統計的モデリング
― 知識社会の基盤技術 ―

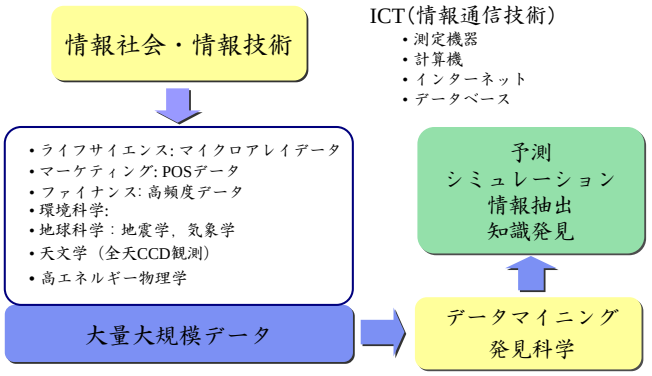
北川源四郎
統計数理研究所

数理システムユーザコンファレンス
2007年11月22日

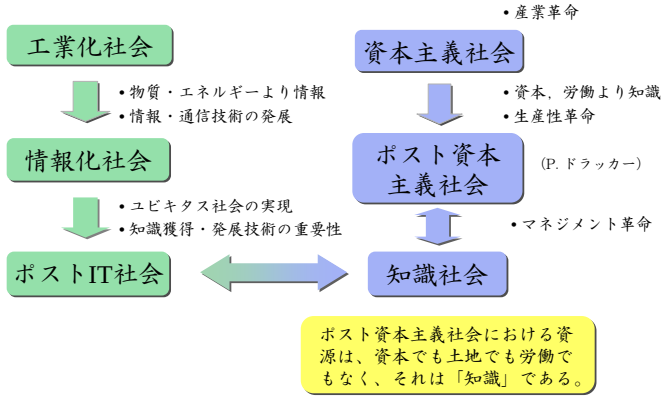
概 略

- 社会と科学研究の変化
  - ポストIT社会・知識社会へ
  - 知識の変化・科学研究の変化
- 統計科学の重要課題
  - 理論科学・実験科学からデータ中心科学へ
  - 情報統合・知識創造
  - 平均から個性へ ：Personalization
- 状態空間モデリング
  - 情報抽出
  - 非線形・非ガウス型フィルタ

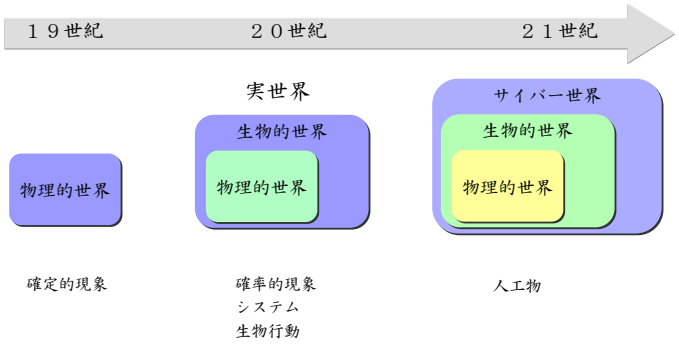
Post-IT時代と大量大規模データ



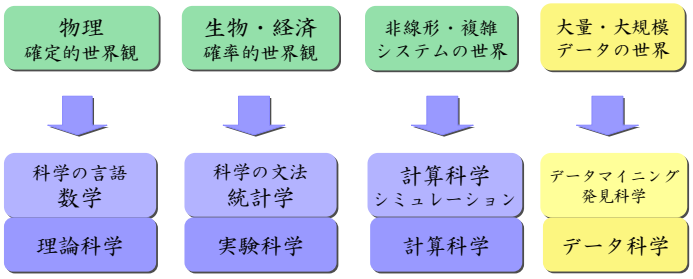
社会の変化



科学研究対象の変化

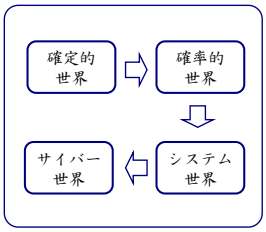


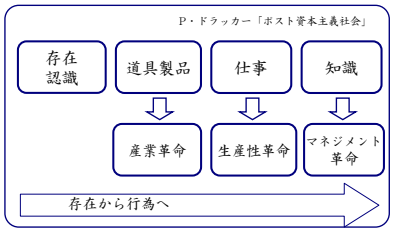
科学的方法論の変遷



知識の変化

普遍の真理 ⇒ 行動に有益な情報





データと研究方法の変化（例）

生物学		統計学
博物学	観察データ	記述統計学
実験科学	少数実験データ	推測統計学
現代博物学？	大量観察データ	データ科学
堀田凱樹		テューキィ、林、柴田・坂内

