Generative Topographic Mapping (GTM) でデータの可視化・回帰分析・モデルの逆解析を一緒にやってみた

明治大学 理工学部 応用化学科 専任講師　　金子 弘昌

2018年5月22日 (火)

第7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
- Website: https://datachemeng.com/
  - 「明治 金子」で検索
- Twitter: @hirokaneko226
- 部屋: 第二校舎D館D409
- E-mail: hkaneko@meiji.ac.jp
- Tel: 044-934-7197
- オンラインサロンはじめました！
  - https://datachemeng.com/onlinesalon/

✓生年月日
- 1985年1月9日 (33歳)

自己紹介

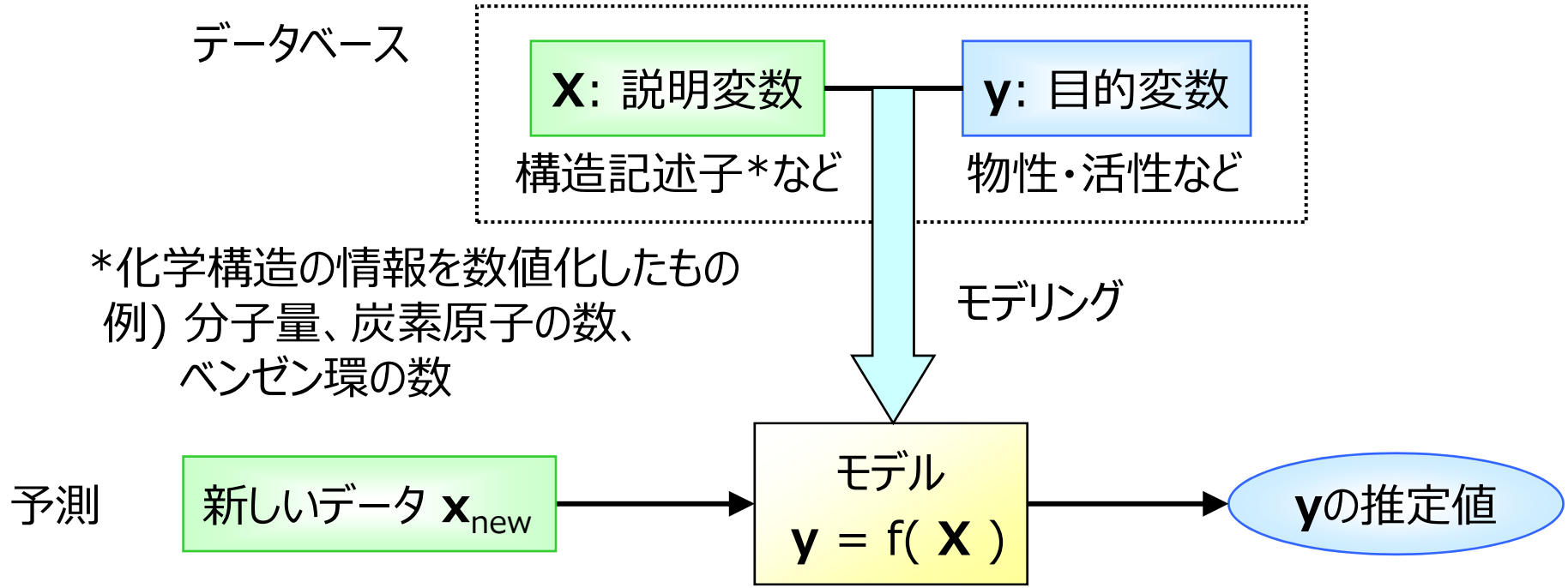
✓出身地
- 栃木県足利市
  - あしかがフラワーパーク
  - 足利学校
  - 相田みつを
  - ココ・ファーム・ワイナリー

✓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	

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

データ解析の一般的な流れ

✓データ収集
- データの前処理

✓データの可視化：主成分分析 (PCA) など

✓モデル構築：サポートベクターマシン・回帰 (SVM, SVR) など
- 今回は回帰分析
- モデルの検証

✓モデルの逆解析
- y から X を推定
- 順解析を繰り返す、とか

一気にデータ解析できないか？

✓データの可視化・モデル構築・モデルの逆解析は
別々に行われている

- たとえば、データの可視化の結果とモデル構築の結果との間に関係があるわけではない
- 別々にやるのは面倒 (?)

