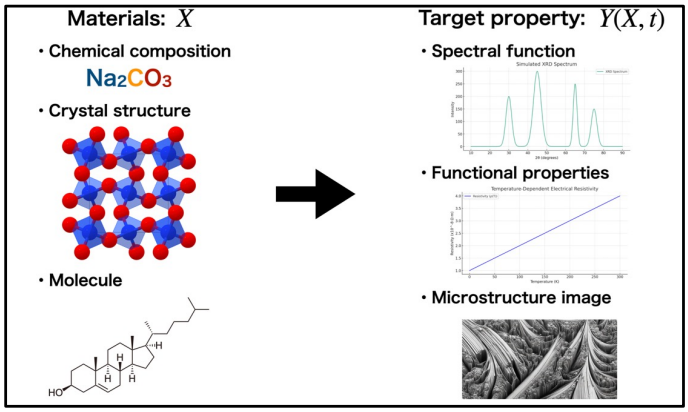
2024年5月24日　統計数理研究所　オープンハウス

関数データのためのベイズカーネル回帰

草場　穫　　先端データサイエンス研究系　特任研究員

1. 研究要旨

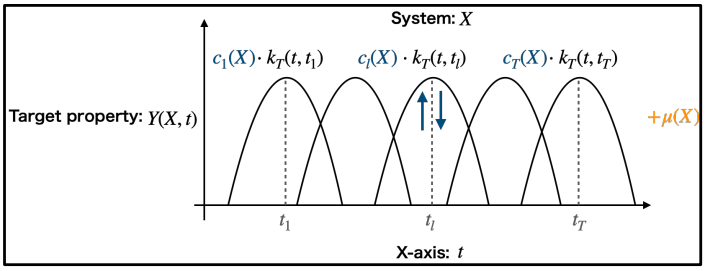
実験やシミュレーションから得られる材料データは、本質的にスカラー値ではなく、スペクトルや分布などの関数値であることが多いが、材料データに関数回帰モデルを適用した研究は少ない。このような材料データへの応用を動機として、我々はカーネル法に基づく関数型出力回帰モデルを提案する。このモデルは、スカラー出力を予測する通常の教師あり学習モデルを個別に学習するのに比べ、関数内の共分散構造を捉えることができ、学習効率と予測精度の向上が期待できる。関数データ解析 (FDA) における既存の function-on-scalar モデルとは異なり、このモデルは共変量にカーネル法を適用することで複雑化することなく、高次元非線形性を自然に取り扱うことができる。またモデルの完全なカーネル化により、再生核カーネルヒルベルト空間における理論的特性を理解することができる。さらに、提案モデルは、ベイズの観点から解析的に導出された関数出力予測分布を提供し、予測関数の予測不確かさの定量化を可能にする。これは材料データを含む実用的なアプリケーションにとって重要である。ポスターでは、人工データに対する提案モデルの予測結果を紹介する。



2. 提案モデルの概要

ベクトル入力値 X から X に依存する関数 $Y(t)$ を予測する回帰問題を考える。我々が提案するモデル (KRFD) では、t 空間に配置された正定値カーネル関数 $k_T(t,t_l)$ の線形結合により関数回帰を行う (カーネル中心 t_l は観測関数データの測定地点に設置する)。各カーネル関数 $k_T(t,t_l)$ の高さは X に依存する係数関数 $c_l(X)$ によって調整され、これにより X に依存する関数の表現を行う。さらに、KRFDでは係数関数 $c_l(X)$ は X 空間上の正定値カーネル $k_G(X,X_k)$ の線形結合によって表現される (カーネル中心 X_k は訓練データ点上に設置する)。また、材料データで頻繁に見られる系 (X) ごとの出力関数のシフトに対応するため、X 依存の定数項として $\mu(X)$ を追加した。$\mu(X)$ も同様に X 空間上の正定値カーネル $k_M(X,X_m)$ の線型結合によって表現される。これらをまとめると、KRFDモデルは未知パラメータに関して線形モデルの形で表現される。非線形性はカーネルを通してのみモデルに組み入れられる。KRFDモデルの概念図と数式表現は下図に示される。

