Joint paper on discriminant analysis of the confinement database

A. Kus, A. Dinklage

Max-Planck-Institut für Plasmaphysik, IPP-EURATOM Association, Greifswald, Germany

CWGM 8.0 @NIFS

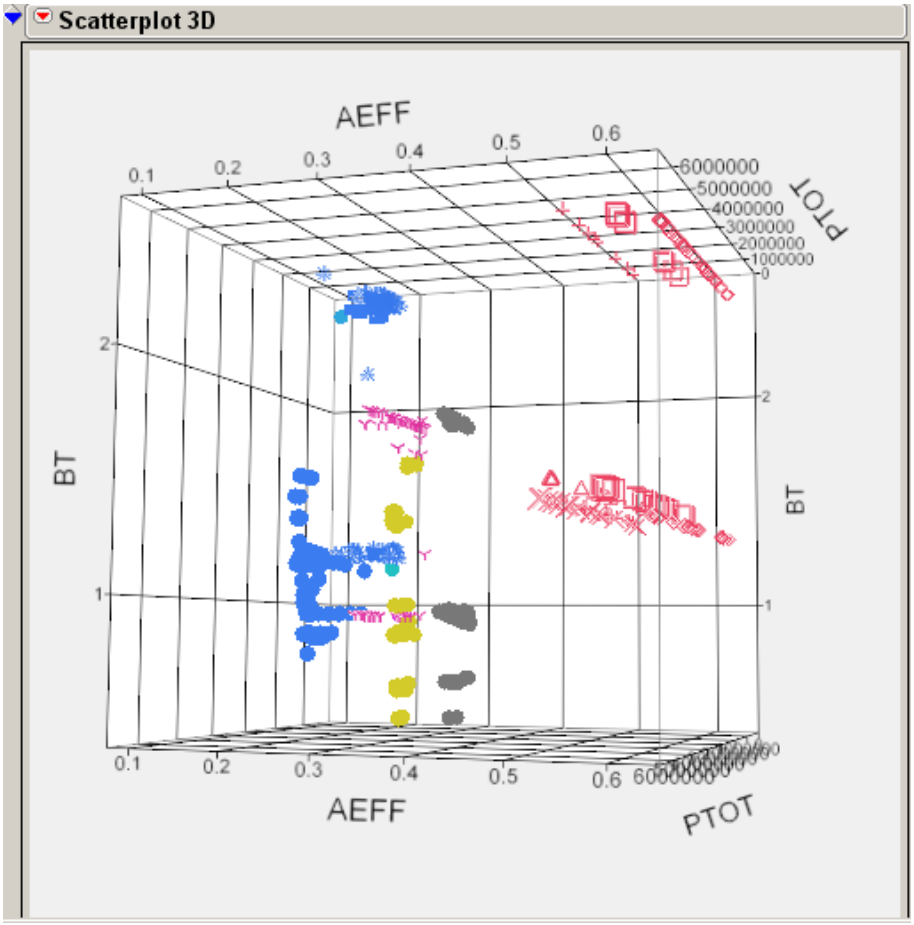
16 – 17 Mar. 2011

Identification of variables causing clustering in the global energy confinement data by use of discriminant analysis

A. Kus[1], A. Dinklage[1], E. Ascasibar[2], C.D. Beidler[1], A.A. Beletskii[3], B. D. Blackwell[4], T. Estrada[2], H. Funaba[5], J. Geiger[1], J.H. Harris[4,6], C. Hidalgo[2], M. Hirsch[1], D. Lopez-Bruna[2], A. Lopez-Fraguas[2], H. Maaßberg[1], T. Minami[5], T. Mizuuchi[7], S. Murakami[7], N. Nakajima[5], S. Okamura[5], D. Pretty[4], M. Ramisch[8], S. Sakakibara[5], F. Sano[7], U. Stroth[1], Y. Suzuki[5], Y. Takeiri[5], J. Talmadge[9], K. Thomsen[10], V. Tribaldos[2], Yu. A. Turkin[1], K.Y. Watanabe[5], A. Weller[1], R. Wolf[1], H. Yamada[5], M. Yokoyama[5]

