変数選択手法っていろいろあるけど何を使えばいいの？

明治大学 理工学部 応用化学科 専任講師　　金子 弘昌

2017年5月16日 (火)

第5回ケモインフォマティクス若手の会@渋谷ヒカリエ

自己紹介

✓明治大学 応用化学科 応用化学科 専任講師　金子 弘昌

- データ化学工学研究室
- 部屋: 第二校舎D館D409
- E-mail: hkaneko@meiji.ac.jp
- Tel: 044-934-7197

✓生年月日

- 1985年1月9日 (32歳)

- 同い年の芸能人
  - 綾瀬はるか、松山ケンイチ、松下奈緒、TAKAHIRO、木村カエラ、速水もこみち、島袋寛子　など

自己紹介

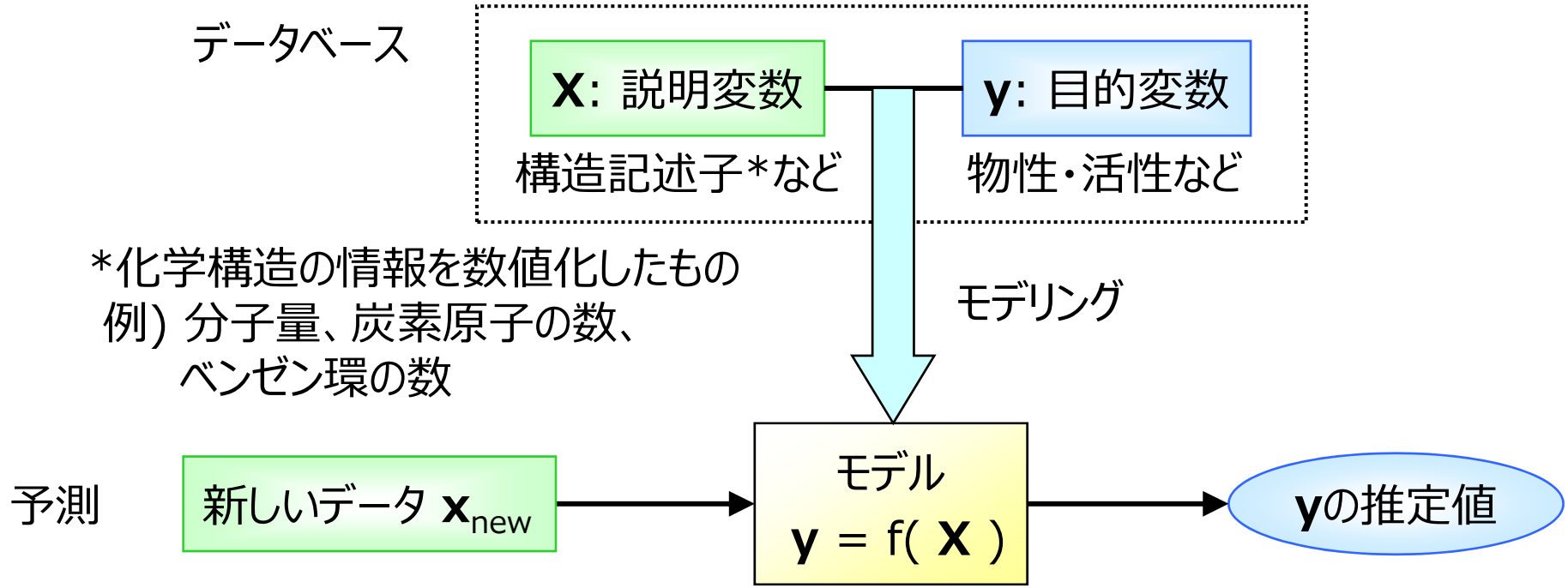
✓出身地
- 栃木県足利市
  - あしかがフラワーパーク
  - 足利学校
  - 相田みつを
  - ココ・ファーム・ワイナリー

