簡便エミュレーションによる実験計画の高速化

樋口知之

情報・システム研究機構　統計数理研究所

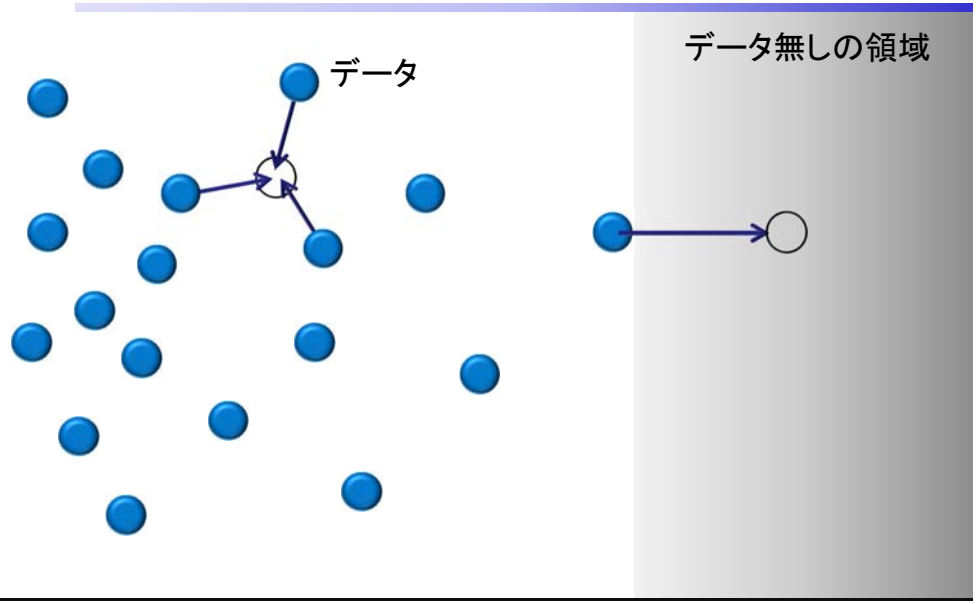
アウトライン

1. データ同化
   a. 基本概念と目的
   b. アルゴリズムの基本
   c. 設計技術の革新
   d. DAセンターによる応用事例
2. 簡便エミュレータ
   a. 必要な理由と動向
   b. 構成法

3. スパース回帰：LASSO, CS, NMF
4. GPR: Gaussian Process Regression

1.a 基本概念と目的

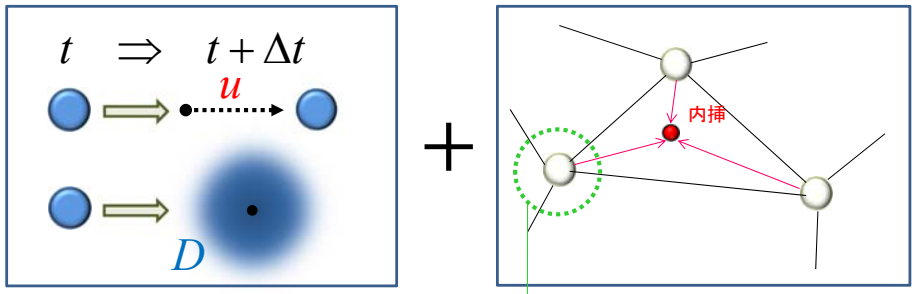
内挿と外挿問題



いろいろなビジネス展開が可能

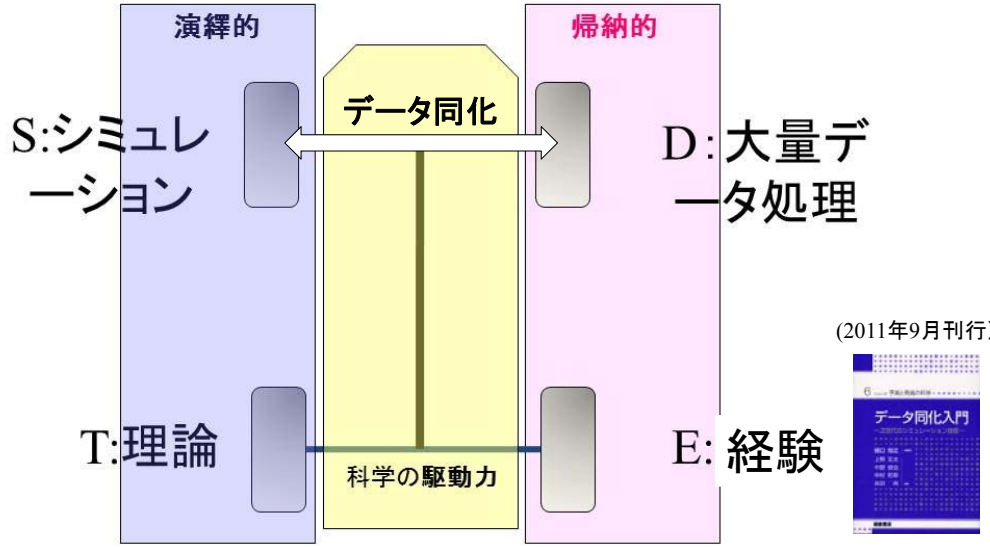
移流と拡散 ＋ **クラウドソーシング** ＝ 予測能力

フォワード計算モデル　　現況を捉える認識力 スマホ(とGPS)



統計数理研究所

つなぐ：データ同化



統計数理研究所

1.b アルゴリズムの基本

統計数理研究所

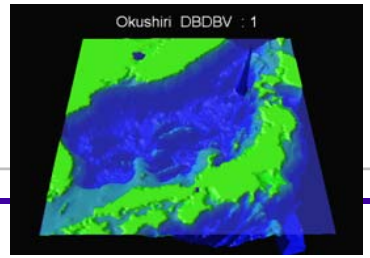
シミュレーションモデルの構成（1）

PDE to approximate real physical system
(continuous time/space)

$\frac{\partial x}{\partial t} = cx^2 + \cdots$　PDE : Partial differential equation

