物質科学におけるデータ同化手法開発と応用

奈良先端科学技術大学院大学

原嶋 庸介

物質科学：実験と理論

実験　　理論

- 学理の追求
- 物質の発見

個々に実施して比較検証するだけではなく、
相乗効果を狙いたい

材料開発における実験と第一原理計算の関係

- 実験と第一原理計算では長所と短所が相補的
  - 実験: データ数が少なく、偶然誤差の影響が大きくなる
  - 第一原理計算: データ数が多く、偶然誤差(不純物の配置など)の影響を小さくできる一方で、理論誤差がある

- 実験と第一原理計算のデータを統合して、物性値の予測精度を向上させたい

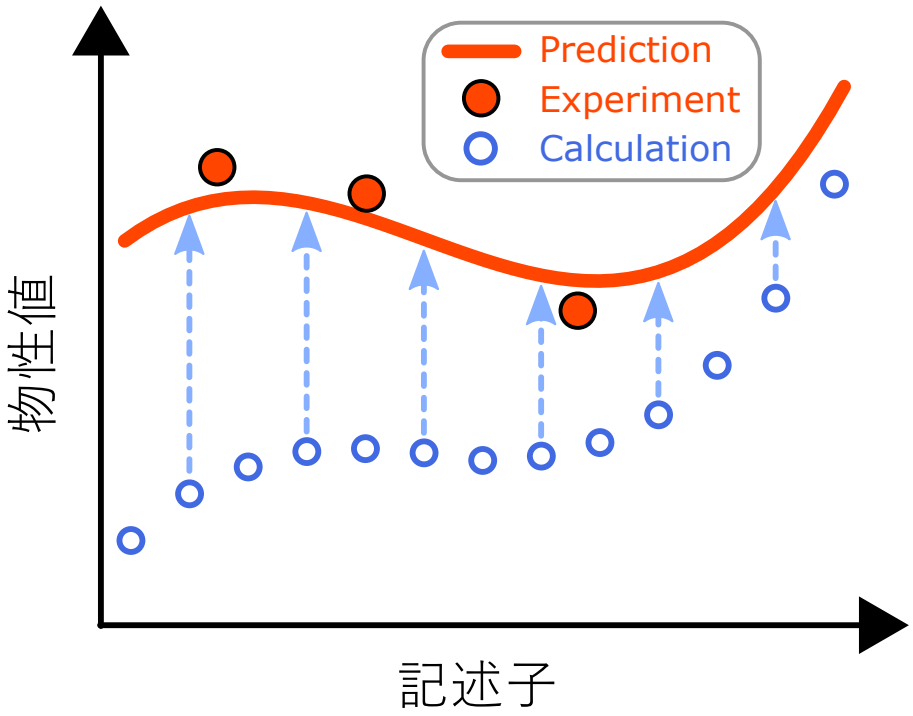
➢データ同化

データ同化とは?

- 実験と数値計算を統合(同化)させて、予測精度を上げる
- "***Data assimilation*** *is a mathematical discipline that seeks to optimally combine theory (usually in the form of a numerical model) with observations.*" (https://en.wikipedia.org/wiki/Data_assimilation)
- 現在では主に気象予報などで時系列データを扱う手法として使われている

➢物質科学の事情を考慮した新しいアルゴリズムが必要

実験とシミュレーションのデータにあるギャップ



- 計算値が実験をよく再現するのが理想

$$y_{\mathrm{expt}}(x) = y_{\mathrm{comp}}(x)$$

実際には系統誤差が生じている
例)平均場近似の過大評価など

$$y_{\mathrm{expt}}(x) = C y_{\mathrm{comp}}(x) + \mathcal{R}(x)$$

- サンプル抽出点(記述子)が
実験とシミュレーションで一致しない
  - ➢ 差が直接計算できない

これらの問題を回避したデータ同化手法を開発した

Y. Harashima, *et al.*, Phys. Rev. Materials, **5**, 013806 (2021).

Multivariate Gaussian modelで表現する

確率変数が正規分布によって分布していると仮定する

$$p(z;\Lambda) = \sqrt{\frac{|\Lambda|}{(2\pi)^d}} \exp\left(-\frac{1}{2} z^T \Lambda z\right)$$

記述子と目的変数の両方を確率変数と考える

$$z^T = (\underbrace{1, x_1, x_1^2, x_1 x_2, \cdots}_{\text{記述子}}, \underbrace{y_{\text{comp}}, y_{\text{expt}}}_{\text{目的変数}})^T$$