✓データの可視化・モデル構築・モデルの逆解析をつなげて
一気にできないか？

- データセットは一つなので、この方が効率的

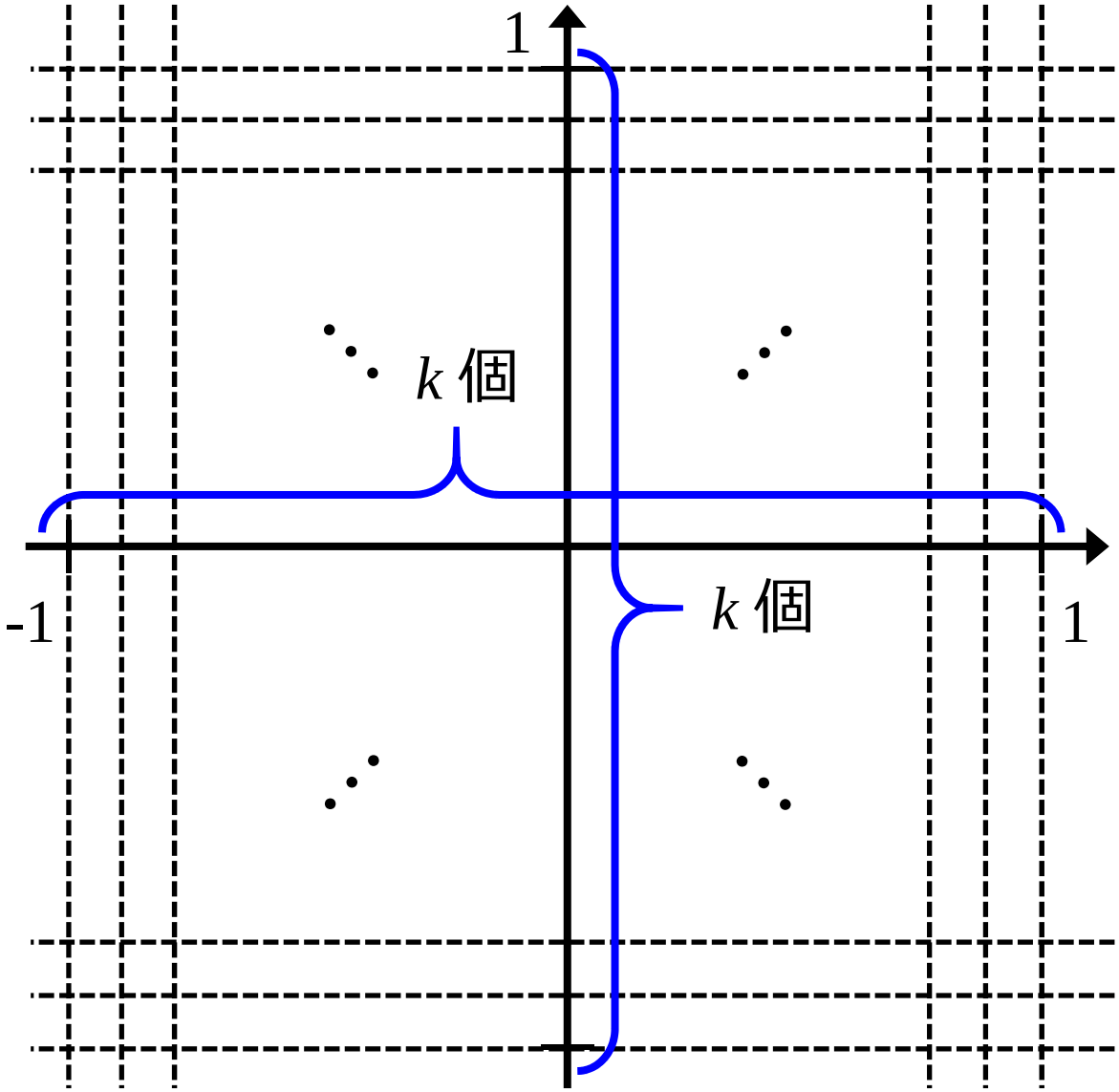
GTMに着目

✓Generative Topographic Mapping (GTM) [1]

- データを可視化・見える化するための非線形手法
- 主成分分析などとは異なり、はじめに二次元平面の座標を作り、それを実際の多次元空間のサンプルに合わせ込む
- ゴム状のシート (二次元平面) を曲げたり伸び縮みさせたりしながら、多次元空間にあるサンプルを通るようにシートを置き、そのシートにサンプルを射影するような手法
- 自己組織化マップ (Self-Organizing Map, SOM) のいろいろな問題点を解決した、上位互換の手法
- ２次元平面において近いサンプル同士は、多次元空間においても近いことが補償されている
- 詳細は https://datachemeng.com/generativetopographicmapping/

[1] C.M. Bishop, M. Svensén, C.K.I. Williams,Neural Comput. 1998, 10, 215–234.

