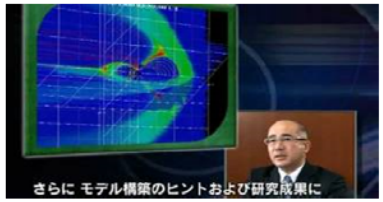
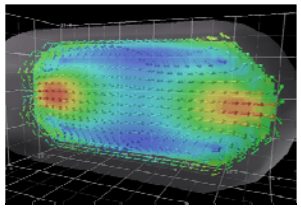
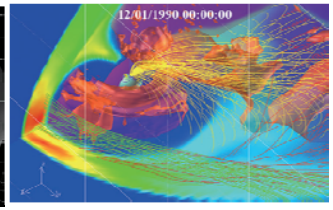
データ同化による エミュレータ・デザイン学の創設

樋口知之　（情報・システム研究機構　統計数理研究所）



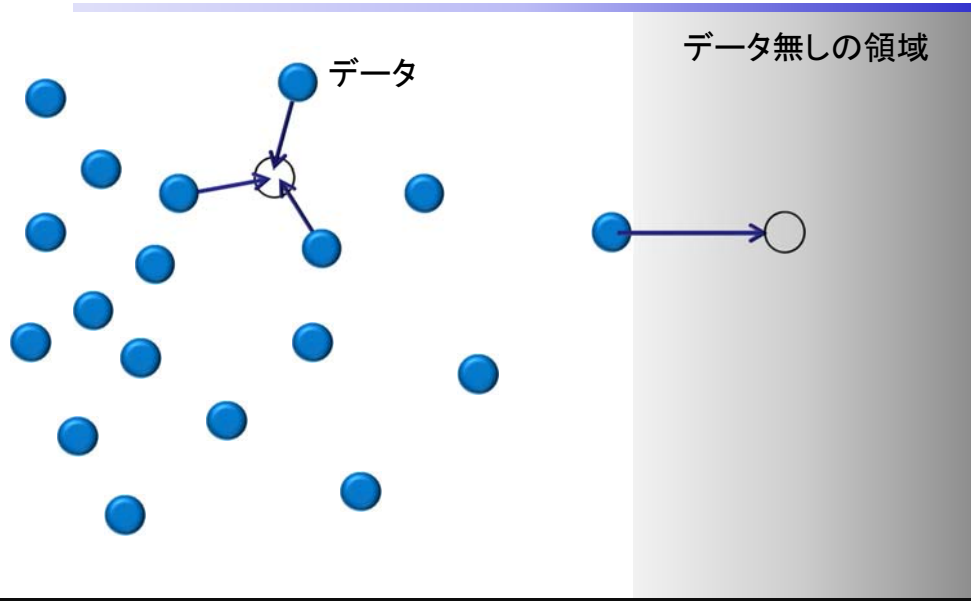
アウトライン

何度も講演

- 内挿と外挿
- データ同化
- 実例：　地震音波とバイオイメージング

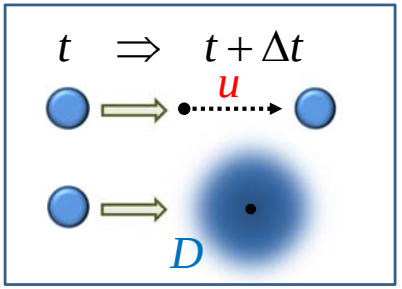
- エミュレータ
- スパース回帰（気象予測）
- オンライン機械学習
- GPR: ガウス過程回帰