記述子の値が与えられた場合の目的変数の条件付き確率を考える

$$p(y|x;\Lambda) = \sqrt{\frac{|\Lambda_{yy}|}{(2\pi)^2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Lambda_{yy}(y-\mu)\right)$$

$$\mu = -(\Lambda_{yy})^{-1}\Lambda_{yx}x \longleftarrow \text{これが予測関数}$$

Gaussian modelと線形回帰モデルの関係

- 線系回帰モデル:

$$y = Wx$$

- Gaussian model:

$$y = -(\Lambda_{yy})^{-1}\Lambda_{yx}x$$

- 係数行列:

$$W = -(\Lambda_{yy})^{-1}\Lambda_{yx}$$

- Λ が与えられれば予測モデルが得られる

尤度と推定法

- 目的変数に関する尤度:

$$L(\Lambda) = \sum_{n=1}^{N} \ln p(y_n|x_n;\Lambda)$$

$$p(y|x;\Lambda) = \sqrt{\frac{|\Lambda_{yy}|}{(2\pi)^2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Lambda_{yy}(y-\mu)\right)$$

$$\mu = -(\Lambda_{yy})^{-1}\Lambda_{yx}x$$

- $\Lambda_{yy}, \Lambda_{yx}$を$L(\Lambda)$について最適化することで予測モデルを立てる

欠測データの取り扱い

- 欠測データの変数は積分消去する

$$
\begin{aligned}
p(y_1) &= \int dy_2 \; p(y_1, y_2) \\
&= \sqrt{\frac{|\bar{\Lambda}_{11}|}{2\pi}} \exp\left(-\frac{1}{2}(y_1 - \mu_1)\bar{\Lambda}_{11}(y_1 - \mu_1) \right) \\
\bar{\Lambda}_{11} &= \Lambda_{11} - \Lambda_{12}(\Lambda_{22})^{-1}\Lambda_{21}
\end{aligned}
$$

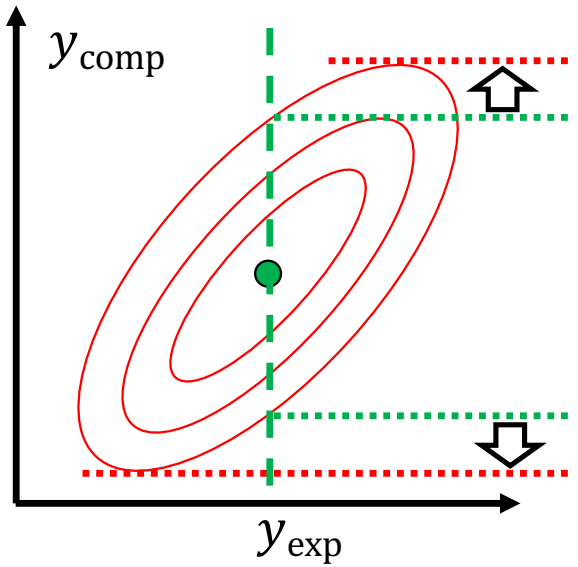
- 積分消去後の確率で定義した尤度を完全尤度という
- 完全尤度の最適化からΛを決定する

K.Takai and Y. Kano, Comm. Stat. **42**, 3174 (2013).
高井啓二 et al. 欠測データの統計科学 医学と社会科学への応用, 岩波書店 (2016).

欠測データは回帰への寄与を軽減

記述子	シミュレーション	実験
1.0	933.718	1000
0.9	945.811	994
0.8	962.004	nan
0.7	981.454	965
0.6	1002.318	nan
⋮	⋮	⋮