KRFDモデルは、先行研究モデル [1] の完全なカーネル化と見做すことができ (先行研究では係数関数はニューラルネットワークでモデル化されていた)、これによってモデルが単純化されるため、解析的最適解の導出、ベイズ化、モデルの理論的考察を可能にする。KRFDモデルの形式は、入力ペア (X_i,t_j) に対する再生核ヒルベルト空間の観点から、separable kernels [2] の仮定の下で自然に導かれる (詳細省略)。



Kernel Regression for Functional Data (KRFD)

$$Y(X,t) = f(X,t) + \mu(X)$$

Variation term: $f(X,t) = \sum_{l=1}^{T} c_l(X) k_T(t,t_l),\ c_l(X) = \sum_{k=1}^{N} \theta_{kl} k_G(X,X_k) \rightarrow f(X,t) = \sum_{k=1}^{N}\sum_{l=1}^{T} k_G(X,X_k)\theta_{kl}k_T(t,t_l)$

System-dependent constant term: $\mu(X) = \sum_{m=1}^{N} c_m k_M(X,X_m)$

Training data
$X_i\ (i=1,\dots,N) \in \mathbb{R}^p$
$t_j\ (j=1,\dots,T) \in \mathbb{R}^q$
$Y(X_i,t_j) \in \mathbb{R}$

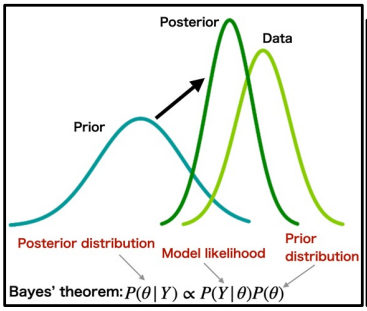
[1] Iwayama et al., Functional Output Regression for Machine Learning in Materials Science (2022).
[2] Ciliberto, Carlo, et al., Convex learning of multiple tasks and their structure (2015).
[3] Akiba, Takuya, et al., Optuna: A next-generation hyperparameter optimization framework (2019).

3. 提案モデルの推定

本研究において、KRFDモデルはベイズモデルとして扱われる。ベイズモデルではベイズの定理に従って未知パラメータが推定される (左下図参照)。この枠組みでは、モデルの未知パラメータは変数では無く、事前分布 (prior distribution) を持つ確率変数として取り扱われる。これをモデル尤度 (model likelihood) を介して、ベイズの定理に従って事後分布 (posterior distribution) に更新することが、ベイズの枠組みにおけるモデル訓練 (推定) に相当する。そのためには、モデル尤度と未知パラメータの事前分布を用意する必要がある。モデル尤度は観測データの生成過程についての仮定 (KRFDモデル+独立同一な正規分布観測ノイズ, 詳細は下の数式参照) から自然に設計することができる。事前分布は右下図に示されるように与えた (KRFDモデルの関数複雑性に対して X 空間と t 空間別々に正則化できるよう設計した)。これらの確率分布に、ベイズの定理とガウス分布に対して知られている公式を適用することで事後分布が解析的に導出された (詳細省略)。また、新しいデータ点 (X_{new},t_{new}) に対する予測分布は事後分布をKRFDモデルにあってはめることで解析的に導出された。

Assumption for observed data

$$Y(X_i,t_j) = \sum_{k=1}^{N}\sum_{l=1}^{T} k_G(X_i,X_k)\theta_{kl}k_T(t_j,t_l) + \sum_{m=1}^{N} c_m k_M(X,X_m) + \epsilon_{ij},\quad \epsilon_{ij} \overset{i.i.d}{\sim} N(0,\sigma^2),\quad (i=1,\dots,N,\ j=1,\dots,T).$$



Model likelihood

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{c},\sigma^2) = N(\boldsymbol{G}\otimes\boldsymbol{T}\cdot\boldsymbol{\theta} + \boldsymbol{M}\otimes\boldsymbol{1}_T\cdot\boldsymbol{c},\ \sigma^2\boldsymbol{I}_{NT})$$