内挿と外挿問題

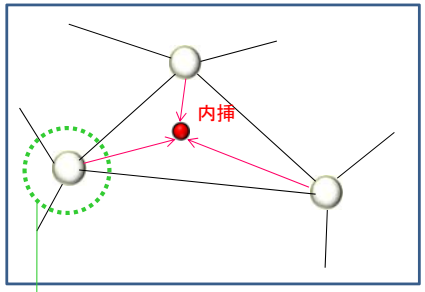


いろいろなビジネス展開が可能

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）

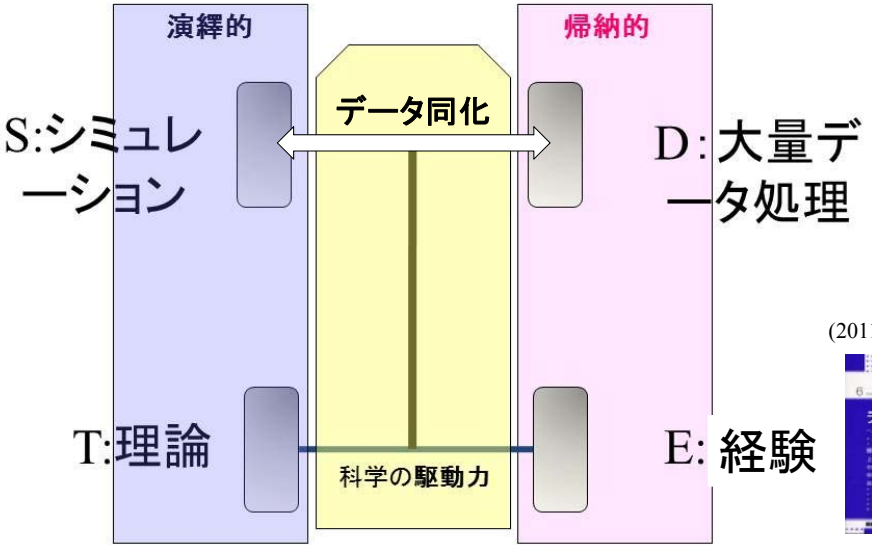


＋



情報の不確実性
（多人数からのレポート：多様なノイズの混在）

つなぐ：データ同化



(2011年9月刊行)

データ同化の目的：気象・海洋学の観点から

[1] 予報を行うための最適な初期条件を求める。これは既に、現業の天気予報で実用化されていることである。

[2] シミュレーションモデルを構成する際の最適な境界条件を求める。連成現象を取り扱う際の適応的な境界条件設定もこの作業に含まれる。

[3] スケールが異なるシミュレーションモデル間の橋渡しを行うスキーム内に含まれる諸パラメータの最適な値を求める。経験的に与えられるモデル内のパラメータ値の検証も一つの具体例である。

[4] シミュレーション（物理）モデルにもとづいた、観測されていない時間・空間点における観測値の補間を行う。この作業は再解析データセットの生成とも呼ばれる。このデータセットから新しい科学的発見をもくろむ。 ダウンスケーリング

[5] 時間・経費を節約できる効率的な観測システムを構築するための仮想観測ネットワークシミュレーション実験や感度解析を行う。

（参考文献： 蒲地 他、「統計数理」、54(2), 223-245, 2006.）

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

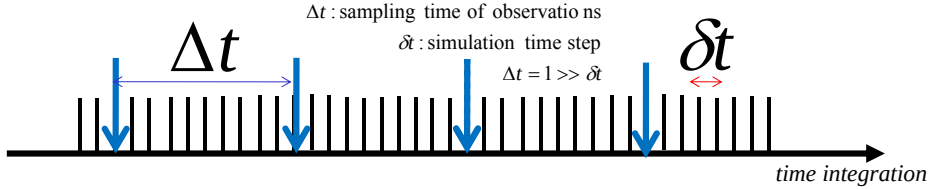
$$\begin{cases} \boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t), & \boldsymbol{v}_t \sim p(\boldsymbol{v} \mid \boldsymbol{\theta}_{\mathrm{sys}}) \\ \boldsymbol{y}_t = h_t(\boldsymbol{x}_t, \boldsymbol{w}_t), & \boldsymbol{w}_t \sim p(\boldsymbol{w} \mid \boldsymbol{\theta}_{\mathrm{obs}})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

