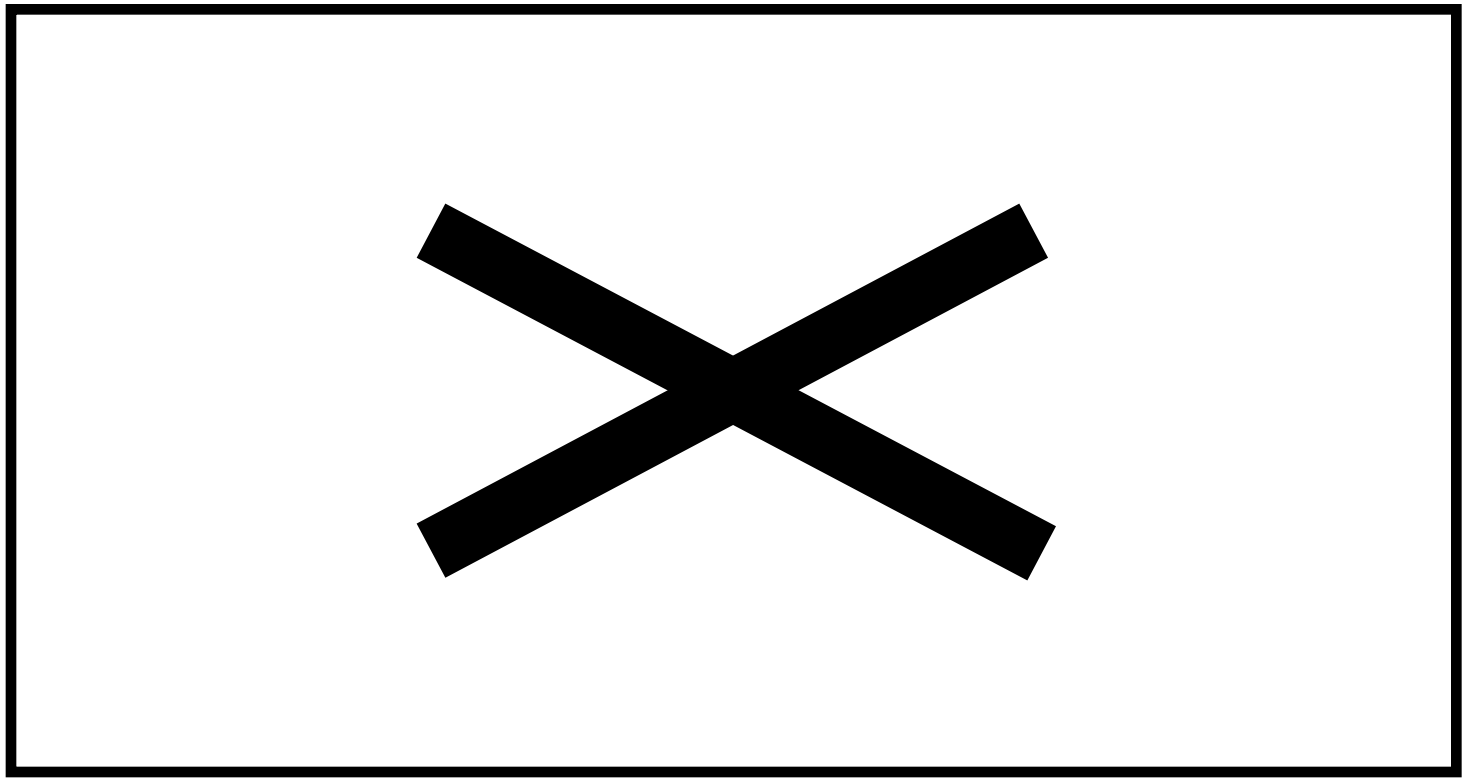
Wssa for the 16 variables



