計算工学講演会論文集　Vol.30 (2025年6月)　　日本計算工学会

潜在変数によるサロゲートモデル近似能力の飛躍的発展

Drastic Improvement of Surrogate Model with Latent Variable Approach

宮田悟志
Satoshi Miyata

博(工) ダッソー・システムズ（株）（〒141-6020 東京都品川区大崎2-1-1 ThinkPark Tower, E-mail:satoshi.miyata@3ds.com)

Latent variable approach for surrogate modeling, which has significant impacts on simulation-based design optimizations and system's approaches, is argued. Concept of latent variable and its application to machine learning are briefly introduced: Gaussian Process Latent Variable Model (GPLVM). Benchmark tests with three learning data sets and deep engineering optimization application demonstrate excellence of GPLVM surrogate models, comparing with those from traditional surrogate approaches(RSM, RBF).

Key Words : *Surrogate Model, Latent variable, Gaussian Process, GPLVM, Optimization*

1.　はじめに

実験や試作の代替としてCAEに代表される数値シミュレーションを使うことは製品設計の有効な方策として今日定着している．またそのさらなる活用法として，一連のシミュレーションをワークフローとして定義し，システム工学的な手法（要因効果分析，最適化，トレードオフ分析，信頼性解析，ストキャスティック解析等）で廻すことも目新しいことではなくなった．しかしワークフローのシステム工学的方法の活用は，その実行時間から制限を受ける．非線形特性を評価するためには，検討因子数Nに対してその多項式オーダー（例えば$O(N^2)$～$O(N^3)$）回数のシミュレーション実行が必要となることが普通であるが，これは通常の製品開発サイクルで実行可能とは限らない．たとえば検討因子が10個の場合，100回～1000回のシミュレーション実行回数となるが，この規模の計算を1日程度で完了できるのは，ごく小規模なシミュレーションモデルに限られる．この理由から最適設計の分野では「近似モデル」や「サロゲートモデル」（本稿ではサロゲートモデルと呼称）と呼ばれる代替評価手法が古くから試みられてきたが，一般性を有し広範なシミュレーションに適用可能なサロゲートモデルの開発は，現在も課題である．そこで本稿では，近年の機械学習分野の研究成果を反映した，潜在変数モデルのサロゲートへの適用の有効性を論じる．具体的には，ガウス過程に潜在変数を導入したモデル：Gaussian Process Latent Variable Model（GPLVM）について述べるが，RSMやRBFといった線形サロゲートモデルに比べ顕著な近似能力と汎化能力の向上が得られることを示す．

2.　潜在変数

潜在変数（Latent Variable）は，現象の生成機構に内在する変数として定義される．ある現象$x_i, i=1,2,...$が観測されていて，それがパラメータθを持つ確率分布から生成されたものであると考えることは

$$\prod_i p(x_i;\theta) \tag{1}$$

と表記される．そしてこの生成モデルの決定は，たとえば

$$\arg\max_\theta \prod_i p(x_i;\theta) \tag{2}$$

とモデル中のパラメータθを選ぶことである（最尤推定）．ここにおける論点は，観測されている変数x_iだけが現象を生起させているかどうかは断定できない場合が多いという点である．隠れた変数z_j, $j=1,2,...$が存在していて，現象の生成モデルを

$$\prod_{i,j} p(x_i, z_j; \theta) \tag{3}$$

と考えた方が，観測される現象の振る舞いをよく記述できる場合もある．ただし，この隠れた変数はそもそも素性が事前には知られていないいから，所与のデータから推定して生成モデルに組み入れる必要がある．端折った表現をすれば，これが潜在変数である．当然，潜在変数を含む生成モデルの推定は (2)式のような単純なものとはならず，EMアルゴリズムや変分ベイズ（変分推論）法といったより複雑な手法が必要になるが，これらが上手く機能すれば生成モデルの現象記述力は大きく向上する．

3.　ガウス過程，サロゲートモデル，潜在変数

(1) ガウス過程とサロゲートモデル

ガウス過程（Gaussian Process）は，システムの出力yがガウス分布（正規分布）に従うとする確率モデルである．

$$y = f(\boldsymbol{x}),\ y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4}$$

yが確率変数であるから，これを与える$\boldsymbol{x}$も確率変数である．サロゲートモデルを一般に考える場合，入力値$\boldsymbol{x}$に対する出力値yを得ることが基本であるから，(4)式の平均値$\boldsymbol{\mu}$の部分だけで考え，近似の不一致は誤差とする

$$\mu \equiv \hat{y} = \hat{f}(\boldsymbol{x}),\ \varepsilon \equiv y - \hat{y} \tag{5}$$