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

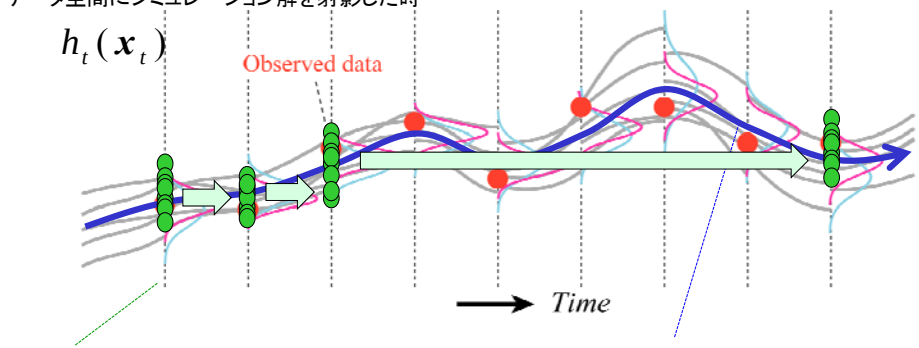
Δt : sampling time of observations
δt : simulation time step
$\Delta t = 1 >> \delta t$



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）

統計数理研究所

Our interest in a statistical inference by DA

■ State Estimation and forecasting

$$p(x_t \mid y_{1:t}) \quad or \quad p(x_t \mid y_{1:T})$$

状態（時刻依存）変数の推定

■ Parameter Estimation in a simulation model

$$p(\theta \mid y_{1:T}) \propto p(y_{1:T} \mid \theta) \cdot p(\theta)$$

モデルパラメータの推定

$\theta \Rightarrow \mathbf{x}_0$ 4DVar

統計数理研究所

Japan, An Earthquake-Prone Country (Nagao et al., 2011)

Nagao

acoustic wave

Application to analyze an acoustic wave which is generated by earthquake followed by Tsunami.

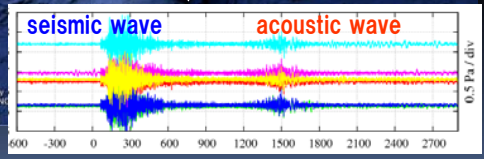
Japan is an earthquake-prone country, so our government has been promoting many researches related to earthquake prediction and earthquake disaster prevention.

When each earthquake occurred, many **micro-barometer sensors** located at distances of hundreds or thousands of kilometers from the epicenter observed "the sound of the earthquake" in the form of atmospheric pressure perturbations.

In the case of the Iwate-Miyagi Nairiku Earthquake, for example, micro-barometer sensors 417 km away from the epicenter observed acoustic waves excited by the earthquake.

Because acoustic velocity is much faster than the propagation velocity of tsunami, such acoustic signals may be useful in predicting the magnitude of approaching.

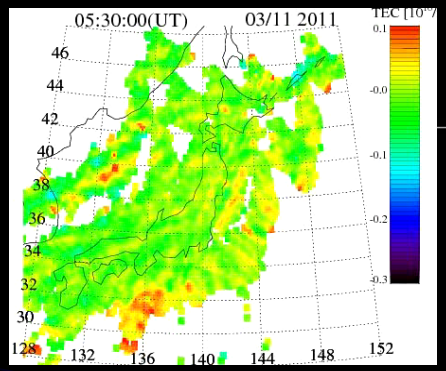
observation



統計数理研究所
データ同化研究開発センター

Ionospheric Disturbances Excited by the 2011 Great East Japan EQ

Nagao



GPS Observation

(Website of Kyoto University)

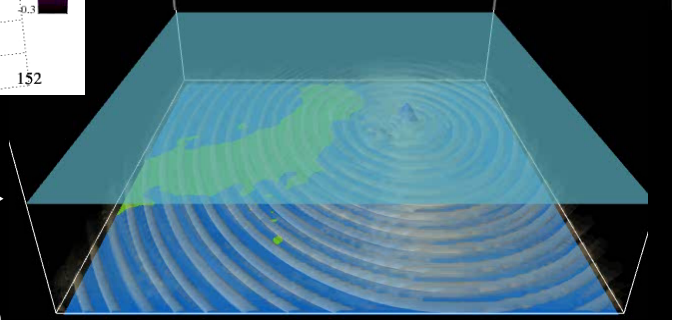
Observation (GPS-TEC)

From website of Kyoto Univ.