GTMの概要



✓二次元マップ

✓各グリッドに正規分布

✓データセットに
正規分布の重ね合わせ
がフィットするように
マッピング

✓データセットは、
正規分布の重ね合わせ
で表現される

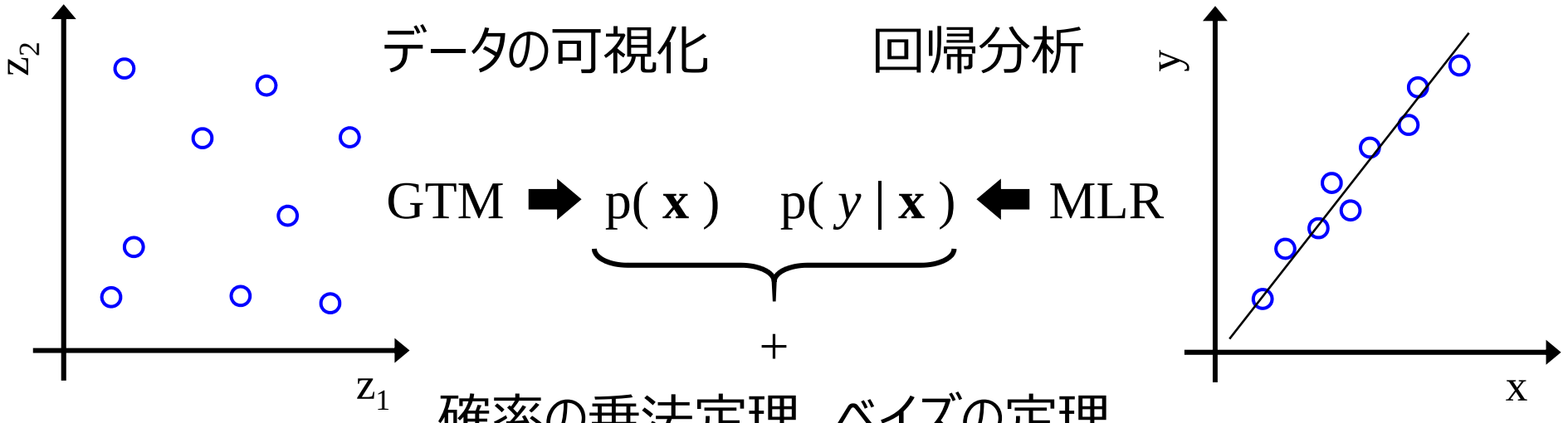
GTMで回帰分析・モデルの逆解析も！

- ✓Generative Topographic Mapping-Multiple Linear Regression (GTM-MLR)
- ✓Generative Topographic Mapping Regression (GTMR)

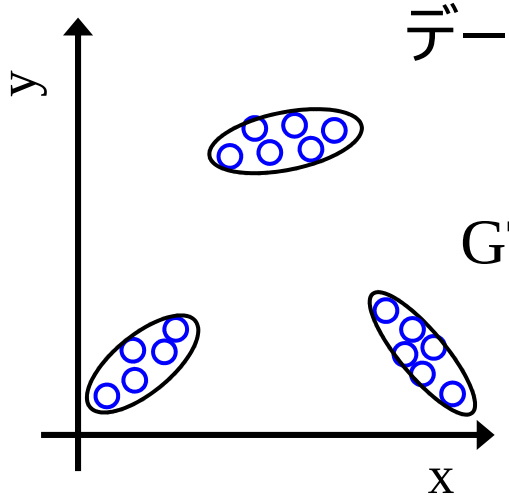
を提案、論文投稿中

- ✓GitHub: https://github.com/hkaneko1985/gtm-generativetopographicmapping にてPython, MATLABコードを公開

GTM-MLR


z_2
z_1
データの可視化
回帰分析
y
x
GTM
p(**x**)
p(y | **x**)
MLR
+
確率の乗法定理、ベイズの定理
p(**x** | y)
モデルの逆解析