[1]Max-Planck-Institut für Plasmaphysik, Euratom Assoc., Garching & Greifswald, Germany,
[2]Laboratorio Nacional de Fusión, Asociación Euratom/CIEMAT, Madrid, Spain,
[3]Institute of Plasma Physics, Kharkov, Ukraine,
[4]Australian National University, Canberra, Australia,
[5]National Institute for Fusion Science, Toki, Japan,
[6]Oak Ridge National Laboratory, Oak Ridge, USA,
[7]Kyoto University, Kyoto, Japan,
[8]Universität Stuttgart, Stuttgart, Germany,
[9]University of Wisconsin, Madison, USA,
[10]European Commission, DG Research J.4, Brussels, Belgium

Example view of the 14 ISS04 subgroups



Questions:

1) Do exist any "natural" subgroups in the data? **CLUSTER ANALYSIS**
2) Which quantities cause the clustering? **DISCRIMINANT ANALYSIS**

Findings so far (a preliminary ISS04 datasets analysis)

- Cluster analysis (in the 7-dim. space spanned by LOG_TAU, LOG_A, …, LOG_I) reveals the existence of cohesive subgroups (clusters) in the data

- Clusters differ from ISS04 subgroups but resemble ISS04 grouping roughly

- Saturation in regression coefficients when increasing the number of clusters
  a. No clustering (all data in one cluster/one group): differences to ISS04
  b. There ist a kind of saturation in the number of clusters: a regression analysis on data divided into ca. 11-18 clusters provides results similar to ISS04 scaling formula

Reminder: **Linear discriminant function analysis**

Linear discriminant function analysis
aims at determination of linear combinations of independent variables (*predictors*) that discriminate among the *categories* of the grouping variable.

A linear combination of predictor variables X_1, …, X_p, called discriminant function is constructed such that it assigns its values into two subgroups that differ as much as possible.

In the most simple case, with two subgroups (e.g. L/H mode data), there exists only one discriminant function

$$D = b_1 X_1 + b_2 X \ + \ldots + b_p X_p$$ (for data previously standardized)

A model must be correctly defined (all relevant variables involved, e.g. using „stepwise variable selection“ procedure).

Reminder: **Linear discrim. function analysis (cont.)**

In a general case the maximal number of discriminant functions is equal

- number of subgroups minus one, or
- number of used variables,

whatsoever is smaller.

Properties of discriminant functions:

- All discriminant functions are pairwise orthogonal (uncorrelated)
- Viewing the coefficients ***b***'s one can see how the predictor variables contribute to the discrimination: the larger the ***b*** the larger the contribution (This is valid only for standardized data.)

Implementation of discriminant analysis

- Learning process

  Determination of discriminant functions (model development):

  Which quantities decide about the affiliation of an observation to the particular group (*what distinguish high-beta discharges from others*?)

- Prediction

  Application (after a discrim. model was developed):