のようなモデル化が普通であるが，ガウス過程で $y = f(\boldsymbol{x})$ を考える場合には，正規確率分布の共分散成分 $\boldsymbol{\Sigma}$ もサロゲートモデルのパラメトリックな定義に折り込むことが特徴である．よって $\boldsymbol{\Sigma}$ の与え方がガウス過程のモデル化におけるポイントとなる．共分散成分は2点 $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ の関係として定義される量であるが，これを

$$\Sigma_{i,j} \equiv k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{\theta_2}\right) \tag{6}$$

のようなカーネル関数 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ でモデル化する．ただし関数形は，正値性や対称性といった幾つかの条件を満たせば他は任意であり，様々なカーネル関数が存在する．また，$\boldsymbol{\Sigma}$ を2点 $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ の関数でモデル化するという定式化は Kriging 近似と同一である．実際，ガウス過程によるサロゲートモデルは，今日では Kriging 近似の一般化（カーネル関数とベイズ推定的な視点による）として位置づけられている．

(2) GPLVMs

ところで，潜在変数の概念とその利用は機械学習に固有のものではなく，20世紀から行われて来た．主成分分析（PCA）や因子分析（FA）などがそれであり，高次元データをより低次元の特徴空間で表現する，特徴抽出と次元削減の目的で使用されて来た．その古典的なアプローチでは，観測データの共分散行列を考え，その固有ベクトルを固有値の大きい側から選択することで特徴変数を構成する．観測変数空間から潜在変数空間の構成は線形写像にもとづいている：

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{T\&B.1}$$

上式においては $\mathbf{t}$ が観測変数ベクトル，$\mathbf{x}$ が潜在変数ベクトルであり，$\mathbf{W}$ がその線形写像行列である（数式の番号は文献[1]の其れをそのまま引用し，他の数式と区別するために接頭辞"T&B"を付加した）．TippingとBishop は，この過程を確率モデルとして捉え

$$\mathbf{t} \mid \mathbf{x} \sim N(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{T\&B.2}$$

$$\mathbf{t} \sim N(\boldsymbol{\mu}, \mathbf{C}),\ \mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} \tag{T\&B.3}$$

(T&B.1)における観測変数行列の最大固有値の選択は，観測変数 $\mathbf{t}$ についての尤度最大化と等価であることを示した[1]．Lawrence はこの仕事をさらに進め，ガウス過程と(6)式タイプのカーネル関数を用いることで，これを非線形写像に拡張した[2].

$$p(\mathbf{Y} \mid \mathbf{X}, \beta) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}})\right)}{(2\pi)^{DN/2}|\mathbf{K}|^{D/2}} \tag{L.1}$$

ただし(L.1)では $\mathbf{y}$ が観測変数ベクトル，$\mathbf{x}$ が潜在変数ベクトルであり，これらの大文字表記 $\mathbf{Y}, \mathbf{X}$ はN個の標本点からなる $\mathbf{y}, \mathbf{x}$ の行列を意味する．また，D は $\mathbf{y}$ の次元，$\mathbf{K}$ はカーネル関数を要素とする行列である．(L.1)の尤度を最大化することで潜在変数 $\mathbf{X}$ が得られるが，これは共役勾配法を使用した非線形最適化により計算される．GPLVM（Gaussian Process Latent Variable Model）という名称がここで初めて提案されるが，Lawrenceは単一のモデルをGPLVMと定義するのではなく，ガウス過程による非線形写像の尤度を最大化するモデルの総称として使用している．具体的に[2]で論じられているのはPCAのGPLVMである．ところで，[2]は重要な文献であるが，非常にテクニカルな内容で一般のシミュレーション利用者には難しいかもしれない．[3]の第7章に分かり易い解説があるので，本稿の読者にはこちらをお勧めする．

4. ベンチマーク問題

(1) サロゲート対象と最適設計問題

図1にサロゲート対象となるシミュレーションモデルを示す．乗用車サスペンションの構成部品で，FEMシェル要素でモデル化されたプレス成型材である．8つの部位のシェル厚さが設計変数（観測変数）である．この部品が走行状態で入力される荷重に対して十分に剛，かつ軽量であることが設計目標である．重量最小化 - 剛性制約とする最適化問題（式(7)-式(9)）としてこれを満足する板厚分布を計算するが，最適解のロバスト性を保証するために，負荷荷重に正規確率分布に従う変動（$\sigma = 0.1 * \mu$）を導入し，この条件下で制約条件満足に対して信頼度を満たすことを課す（式(10)-式(12)）．

$$\text{Minimize} \quad Weight \tag{7}$$

$$max.elastic\ strain : LE\_max \le 0.00145 \tag{8}$$

$$max.deformation : U\_max \le 0.5 \tag{9}$$

$$\text{Rel\_LB} \le reliability\_LE\_max \tag{10}$$

$$\text{Rel\_LB} \le reliability\_U\_max \tag{11}$$

$$\text{Rel\_LB} = \{NA, 0.9, 0.99, 0.999, 0.9999, 0.99999\} \tag{12}$$

この問題設定は，いわゆる信頼性最適化（Reliability-based Optimization，RBO）であり，通常の最適化計算×各設計点における信頼度評価，で計算が進む．荷重条件は図1に示すように，並進成分が3つ，回転成分が3つの計6成分であるから，Mean value法による信頼度評価では各設計点において7回のシミュレーション実行が行われる．つまりこのRBOは通常の最適化の7倍の計算量を要する問題設定となる．信頼度の評価方法は複数存在するが，それらの中でMean value法が最小のシミュレーション実行数であり，最適化過程で併用する信頼度評価法として実用性が高い．この理由から本稿では Mean value法を使用した．

