簡便エミュレーションによる実験計画のスマート化

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
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
4. GPR: Gaussian Process Regression

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)

$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1})$

代入計算

(time varying)

Boundary conditions $\boldsymbol{v}_t$

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

$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t)$

State Vector

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

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

$$\begin{cases} \boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t), \quad \boldsymbol{v}_t \sim p(\boldsymbol{v} \mid \boldsymbol{\theta}_{\mathrm{sys}}) \\ \boldsymbol{y}_t = h_t(\boldsymbol{x}_t, \boldsymbol{w}_t), \quad \boldsymbol{w}_t \sim p(\boldsymbol{w} \mid \boldsymbol{\theta}_{\mathrm{obs}}) \end{cases}$$

気象・海洋のデータ同化の枠組み

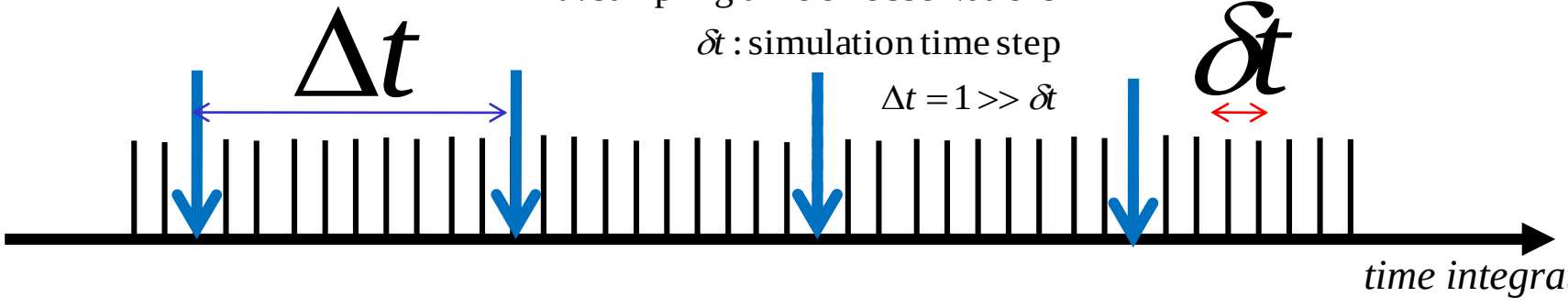
$$\boldsymbol{y}_t = H_t \boldsymbol{x}_t + \boldsymbol{w}_t, \boldsymbol{w}_t \sim N(0, R_{\mathrm{obs}})$$

Observation model
Measurement model
観測モデル

Δt : sampling time of observations

δt : simulation time step

$\Delta t = 1 >> \delta t$



ベイズの反転公式

x：興味のある対象

y：データ

順解析：シミュレーション等

ベイズの反転公式

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{\sum p(y \mid x)p(x)}$$

逆解析

ベイズの定理。等号の右側と左側で，赤と青で示した変数部分の縦棒との相対関係が反転していることがわかる。この事実により，ベイズの反転公式と呼ばれる。

“連成”ベイズ（ベイズシミュレーション）の普及

■ベイズモデルの階層がますます高層化　(HDP: 階層ディリクレプロセス)