Prior distributions

$$p(\boldsymbol{\theta}) = N\left(\boldsymbol{0},\ \left(2\lambda_a(\boldsymbol{G}\otimes\boldsymbol{T}^2) + 2\lambda_b(\boldsymbol{G}^2\otimes\boldsymbol{T}) + 2\lambda_c(\boldsymbol{G}\otimes\boldsymbol{T})\right)^{-1}\right)$$

$$p(\boldsymbol{c}) = N\left(\boldsymbol{0},\ \left(2T\lambda_d\boldsymbol{M}\right)^{-1}\right)$$

$$p(\sigma^2) = IG(\alpha,\beta)$$ ←prior for the variance of observation noise

By known math formulas for Gaussian distributions

Posterior distributions
→ Prediction distribution on (X_{new},t_{new})

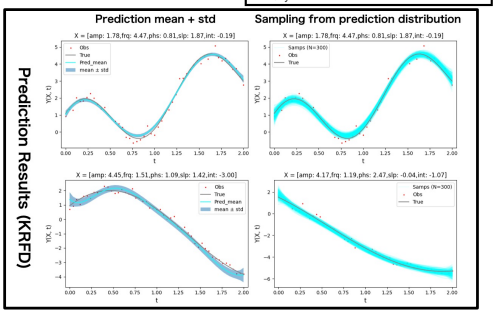
4. 人工データに対する適用例

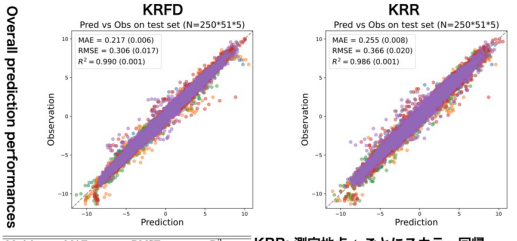
人工データは、真のデータ生成過程として入力値 X に依存する正弦波+直線の関数を持ち、それらに正規分布のノイズを加えたものとした (下の数式参照)。入力値 X は傾きや周波数など物理的に意味を持つパラメータをベクトルとしてまとめたものである。t 空間で均等に配置された測定地点は51個用意された。入力値 X はランダムに1,000個生成され、データ生成過程より各入力値に対応する関数出力データも生成された。このデータの25%をテストセットとし、残りを訓練データとした。KRFDのハイパーパラメータは事前分布の形 (=正則の強さ) を決める λ_aなどのパラメータとカーネル $k_G(X,X_k)$等のカーネル幅を決めるパラメータからなり、これらは訓練データ上のクロスバリデーションロスをoptuna [3] によってベイズ最適化することで決定された。テストセット上での予測結果を以下に示す。予測平均と予測標準偏差を求めた例と予測分布から関数サンプリングを行った例を両方図示した。全体的な予測性能評価は一番下にまとめた。Linear function-on-scalar 回帰 (LFOS) とカーネルリッジ回帰 (KRR) による予測も行い、KRFDと予測精度を比較した。

Data generation process: $y(t) = a\sin(bt+c) + dt + e + \epsilon,\quad \epsilon \overset{i.i.d}{\sim} N(0,\sigma^2).$

a : amplitude　　d : slope
b : frequency　　e : intercept
c : phase　　ϵ : noise

Training data
$X_i = (a_i,b_i,c_i,d_i,e_i)^T,\quad (i=1,\dots,N) \in \mathbb{R}^5$
$t_j\ (j=1,\dots,T) \in \mathbb{R}$
$Y(X_i,t_j) \in \mathbb{R}$





Models	MAE	RMSE	R^2
KRFD	0.217 (±0.006)	0.306 (±0.017)	0.990 (±0.001)
KRR	0.255 (±0.008)	0.366 (±0.020)	0.986 (±0.001)
LFOS	1.278 (±0.036)	1.663 (±0.047)	0.718 (±0.018)

KRR: 測定地点 t_l ごとにスカラー回帰

LFOS: $Y(X,t) = \beta_0(t) + \sum_{j=1}^{p} x_j\beta_j(t)$ *$X = (x_1,\dots,x_p)^T$