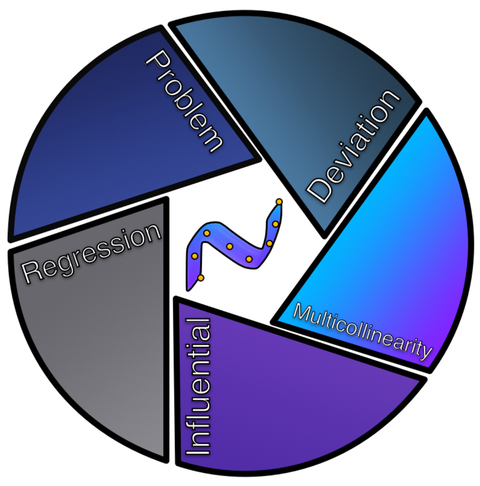
IPB98: Robust regression methods motivated by deep analysis of scaling multicollinearity

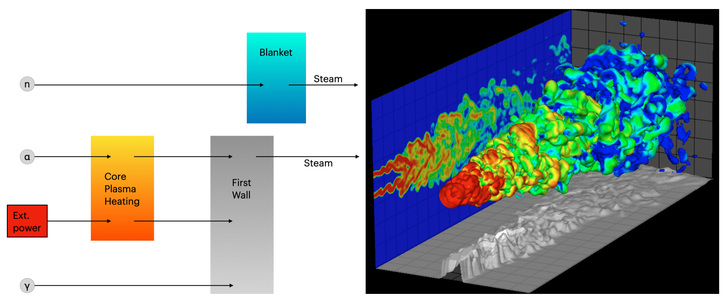
J. Hall[1], G. Verdoolaege[1]
[1]Department of Applied Physics, Ghent University, Belgium;
Correspondence to: joseph.hall@ugent.be

GHENT UNIVERSITY | FACULTY OF ENGINEERING AND ARCHITECTURE
*inf*usion

The well-known IPB98 scaling law for the energy confinement in tokamak H-mode plasmas has recently been revised. A considerably larger data set was used for estimating the scaling, including data from devices with fully metallic walls (JET and ASDEX Upgrade). In the new scaling, the dependence of the confinement time on several predictor variables turns out to be rather weak. Nevertheless, one key difference with the '98 scaling is the significantly weaker dependence on machine size, from quadratic to slightly stronger than linear. This work has been aimed at understand the origin of the reduced size scaling.



Problem: Reduction in regression coefficient αR



Fusion power balance depiction 1) High energy neutrons from the fusion reaction providing the largest contribution to output energy **2)** α particles and external heating power contributing to the internal energy of the plasma which in turn contribute to the output energy by **convective** and **diffusive** processes mediated by turbulent processes (as seen in figure b) **3)** Bremsstrahlung radiation

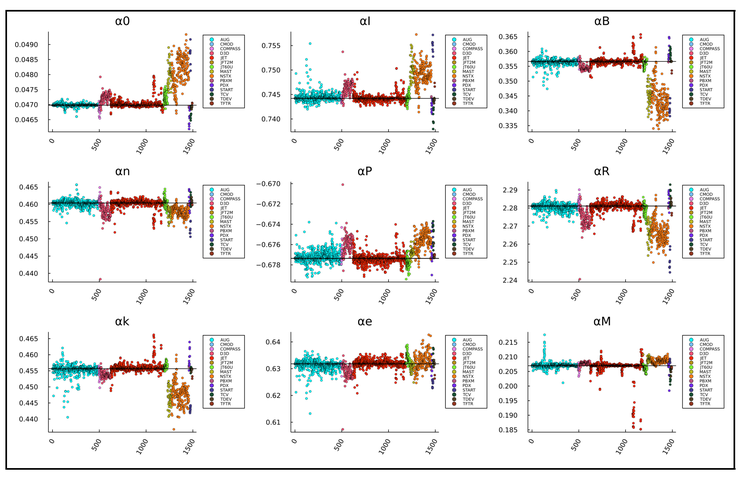
The IPB98 and ITPA20-IL scaling laws both implement a log-linear regression model

$$\tau_{\mathrm{E,th}} = \alpha_0 I_{\mathrm{p}}^{\alpha_I} B_{\mathrm{t}}^{\alpha_B} \bar{n}_{\mathrm{e}}^{\alpha_n} P_{\ell,\mathrm{th}}^{\alpha_P} R_{\mathrm{geo}}^{\alpha_R} \kappa_{\mathrm{a}}^{\alpha_\kappa} \epsilon^{\alpha_\epsilon} M_{\mathrm{eff}}^{\alpha_M}$$