■ノンパラメトリックベイズ法の浸透

・ディリクレプロセスの利用が自然言語の分野ですすんだ。

・モデル選択からモデル混合へ。

■ミクロからマクロへ生成モデルの連続結合（“連成”）による、いわばフォワード推論がベイズ計算の主流

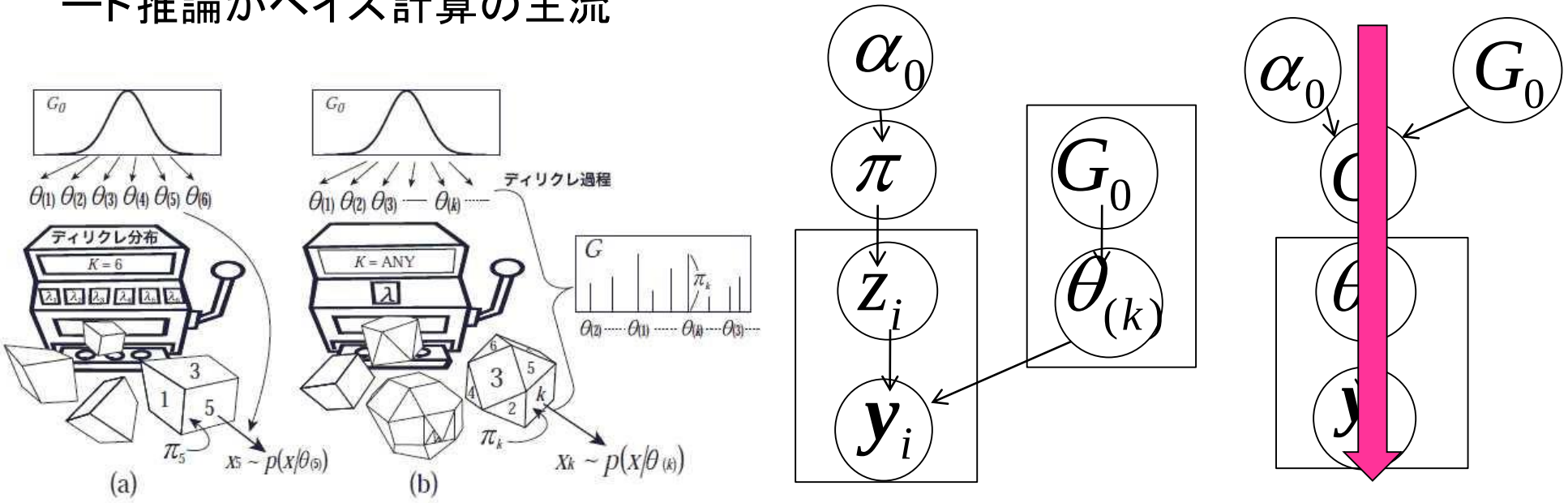


Fig. 2.　ディリクレ分布／ディリクレ過程サイコロ生成器

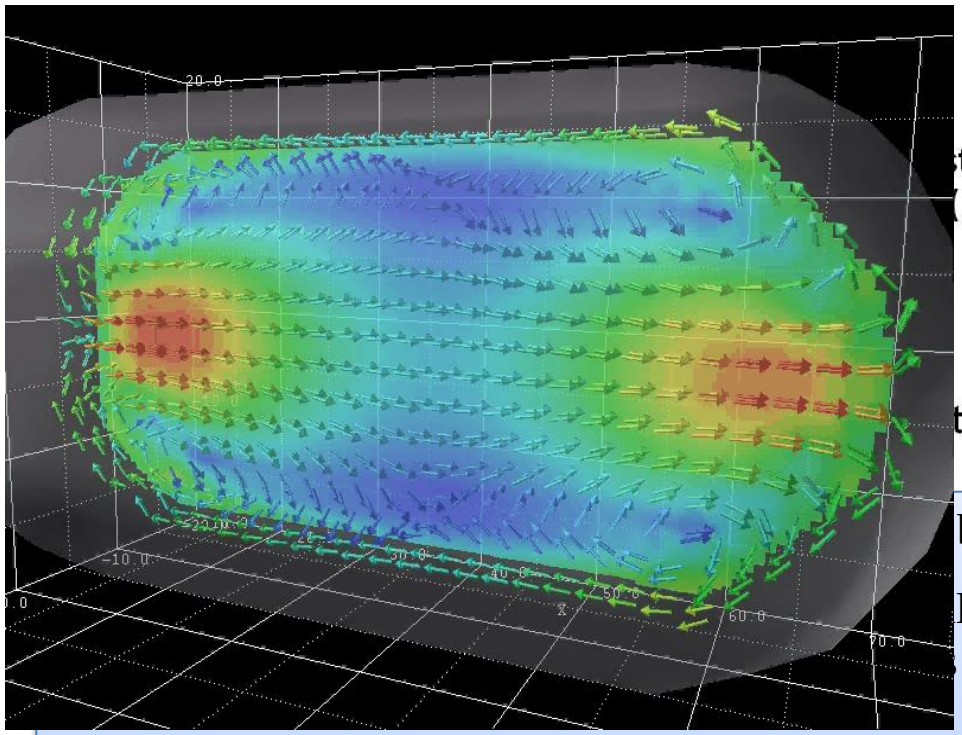
上田＆山田 (応用数理, 2007)から

K-mixture model vs. DP mixture model

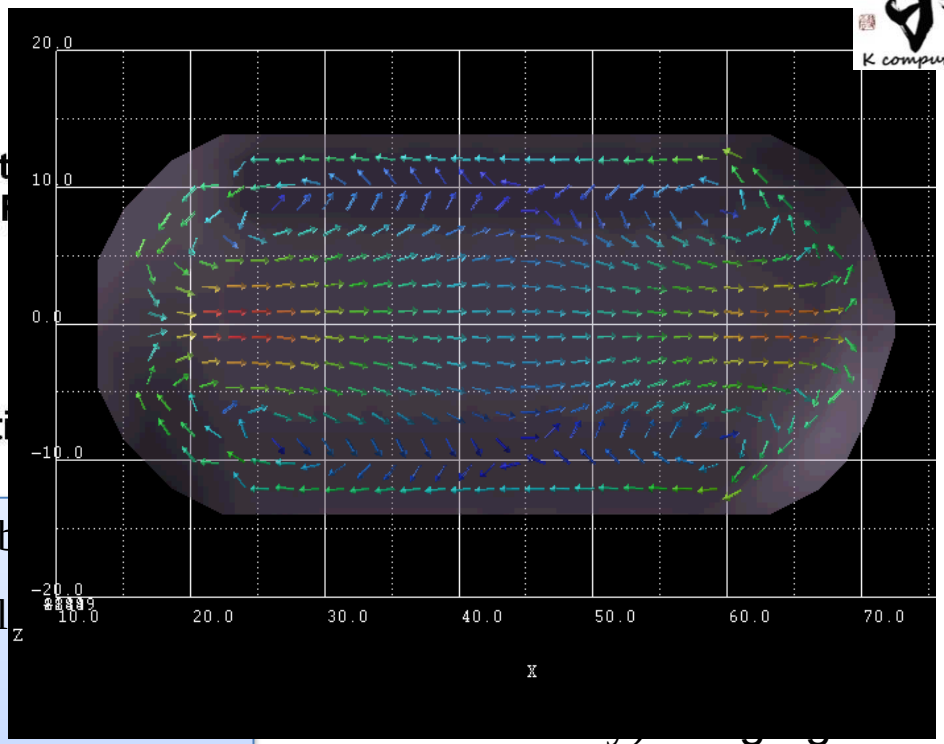
Data Assimilation on Intracellular Fluid in C.elegans embryo
(Collaboration with Dr. Kimuta at NIG)

K computer

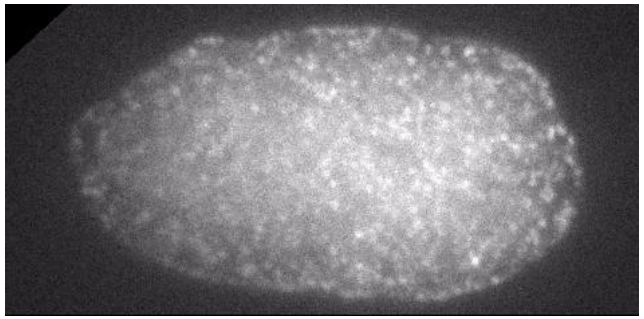
Assimulation



Observation



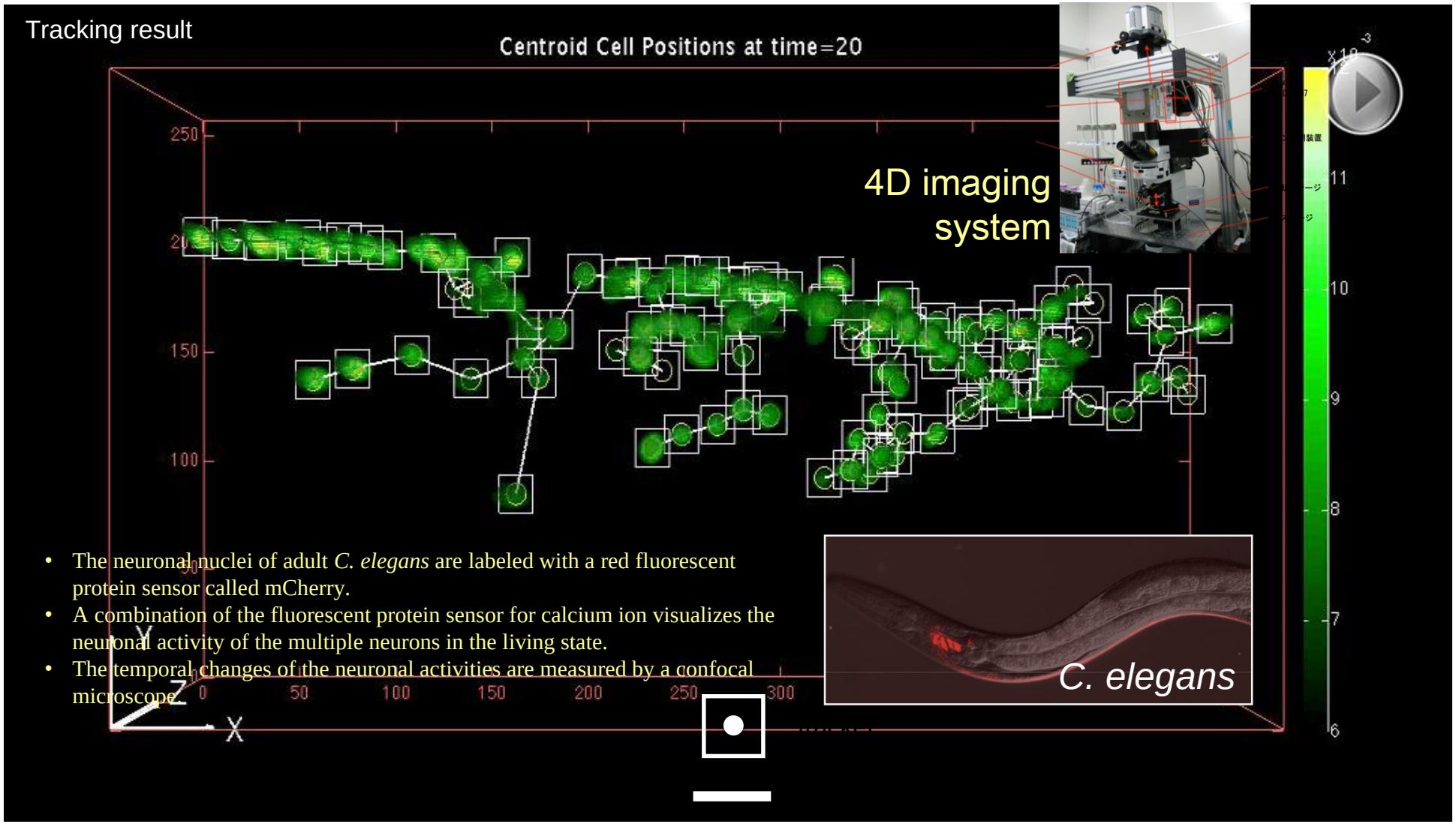
We assume that proteins "**myosin**" in the cell wall drive physically such intra-cellular flow, and we perform data assimilation by combining a numerical simulation that solves simultaneously these physical equations and the observation obtained by the image analyses.