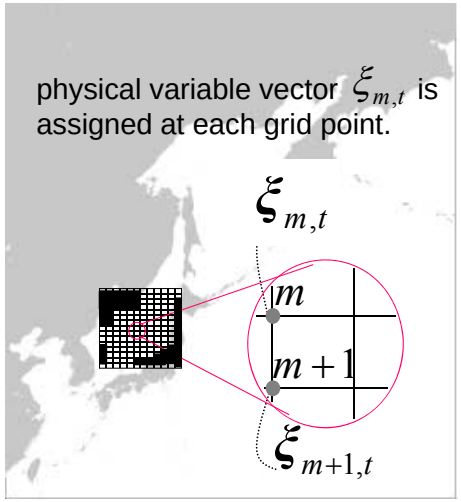
Discrete simulation model
(discrete time/space, FDE)

Suppose a case where we conduct a two-dimensional simulation experiment for understanding the flow of shallow water such as the tsunami.



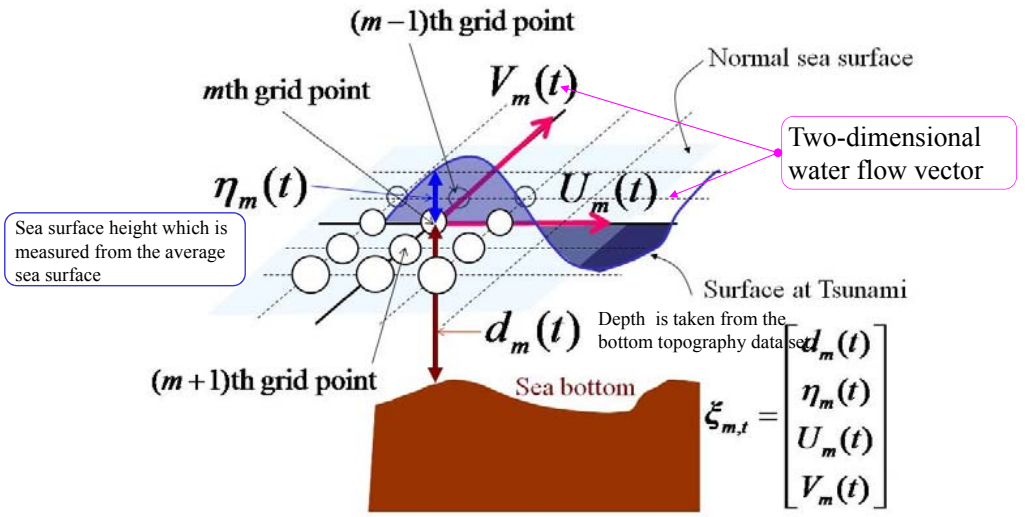
（Tsunami simulation model in Japan Sea）

physical variable vector $\xi_{m,t}$ is assigned at each grid point.



統計数理研究所

シミュレーションモデルの構成 (2)



システムモデルとしてのシミュレーションモデル

(simplified meteorological model around Japan)

PDE to approximate real physical system (continuous time/space)

$\frac{\partial x}{\partial t} = cx^2 + \cdots$ PDE : Partial differential equation

Discrete simulation model (discrete time/space, FDE)

$x_t = f_t(x_{t-1})$

代入計算

(time varying) Boundary conditions v_t

Nonlinear state space model (discrete time/space, stochastic(SDE))

$x_t = f_t(x_{t-1}, v_t)$

State Vector

$$x_t = \begin{bmatrix} \xi_{1,t} \\ \vdots \\ \xi_{m,t} \\ \xi_{m+1,t} \\ \vdots \\ \xi_{M,t} \\ \theta \end{bmatrix}$$

データ同化と一般状態空間モデル

State Vector (Simulation variables)

システムモデル
Stochastic simulation model

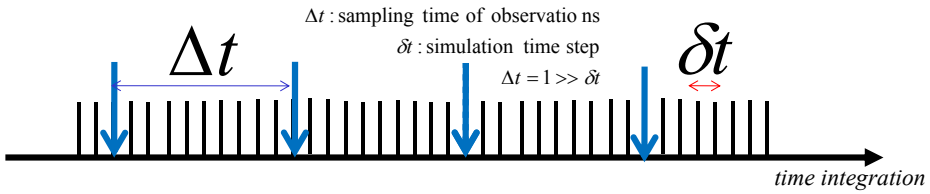
$$x_t = f_t(x_{t-1}, v_t), \quad v_t \sim p(v \mid \theta_{\text{sys}})$$

$$y_t = h_t(x_t, w_t), \quad w_t \sim p(w \mid \theta_{\text{obs}})$$

気象・海洋のデータ同化の枠組み

$y_t = H_t x_t + w_t, \; w_t \sim N(0, R_{\text{obs}})$

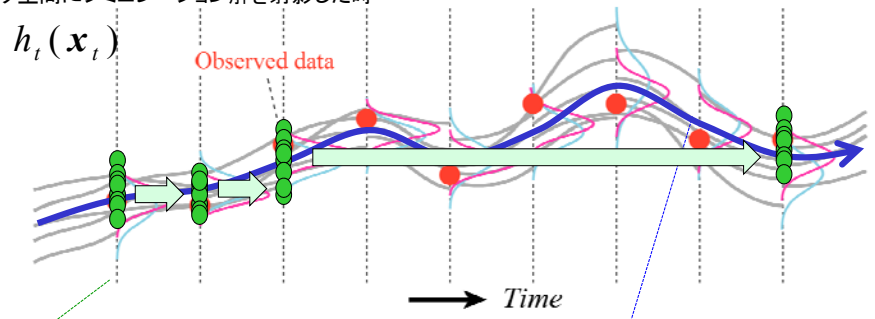
Observation model
Measurement model
観測モデル



逐次（アンサンブル）vs. 非逐次（最適パス）

データ同化のイメージ

データ空間にシミュレーション解を射影した時



逐次（オンライン）型： 集団の時間発展を追う。つまり、**Swarm Filter**

代表例： EnKF (Ensemble Kalman Filter)

非逐次（オフライン）型：**ベスト初期値**をもつパスを求める

代表例： 4次元変分法（Adjoint法）

マルチエージェントシミュレーションへの応用

Cyber-physical systems (CPS) are engineered systems that are built from and depend upon the synergy of computational and physical components.