GTMR



データの可視化

GTM ➡ $p(\mathbf{x}, y)$ +

確率の乗法定理
ベイズの定理

回帰分析

$p(y \mid \mathbf{x})$

$p(\mathbf{x} \mid y)$

モデルの逆解析

✓GTMで同時確率分布を計算するときには、X, y は区別していない

✓y が複数あっても、yの間の関係を考慮したモデル構築・
モデルの逆解析ができる

数値シミュレーションデータで検証

✓線形

✓非線形

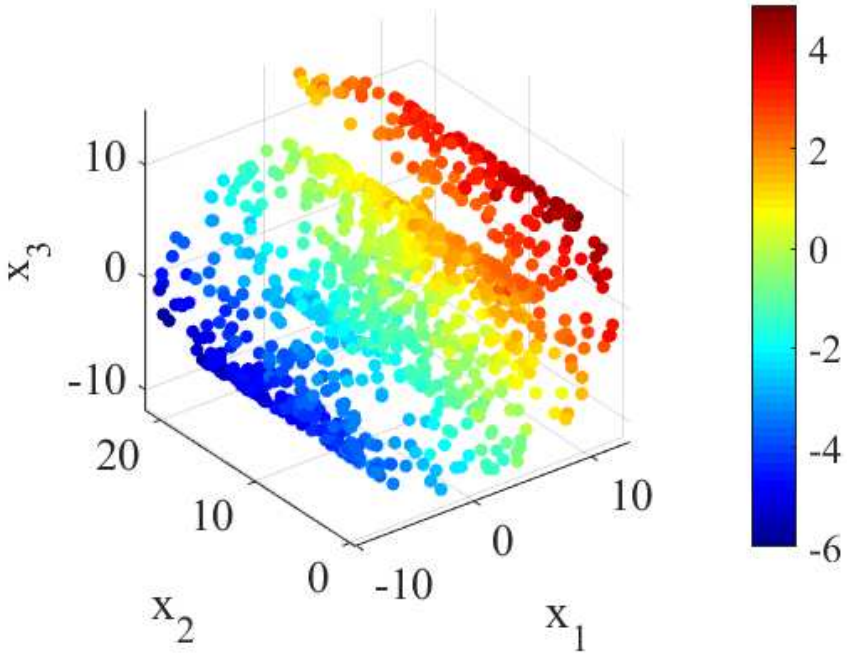
線形の数値シミュレーションデータ

✓線形

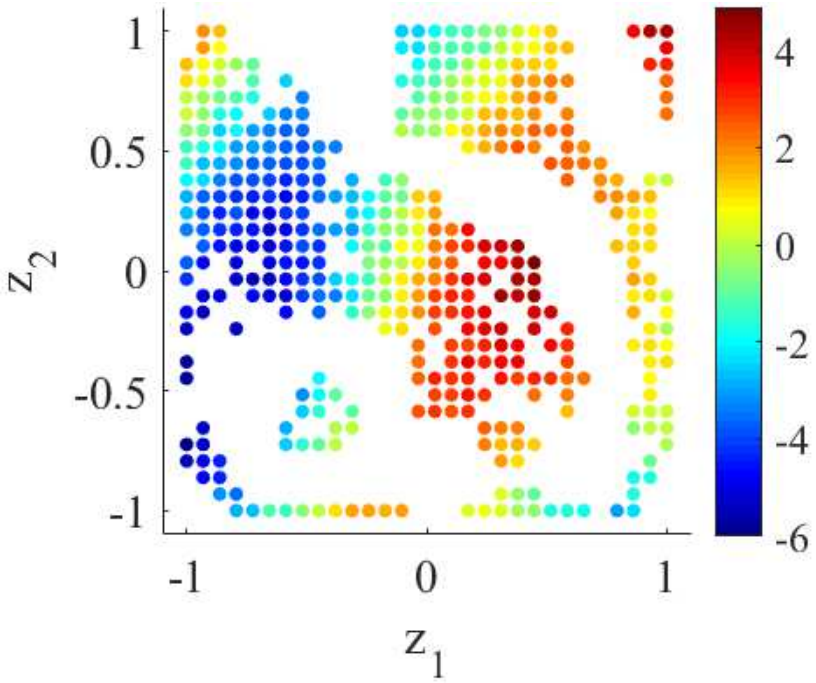
- Swiss roll
  - sklearn.datasets.make_swiss_roll