2011/03/11 05:46:18

In the case of the Great East Japan Earthquake, the dense array of GPS receivers in Japan clearly observed the acoustic waves causing oscillations in the ionosphere at the altitude of a few hundred kilometers.

ionosphere →



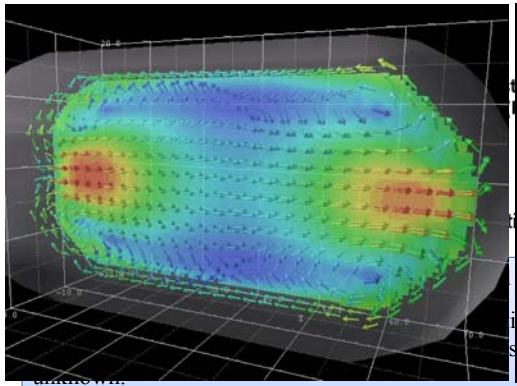
Assimilation

統計数理研究所
データ同化研究開発センター

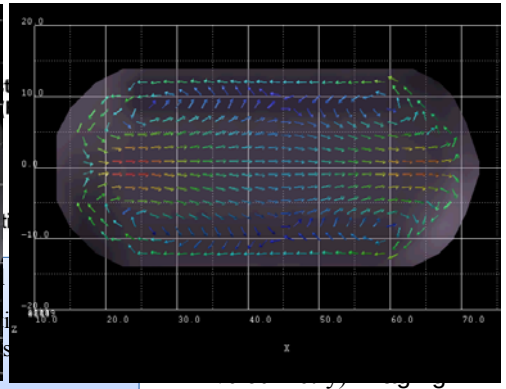
Data Assimilation on Intracellular Fluid in C.elegans embryo
(Collaboration with Dr. Kimuta at NIG)

Nagao

Assimilation



Observation

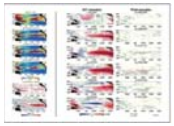


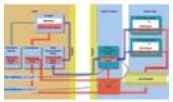
We assume that proteins "**myosin**" in the cell wall drive physically such intra-cellular flow, and we perform data assimilation by combining a numerical simulation that solves simultaneously these physical equations and the observation obtained by the image analyses.

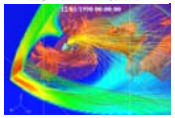
データ同化研究開発センター

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

Video (10min) uploaded to YouTube
Introduction to R&D Center for Data Assimilation



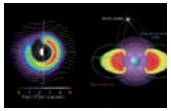


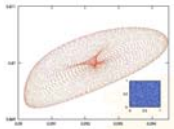


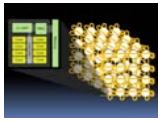
Research Projects

Methods:

- Applications of Sequential Data Assimilation Methods
  - Meteorology
  - Space Science
  - Life Science
- Advanced Monte Carlo algorithm and Applications
- Statistical Algorithms Dedicated to Massively Parallel Computers







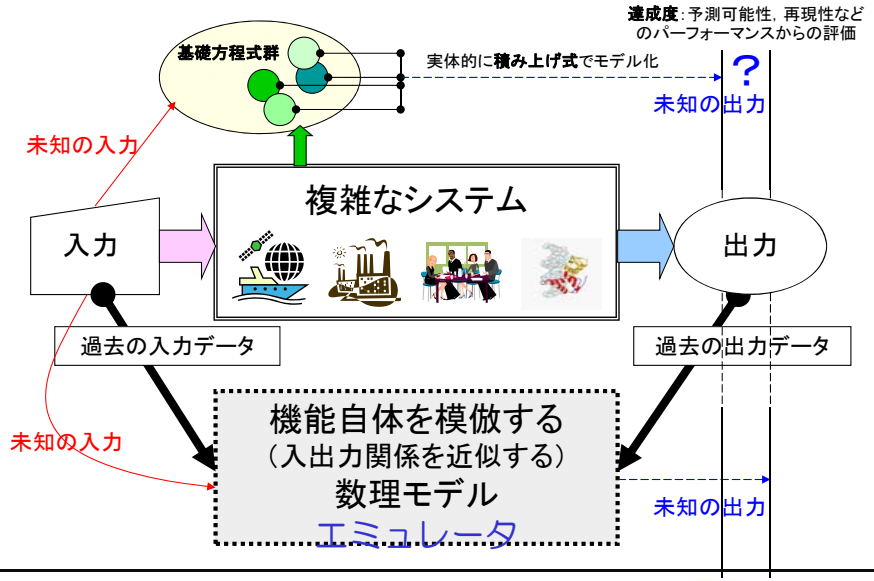
Hardware:

- Hardware Random Number Generators

Computational Services:

- Statistical Analysis System for Massively Parallel Computing Enviro
- Cloud Computing Services
- Visualization Software

機能のモデル化：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



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

機能のモデル化：“見よう見まね”を科学する

- “見よう見まね”のプロセスを加速する。

  一体系化されていない研究分野において有効

- “見よう見まね”による完成の域がお手本を超える。

  一贋作が“本物”を超える

日本人は伝統的に“贋作”を心から嫌う傾向が強いが、同じく漢字文化圏にある中国人で心ある人は、過去の文物に目を転じるとき、それがたとえ”贋作”であろうと判断できても、”本物”を超える出来栄えであるならば、自分の心と目を満足させるため、”本物”以上にその”贋作”を尊重して手に入れると伝えられている。

“贋作”と”本物”の、どちらが本物か実は分からない

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

気象予測：気象エミュレータ

観測データベクトル（数千～1万次元）

シミュレーション変数の計算結果（10万～100万次元）
同化の結果を使う場合もある

サンプル

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

ある地点の降雨量（温度）

（スパース）線形回帰

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

スパース学習

予測

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

シミュレーション変数ベクトルの新しい入力

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

$\mathbf{x}_{t-1}^+$を多数シミュレートし$y_{i,t}^+$の
アンサンブル予測を行なう

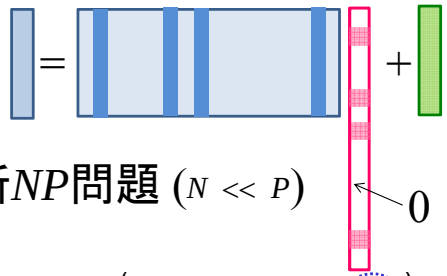
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

減らす1. スパース（疎性を利用した）最適化

例：多変量回帰

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

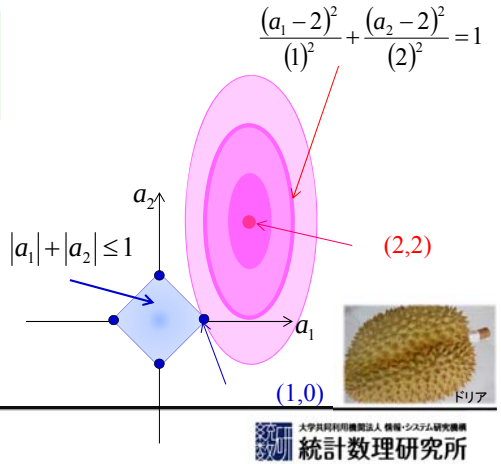
$$\mathbf{y} = X\mathbf{a} + \mathbf{e}$$



新NP問題 ($N << P$)

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

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



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

減らす2. 行列、テンソル分解

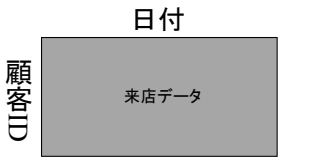
ID付きPOSデータ

i軸：顧客ID

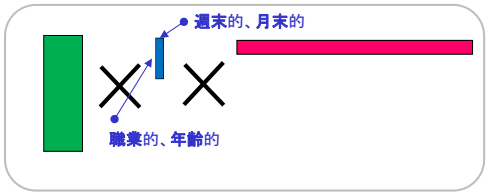
k 軸：日時

y_{ijk}=0,1,2…

j 軸:商品



だいたい似かよるように
代数演算のみで分解



『バラバラを癖で束ねる』を実現

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

ストリーミング計算

データは現場で産出。移動は困難。

Edge-Heavy Data (PFI, 丸山@ISM)