データ同化研究開発センター
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Cell tracking

Tracking result



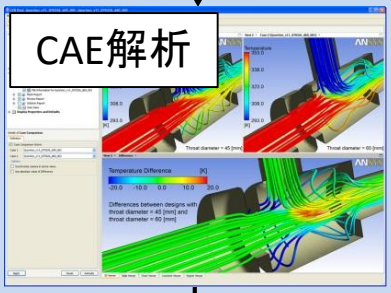
- The neuronal nuclei of adult *C. elegans* are labeled with a red fluorescent protein sensor called mCherry.
- A combination of the fluorescent protein sensor for calcium ion visualizes the neuronal activity of the multiple neurons in the living state.
- The temporal changes of the neuronal activities are measured by a confocal microscope.

Quantitative analysis of neural activities with Ca2+ imaging of C.elegans whole brain

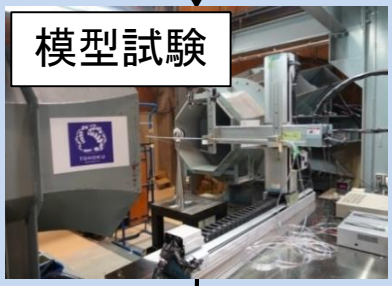
（Thanks to Dr. Tokunaga, 徳永＠統数研＆九工大から借用、改変）

データ同化による設計技術の革新

CAE解析



模型試験



性能評価

性能評価

得られた結果の統合
（人為的プロセス）

再設計/
最適設計

設計プロセスへのCAE導入により
試験コストの低減を実現してきた

データ同化
の導入

CAE解析

センサーネットワークを
擁した超サイバー試験

データ同化を生かした設計プロセス

**統一的視点で
の性能評価**

**ビッグデータ
統合解析**

**CAEモデル高
度化・最適化**

**試験システム
の高度化**

再設計/
最適設計

データ同化による
CAE解析と模型試験
の統合的アプローチ

データ同化による設計プロセスで利用さ
れる試験・CAEデータの信頼性・精度向
上

- **時間コストの低減**
- **製品の高度化による競争力強化**

CAE: Computer-Aided Engineering

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、ギプ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

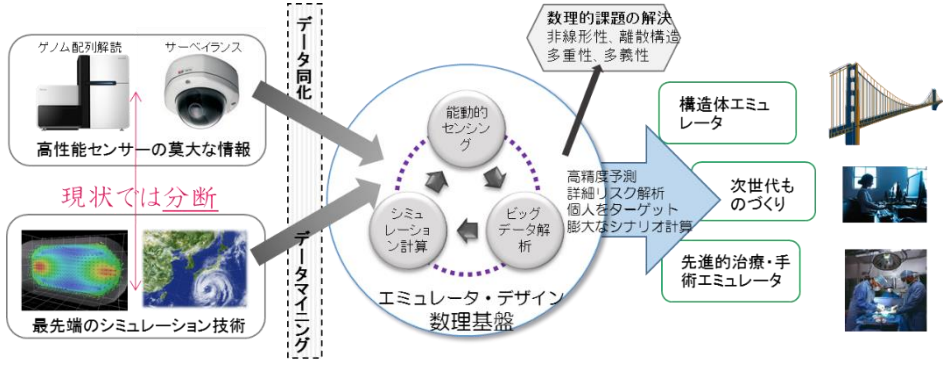
次元削減を目的としたスパース回帰

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.