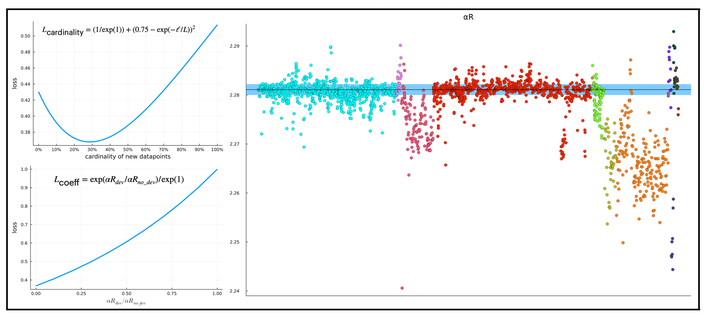
	α0	αI	αB	αn	αP	αR	$\alpha\kappa$	$\alpha\epsilon$	αM
DB2P8	0.056	0.93	0.15	0.41	-0.69	1.97	0.78	0.58	0.15
ITPA20-IL	0.067	1.29	-0.13	0.147	-0.64	1.19	0.67	0.56	0.3

Deviation: Determining most influential data points in the reduction of αR

- Use the initial database regression values as the baseline.
- Separately add, point by point, the new data
- Re-evaluate the regression each time and plot the changes in coefficient values.



Deviating points: Each subplot shows a separate coefficient of the regression, colour coded by machine. x-axis: data entry index, y-axis: coefficient value



a) Loss function elements: Reduction of the cardinality is one of the targets therefore the loss function is designed to have a minimum at just below a third of the overall new data points. **b) The optimised allowed deviation:**

Aim: Determine the smallest set of points which cause the largest difference in αR

Multicollinearity: Resulting in too many degrees of freedom

The issue of multicollinearity between the predictor variables results in a loss of physical understanding of the coefficients which can deviate widely.

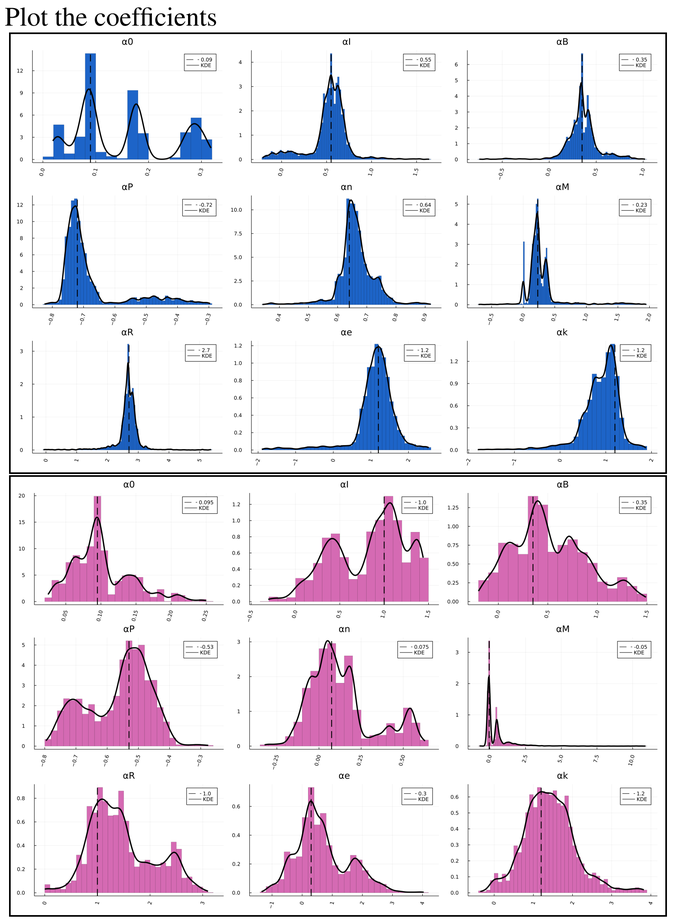
Scan the QR codes

Left: uncorrelated predictor variables, such that the data-points span the whole space providing a semi-stable regression hyperplane surface; **right:** highly correlated predictor variables reducing the effective degrees of freedom from three to two such that the regression hyperplane is unstable in the dimension perpendicular to the correlated data.