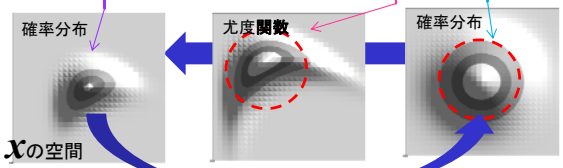
M2Mの増大
↓
トラフィックの大渋滞

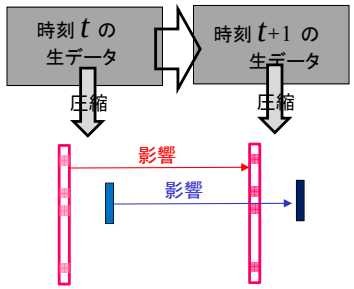
✓ パラメータ(w)オンライン学習法
✓ スパース学習

FOBOS: Forward Backward Splitting
RDA: Regularized Dual Averaging

ベイズの定理と情報循環

$$p(\underline{x} \mid y) = \frac{p(y \mid x) \cdot p(x)}{p(y)} \propto p(y \mid x) \cdot p(\underline{x})$$





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

オンライン計算 in 機械学習

パラメータ(a)オンライン学習法

FOBOS: Forward Backward Splitting
RDA: Regularized Dual Averaging

一般の学習器

$$\min_{\mathbf{w}} \ell(X\mathbf{a}) + \psi(\mathbf{a}) \qquad \{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^{T}$$

時刻 t の正誤(ある, なし)
時刻 t の画像データ

$$\mathbf{a}_{t+1} = \min_{\mathbf{a}} \left\{ \left\langle \left| \nabla_{\mathbf{a}} \ell(\mathbf{y}_t, \mathbf{x}_t \cdot \mathbf{a}) \right|_{\mathbf{a}=\mathbf{a}_t}, \mathbf{a} \right\rangle + \psi(\mathbf{a}) + \frac{1}{2\eta_t} \|\mathbf{a} - \mathbf{a}_t\|^2 \right\}$$

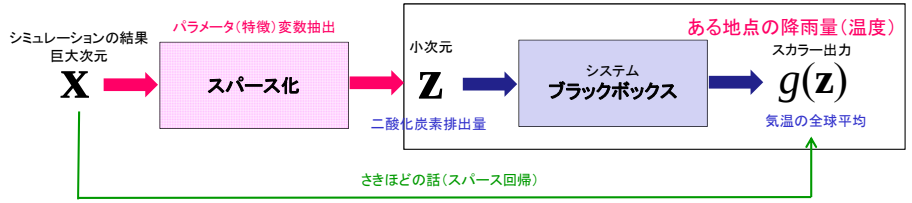
降下法

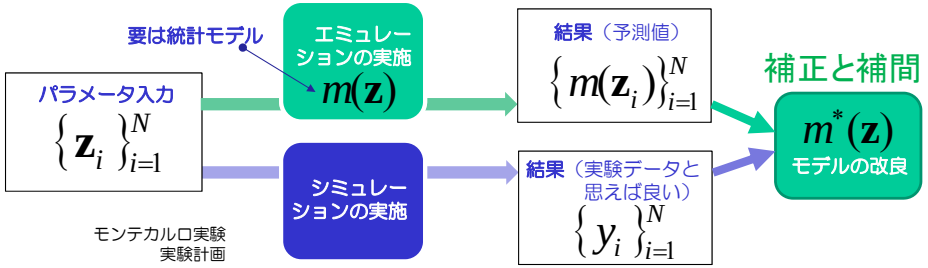
スパース性

Fused LASSO (Tibshirani et al., 2005)のオンライン版

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

エミュレータ：
シミュレーションパラメータを入力とする出力値を模倣





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

エミュレータ：統計モデル（要は、線形回帰モデル＋GP）

システムモデル

Gaussian Process: g は無限次元(連続変数)