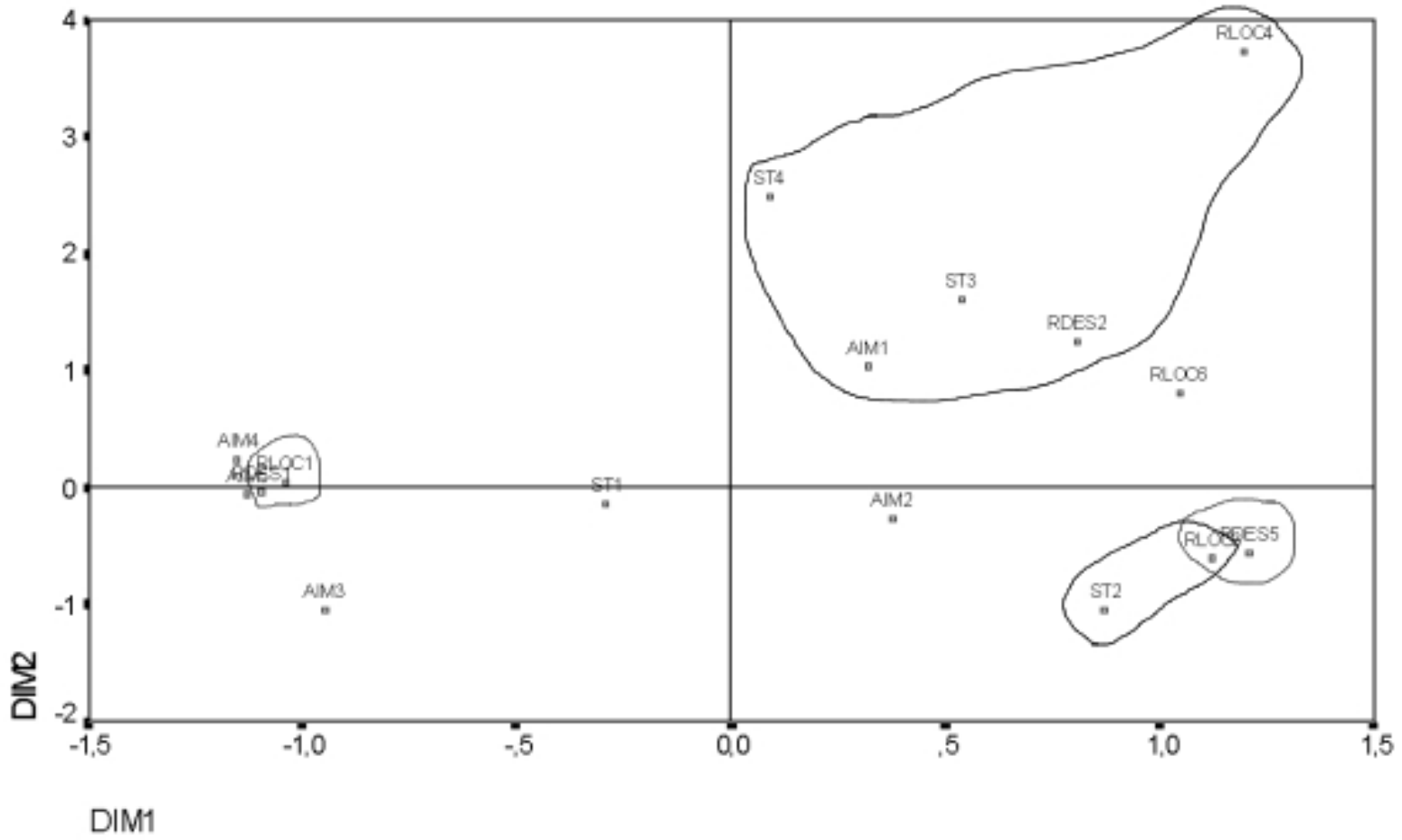
Factorial space 1x2 for the 16 variables