$$y = 0.3x_1 - 0.2x_2 + 0.1x_3$$

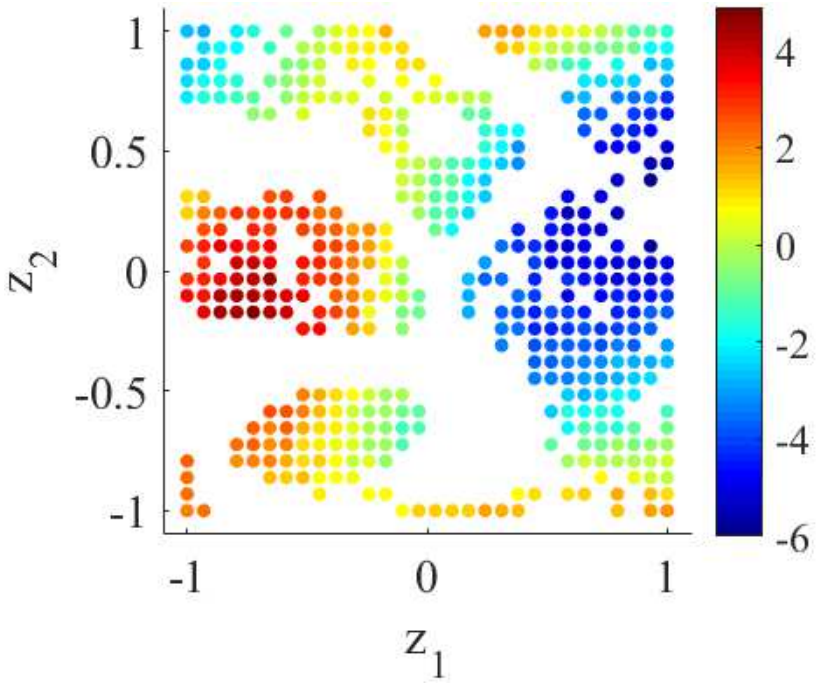
- 正規分布に従うノイズを追加
  - トレーニングデータ: 500
  - テストデータ: 500



可視化の結果 (線形)