✓経歴
- 足利高校
  → 東京大学
  → 東京大学大学院 修士課程
  → 東京大学大学院 博士課程
  → 東京大学大学院 助教
  → 明治大学へ

✓趣味
- ソフトテニス
- ジョギング
- 読書 (マンガ含む)
- 映画鑑賞

✓家族
- 妻１人、娘１人の３人家族

QSPR・QSAR



例) **X**: 2変数
データ数: 3
線形モデル

	x_1	x_2	y
データ1	1	2	5.1
データ2	2	1	3.9
データ3	3	3	9.2

回帰モデル

$y = x_1 + 2x_2 +$ 誤差

エクセルのファイルだとデータはこんな感じ

y　　X

	logS	MinAbsPa	NumRadic	HeavyAtor	MaxAbsES	MaxAbsPa	MaxEStat	MinPartia	ExactMolV	MolWt	NumValer	MinEState	M
CC(N)=O	1.58	0.21379	0	54.028	9.222222	0.369921	9.222222	-0.36992	59.03711	59.068	24	-0.33333	0
CNN	1.34	0.001725	0	40.025	4.597222	0.271722	4.597222	-0.27172	46.0531	46.073	20	1.652778	1
CC(=O)O	1.22	0.299685	0	56.02	9	0.481433	9	-0.48143	60.02113	60.052	24	-0.83333	0
C1CCNC1	1.15	0.004845	0	62.051	3.222222	0.316731	3.222222	-0.31673	71.0735	71.123	30	1.25	
NC(=O)NO	1.12	0.335391	0	72.023	9.229167	0.349891	9.229167	-0.34989	76.02728	76.055	30	-0.93981	0
OCC(O)CO	1.12	0.100047	0	84.03	8.166667	0.393593	8.166667	-0.39359	92.04734	92.094	38	-0.9537	0
CC(=O)N(C)C	1.11	0.218425	0	78.05	10.06944	0.349064	10.06944	-0.34906	87.06841	87.122	36	0.092593	0
c1ccnnc1	1.1	0.049569	0	76.058	3.534722	0.159176	3.534722	-0.15918	80.03745	80.09	30	1.638889	1
c1cncnc1	1.1	0.114757	0	76.058	3.673611	0.244832	3.673611	-0.24483	80.03745	80.09	30	1.5	
OCC(O)C(O)C(O)C	1.09	0.110579	0	168.06	8.95662	0.393579	8.95662	-0.39358	182.079	182.172	74	-1.66931	0
CC(N)CC(=O)O	1.08	0.304406	0	94.049	9.729722	0.481188	9.729722	-0.48119	103.0633	103.121	42	-0.83796	0
C1CNCCN1	1.07	0.007723	0	76.058	3.222222	0.314206	3.222222	-0.31421	86.0844	86.138	36	1.138889	1
C1CCNCC1	1.07	0.004891	0	74.062	3.284722	0.316733	3.284722	-0.31673	85.08915	85.15	36	1.25	
Oc1ccncc1	1.02	0.118146	0	90.061	8.592222	0.50785	8.592222	-0.50785	95.03711	95.101	36	0.259259	0
Oc1ccccn1	1.02	0.210156	0	90.061	8.521111	0.493268	8.521111	-0.49327	95.03711	95.101	36	0.071759	0
O=C(O)CCCC(=O)	1	0.302856	0	124.051	9.787862	0.48123	9.787862	-0.48123	132.0423	132.115	52	-0.94792	0
CN1CCOCC1	1	0.05935	0	90.061	5.098194	0.378793	5.098194	-0.37879	101.0841	101.149	42	0.913194	0
COC(N)=O	0.97	0.403708	0	70.027	9.368056	0.453034	9.368056	-0.45303	75.03203	75.067	30	-0.74537	

http://datachemeng.wp.xdomain.jp/pythonassignment/ からダウンロード可能

回帰モデルの推定性能を上げたい！

✓構造記述子の選択 (変数選択)

- ノイズ・雑音のような変数を削除することでモデルの推定性能向上
- 単純に変数の数を減らしたい人もいる

✓外れサンプルの削除

✓オーバーフィッティング、アンダーフィッティングの解消

変数選択の方法　モデリング不要

✓ランダムに選択

✓似ている変数の組の１つを削除

- 相関係数の高い (0.9, 0.99とか以上の) 変数の組の１つを削除

変数選択の方法　モデリング必要

✓PLS-β
- PLS(Partial Least Squares)の標準回帰係数の絶対値の小さい変数を削除

✓PLS-VIP
- PLS後のVIP (Variable Importance in Projection) の値が小さい変数を削除

✓LASSO (Least Absolute Shrinkage and Selection Operator)
- Yの誤差と一緒に回帰係数の値も小さくすることで、いくつかの回帰係数が0になることを利用