目的変数(y_{comp}, y_{exp})の分布の模式図



- 欠測によりデータの分布幅が広がる

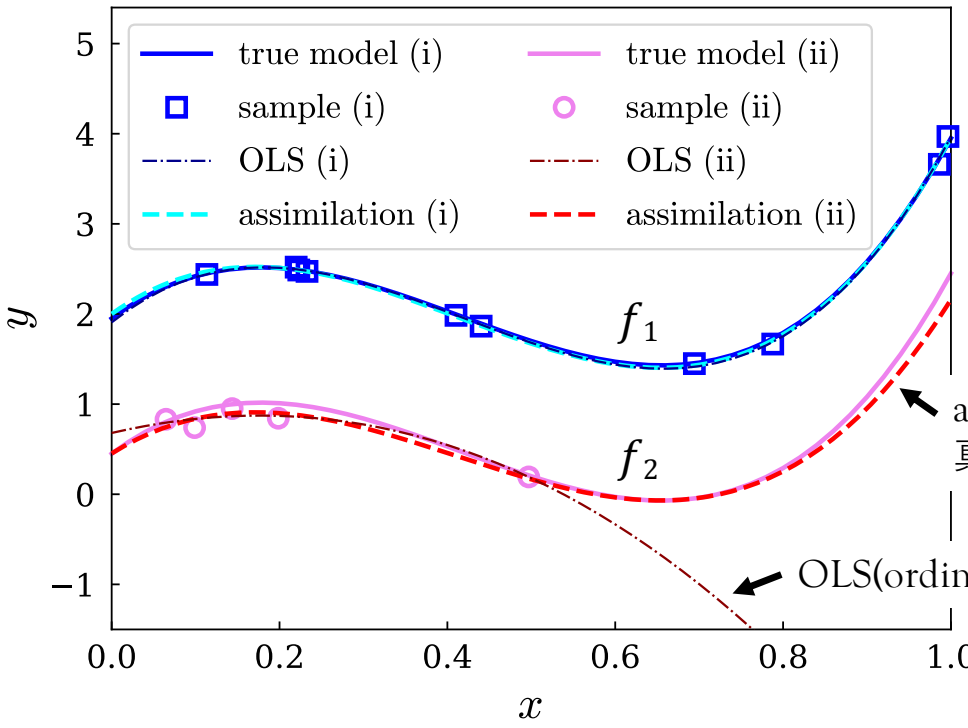
➢ データの重みが軽くなる

デモンストレーション

真のモデルを用意

$f_1(x) = 2 - x + 5(x - 0.7)^2 + 20(x - 0.5)^3$

$f_2(x) = f_1(x) - 1.5$ ← 定数だけ異なる



- model(i): 10サンプル; error小 (0.03), x=[0.0,1.0]
- model(ii): 5サンプル; error大 (0.1), x=[0.0,0.5]
- model(ii)とmodel(ii)で独立にサンプリング

- 3次の多項式によるfitting
- model(ii)のfitは1,2,3次の項は制限する
- OLS fit (model毎のfitting)と比較

assimilation(同化)はデータがないところでも
真のモデルをよく表現できている

OLS(ordinary least square)だと予測がうまくいかない

➢ 本手法では他方のサンプルを使って
擬似的に外挿が可能になる

例) 永久磁石化合物の探索

- 現在の永久磁石化合物の代表例:

結晶型: $\mathrm{Nd_2Fe_{14}B}$

- 実際は、

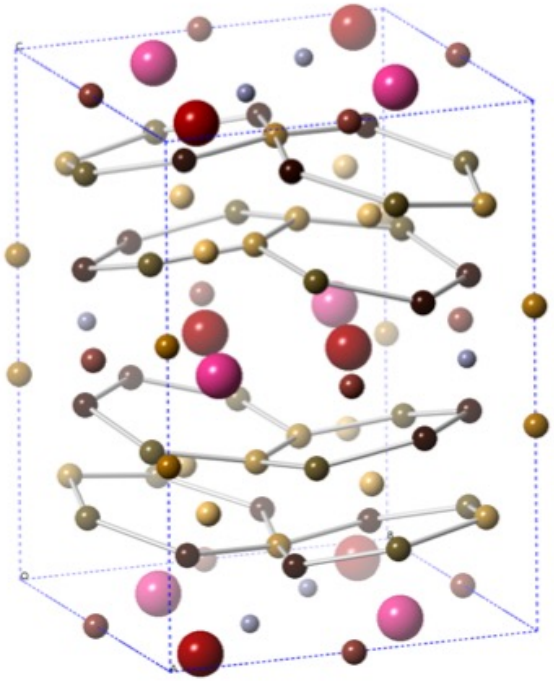
$$(\mathrm{La, Ce, Pr, Nd, \cdots})_2(\mathrm{Fe, Co, Ni, \cdots})_{14}(\mathrm{B, C, N})$$