また式(12)に示したように，信頼度は無設定 NA から99.999% までの6水準を設定している．これは信頼性最適化をトレードオフ分析的に外側から廻すことを意味する．結局，通常の最適化の7倍×6倍＝42倍が必要なシミュレーション実行回数となる．サロゲートモデル導入による

Design Parameters

	Lower Bound	Initial Value	Upper Bound
A_PIPE	2.0	4.0	5.0
G_PIPE	1.0	3.0	4.0
LWR_1	1.0	2.6	4.0
LWR_2	1.0	2.6	4.0
LWR_3	1.0	2.6	4.0
UPR_1	1.0	2.3	4.0
UPR_2	1.0	2.3	4.0
UPR_3	1.0	2.3	4.0

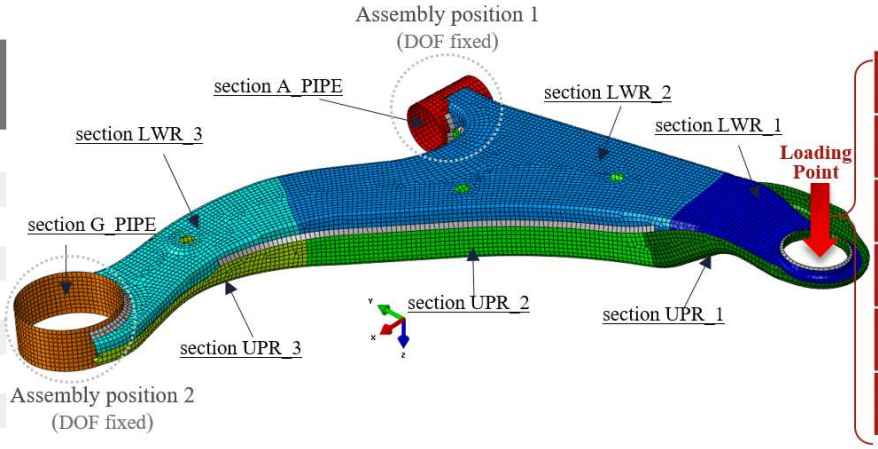

図-1 シミュレーションモデルと検討因子

評価の高速化が必要となる問題設定である．しかし，設計解検討法として漏れのない完備なものを考えるなら，これは必須アイテムからなる最小側の構成である．このボリュームのシミュレーション数を廻せるか否かが，ワークフローが高い知見をもたらす基盤となるのか，構築に手間のかかるショーケースで終わるかの分かれ目である．そして，この方法論的課題を克服するために，サロゲートモデルが存在する．

(2) サロゲートモデルの構築

GPLVMに基づくサロゲートモデルの有効性を確認するために，対照としてRSMとRBFによるサロゲートモデルを作成した．これら3種のサロゲートモデルは，ダッソー・システムズ社の **3D** EXPERIENCE® Platform [4]により作成した．**3D** EXPERIENCE® 版 GPLVM の仕様詳細は非公開であるが，RSMは一般的な多項式回帰モデル（本稿の計算では，最高次数は6，冗長項をモデルから除外するために項選択を使用），RBFは[5]の定式化に基づくShape function型であり（Neural Network型ではない），基底関数の拡がりを制御するパラメータを1つ持つ．

学習データは，8つの板厚と6つの荷重成分から成る14次元空間から，標本数が300点，500点，1000点の3ケースを設定した．最適ラテン超方格計画を図1に示した問題空間（変数の上下限値の直積空間）に適用することで必要な学習データを生成した．

以上の，3種の手法×3種の学習データにより，合計9つのサロゲートモデルを作成して比較検討を行った．

5. ベンチマーク結果

(1) サロゲートモデル

まず生成されたサロゲートモデルの概要として，図2に300点での結果を，図3と図4 に，500点と1000点の場合の結果をそれぞれ示す．これら散布図の横軸は学習データにおける応答値，縦軸はサロゲートモデルの予測する応答値である．ベンチマークRBO問題のLE_maxとU_maxそれぞれについて，実値と予測値の相関を表示している．また図中には自由度調済み整決定係数 R^2_{adj} も表示している．目的関数であるWeight については，シェル要素の厚さを設計変数とするこの問題設定では，1次多項式で十分に近似されモデル推定の是非を論じる必要がないため，掲載はしていない．300点で生成した LE_max の RSMとRBFの結果，500点と1000点のRBF の LE_max で少し点のバラツキが見られる以外は，どれも高い一致を示している．実際，これら良好な結果の重相関係数値は 0.95～0.99であり，通常の実験データ解析からすれば異常な高さであるが，機械学習結果の評価基準として見れば普通である．