変数選択の方法　たくさんモデリング必要

✓Stepwise

- 一つずつ変数を追加・削除を繰り返して、ある指標の値が大きくなるように変数選択

✓GAPLS

- 遺伝的アルゴリズム(Genetic Algorithm, GA) とPLSとを組み合わせた手法、ある指標の値が大きくなるように変数選択

- ある指標・・・クロスバリデーション後の r^2 (r^2_{CV}) とか

で、どれを使えばいいの？

✓ランダムに選択は単純すぎる？

✓PLS-β・PLS-VIPはどっちがいいの？

✓最近LASSOをよく見るけどどうなの？

✓GAPLSは時間かかりそうだけどよさげ？指標次第？

調べてみました

✓QSPR: 1290個の化合物に関するlogS[1]

✓QSAR(toxicity): 1,093 化合物の T. Pyriformis に対する 50%阻害濃度 ($pIGC_{50}$) [2]

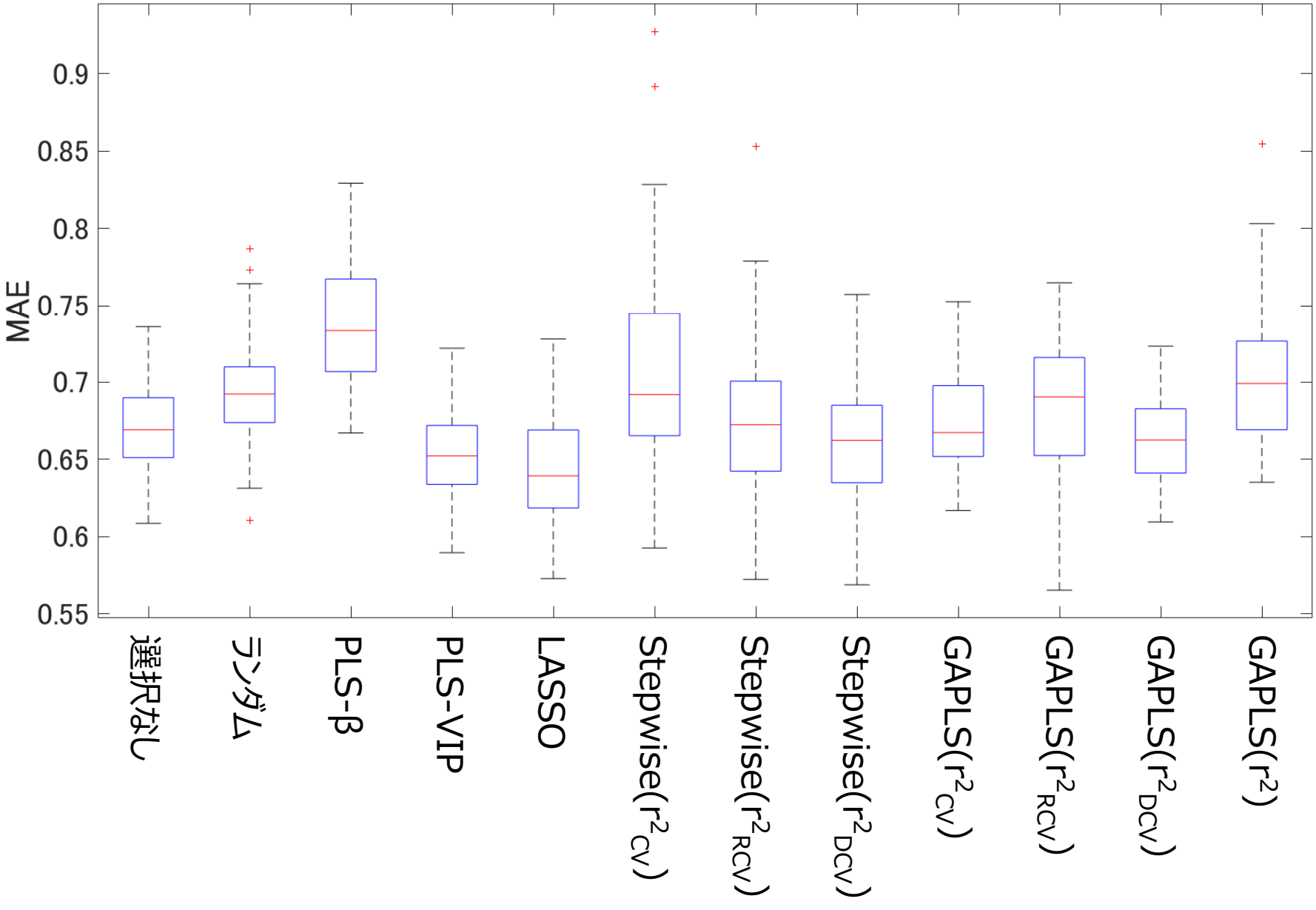
- 構造記述子: RDKit[3]で計算した 206記述子

- モデル構築用サンプル数: 30, 100, 500
- ランダムにサンプルを選択
  - それ以外のサンプルがモデル検証用サンプル

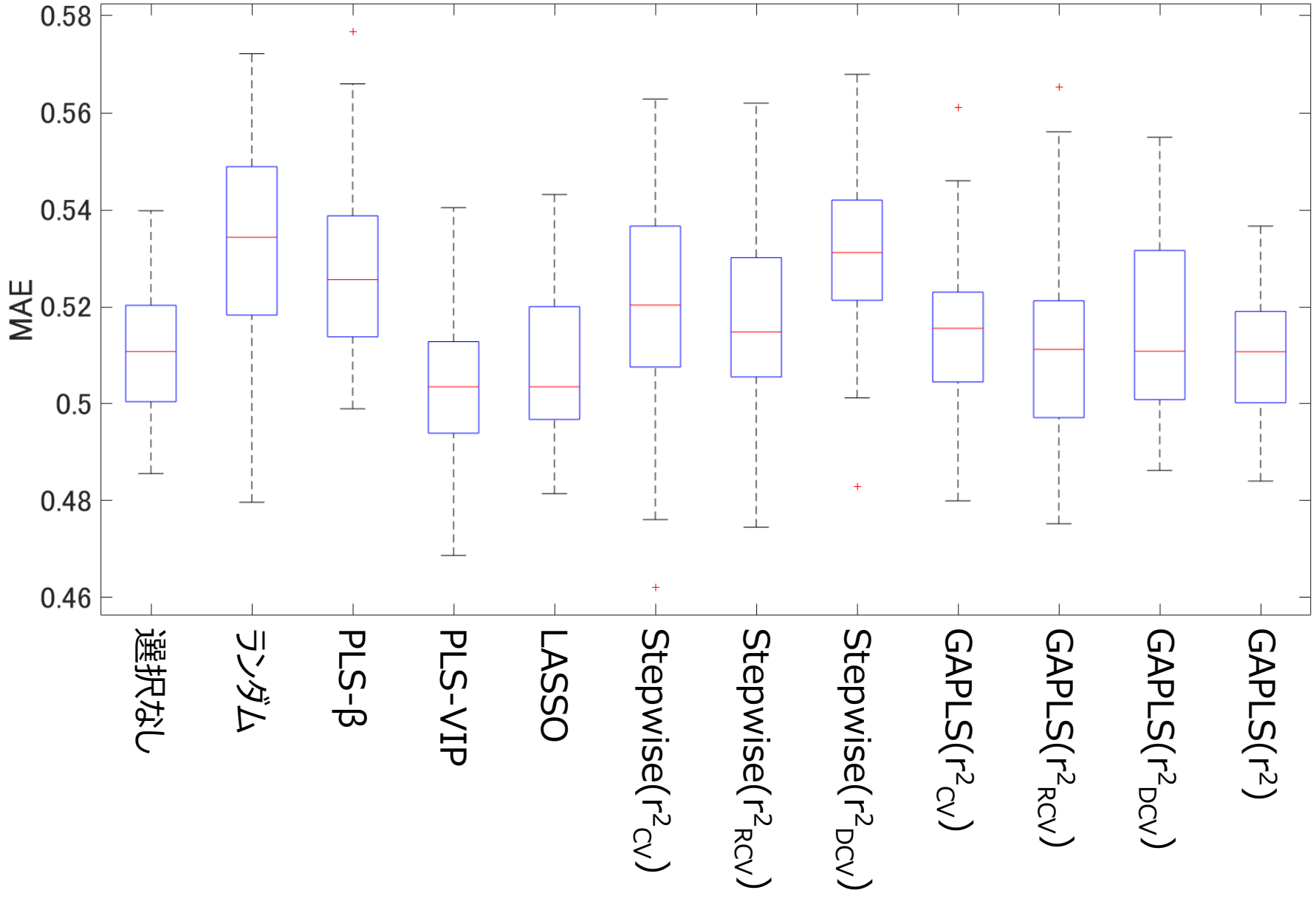
- 50回繰り返して、モデル検証用サンプルの MEA を比較
  - MAE (Mean Absolute Error): 誤差の絶対値の平均

[1] T.J. Hou, K. Xia, W. Zhang, X.J. Xu, J. Chem. Inf. Comput. Sci., 44, 266, 2004.
[2] http://www.cadaster.eu/node/65
[3] http://www.rdkit.org/