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

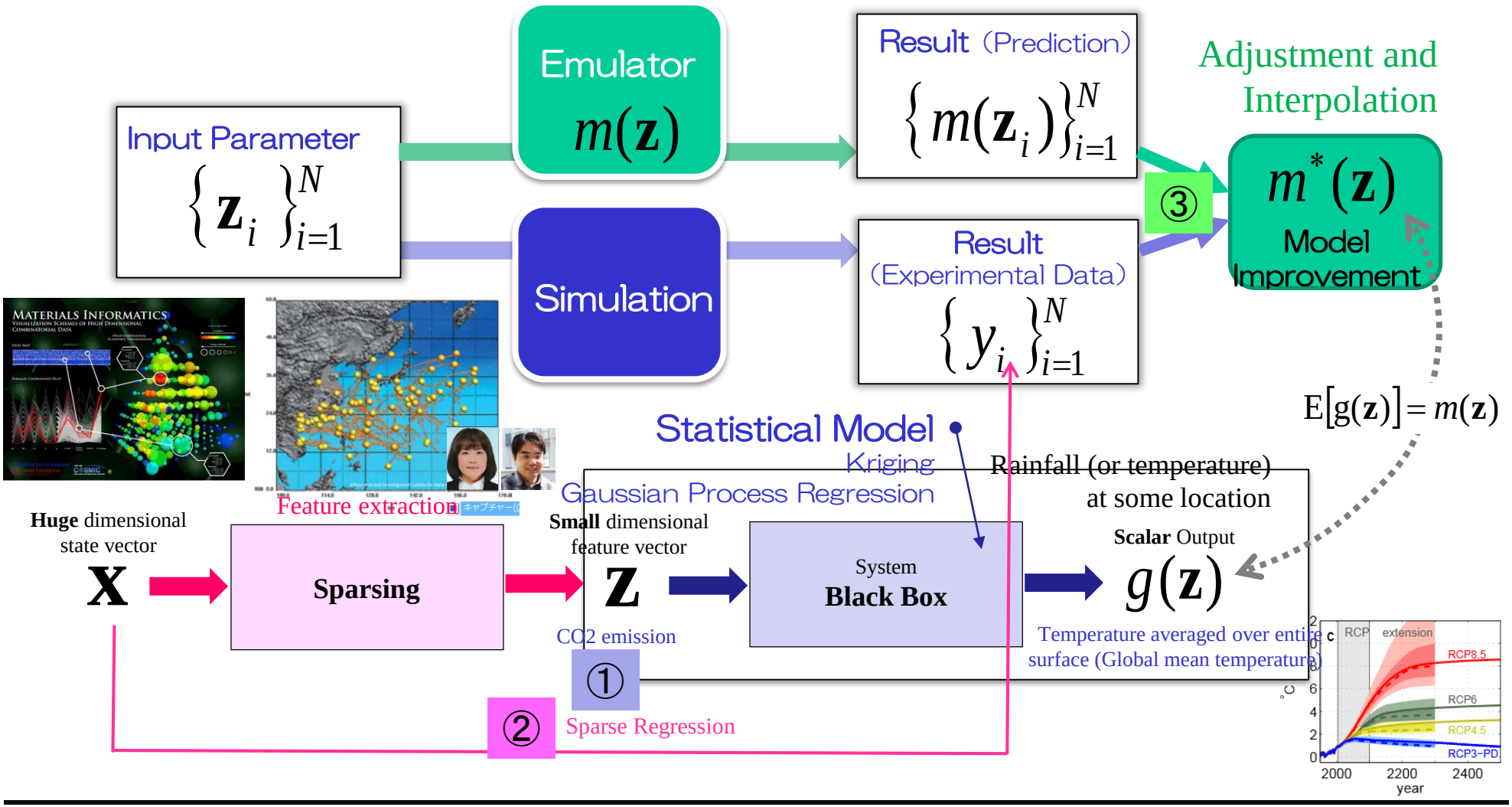
- 高い計算コスト(オンライン計算が難しい)
- 不確実性の取り扱い方

“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。”

Simulation → Data Assimilation → Emulator :

Monte Carlo Experiments
→Experimental Design, UQ

Statistical model for predicting an output given input parameters



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

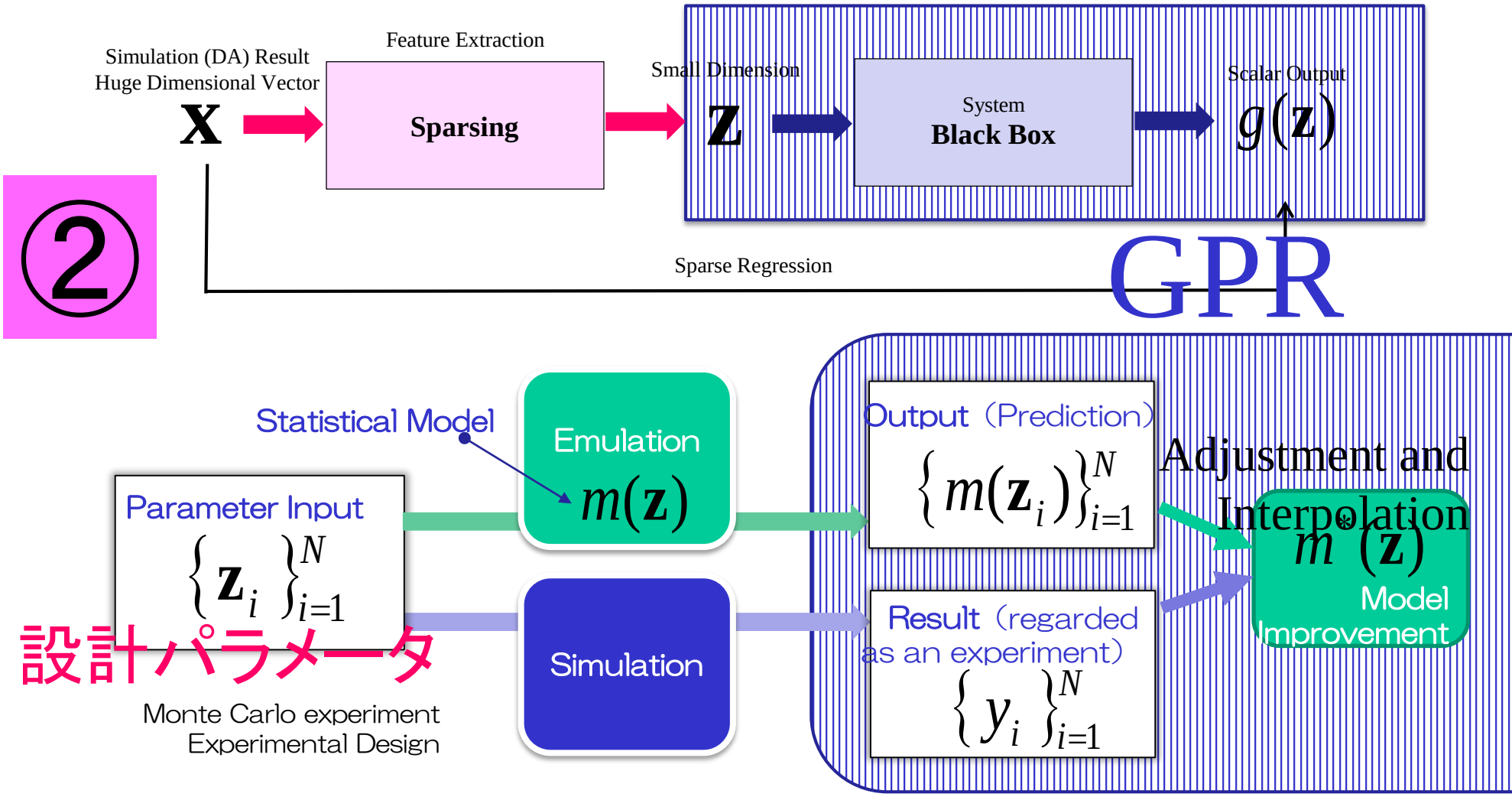
2. シミュレーションを複数回走らせる。

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。

Emulate an output of simulation given parameters



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

②

Observation Vector $(10^3 \sim 10^5)$

Simulation Result $(10^4 \sim 10^7)$ (or Assimilation Result)