次に，図5，図6，図7に生成された応答のサーフェスを示す．それぞれ同様に，300点，500点，1000点の結果である．RBO問題の制約条件：(8)式と(9)式の満足の可能性を図のコンター色・コンター線から概観できるように軸変数を選んでいて，散布図による相関観察では判らない，より実際的な情報が表示されている．特定のコンター色（線）がRBO問題の制約臨界に対応するが，その幾何的な形状は学習データ数とサロゲートの方法で明らかな差が存在することがわかる．類似性を感じるのは $R^2_{adj} \geq 0.98$ を持つサロゲート結果間のみである．有制約最適化問題の最適解はこの制約臨界に現れるので，図5，図6，図7のコンター表示におけるサロゲート結果の差異は，得られる最適解の差異と同義である．また，RBOにおける信頼性制約条件 (10)式，(11)式は応答曲面の勾配から計算される．そのため図5，図6，図7のコンター表示の差異は，信頼度評価においては，より一層拡大されて影響する．

これら全ての図から，GPLVMによるサロゲート結果は，全ての学習データについて，他のサロゲートモデルに優越していることが分かる．300点という本ベンチマーク問題の最小学習データでも，1000点の学習データから生成されたモデルに比べると量的精度は劣るものの，表現されるサーフェスの幾何的な類似は高く保たれている．これはRSMによるサロゲート結果と比較すると顕著である．RSMの場合，学習データ数を増やすにつれて幾何的な近似度合いは向上するが，300点での結果は良いとは言えない．類似の傾向はRBFによるサロゲートでも見られるが，

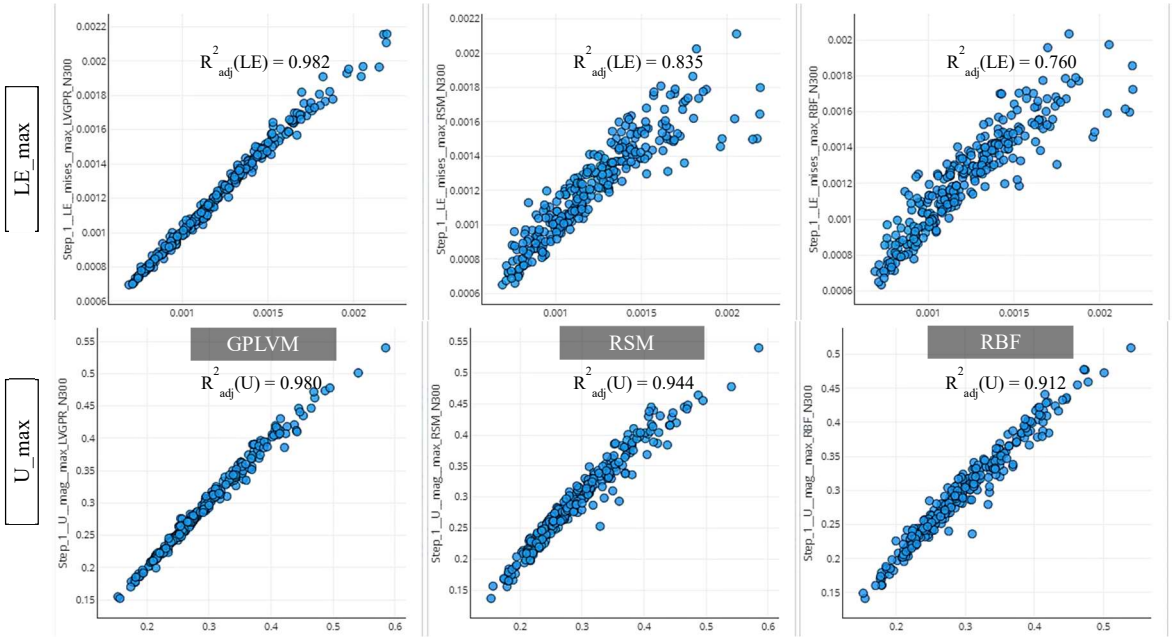

図-2 学習データとサロゲートモデル予測値 (N=300)