The term "cyber-physical systems" refers to the tight conjoining of and coordination between computational and physical resources.

National Science Foundation WHERE DISCOVERIES BEGIN

An embedding of the multi-agent based mode, that is a person-based model to describe each behavior, can be easily achieved by replacing a **grid** with a **person** in the definition of the state vector.

Human Modeling

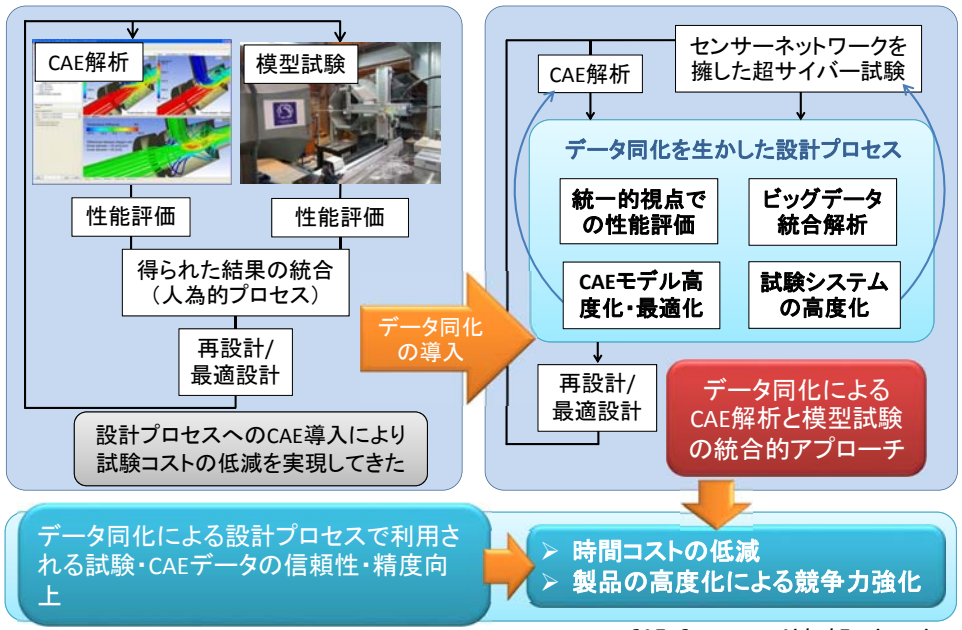
$$\xi_{m,t} = (\text{variables to specify a behavior of the } m\text{-th agent})$$

$$\boldsymbol{x}_t = \begin{bmatrix} \xi_{1,t} \\ \vdots \\ \xi_{m,t} \\ \xi_{m+1,t} \\ \vdots \\ \xi_{M,t} \\ \boldsymbol{\theta} \end{bmatrix} \Leftarrow f_t \left(\begin{bmatrix} \xi_{1,t-1} \\ \vdots \\ \xi_{m,t-1} \\ \xi_{m+1,t-1} \\ \vdots \\ \xi_{M,t-1} \\ \boldsymbol{\theta} \end{bmatrix} \right)$$

$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ **MAS**: Multi-Agent based Simulation

1.c 設計技術の革新

データ同化による設計技術の革新



CAE: Computer-Aided Engineering

データ同化型シミュレーション技術開発によるものづくり・設計の革新
－データ同化技術の工学への応用研究開発－(1/2)

データ同化型シミュレーション技術開発によるものづくり・設計の革新
ーデータ同化技術の工学への応用研究開発ー(2/2)

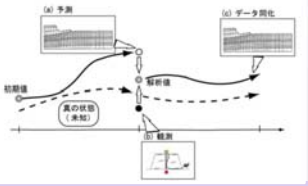
COCN Council on Competitiveness-Nippon
HPC応用研究会提言
産業競争力懇談会
(2012年3月6日)
より抜粋・要約

【効果あるモデリング方法論】
(ア)自然現象を数学モデルに近似するモデルによる誤差 (イ)材料データなど入力データを持つばらつき(構成式, 減衰率, 熱伝導率など)に対応するシーズ技術としては, DA(データ同化)を提言したい。DAは複雑な現象を科学的に理解し, 精度よく予測したいという要求にこたえる技術であり, モデルから複雑現象を再現する演繹アプローチと複雑現象の観測結果からモデルを推測する帰納アプローチが融合した次世代のシミュレーション技術である。これまで主に地球科学の分野において数値モデルの再現性を高めるためにモデルに観測データを埋め込み, 馴染ませることを意図して研究された。モノづくり分野において, なじみが薄い名前ではあるが, たとえば, 電子機器の設計分野では20年も前から実質的に同等の技術が活用されている。

ー現代社会に強く求められるデータ同化技術の研究開発ー

社会インフラ エミュレータ研究事例

地盤工学におけるデータ同化 構造シミュレーション



地盤挙動メカニズム解析と将来予測システムの開発
地盤の変位や応力、間隙水圧の確率分布を得てこれらの確率分布を基礎や土構造物のリスク評価に適用

地盤内応力状態把握 精度検証の実現

次世代ものづくり開発 研究事例

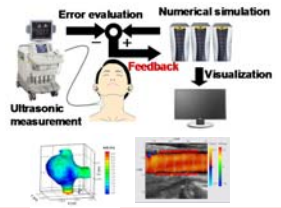
後方乱気流のライダ計測融合 シミュレーション

次世代の飛行機運航システムの開発
仙台空港における実運航機の後方乱気流をライダー計測し3次元の流体シミュレーションに融合

離発着間隔の制限解除 安全時短省エネ運航の実現

先進的治療・手術エミュレータ研究事例

データ同化型血流シミュレーション



超音波計測融合血流解析システムの開発
血流の超音波計測と数値解析を一体化し血流解析システムを開発し、疾患との関係や診断指標について研究

高度診断の実現

1.d DAセンターによる応用事例 （ちょっと脇道）

Application studies carried out by our group for **model** improvements

Research Projects:


- Typhoon trajectory


- Tsunami, Ocean tide


- Auroral phenomena


- 3D structure of ring current

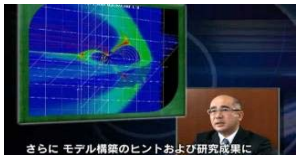

- Acoustic waves
- Intercellular fluid dynamics

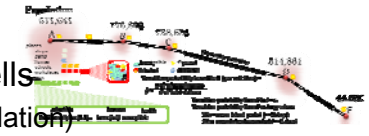

- Genome informatics
- Drug response prediction

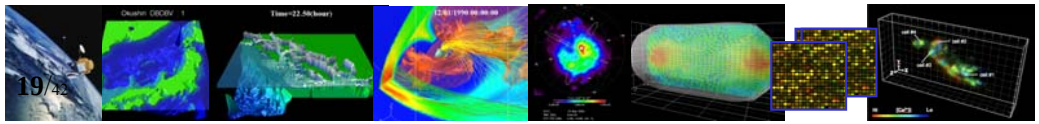

- Neuronal circuits of whole nerve cells


- Influenza Pandemic (Multi agent simulation)

Watch YouTube "Data Assimilation R&D Center at ISM".







簡便エミュレータ
2.a 必要な理由と動向

通常のエミュレータの概念

エミュレータ (Emulator)とは、コンピュータや機械の模倣装置あるいは模倣ソフトウェアのことである。

概要

コンピュータ分野で使われることが多い用語だが、もともとは機械装置全般に使う言葉である。判りやすく言えば、機械を真似る機械である。

語源

エミュレーションやエミュレータは、模倣対象のシステムにおいて、予測できる現象より予測できない現象が支配的である場合に使われる。また、非常に高い安全性が要求される場合にも良く使われる。予測できる現象が支配的な場合や、完全に模倣することが難しい場合はシミュレーション技術を使う。

(Wikipediaより)

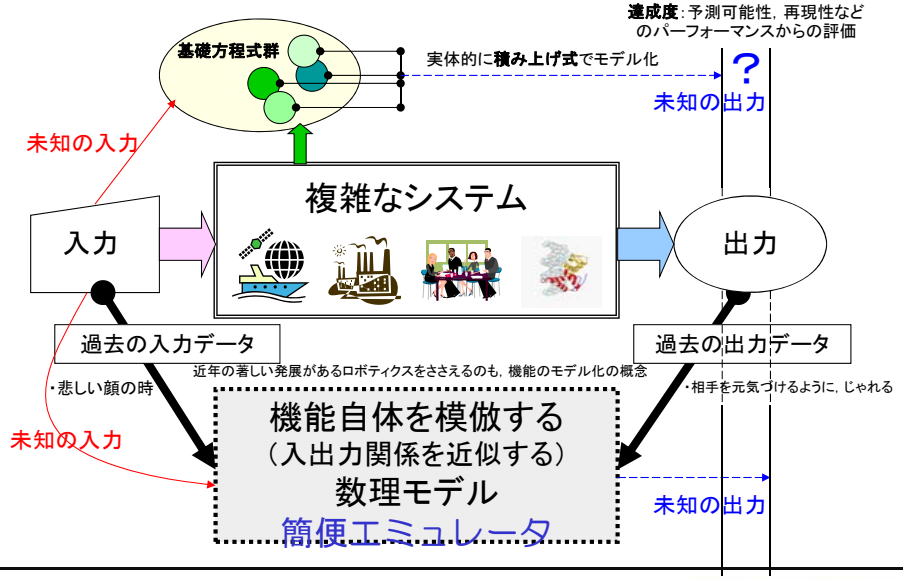
計算コストと予測精度

- 高い計算コスト(オンライン計算が難しい)
- 不確実性の取り扱い方

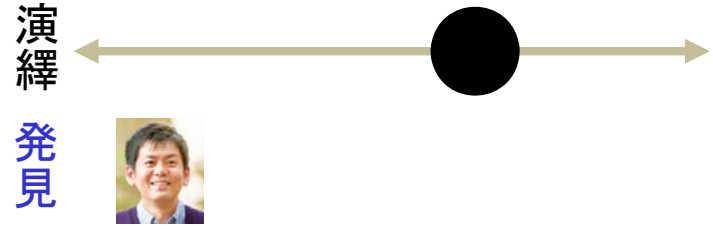
“不確実性を、世界を理解する人間の能力に付随するものではなく、実験に付随するものとしてとらえている。”

機能のモデル化：簡便エミュレータ

高度情報社会におけるユニバーサルな研究課題の表現形



予測と発見



鹿島・京大教授

グレイボックスモデル
(元トヨタ　大畠氏)

統計数理(2006)
第 54 巻 第 2 号 209–210
©2006 統計数理研究所

2006年

特集「予測と発見」

「特集 予測と発見」について

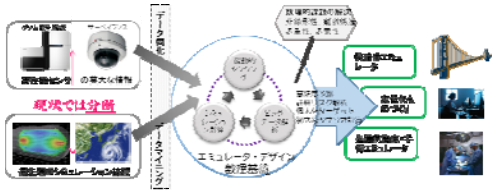
樋口 知之† (オーガナイザー)