希土類元素　　遷移金属元素　　典型元素

- 化学組成はfractionalに与えられる
- 物性値の化学組成依存性の理解は重要課題
- 多次元空間での理解は困難

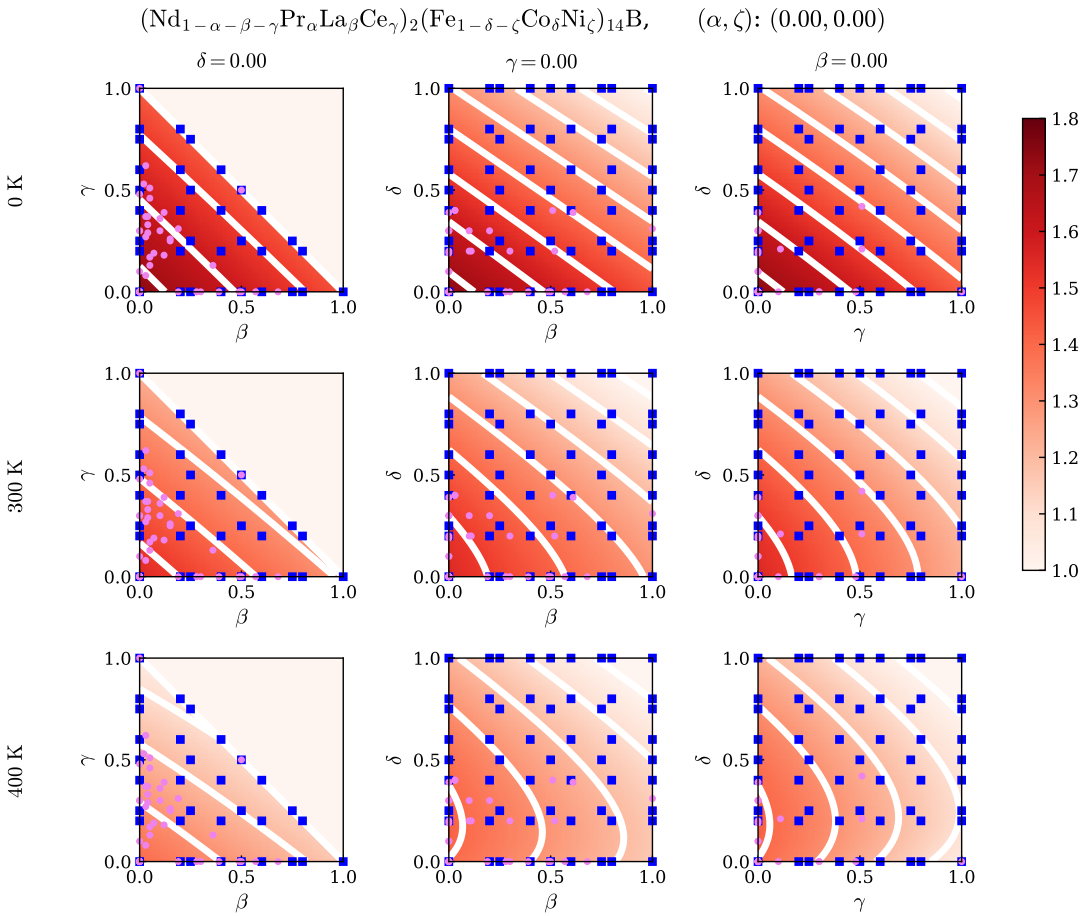
事例: 永久磁石化合物の有限温度磁化予測

$$(\mathrm{Nd}_{1-\alpha-\beta-\gamma}\mathrm{Pr}_{\alpha}\mathrm{La}_{\beta}\mathrm{Ce}_{\gamma})_2(\mathrm{Fe}_{1-\delta-\zeta}\mathrm{Co}_{\delta}\mathrm{Ni}_{\zeta})_{14}\mathrm{B}$$