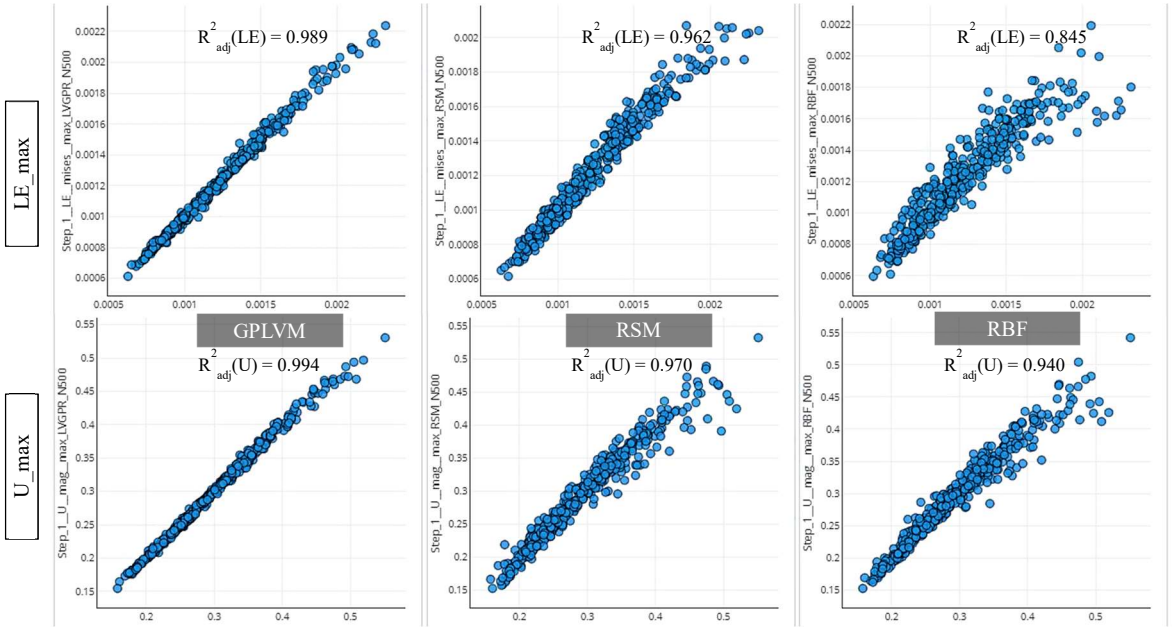

図-3 学習データとサロゲートモデル予測値 (N=500)

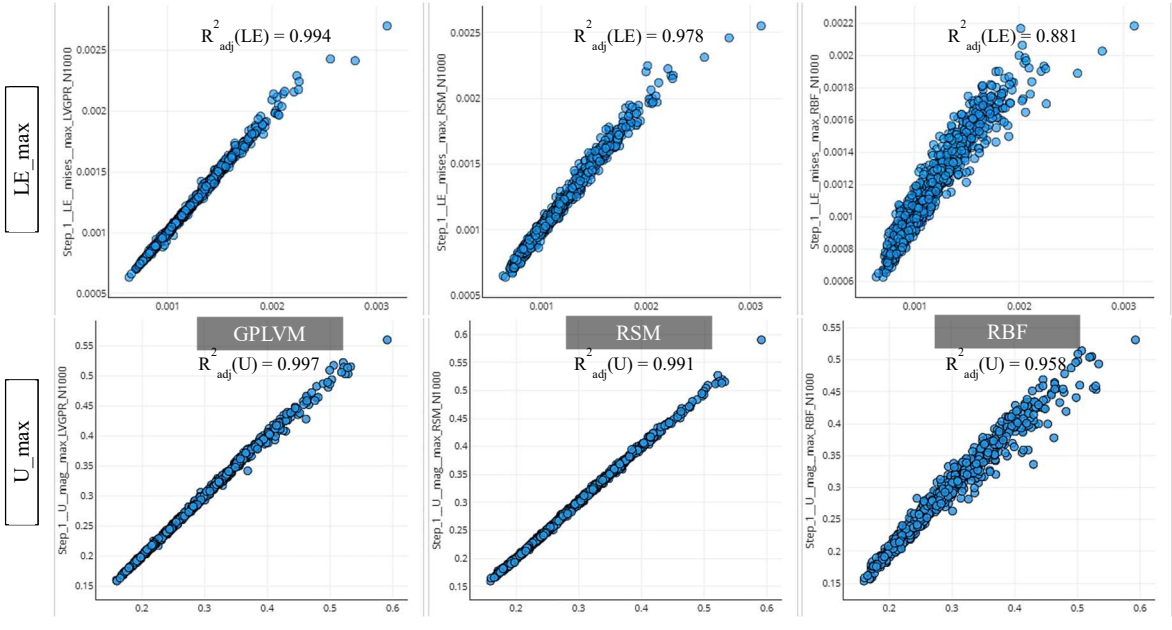

図-4 学習データとサロゲートモデル予測値 (N=1000)