Sample

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

Scalar Variable: Rainfall (Temperature) at some point

Sparse regression

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

For given $\mathbf{z}_i$

Sparse Learning

0

Prediction

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

New input from a simulation system

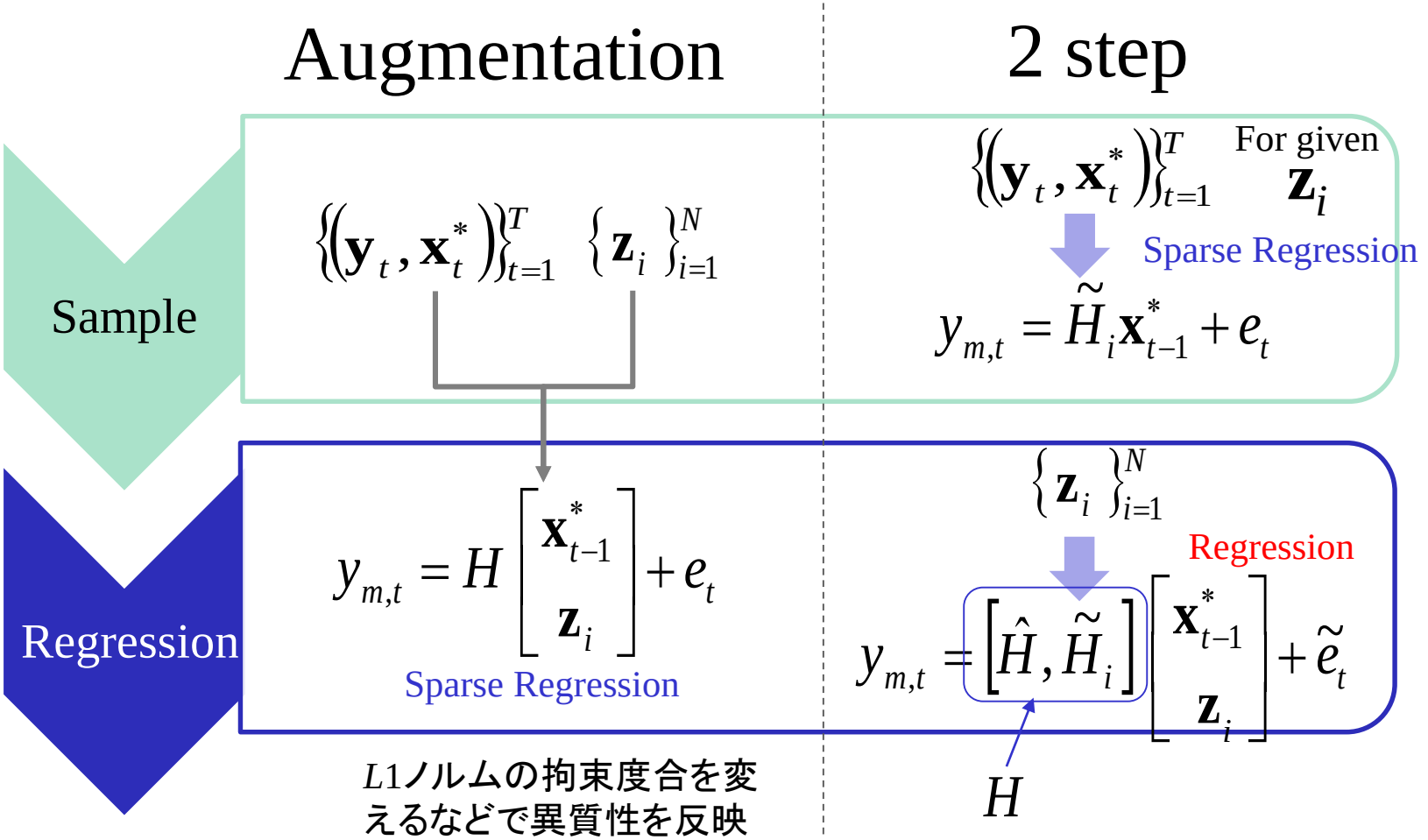
$10^3 \sim 10^5$

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

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

$$p(y^{+}_{m,t} \mid \mathbf{x}^{+}_{t-1}) = \int p(y^{+}_{m,t} \mid \mathbf{x}^{+}_{t-1}, \mathbf{z})\, p(\mathbf{z} \mid \mathbf{x}^{+}_{t-1})\, d\mathbf{z}$$

②



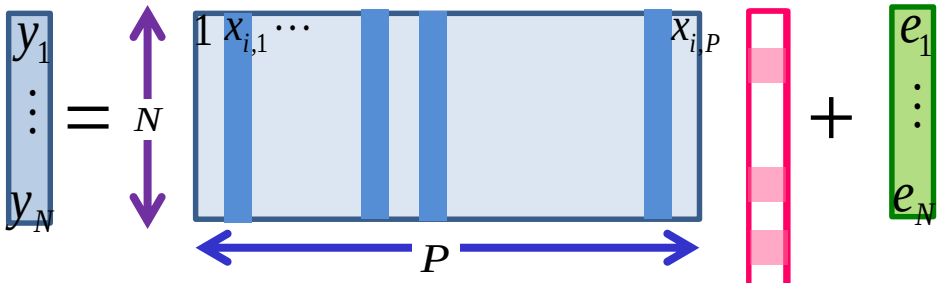
スパース（疎性を利用した）最適化

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 $(N << P)$

0

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

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

解の一意性と解のロバスト化

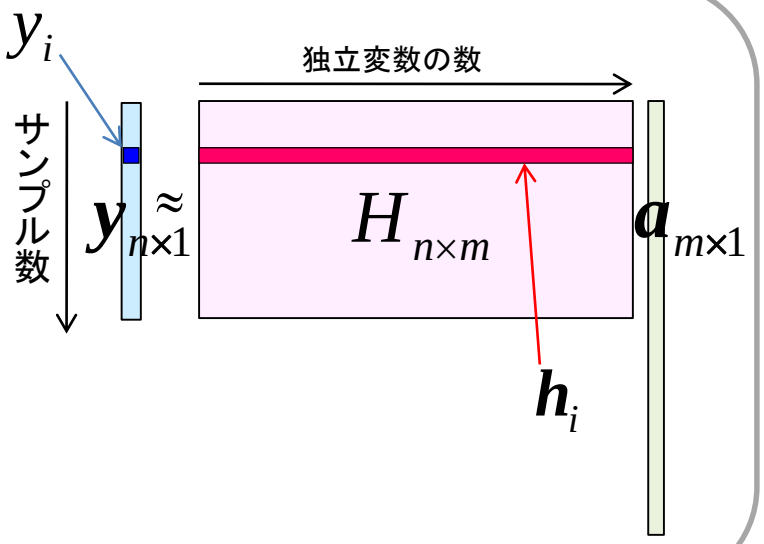
初期のベイズ型逆問題一般型

$$\min_{\boldsymbol{a}} . \left\| \boldsymbol{y} - H\boldsymbol{a} \right\|_2^2 + \left\| U\boldsymbol{a} \right\|_{Q/R}^2$$

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

$$\min_{\boldsymbol{c}} . \left\| \boldsymbol{y} - H\boldsymbol{a} \right\|_2^2 + \lambda \left\| \boldsymbol{a} \right\|_2^2$$