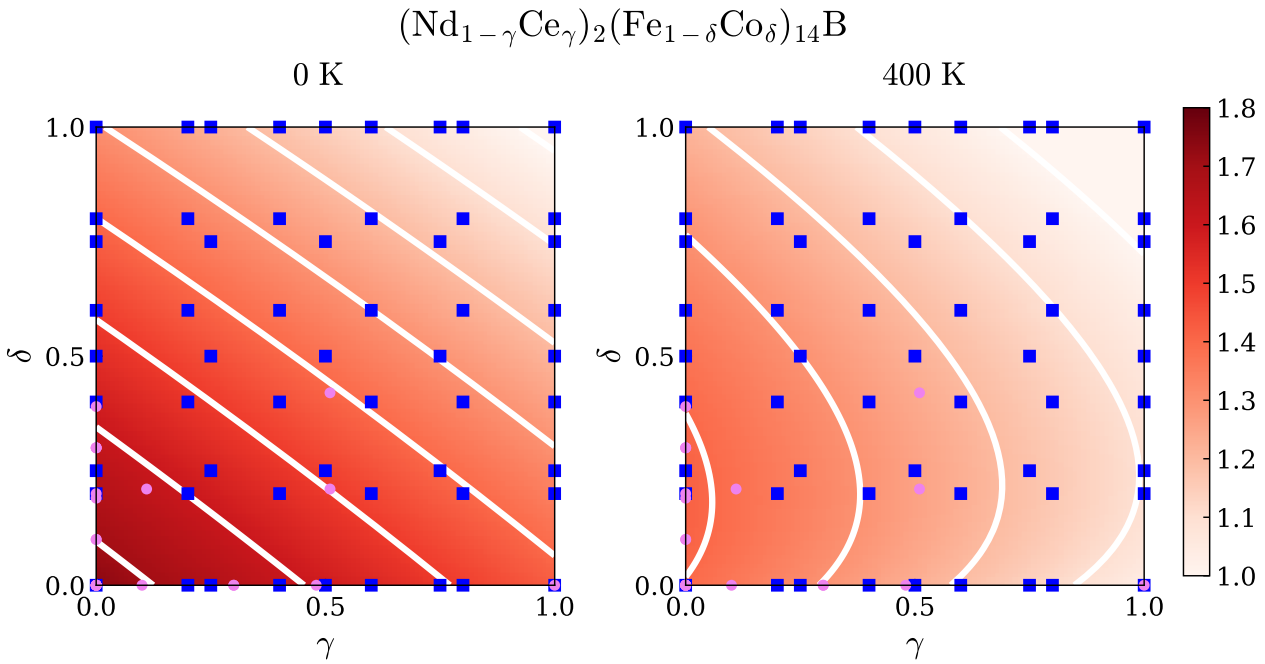
- $\alpha, \beta, \gamma, \delta, \zeta$の自由度に対して任意の温度の磁化の値を予測する