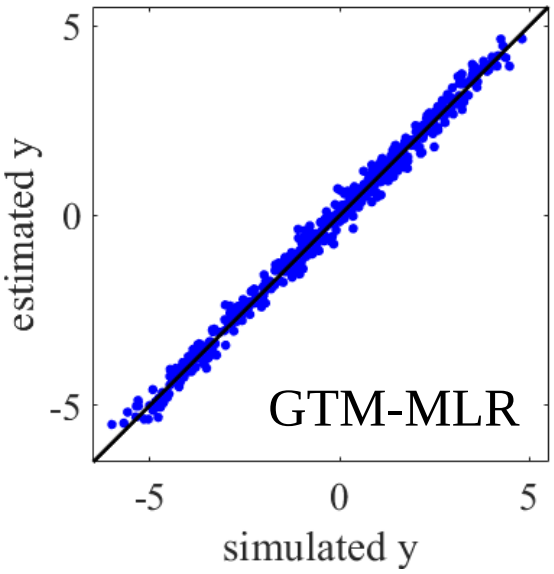
GTM-MLR

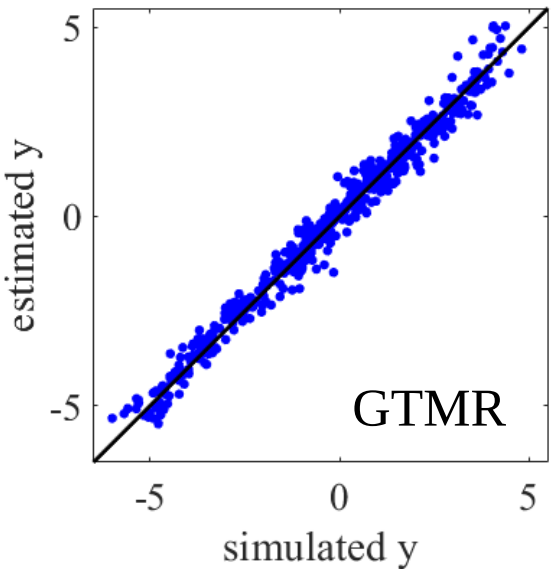


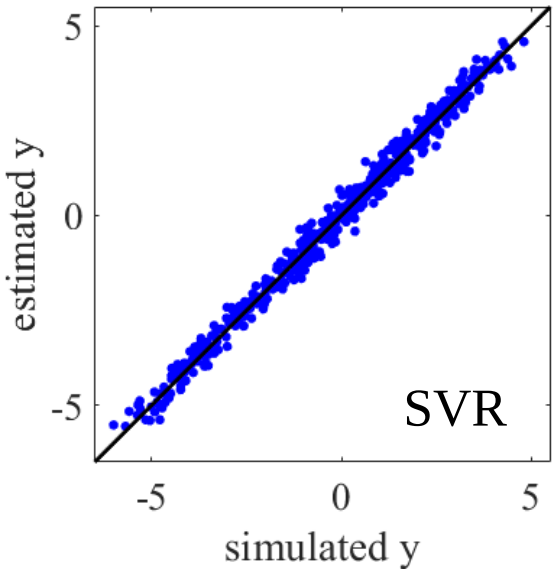
GTMR

回帰分析 テストデータの推定結果 (線形)

	GTM-MLR	GTMR	SVR
r^2	0.991	0.980	0.989
MAE	0.192	0.276	0.209







モデルの逆解析の結果 (線形)

	目標のy	Xの推定値			GTMマップ上のX推定値		実際のy
GTM-MLR	4	11.90	4.22	4.28	0.31	−0.03	4.01
	0	4.85	−0.20	−7.28	−0.03	0.45	0.02
	−5	−5.74	17.51	−7.62	−0.66	−0.17	−5.00
GTMR	4	6.91	3.02	11.32	−0.86	−0.10	4.04
	0	9.59	11.8	−8.41	0.72	1.00	0.01
	−5	−3.67	18.21	−10.28	0.86	0.52	−4.98