UQ: Uncertainty Quantification

- 欧米では、計算機シミュレーション結果の信頼性を具体的に確立するための方法論の研究が急速に熱を帯びてきており、ASME(The American Society of Mechanical Engineers)がVerification and Validation（通常V&Vと呼称）の標準化に大きな力を注いでいる。例えば、2006年には固体力学に対して、2009年には流体力学および熱解析に関する計算機シミュレーションのV&Vが公表されている。
- 欧州においては流体力学の分野で同種の研究活動が2012年から活発化しており、Uncertainty Quantification (UQ) in Industrial Analysis and Design の名のプロジェクト研究が現在進行中である。
- NASAでは、NASA UQ challenge 2014と題して、スパースな限定されたパラメータセットに関するシミュレーションの結果データから、UQをモデル化するコンペを開始した。
- 米国統計コミュニティは、2011-12年に、NSFのサポートを受ける機関SAMSI（Statistical and Applied Mathematical Sciences Institute）にてUQを集中的に研究するプログラムを立ち上げた。
- 米国統計学会はSIAM(Society for Industrial and Applied Mathematics)と共同でJournal on UQの刊行を2014年に開始した。その雑誌の取り扱う主たる分野としてsensitivity analysis, model validation, model calibration, data assimilationの4つがあげられている。最新号の論文（4本掲載）は、感度解析、ガウス過程回帰、モデル較正、ギブスサンプラーの解析のテーマとなっており、ほぼ統計学の範疇である。

重要な技術：
ガウス過程回帰や、その古典版とも言えるクリギング
次元削減を目的としたスパース回帰

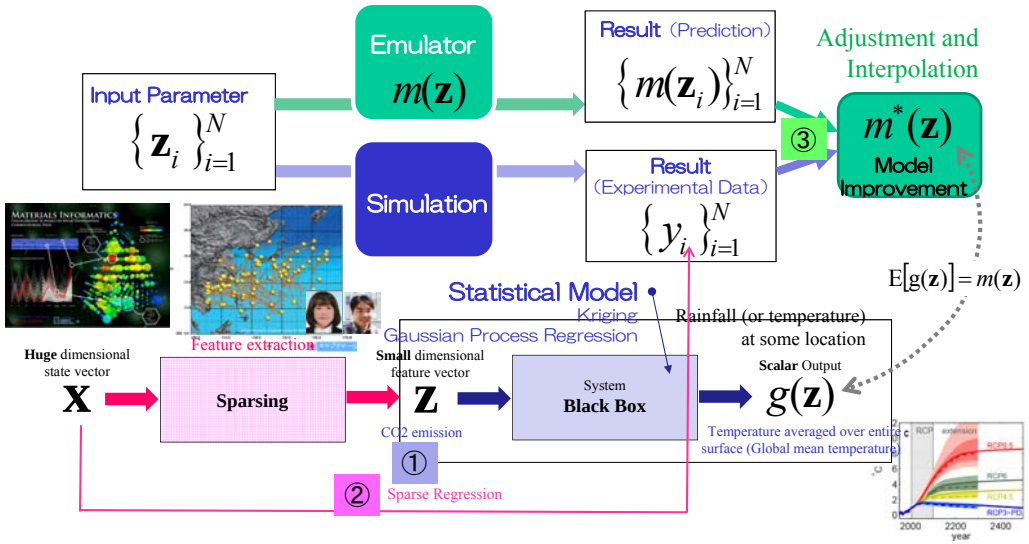
中野慎也、樋口知之、地球科学におけるシミュレーションとビッグデータ —データ同化とエミュレーション—、電子情報通信学会誌、Vol.97(10), pp.869-875, 2014.
樋口知之、中村和幸、データ同化によるオンラインセンシングの高度化, 計測自動制御学会誌, Vol.51(9), 2012.
長尾大道、佐藤光三、樋口知之, マルコフ連鎖モンテカルロ法を利用したトレーサー試験からフラクチャーの物理パラメータを推定する方法, 石油技術協会誌, Vol.78(2), pp.197-209, 2013.
Iba, Y. and Akaho, S., Gaussian process regression with measurement error, IEICE Trans. E93-D(10), 2010.



2.b 構成法

Simulation → Data Assimilation → Emulator :
Statistical model for predicting an output given input parameters