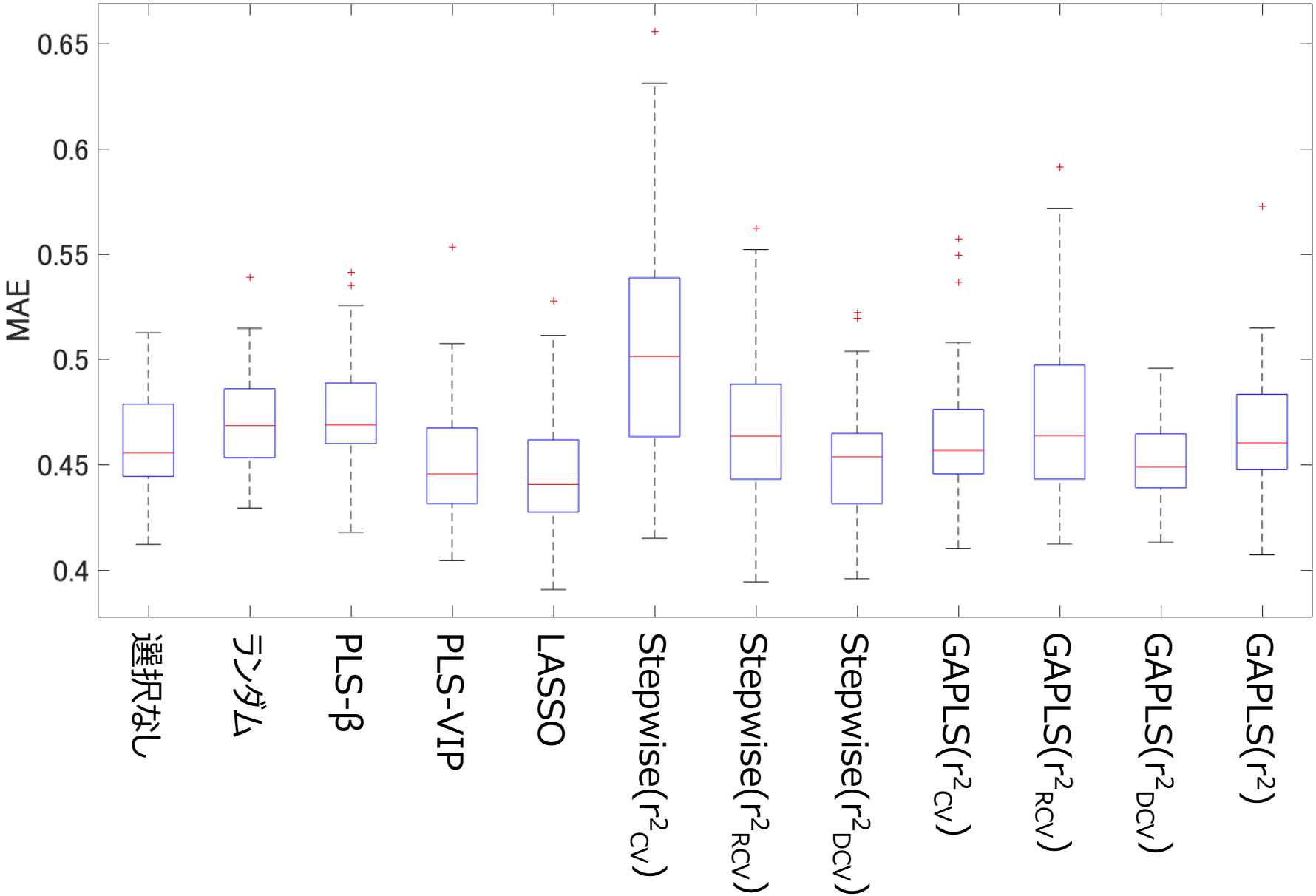
QSPR 100サンプル



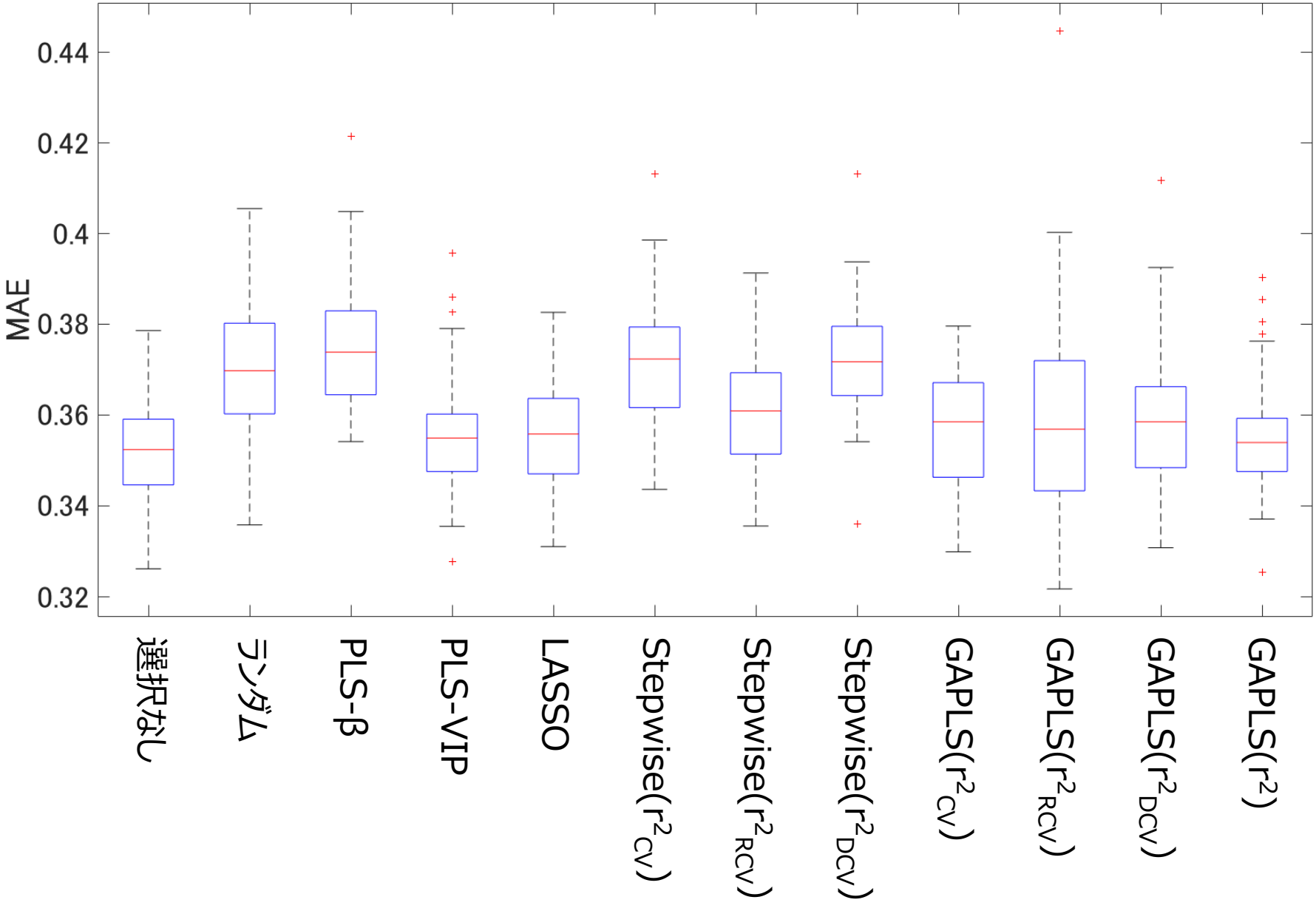
QSPR 500サンプル



QSAR 100サンプル



QSAR 500サンプル



まとめ

✓説明変数を選択しないときと比べて、推定性能の大きな向上は見られなかった

✓特にサンプル数が大きいときに、GAにおける指標の違いによって結果に大きな差異はなかった

- オーバーフィットしそうな r^2 でもサンプルが多いと他の指標とあまり変わらなかった

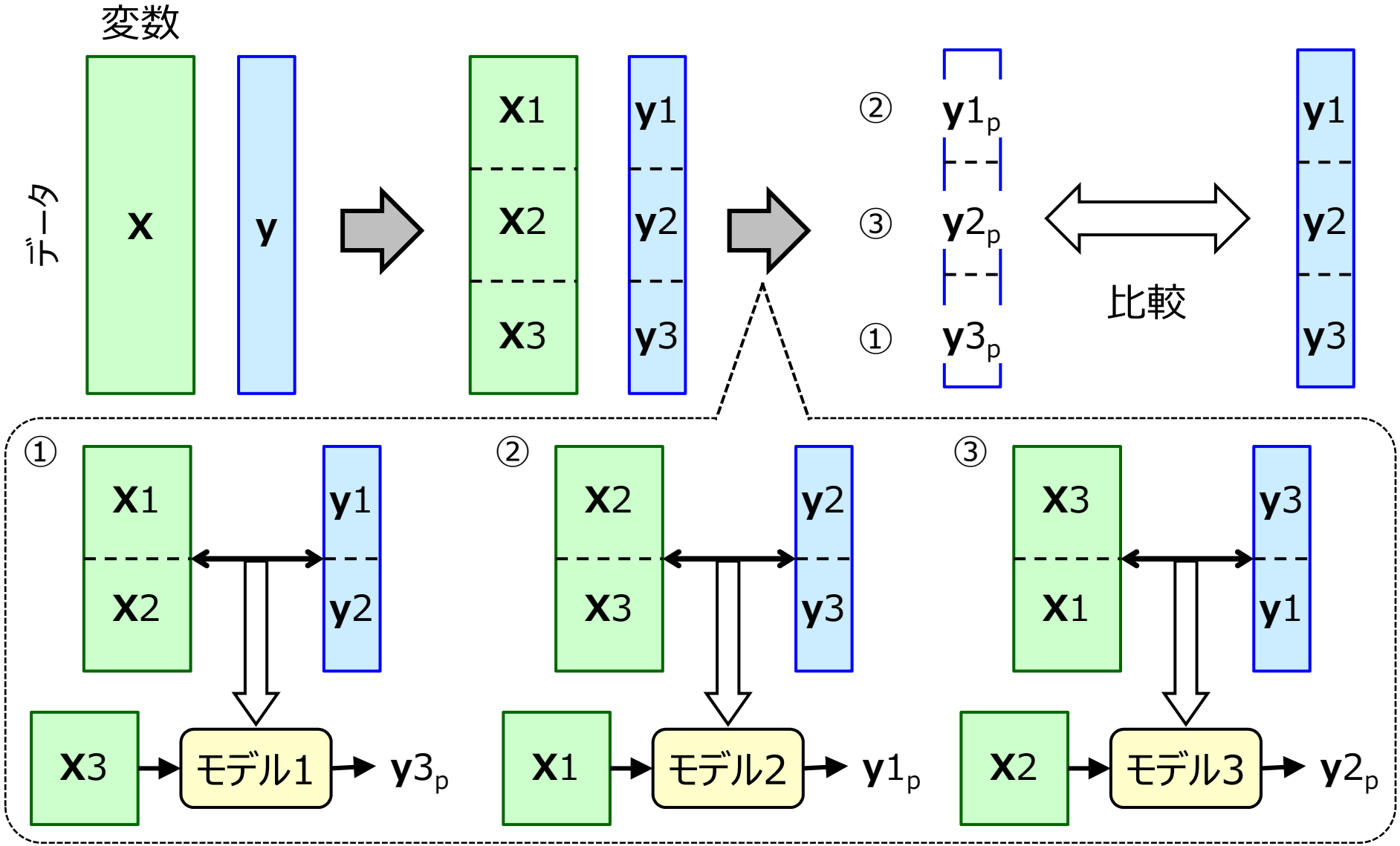
✓PLS-VIP と LASSO が良さそう

補足資料 設定

- ✓ランダム: ランダムに半分選択
- ✓PLS-β, PLS-VIP: 中央値以上を選択
- ✓LASSO: λ…0.1, 0.2, …, 4.9, 5 の中で r^2_{CV} が最大のものを選択
- ✓Stepwise: 変数増減法
- ✓GA: 世代数300、個体数300

補足資料 クロスバリデーション (CV)

✓例) 3-fold クロスバリデーション (Cross-Validation, CV)



補足資料 クロスバリデーション (CV)

✓今回は 5-fold クロスバリデーション を使用

補足資料 クロスバリデーション繰り返し(RCV)

✓クロスバリデーション繰り返し
(Repeated Cross-Validation, RCV) [1]
- クロスバリデーションを繰り返して、r^2_{CV} の平均値に用いる (r^2_{RCV})

✓クロスバリデーションの結果を指標にすると(たとえば r^2_{CV})、
分割の仕方でたまたま良い結果になったり、逆に悪い結果になったりする

✓クロスバリデーションを繰り返し行い、それらを平均化することで、
“たまたま” を防ぐ
- 今回は 30 回

[1] P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometr., 23, 160-171, 2009.