ロバストな解を求めるために

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

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

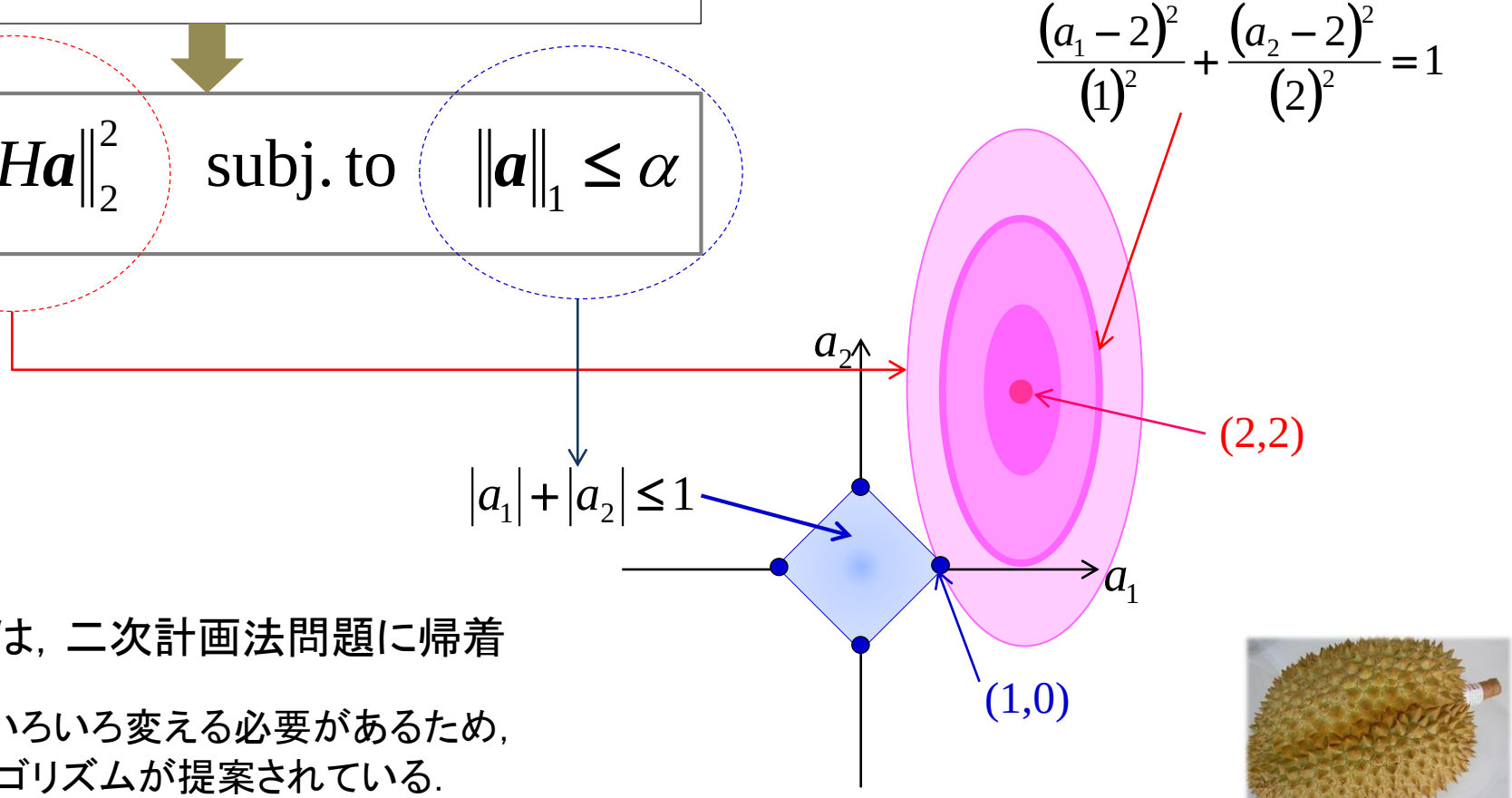
LASSO: Least Absolute Shrinkage and Selection Operator

R.Tibshirani (*JRSS B*, 1996)

■ $m > n$ の場合は解を一意に定められない。

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

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



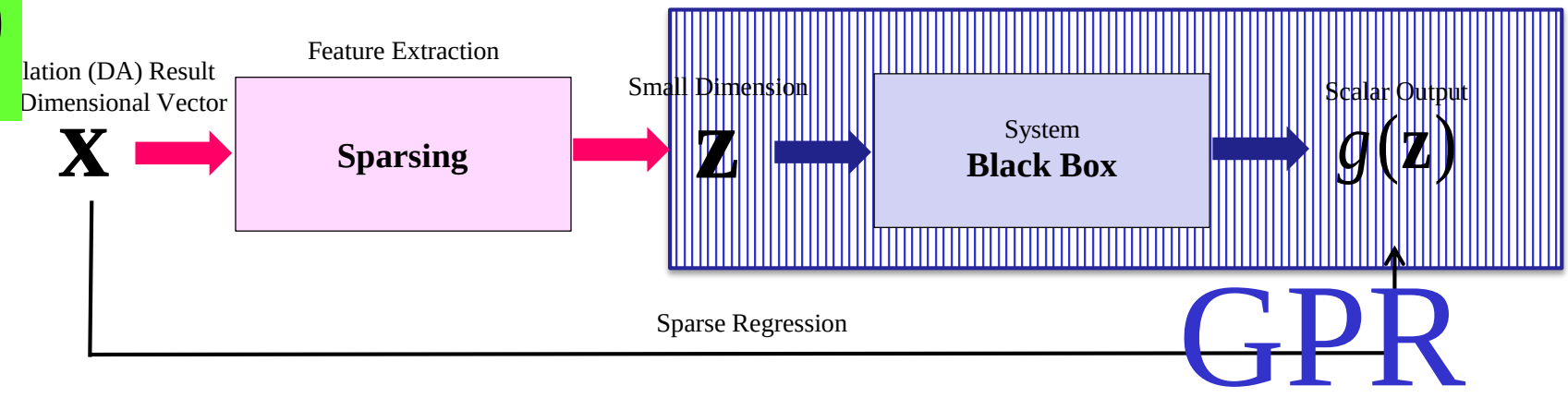
具体的解法は，二次計画法問題に帰着

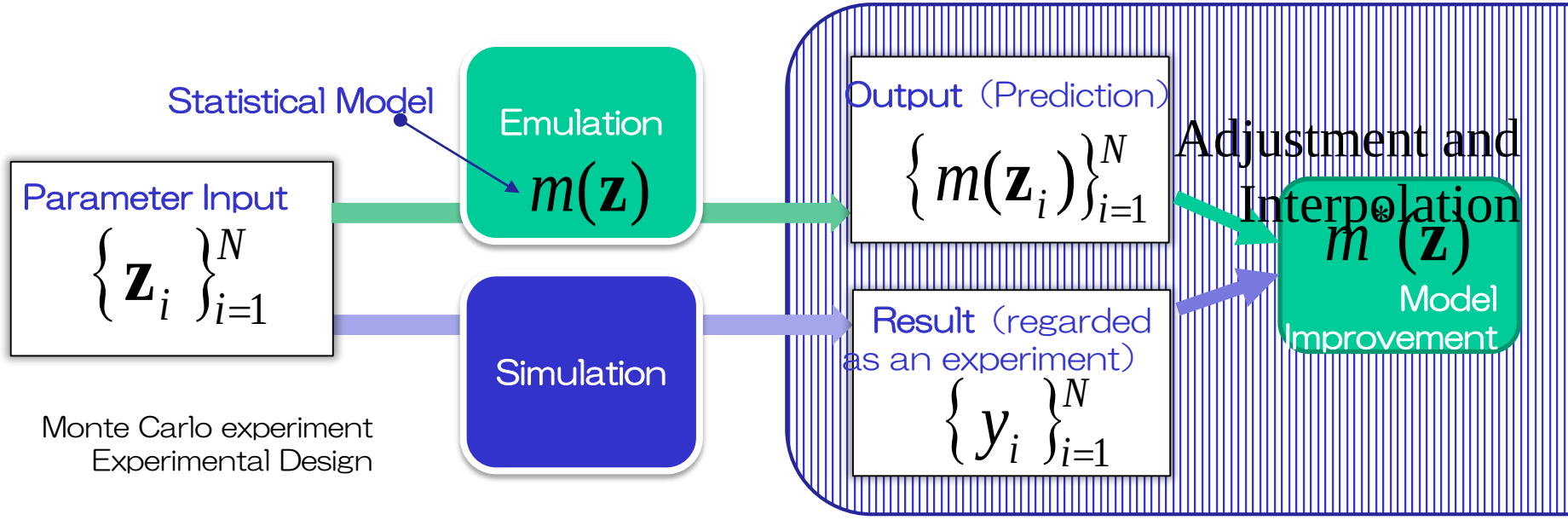
実際にはαをいろいろ変える必要があるため，
効率的なアルゴリズムが提案されている．