Monte Carlo Experiments
→Experimental Design, UQ



エミュレータの設計法

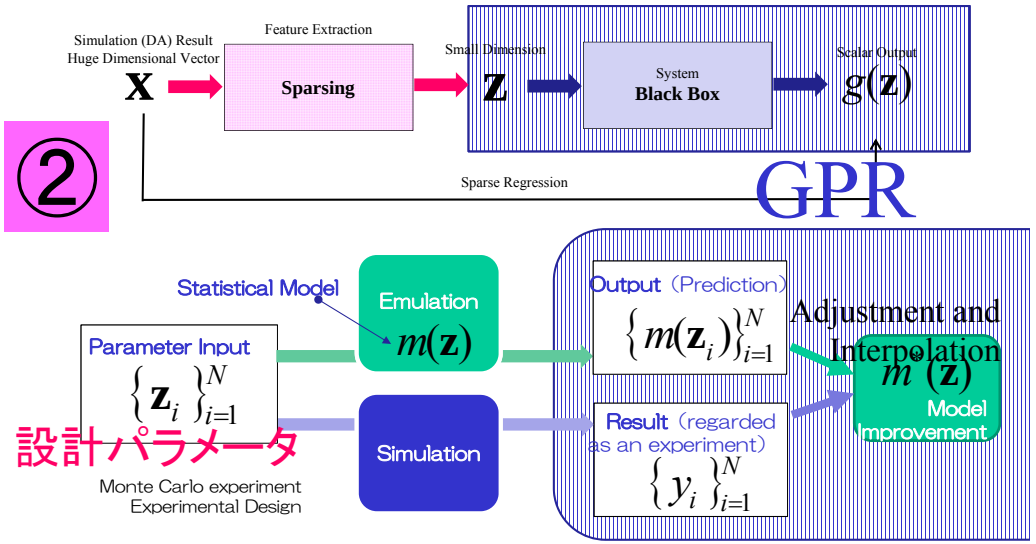
① 1. 目的変数（スカラー値）を決める。

② 2. シミュレーションを複数回走らせる。
3. スパース回帰により、入力パラメータベクトルを同定する。

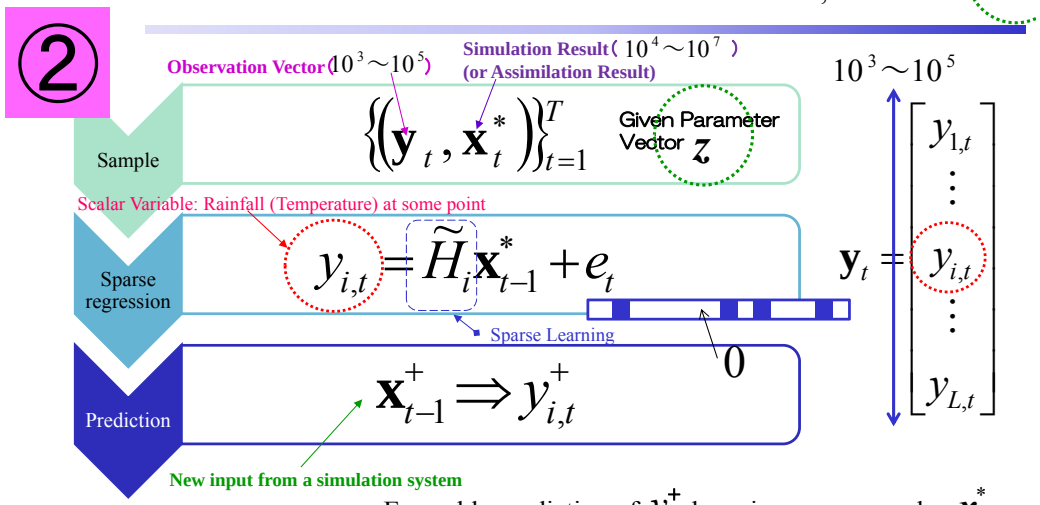
③ 4. GPRにより、出力関数（応答局面）をもとめる。

Emulator and Emulation

Emulate an output of simulation given parameters



Climate Prediction : Climate Emulator $p(y_{i,t} \mid \boldsymbol{x}^{+}_{t-1}, \boldsymbol{z})$



$$\left\{\left(\mathbf{y}_t, \mathbf{x}^{*}_{t}\right)\right\}_{t=1}^{T}$$

$$y_{i,t} = \tilde{H}_i \mathbf{x}^{*}_{t-1} + e_t$$

$$\mathbf{x}^{+}_{t-1} \Rightarrow y^{+}_{i,t}$$

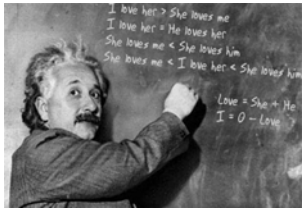
$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{i,t} \\ \vdots \\ y_{L,t} \end{bmatrix}$$

Ensemble prediction of $y^{+}_{i,t}$ by using many samples $\boldsymbol{x}^{*}_{t-1}$

$$p(y^{+}_{i,t} \mid \boldsymbol{x}^{+}_{t-1}) = \int p(y^{+}_{i,t} \mid \boldsymbol{x}^{+}_{t-1}, z)\, p(z \mid \boldsymbol{x}^{+}_{t-1})\, dz$$

3. スパース回帰

LASSO, CS, NMF



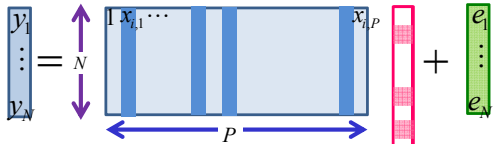
スパース（疎性を利用した）最適化

例：多変量回帰

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots\cdots + a_P x_{i,P} + e_i \quad (i = 1, \ldots, N)$$

$$\boldsymbol{y} = H\boldsymbol{a} + \boldsymbol{e}$$

モデルでは表現できない部分（誤差というのは不適切）



新NP問題 $(N << P)$

$$\boldsymbol{a}* = \min_{\boldsymbol{a}} \left\{ |\boldsymbol{y} - H\boldsymbol{a}|^2 + \lambda |\boldsymbol{a}| \right\}$$

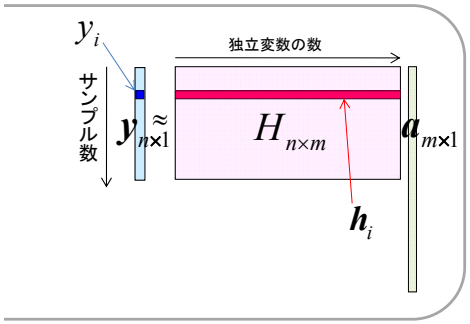
$$\min_{\boldsymbol{a}} |\boldsymbol{y} - H\boldsymbol{a}|^2 \text{ subj. to } |\boldsymbol{a}| \le \alpha$$

解の一意性と解のロバスト化

初期のベイズ型逆問題一般型 $\min_{\boldsymbol{a}} \|\boldsymbol{y}-H\boldsymbol{a}\|_2^2 + \|U\boldsymbol{a}\|_{Q/R}^2$

通常の線形回帰モデル $y_i = \boldsymbol{h}_i \cdot \boldsymbol{a} + w_i$

■ $n > m$ の場合はたいてい大丈夫



■ $m > n$ の場合は解を一意に定められない。

リッジ回帰 $\min_{c} \|\boldsymbol{y}-H\boldsymbol{a}\|_2^2 + \lambda\|\boldsymbol{a}\|_2^2$

ロバストな解を求めるために $\min_{c} \|\boldsymbol{y}-H\boldsymbol{a}\|_2^2 + \lambda\|\boldsymbol{a}\|_1$

$\min_{c} \|\boldsymbol{y}-H\boldsymbol{a}\|_1 + \lambda\|\boldsymbol{a}\|_1$

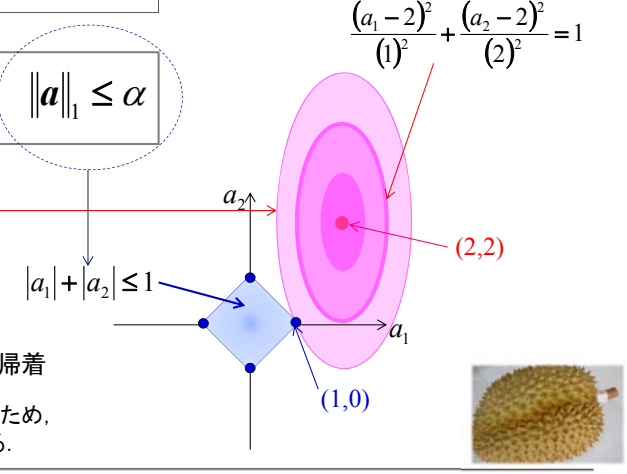
LASSO: Least Absolute Shrinkage and Selection Operator