  Assignment of a new observation to the particular group

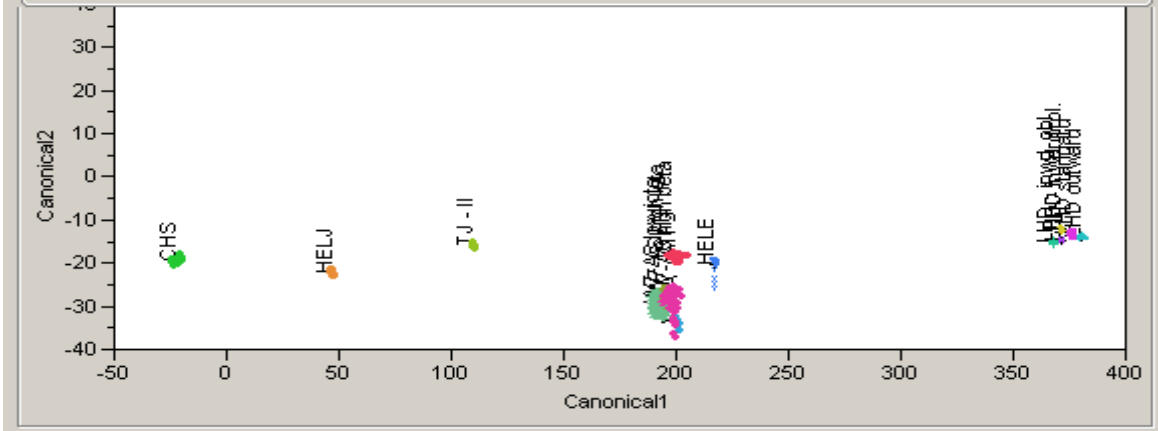
  Examples

  - Plasma physics: an observation with certain parameters is assigned to L or H-Mode
  - Medicine: particular risk factors cause a specific disease
  - (Psychology, war against terrorism, …)

ISS04 subgroups in the first two discrim. dimensions

Discriminant Analysis

Canonical Plot



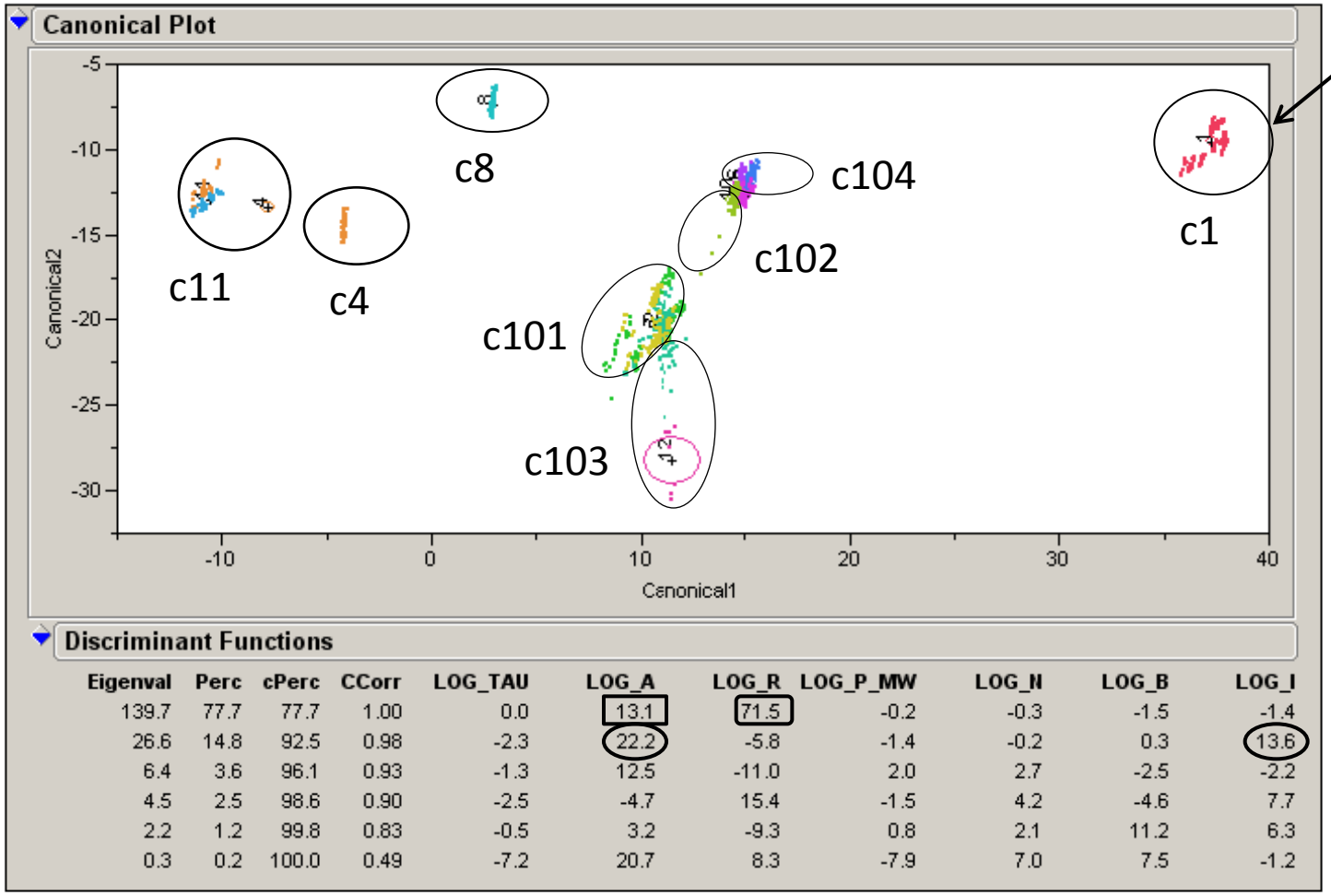
Eigenvalue	Percent	Cum Percent	Canonical Corr
9786.67554	99.4636	99.4636	0.99994891
32.7370848	0.3327	99.7963	0.98506803
17.5219957	0.1781	99.9744	0.97263052
1.66616549	0.0169	99.9913	0.79052483
0.69515156	0.0071	99.9984	0.64037665
0.09475779	0.0010	99.9993	0.2942039
0.06635034	0.0007	100.0000	0.24944317