補足資料 ダブルクロスバリデーション(DCV)

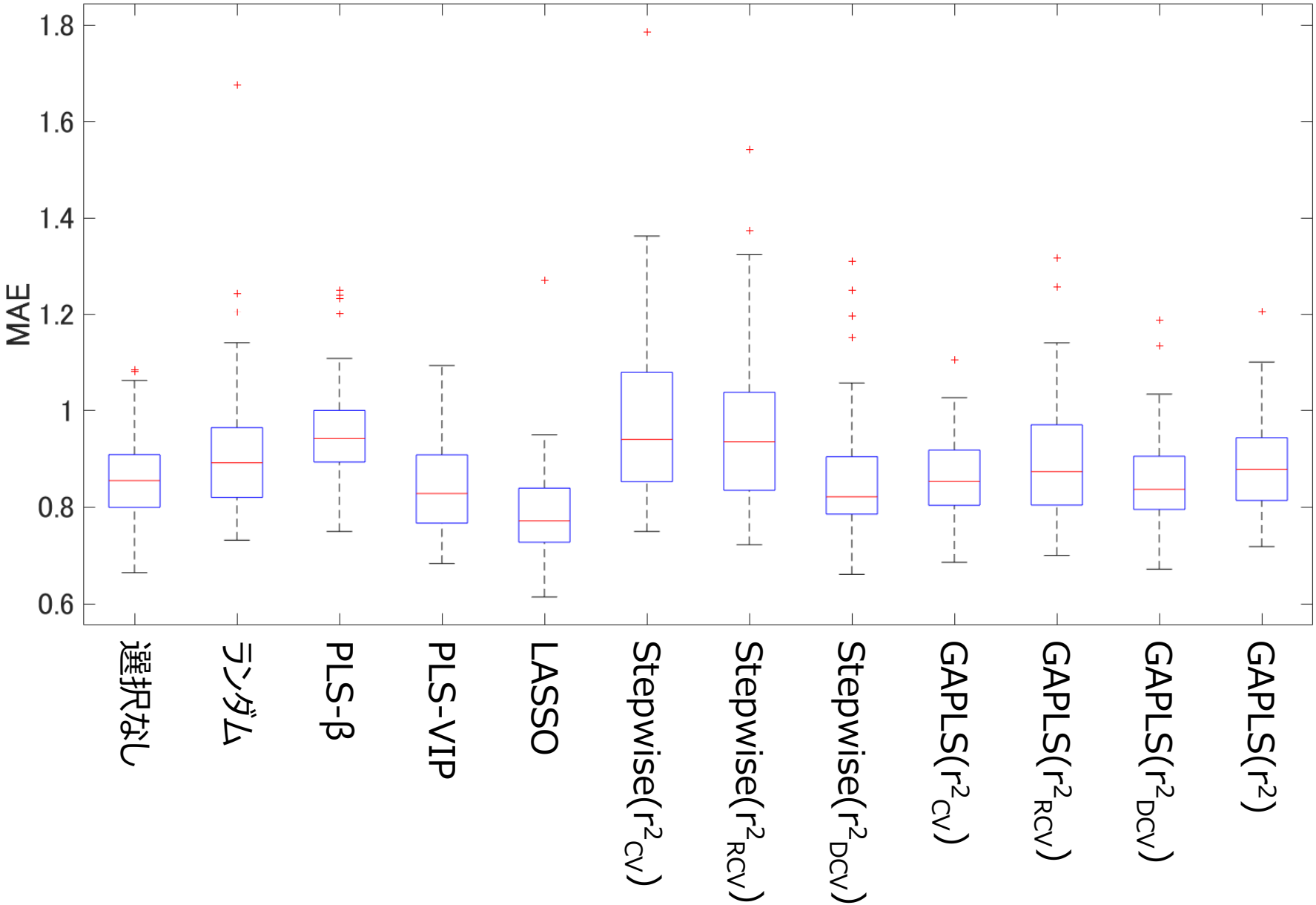
✓ダブルクロスバリデーション (Double Cross-Validation, DCV) [1]
- クロスバリデーションを入れ子にして、二重に行うこと

✓クロスバリデーションの結果を指標にすると(たとえば r^2_{CV})、
オーバーフィッティングを起こす可能性がある
- PLSでクロスバリデーションの結果がよくなるように成分数を選ぶため

✓内側のクロスバリデーションで成分数を最適化し、
外側のクロスバリデーションの結果を指標にする (r^2_{DCV})

[1] P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometr., 23, 160-171, 2009.

補足資料 QSPR 30サンプル



補足資料 QSAR 30サンプル