多次元化学組成空間における任意の温度での磁化

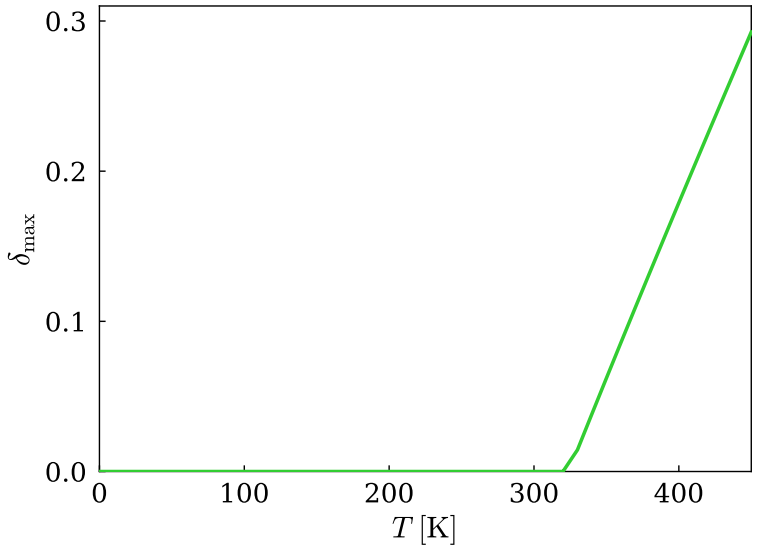


- いろいろな組成で予測可能

非線形な振る舞いの検出: Co濃度依存性



磁化最大値を与えるCo濃度の温度依存性

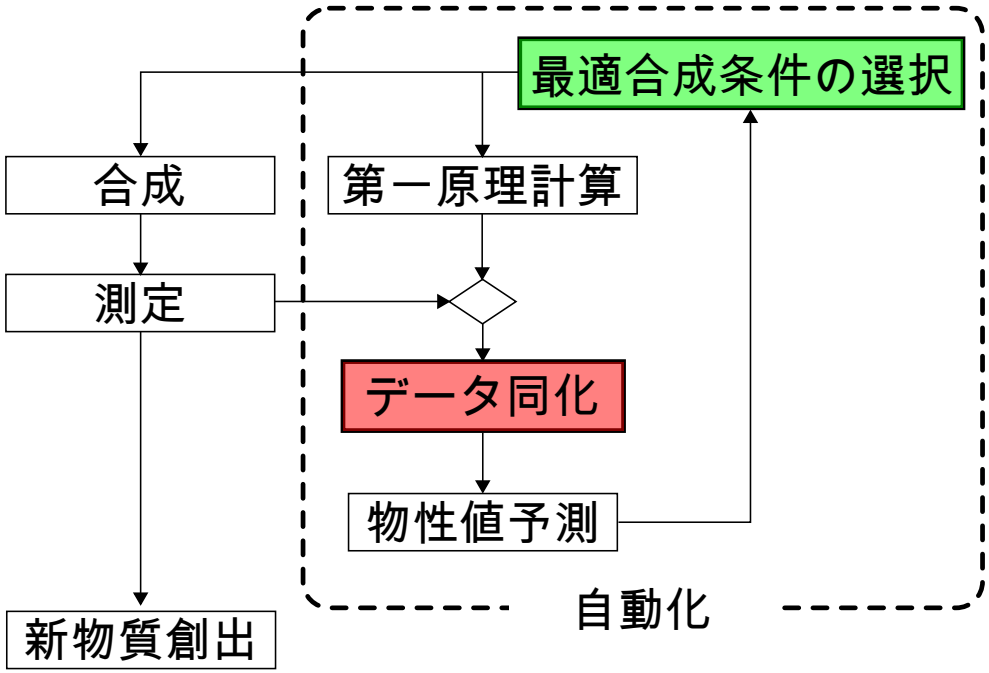


- 低温ではCo濃度によって磁化が単調減少する
- 高温(>320K)では有限のCo濃度で磁化が最大になる

物質探索

- 従来の探索では合成、測定を繰り返す

- シミュレーションデータと同化し、物性値予測モデルを高精度化する
- Bayes最適化により、次の合成条件を提案する



尤度と推定法

- 目的変数に関する尤度:

$$L(\Lambda) = \sum_{n=1}^{N} \ln p(y_n|x_n;\Lambda)$$

$$p(y|x;\Lambda) = \sqrt{\frac{|\Lambda_{yy}|}{(2\pi)^2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Lambda_{yy}(y-\mu)\right)$$

$$\mu = -(\Lambda_{yy})^{-1}\Lambda_{yx}x$$

- $\Lambda_{yy}, \Lambda_{yx}$を$L(\Lambda)$について最適化(先行研究では最尤推定で決めていた)

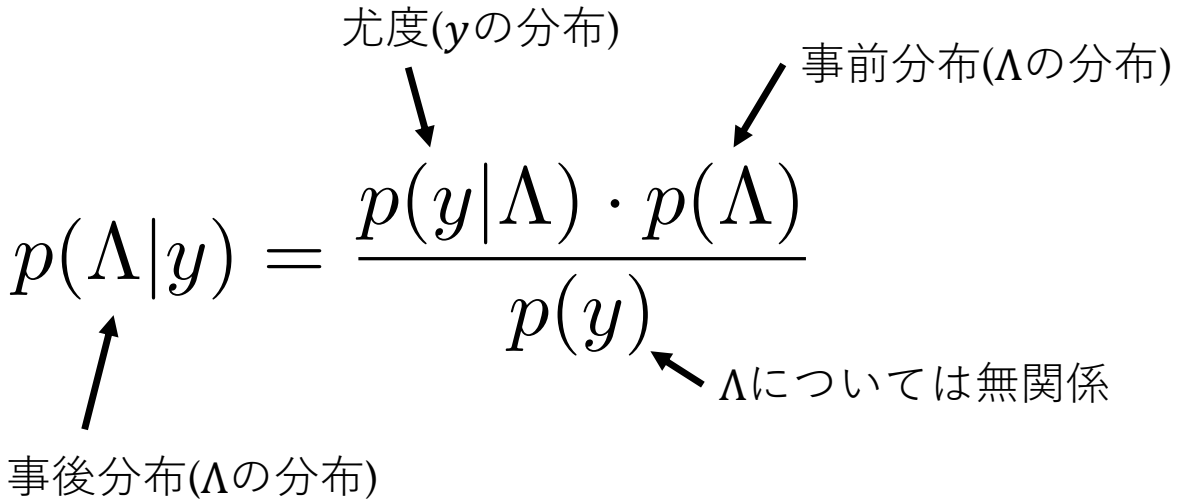
➢ 本研究：Bayes推定の導入

✓ 逐次更新可能：過去の全データが不要

✓ 事後分布(Λの分布)：Bayes最適化

Bayesの定理とBayes推定

Bayesの定理:



最尤推定：尤度をΛについて最大化する

Bayes推定：事後分布をΛについて最大化する

Bayes推定

- 事後分布の漸化式

$$p(\Lambda|\{y_i\}_1^N,\{x_i\}_1^N)=\frac{p(\{y_i\}_1^N|\Lambda,\{x_i\}_1^N)\cdot p(\Lambda|\{x_i\}_1^N)}{p(\{y_i\}_1^N|\{x_i\}_1^N)}=p(y_N|x_N;\Lambda)\cdot p(\Lambda|\{y_i\}_1^{N-1},\{x_i\}_1^{N-1})$$

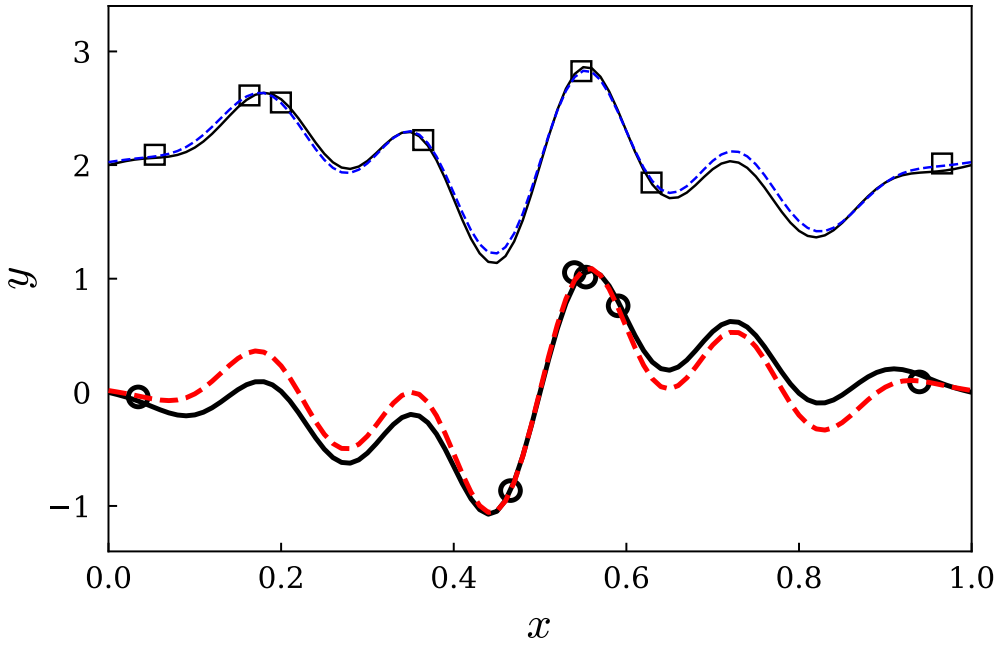
- 事後分布をC^Nで表現可能

$$p(\Lambda|\{y_i\}_1^N,\{x_i\}_1^N)=\sqrt{\frac{|\Lambda_{yy}^0|}{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathrm{Tr}(C^N\Lambda^0)\right)$$

$$C^N_{\gamma_1,\gamma_2}\equiv\sum_{i=1}^N z_{i,\gamma_1}z_{i,\gamma_2}\qquad\Lambda^0\equiv\begin{pmatrix}\Lambda_{xy}(\Lambda_{yy})^{-1}\Lambda_{yx} & \Lambda_{xy}\\ \Lambda_{yx} & \Lambda_{yy}\end{pmatrix}$$

- 事後分布を最大化をするΛ_{yy}とΛ_{yx}を使ってμを構築

多項式以外の基底関数による展開



- sin関数による展開

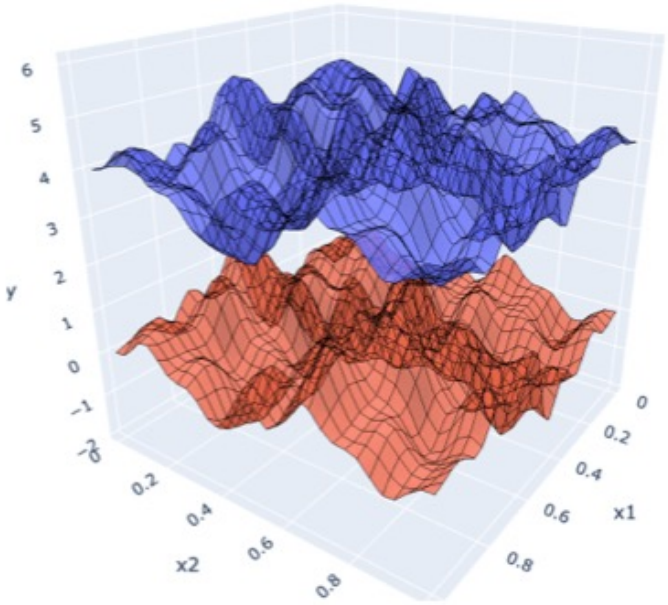
$$y_1 = 2\phi_0(x) + 0.2\phi_1(x) + 0.3\phi_2(x) - 0.2\phi_3(x) + 0.1\phi_4(x) - 0.3\phi_5(x) + 0.2\phi_6(x)$$