R.Tibshirani (*JRSS B*; 1996)

■ $m > n$ の場合は解を一意に定められない。

$\min_{\boldsymbol{a}} \|\boldsymbol{y}-H\boldsymbol{a}\|_2^2 + \lambda\|\boldsymbol{a}\|_1$

$\min_{\boldsymbol{a}} \|\boldsymbol{y}-H\boldsymbol{a}\|_2^2$ subj. to $\|\boldsymbol{a}\|_1 \le \alpha$



具体的解法は，二次計画法問題に帰着

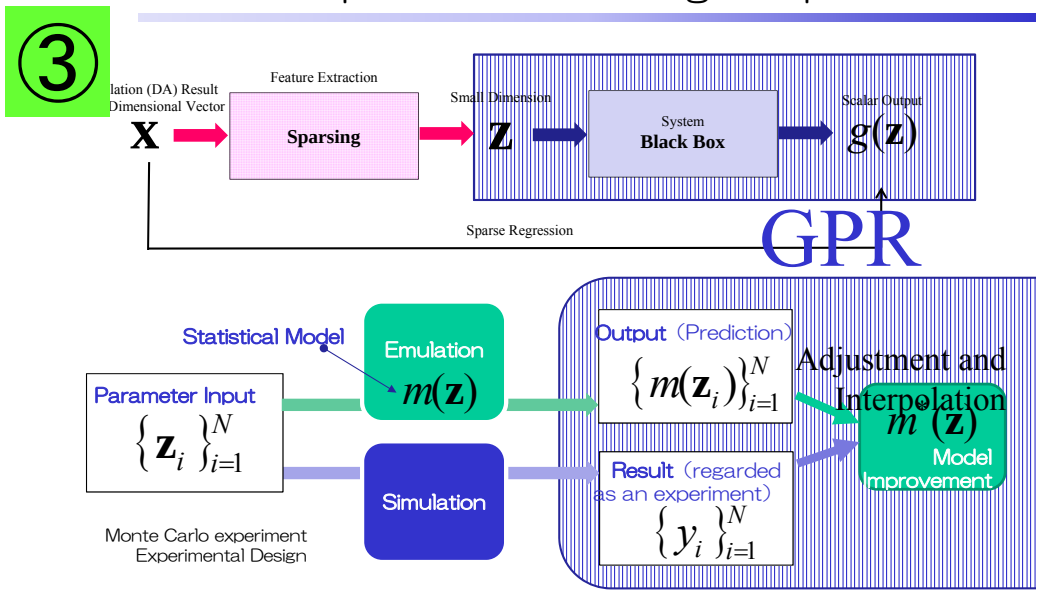
実際にはαをいろいろ変える必要があるため，
効率的なアルゴリズムが提案されている．

LARS: Least Angle Regression (Efron *et al.*, 2004)

4. Gaussian Process Regression

Emulator and Emulation

Emulate an output of simulation given parameters



Emulator : Statistical Model (Linear Regression+GP)

System Model

Gaussian Process : g is a continuous function of $\mathbf{z}$

$$p(g)=p(g(\mathbf{z}))=N(m(\mathbf{z}),k(\mathbf{z},\mathbf{z}')\cdot\tau^2)$$

Regression Coefficient

$$\mathrm{E}[g(\mathbf{z})]=m(\mathbf{z})=H\begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}\quad \mathrm{Cov}[g(\mathbf{z}),g(\mathbf{z}')]=k(\mathbf{z},\mathbf{z}')\cdot\tau^2$$

Statistical model is given — Covariance function — Kernel function : Mutual Distance between input parameters — Common variance

Kernel function

$$k(\mathbf{z},\mathbf{z}')=\exp\{-(\mathbf{z}-\mathbf{z}')^T R(\mathbf{z}-\mathbf{z}')\}$$

Observation Model

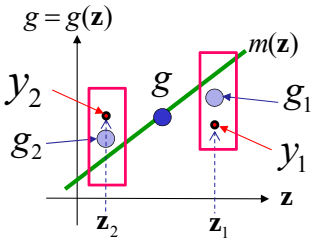
$$y_i=g_i+e,\quad e\sim N(0,\sigma^2)_{(i=1,\cdots,N)}$$

Posterior Distribution

g is a continuous function

$$p(g\mid Y)\propto p(Y\mid g)p(g)$$

$$Y^{\mathrm{T}}=[y_1,\cdots,y_N]$$



統計数理研究所

Emulator : Design of Kernel function

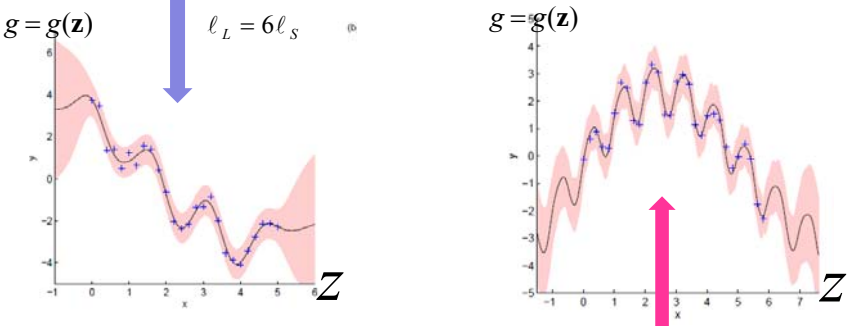
$$\mathrm{Cov}_y[g(\mathbf{z}),g(\mathbf{z}')]=\tau^2k(\mathbf{z},\mathbf{z}')+\sigma^2\delta(\mathbf{z}-\mathbf{z}')$$

$$\mathrm{E}_y[g(\mathbf{z})]=m(\mathbf{z})$$

$$\tau_1^2\exp\left[-\frac{(z-z')^2}{2\ell_L^2}\right]+\tau_2^2\exp\left[-\frac{(z-z')^2}{2\ell_S^2}\right]+\sigma^2\delta(z-z')$$

Kernel function is NOT related to a mean function m(z).