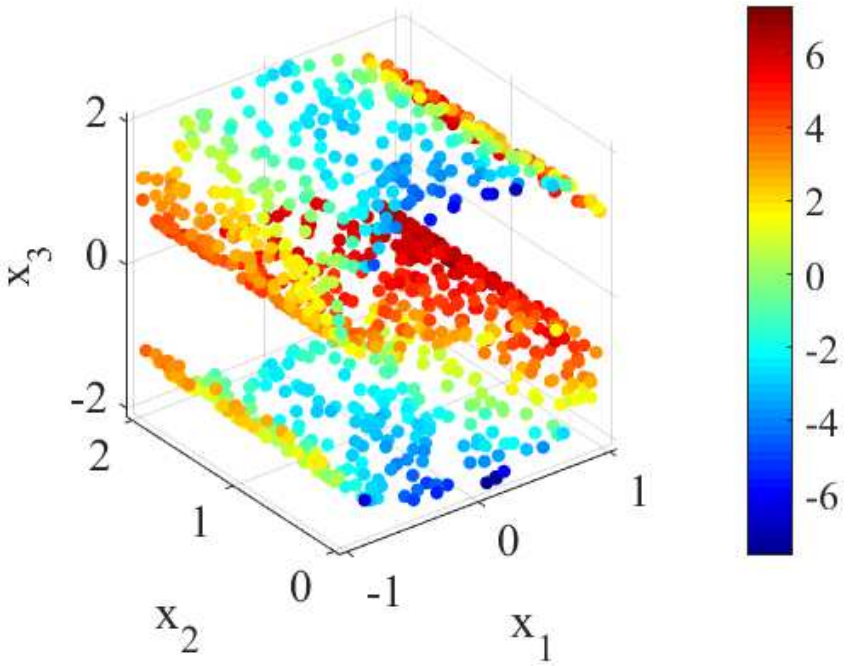
非線形の数値シミュレーションデータ

✓非線形

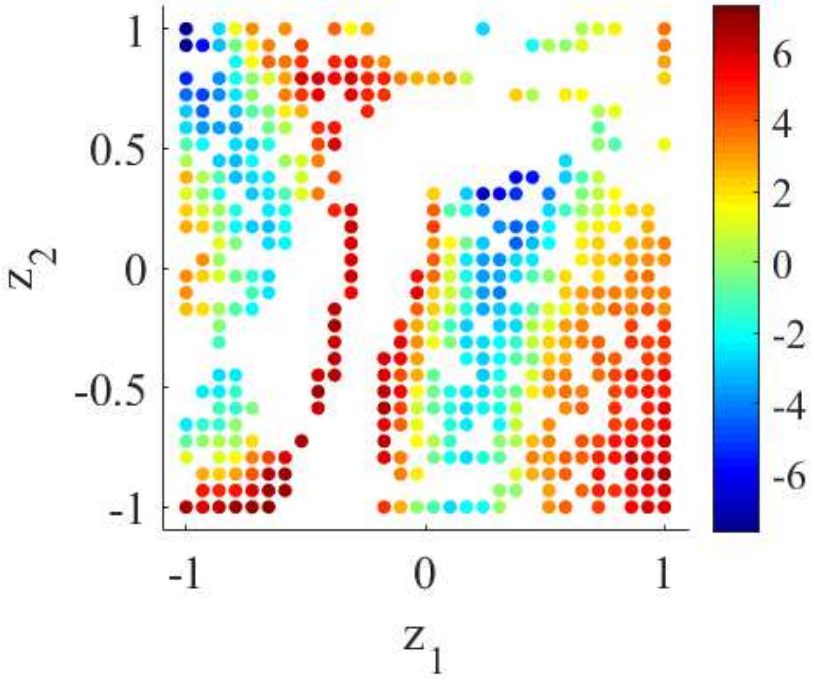
- S-curve
  - sklearn.datasets.make_s_curve

$$y = \arcsin\left(x_1\right) + \log\left(x_2\right) - 0.5{x_3}^4 + 5$$

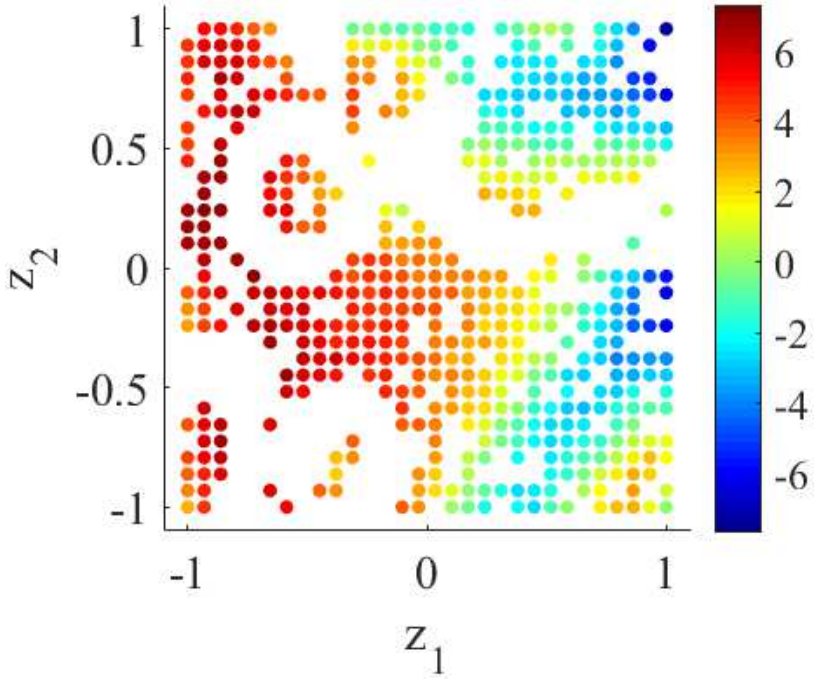
- 正規分布に従うノイズを追加
  - トレーニングデータ: 500
  - テストデータ: 500



可視化の結果 (非線形)