$$y_2 = y_1 - 2\phi_0(x) - 0.6\phi_1(x)$$

$$\phi_n(x) \equiv \sin(2\pi n x)$$

Bayes最適化：データ同化 VS 通常の最小二乗法

- 2次元空間, (x_1, x_2)
- 7基底/次元, $(\phi_n(x) = \sin(2\pi n x), n = 0,1, \dots, 6)$
- 交差項, $(\phi_n(x)\phi_{n'}(x))$
- 7x7=49個のfitting parametersに対応



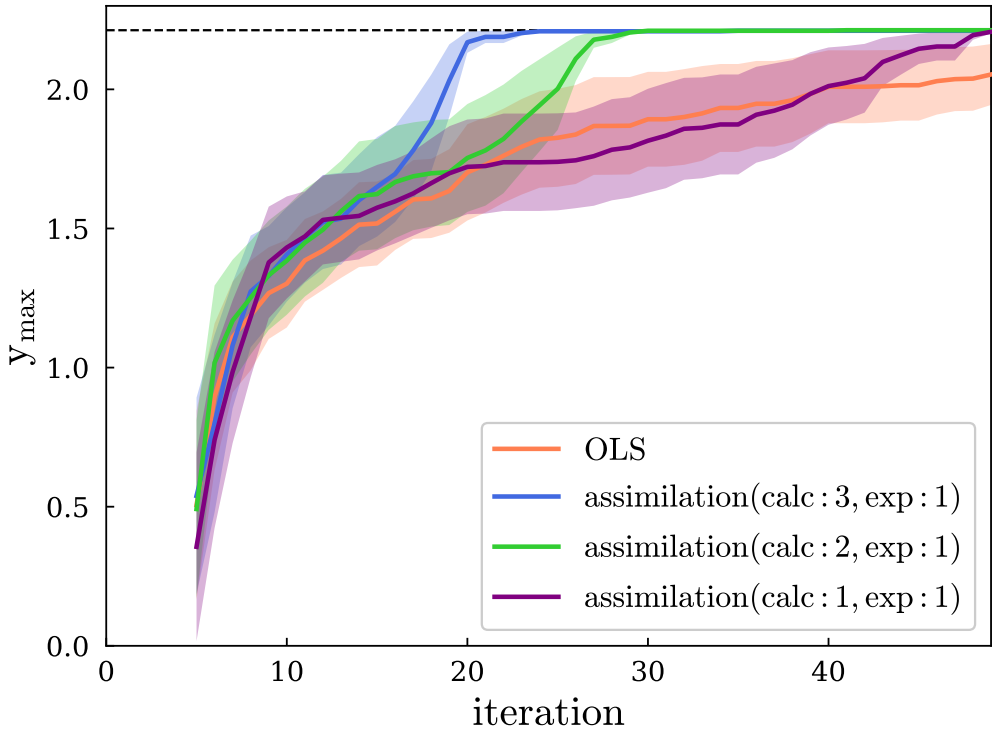
iteration

1. 2つのモデルを使用 (model AとB)
2. それぞれのモデルで5データを取得
3. データから予測モデルを導出
4. 候補点(x_1, x_2)をmodel Bの最大値を与えるように選択
5. m 個の候補点(x_1, x_2)をmodel Aの不確定性が最大になる点で選択

データ同化：3.+4.+5.
最小二乗法：3.+4. only

Bayes最適化：データ同化 VS 通常の最小二乗法

yの最大値の探索



色付きの幅は揺らぎ

$$\epsilon_{\mathrm{u}} = \sum_i \theta(yi - \bar{y}) \max(yi - \bar{y}, 0)$$

$$\epsilon_{\mathrm{l}} = \sum_i \theta(\bar{y} - y_i) \min(\bar{y} - y_i, 0)$$

iの和はサンプルデータ関してとる

- モデルA(calc)が確立されたとき（~49iteration）、モデルB(exp)の最大値が突然発見される

➢ シミュレーションによる探索は最大値を目標にするのではなく、モデルの確立を目指すのが効率的