Scoring Coefs

	LOG_TAU	LOG_A	LOG_R	LOG_P_MW	LOG_N	LOG_B	LOG_I
Canon1	2.5939447	-2.579194	658.55786	1.2886734	-1.437517	0.0729158	-1.676246
Canon2	-1.839636	29.954973	-9.676908	-0.395201	-0.116078	-0.264218	12.217464
Canon3	-0.05654	-21.58654	19.499656	-0.717334	0.0629623	1.5594034	16.824264
Canon4	0.7708655	-5.321855	-1.285584	3.4616421	0.5046022	-4.15697	3.7582824
Canon5	-7.781142	16.766728	14.81657	-6.981952	5.8364484	2.1651982	-2.753825
Canon6	-6.165648	15.325982	1.2215169	-3.662851	4.7953224	11.457827	0.9472032
Canon7	4.1014277	-1.649017	-3.113516	-1.775963	1.7945361	-4.366815	3.070125

5-6 larger groups, determined by LOG_R

ISS04 data clusters along the first two discrim. functions



Discriminant Functions

Eigenval	Perc	cPerc	CCorr	LOG_TAU	LOG_A	LOG_R	LOG_P_MW	LOG_N	LOG_B	LOG_I
139.7	77.7	77.7	1.00	0.0	13.1	71.5	-0.2	-0.3	-1.5	-1.4
26.6	14.8	92.5	0.98	-2.3	22.2	-5.8	-1.4	-0.2	0.3	13.6
6.4	3.6	96.1	0.93	-1.3	12.5	-11.0	2.0	2.7	-2.5	-2.2
4.5	2.5	98.6	0.90	-2.5	-4.7	15.4	-1.5	4.2	-4.6	7.7
2.2	1.2	99.8	0.83	-0.5	3.2	-9.3	0.8	2.1	11.2	6.3
0.3	0.2	100.0	0.49	-7.2	20.7	8.3	-7.9	7.0	7.5	-1.2

Cluster c1
LHD

c4
CHS (68), HELJ

c8
TJ-II

c11
CHS (128)

c101
W7-A,
W7-AS h.+l. iota

c102
HELE

c103
W7-AS h. beta

c104
ATF

- Geometry is crucial for the discrimination (thesis valid only for the used model)
- 92.5 percent of the total variation happens in plane spanned by the first two discriminant functions
- Cluster c103 (W7-AS high-beta) overlaps other W7-AS data (An additional/latent discrim. factor exist?)

To do till EPS:

- Define a subset of „suspected“ variables that may cause clustering
  An automatic procedure is exponentially time consuming:
  10 variables → 1023 (= 2^10-1) combinations (models),
  20 variables → 1 048 575 combinations,
  104 numerical variables in ISHCDB_25 → 2*10^31 possible models!

- Select a dataset for analysis
  Which group of data should indentified at first?
  (High beta?)

- Perform analysis