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

Emulate an output of simulation given parameters

③



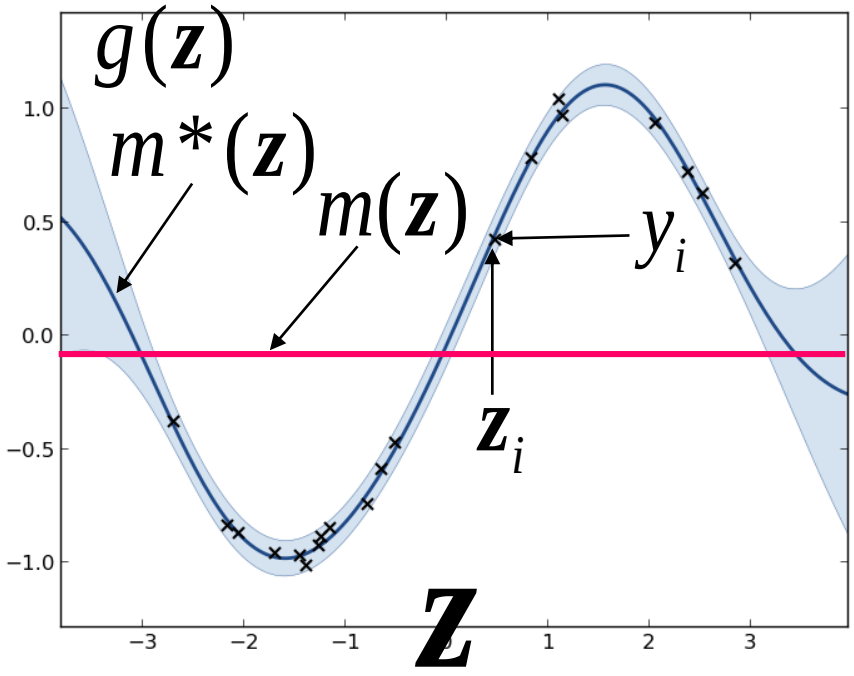


Emulator：Gaussian Process Regression (GPR)

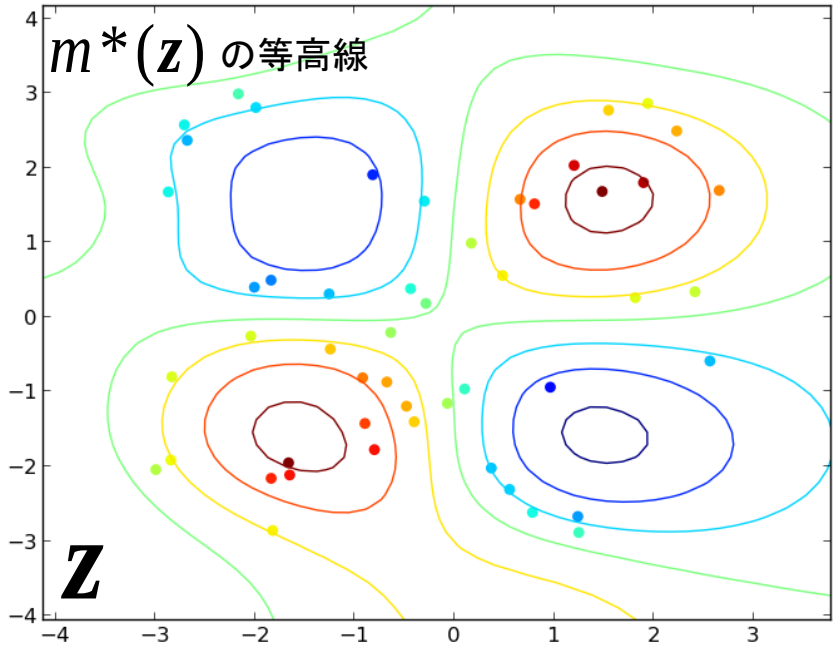
要は内挿の一手法

データ $\{y_i\}_{i=1}^N$ ：シミュレーション結果
入力　$\boldsymbol{z}$　：設計パラメータ（多次元）
出力　$g(\boldsymbol{z})$　：目的変数を出力とする確率的応答関数

1次元



2次元



http://gpytest2.readthedocs.org/en/latest/tuto_GP_regression.html

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

$$\mathrm{E}[\mathrm{g}(\mathbf{z})] = m(\mathbf{z}) = H \cdot \begin{pmatrix} \mathbf{x}^*_{t-1} \\ \mathbf{z} \end{pmatrix}$$

Statistical model is given

$$\mathrm{Cov}[g(\mathbf{z}), g(\mathbf{z}')] = k(\mathbf{z},\mathbf{z}') \cdot \tau^2$$

Covariance function

Kernel function

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

Kernel function : Mutual Distance between input parameters

Common variance

Observation Model

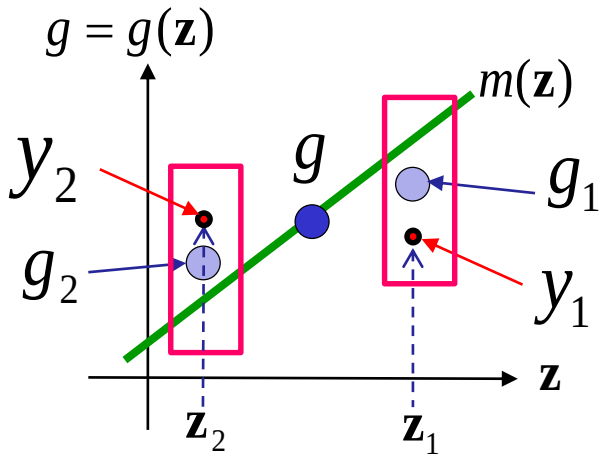
$$y_i = g_i + e, \quad e \sim N(0, \sigma^2) \quad (i = 1, \cdots, N)$$

Posterior Distribution

g is a continuous function

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

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



Emulator : Design of Kernel function

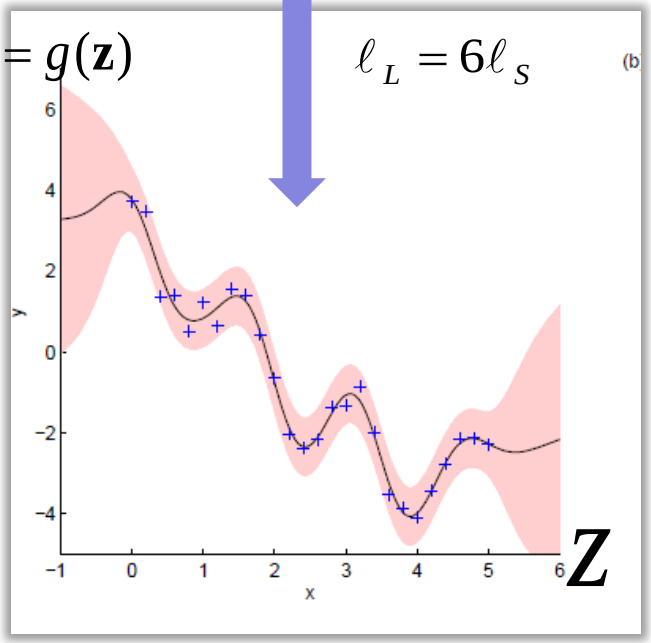
$$\mathrm{Cov}_y\left[g(\mathbf{z}), g(\mathbf{z}')\right] = \tau^2 k(\mathbf{z}, \mathbf{z}') + \sigma^2 \delta(\mathbf{z} - \mathbf{z}')$$