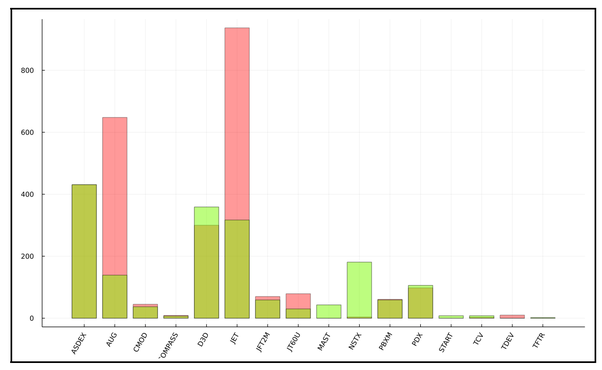
	lnτ	lnI	lnB	lnn	lnP	lnR	lnκ	lnϵ	lnM
lnτ	1.0	0.92	0.35	-0.01	0.69	0.76	0.57	0.64	0.52
lnI	0.85	1.0	0.39	0.19	0.85	0.61	0.68	0.79	0.59
lnB	0.37	0.34	1.0	0.36	0.27	0.01	-0.07	0.25	0.26
lnn	-0.15	0.08	0.27	1.0	0.23	-0.49	0.33	0.18	0.27
lnP	0.55	0.82	0.34	0.2	1.0	0.58	0.67	0.67	0.57
lnR	0.82	0.69	0.37	-0.36	0.62	1.0	0.32	0.28	0.26
lnκ	0.28	0.5	-0.21	0.25	0.46	0.06	1.0	0.44	0.5
lnϵ	0.02	0.3	-0.54	0.07	0.18	-0.27	0.5	1.0	0.47
lnM	0.23	0.33	0.18	0.32	0.27	0.01	0.33	0.22	1.0

Powerset Regression

- Subdivide database by machine
- Regress on the powerset of the subdivided datasets
- Plot the coefficients



Power regression: a) Data optimised to not deviate significantly from the baseline αR; **b)** Data optimised to reduce the coefficient αR the most



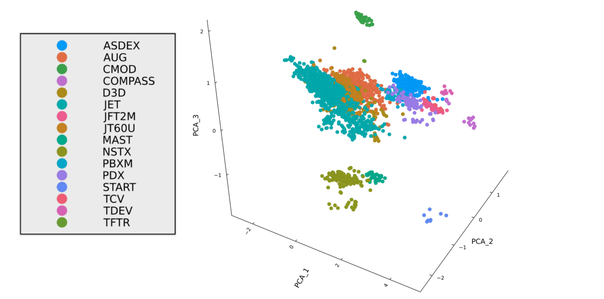
Tokamak representation: green:) Deviation data; **red:)** No deviation data

- Both groups represent all tokamaks equally with the exception of AUG and JET
- Deviation data clearly fluctuates the most widely.
- We expect this occurs due to the multicollinearity
- It seems we can conclude the reduction in αR is related with the multicollinearity of the data.

Aim: Determine methods that can tackle the issue of multicollinearity.

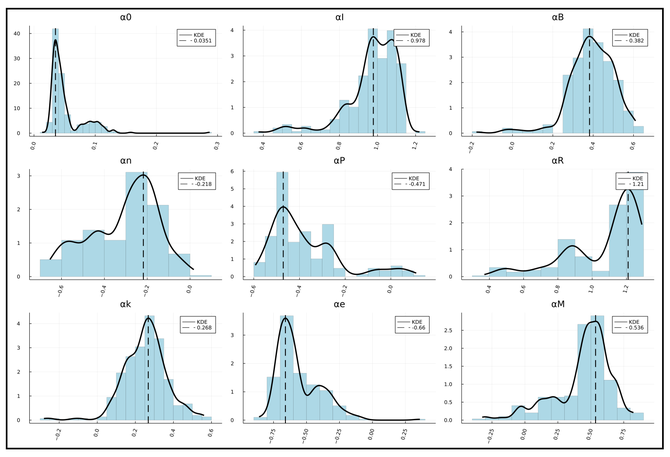
Robust regression: Eliminating outlier and multicollinearity effects.

Principal component Regression



Determining the best number of dimensions is then found by reducing the **coefficient of variation** of the powerset regression:

$$\sum_{i=1}^{n} \frac{\sigma(\alpha_i)}{\mu(\alpha_i)}$$



	α0	αI	αB	αn	αP	αR	$\alpha\kappa$	$\alpha\epsilon$	αM
DB2P8 (5)	0.0452	1.33	-0.151	0.147	-0.601	1.16	0.517	-0.512	0.341
STDDB4V5 (4)	0.0351	0.978	0.382	-0.218	-0.471	1.21	0.268	-0.66	0.536
SELDB5 (4)	0.037	1.1	0.39	-0.22	-0.59	1.1	0.27	-0.59	0.55

References

1) Verdoolaege, G., Kaye, S.M., Angioni, C., Kardaun, O.J.W.F., Maslov, M., Romanelli, M., Ryter, F., Thomsen, K., ASDEX Upgrade Team, EUROfusion MST1 Team, JET Contributors, 2021. The updated ITPA global H-mode confinement database: description and analysis. Nucl. Fusion 61, 076006. https://doi.org/10.1088/1741-4326/abdb91
2) Freidberg, J. (2007). Plasma Physics and Fusion Energy. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511755705
3) ITER Physics Expert Group on Confinement and Transport et al 1999 Nucl. Fusion 39 2175. https://iopscience.iop.org/article/10.1088/0029-5515/39/12/302