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

回帰係数

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

統計モデルは所与

Covariance function

カーネル関数

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

カーネル関数：入力値(パラメータ)間の距離

共通の分散

観測モデル

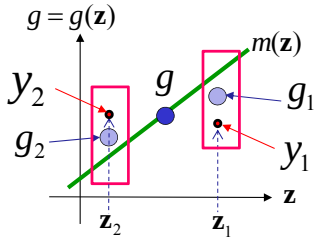
$$y_i = g_i + e, \quad e \sim N(0, \sigma^2) \quad (i = 1, \cdots, N)$$

事後分布

g 連続変数

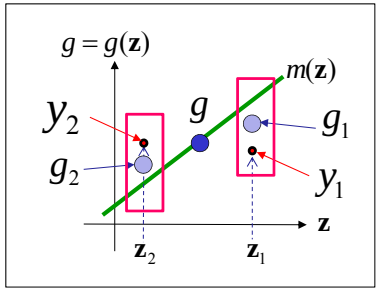
$$p(g \mid Y) \propto p(Y \mid g) p(g)$$

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

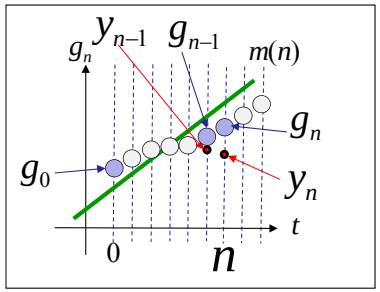


大学共同利用機関法人 情報・システム研究機構 統計数理研究所

エミュレータ：時系列のトレンドモデルとほぼ同じ



g が離散だったら



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

カーネル関数は定数値

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

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

エミュレータ：補正（キャリブレーション）と補間

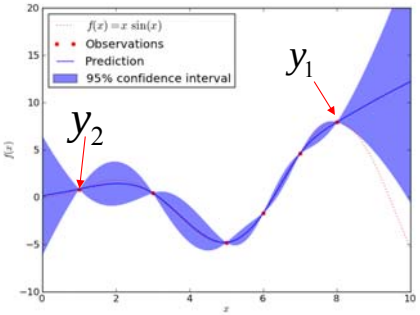
$\frac{\sigma^2}{\tau^2} \to 0$ の場合　*簡単のため*

事後分布

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

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

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



入力値とのずれの程度

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

データと統計モデルの差　　規格化ファクター

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

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

データがもたらす情報により不確実性は必ず減少する

グラム行列

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

統計数理研究所

エミュレータ：フルベイズへ

$\frac{\sigma^2}{\tau^2} \to 0$ の場合　*簡単のため*

事後分布

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

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

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

ガンマ分布

自由度 $N - \lambda$ の t 分布

エミュレータ　サンプル数　回帰モデル部分　$\lambda = \dim(\mathbf{z}) + 1$

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、to appear

Rasmussen, C. and C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

統計数理研究所

エミュレータ：カーネル関数の設計（問題毎に与える）

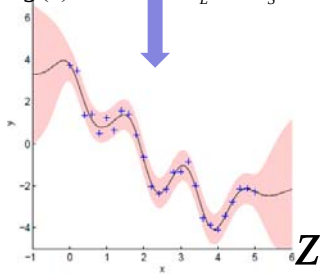
$$\mathrm{Cov}_y[g(\mathbf{z}), g(\mathbf{z}')] = \tau^2 k(\mathbf{z},\mathbf{z}') + \sigma^2\delta(\mathbf{z}-\mathbf{z}')$$

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

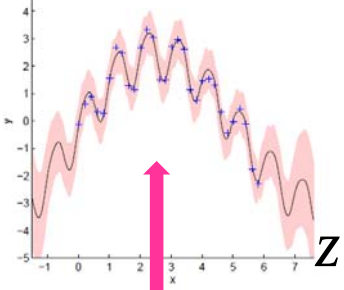
こちらの構造ではない！

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

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



Z



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

データ同化とエミュレータ

データ同化

a. 行列分解
b. ローカル化

機械学習を用いた内挿法

I. スパース回帰
II. オンライン学習
III. カーネル法

横幹連合会誌「横幹」Vol.8,No1（2014年4月15日）
ミニ特集「数理科学の展開とその体制」
『統計数理の誕生とその広がり』 樋口知之

統計数理研究所