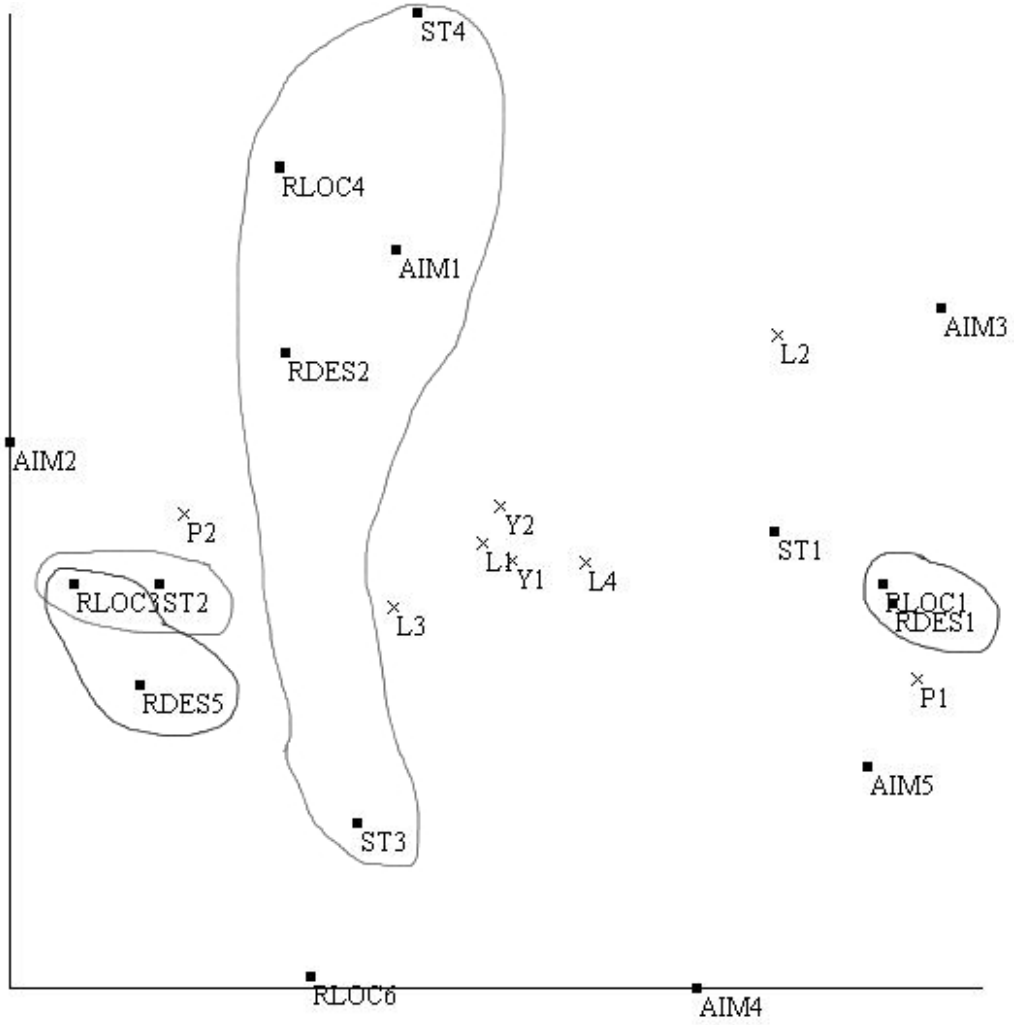
Contributing variables on the two first dimensions (anacor16)



Factorial space 1x2 with contributing points (16)



Wssa with contributing point on anacor16



Main results

- Common points:
  - Increasing of the fit / explained variance when the number of variables is reduced
  - Opposition of variables on the first factor / axis
- Differences:
  - Opposition of variables not always conserved
  - Circle representation vs cross representation

Toward an explanation

- The greatest part of the ‘variance’ of the data is similarly represented on a first dimension
- The ‘correlations’ for the remaining dimension are:
  - Independent in the anacor
  - Interdependent in the wssa
- In a two or three dimensions solution, the last one or the last two ones report(s)
  - A part of the remaining ‘variance’ in the anacor
  - All the remaining ‘variance’ in the wssa
- Those two last points explain together the differences we found (non conservation of oppositions and different shapes of representation)

Conclusions

- Importance of the data coding
- The structural aspect of data