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

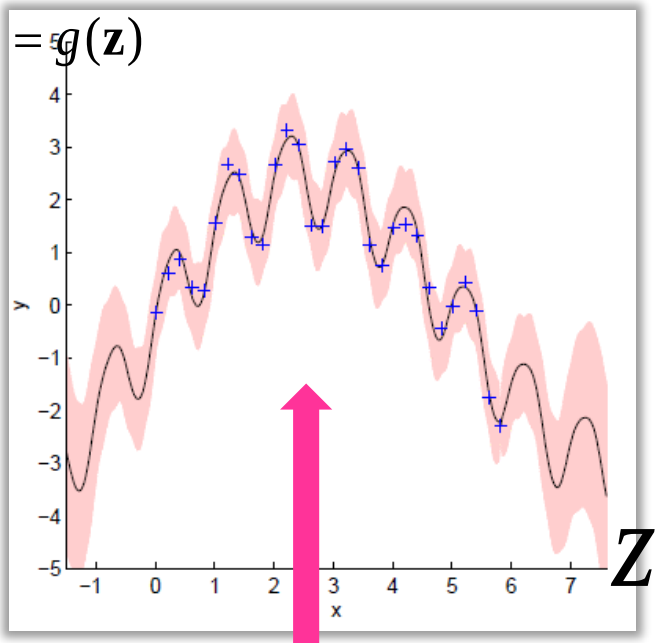
$$\tau_1^2 \exp\left[-\frac{(z-z')^2}{2\ell_L^2}\right] + \tau_2^2 \exp\left[-\frac{(z-z')^2}{2\ell_S^2}\right] + \sigma^2 \delta(z-z')$$

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

$g = g(\mathbf{z})$ $\quad \ell_L = 6\ell_S$



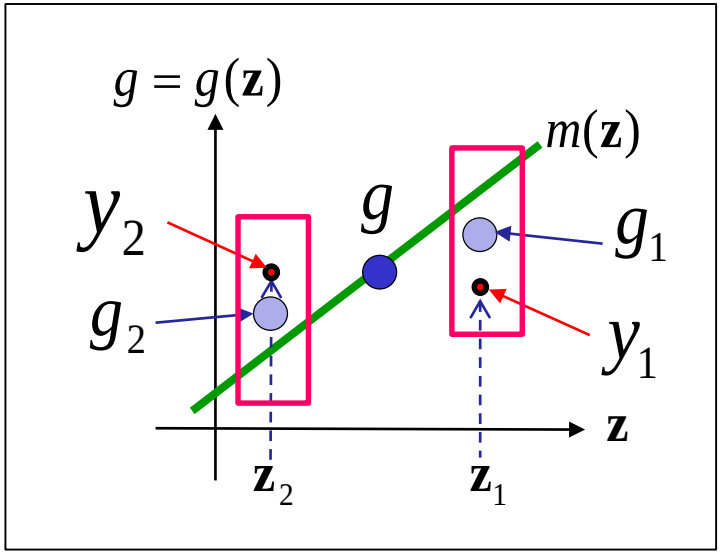
$g = g(\mathbf{z})$



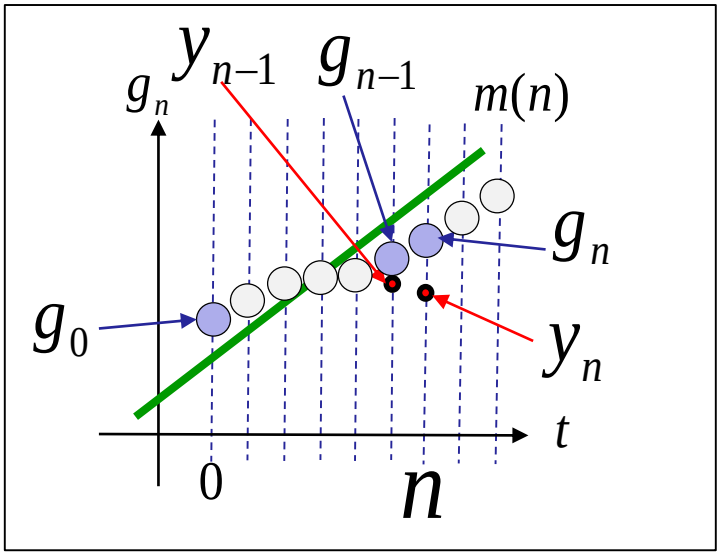
Gaussian Processes for Regression:A Quick Introduction
M. Ebden, August 2008

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

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

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

Emulator : Online Adjustment（Calibration）and Interpolation

For simplicity

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

Posterior distribution

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

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

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

Difference between input $\mathbf{z}$ and some points

nce between Data Statistical Model

Normalization factor

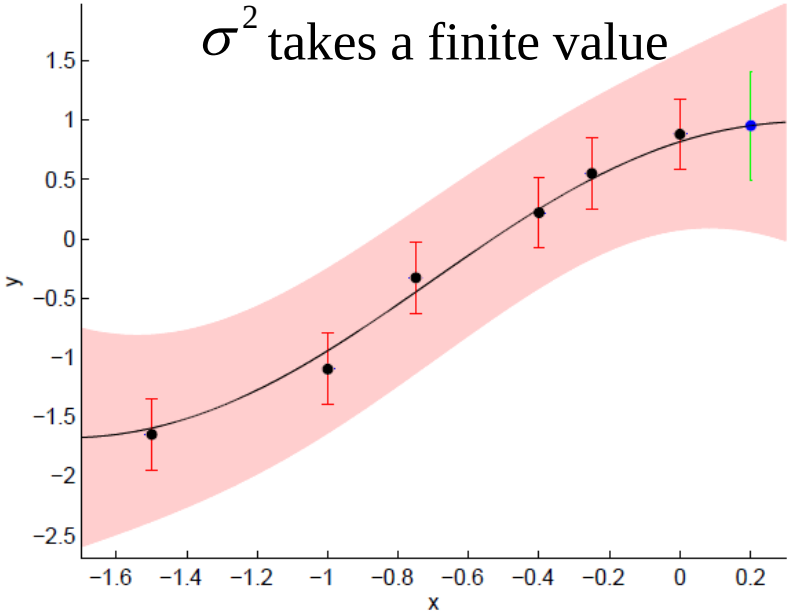
$$\mathbf{z}, \mathbf{z}_1), \cdots, k(\mathbf{z}, \mathbf{z}_N)]$$

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

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.

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

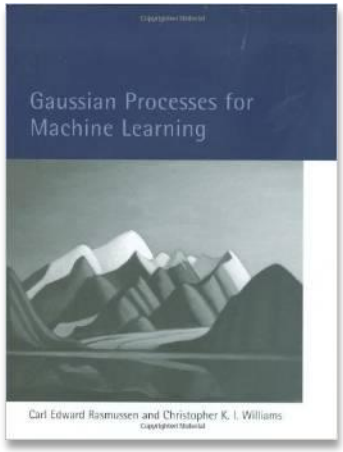
Number of samples

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

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

Emulator

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。