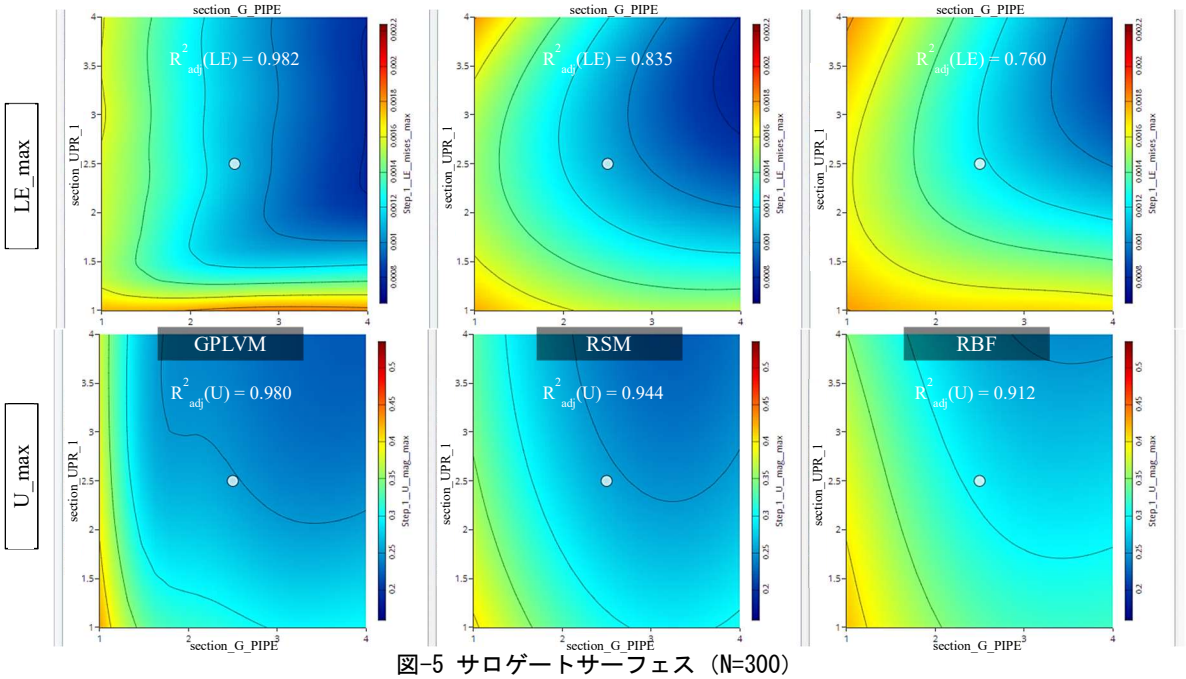

図-5 サロゲートサーフェス（N=300）

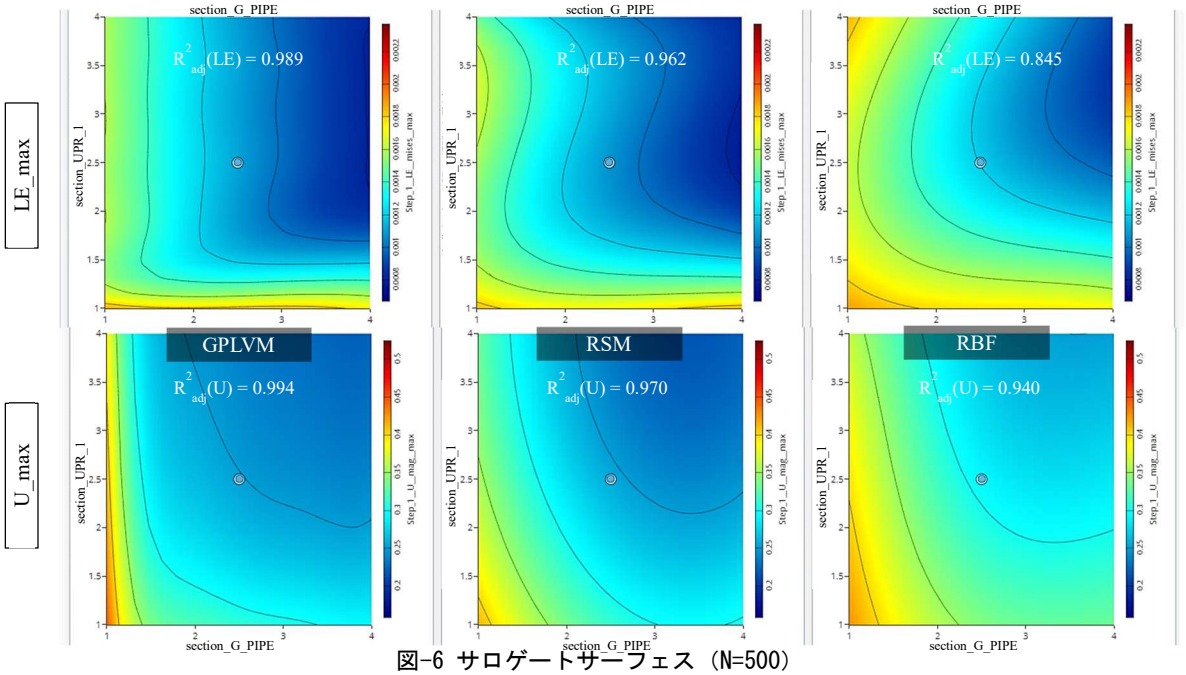

図-6 サロゲートサーフェス（N=500）

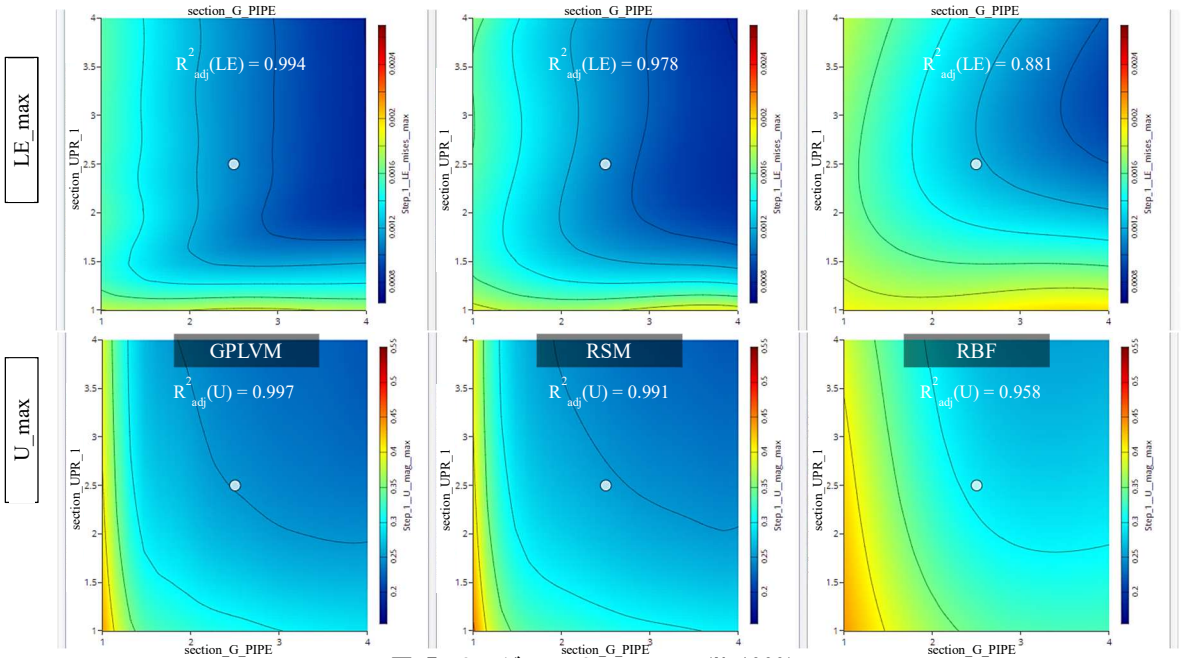

図-7 サロゲートサーフェス（N=1000）

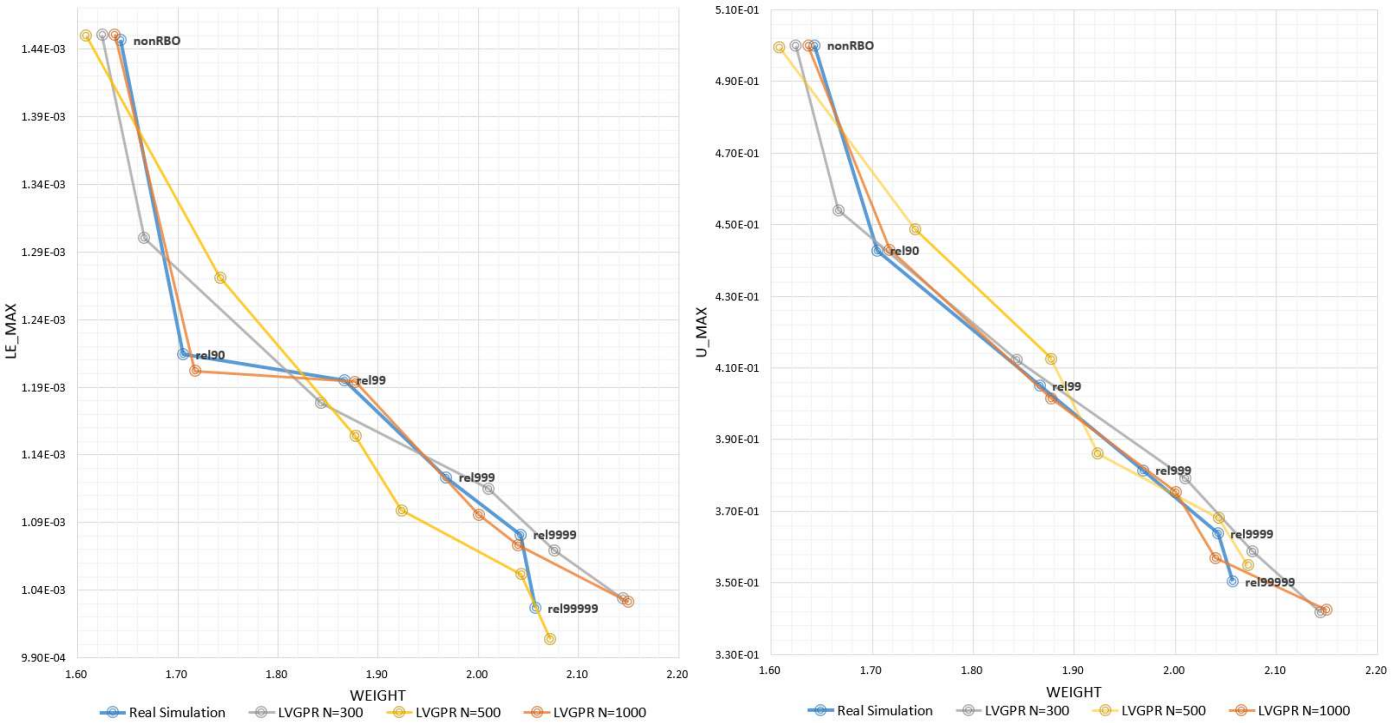

図-8 RBO×トレードオフ解析の結果 GPLVM

このベンチマーク問題ではRBFの結果はRSMに比べて明らかに劣っている．

また同一の学習データ数で比較しても，GPLVMがRSMやRBFより R^2_{adj} で劣るケースは存在しない．

(2) サロゲートモデル上でのRBO

GPLVMのサロゲートモデル上でRBOを6水準実施した結果を図8に示す．本節における図示では，サロゲートに対する対照として，サロゲート非使用の結果を導入している．横軸が目的関数であるWeight，縦軸を制約条件であるLE_max および U_maxとして，信頼度の変化に対するこれらの間のトレードオフ関係を表示している．グラフ左上が信頼度無要求(noRBO)，右下に移るにつれて信頼度への要求が上がって行く．満足可能な制約条件値と最適値の間には明確な相反関係が存在する．RSMとRBF上でも同様のRBOを実施しているが，GPLVMがRSMとRBFに総じて優越することは前節で示したので，本稿では紙面の制約からGPLVMの結果のみ掲載した．