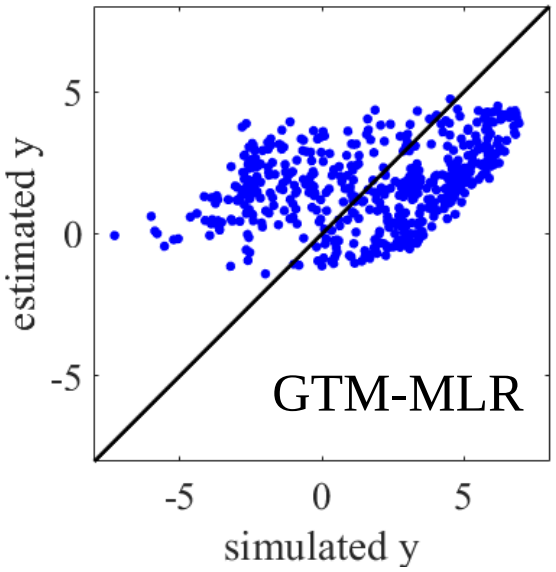
GTM-MLR

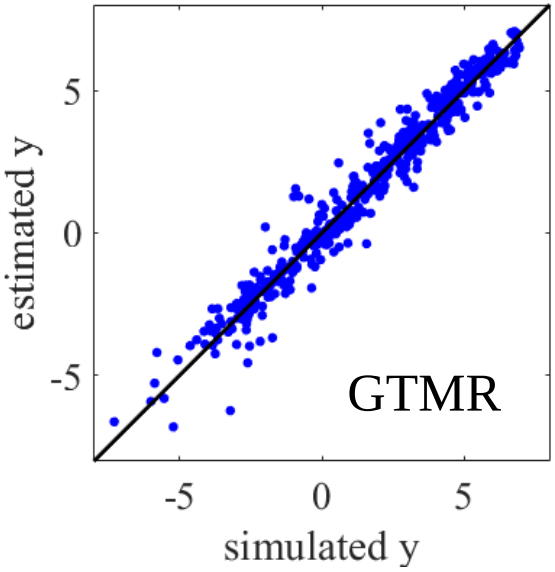


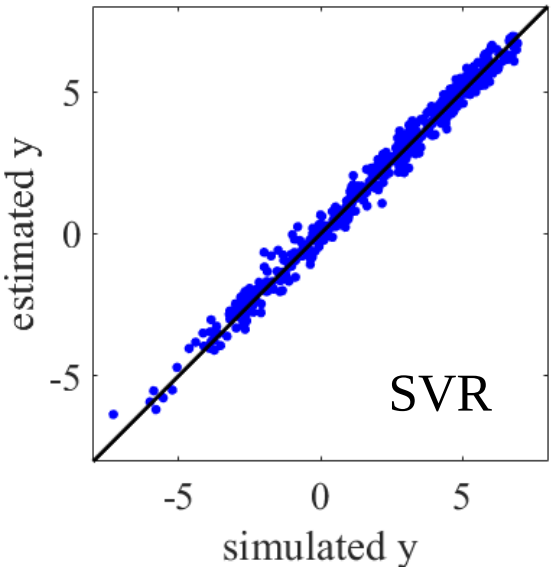
GTMR

回帰分析 テストデータの推定結果 (非線形)

	GTM-MLR	GTMR	SVR
r^2	0.127	0.964	0.987
MAE	2.58	0.433	0.281







モデルの逆解析の結果 (非線形)

	目標のy	Xの推定値			GTMマップ上のX推定値		実際のy
GTMR	6	0.70	1.32	−0.31	−0.59	−0.17	6.1
	−1	−0.59	1.08	−1.82	0.52	0.52	−1.0
	−5	−0.31	0.07	1.94	0.93	−0.10	−5.1

QSPR、QSARへの応用

✓化合物データの収集

✓記述子の計算

✓GTM-MLR, GTMRによるモデリング

✓物性もしくは活性の目標値を設定

✓GTM-MLRモデル、GTMRモデルの逆解析により目標値になりうる記述子の値を推定

✓記述子の推定値にもとづいて化学構造の探索

✓詳しくは論文が掲載されましたらで、、、
もしくは金子研オンラインサロンで

- https://datachemeng.com/onlinesalon/

まとめ

✓データの可視化、回帰分析、モデルの逆解析をつなげる、
GTMに基づく手法を開発した

✓GTM-MLR
- 回帰モデルの性能はMLRモデルに依存
- X と y の間の非線形性に対応できない

✓GTMR
- y が複数のときでも対応可能
- X と y の間の非線形性に対応できる
- ノイズに弱い

補足資料 ハイパーパラメータ候補

✓マップサイズ: 30×30

Hyperparameter	Candidate
Number of RBFs $p^{0.5}$	2, 4, ..., 18, 20
Variance of each RBF σ	2^{-5}, 2^{-4}, ..., 2^{2}, 2^{3}
Regularization coefficient λ	0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}

補足資料 ハイパーパラメータの決め方

✓GTM-MLRにおけるGTM

- k3n-error [1] が最小になるように
  - k3n-error: https://datachemeng.com/k3nerror/

✓GTMR

- 2-fold クロスバリデーション後の r^2 が最大になるように

[1] H. Kaneko, Chemom. Intell. Lab. Syst., 176, 22-33, 2018.

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)