$g=g(z)$ $\quad \ell_L=6\ell_S$ $\quad g=g(z)$

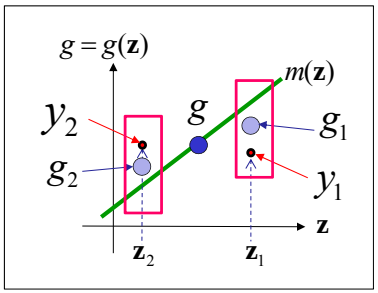


Gaussian Processes for Regression:A Quick Introduction
M. Ebden, August 2008

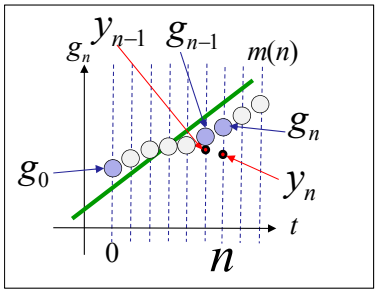
$$\tau_1^2\exp\left[-\frac{(z-z')^2}{2\ell^2}\right]+\tau_2^2\exp\left[-2\sin^2(\nu\pi(z-z'))\right]+\sigma^2\delta(z-z')$$

統計数理研究所

Emulator : Similar structure to SSM if discretizing



If g is defined only on the discrete point



$$\begin{cases} g_n=m(n)+g_{n-1}+v_n,\quad v_n\sim N(0,\tilde{\tau}^2)\\ y_n=g_n+e,\quad e\sim N(0,\sigma^2)\end{cases}$$

Kernel function should positive definite

$$k(n,n-1)=\exp\{-(1)^T R(1)\}$$

統計数理研究所

Emulator : Online Adjustment (Calibration) and Interpolation

For simplicity

In a case for $\dfrac{\sigma^2}{\tau^2}\to 0$

Posterior distribution

$$p(g(\mathbf{z})\mid Y)\sim N(m^*(\mathbf{z}),k^*(\mathbf{z},\mathbf{z}')\tau^2)$$

$$M^{\mathrm{T}}=[m(\mathbf{z}_1),\cdots,m(\mathbf{z}_N)]$$

$$m^*(\mathbf{z})=m(\mathbf{z})+(Y-M)K^{-1}\mathbf{t}(\mathbf{z})$$

Difference between input z and some points

nce between Data Statistical Model — Normalization factor

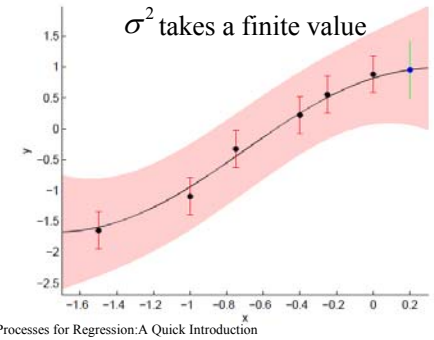
$$\mathbf{z},\mathbf{z}_1),\cdots,k(\mathbf{z},\mathbf{z}_N)]$$

$$k(\mathbf{z},\mathbf{z}')-\mathbf{t}^{\mathrm{T}}(\mathbf{z})K^{-1}\mathbf{t}(\mathbf{z}')$$

n from data always reduce uncertainty

$$\begin{bmatrix} k(\mathbf{z}_1,\mathbf{z}_1) & \cdots & k(\mathbf{z}_1,\mathbf{z}_N)\\ \vdots & \ddots & \vdots\\ k(\mathbf{z}_N,\mathbf{z}_1) & \cdots & k(\mathbf{z}_N,\mathbf{z}_N)\end{bmatrix}$$

σ^2 takes a finite value



Gaussian Processes for Regression:A Quick Introduction
M. Ebden, August 2008

Figure 2: The solid line indicates an estimation of y_* for 1,000 values of x_*. Pointwise 95% confidence intervals are shaded.

統計数理研究所

Emulator : Obtain a full Bayes model

For simplicity
In a case for $\frac{\sigma^2}{\tau^2} \to 0$

Posterior distribution

$$p(g(\mathbf{z}) \mid Y, \tau^2, H) \sim N(m^*(\mathbf{z}), k^*(\mathbf{z}, \mathbf{z}')\tau^2)$$

$$p(g \mid Y) = \int p(g \mid Y, \tau^2, H) p(\tau^2) p(H) d\tau^2 dH$$

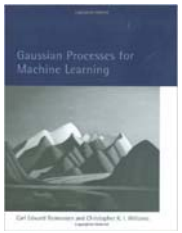
$$p(\tau^2) \propto \frac{1}{\tau^2}, \quad p(H) \propto 1$$

Gamma distribution　　Number of samples

t distribution with a degree of freedom $N - \lambda$

Emulator　　$\lambda = \dim(\mathbf{z}) + 1$

For a part of Regression model



参考文献
J. Sacks *et al*., "Design and analysis of computer experiments," *Statistical Science*, 1989.
M. C. Kennedy and A. O'hagan, "Bayesian calibration of computer models," *J. Roy. Statist. Soc. Ser. B*, 2001.
中野、樋口、"地球科学におけるシミュレーションとビッグデータ―データ同化とエミュレーション―、" 信学会会報、Vol. 97, No.10. 869-875, 2014.
Rasmussen, C. and C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

エミュレータの設計法（再掲）

① 1．目的変数（スカラー値）を決める。

② 2．シミュレーションを複数回走らせる。

3．スパース回帰により、入力パラメータベクトルを同定する。

③ 4．GPRにより、出力関数（応答局面）をもとめる。