図8を見ると1000点から生成したGPLVMの結果は信頼度99.999%の水準を除き，実解析の結果と良く一致している．特に LE_maxの信頼度90%における凹みまで再現している点は注目に値する．学習データを取得した標本空間の次元が14であることを考えると，300点と500点の差異は標本の統計的バラツキに埋もれて断定的に論じ難いが，1000点まで標本点数を上げれば，明らかに質の向上したサロゲート結果が得られると言える．

6. おわりに

学習データ数を3段階に変えてサロゲートモデルの性能評価を行った結果を報告した．適切なサロゲートモデルの構築はシミュレーションによる最適化・システム工学アプローチの課題であり続けて来たが，近年の機械学習分野の発展は，このニッチな領域についても恩恵を与えることが確認された．最新サロゲートモデルであるGPLVMの結果が，一世代前，二世代前のサロゲートモデルである RBF, RSM を上回ることは当然であるが，少数の学習データであっても高度な非線形近似能力を保っていることを確認できたことは非常に価値があると言える．シミュレーションベースの問題解法では学習データの標本数を確保することがとにかく困難であり，近年の機械学習の発展から取り残された印象を持っていたが，その打開の可能性を具体的に示すことができたと考える．

参考文献

[1] M. E. Tipping and C. M. Bishop: Probabilistic principal component analysis. Journal of the Royal Statistical Society, B, 6(3): pp.611–622, 1999.

[2] N. D. Lawrence: Gaussian process latent variable models for visualisation of high dimensional data, In Advances in neural information processing systems, pp.329–336, 2004.

[3] 持橋大地，大場成正：ガウス過程と機械学習（機械学習プロフェッショナルシリーズ），講談社，2019

[4] https://www.3ds.com/ja/3dexperience/

[5] E.J. Kansa: “Motivation for Using Radial Basis Functions to Solve PDE’s”, (Unpublished paper,1999, 現在は ” https://people.clarkson.edu/~gyao/kansa_rbf_pde.pdf” からダウンロード可能