― 現代博物学を超えるために ―

原理主導アプローチとデータ主導アプローチの統合

- 統計的方法
- 逐次フィルタ・平滑化
- データ同化・・・地球環境予測

統計科学は個別的経験を一般的知識に変える

統計科学のグランドチャレンジ

工業化社会 ⇒ 情報化社会 ⇒ 知識社会

２１世紀の社会・学術研究に即した科学的方法論の開発

戦略的課題
予測と知識発見
不確実性とリスク解析
実世界シミュレーション
サービスイノベーション

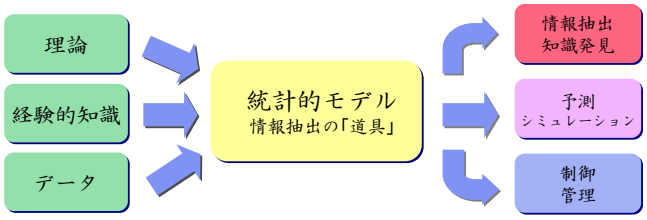
⇒ 統計的問題
情報統合
モデリング
データ同化
個別化

⇒ 知識創造の方法
(科学的方法の融合)

関連する統計的課題

1. 能動的モデリング
2. 新NP問題
3. ベイズモデリング
4. 計算統計学

1. 能動的なモデリング



対象に関するあらゆる知識や解析の目的に応じたモデリング

データからの情報と事前情報の統合

⇒ ベイズモデル

統計的問題の変化

少標本実験・調査データ

⇒ パラメトリック・モデル
検定、モデル選択

Huge data sets
- Bio Science … DNA data, array data
- Environmental Science, Disaster Prevention
- Financial, Marketing ... POS data

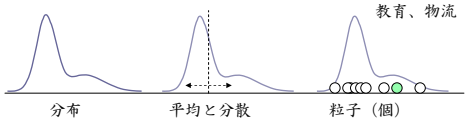
大量データ，複雑システム

⇒ 柔軟なモデリング

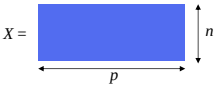
2. Personalization（平均から個人へ）

テーラーメイド創薬
オーダーメイド医療
マイクロマーケティング
サービスイノベーション
教育、物流

●平均から個性へ



● 本質は究極の条件付け



新*np*問題（$n \ll p$）
- マイクロアレイ
- マーケティング

３．ベイズモデリング

- 様々な情報の統合
- Personalization

キーテクノロジーは
ベイズモデリング

ベイズモデリングの発展（1980年代―）

ベイズの定理　18世紀

事前分布の問題
哲学的論争
計算困難性

⇒ 解決

ベイズモデルの隆盛
情報抽出，情報統合，
情報検索

近年の発展
（1）方法論上：事前分布，モデル評価
（2）計算上：　MCMC，MCF
（3）応用上：　生物系統樹推定，統計地震学，データ同化
季節調整，調査データ解析法

４．計算統計学

Real World の解析・予測を目指して

強い仮定
仮想的世界
（硬いモデル）
＋
解析的方法

⇒

弱い仮定
現実的世界
（柔軟なモデル）
＋
計算による方法

- **EM algorithm**
- **Bootstrap**
- **MCMC**（ベイズモデリング）
- **Sequential MCF**（状態空間モデル）

Smoothness Prior

平滑化の問題

$$y_n = f_n + \varepsilon_n, \quad n = 1, \ldots, N$$

y_n　観測値
f_n　未知パラメータ
ε_n　ノイズ（残差）

罰金付最小二乗法

Infidelity to the data
Infidelity to smoothness

$$\min_f \left[\sum_{n=1}^{N} (y_n - f_n)^2 + \lambda^2 \sum_{n=1}^{N} (\nabla^k f_n)^2 \right]$$

ベイズモデルの観点からの選択

Crucial parameter

$$\sum_{n=1}^{N}(y_n - f_n)^2 + \lambda^2 \sum_{n=1}^{N}(\nabla^k f_n)^2$$

$-1/(2\sigma^2)$ をかけて指数をとる

Smoothness Prior

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f_n)^2\right\} \exp\left\{-\frac{\lambda^2}{2\sigma^2}\sum_{n=1}^{N}(\nabla^k f_n)^2\right\}$$

ベイズモデルによる解釈　$\theta=(\lambda^2,\sigma^2)$

$$\pi(f \mid y,\theta) \propto p(y \mid f,\theta)\pi(f \mid \theta)$$

⇒ ABICによる　θ の決定

Akaike (1980)

時系列的解釈と状態空間モデル

$$\sum_{n=1}^{N}(y_n - t_n)^2 + \lambda^2 \sum_{n=2}^{N}(t_n - t_{n-1})^2$$

等価なモデル

$$t_n = t_{n-1} + v_n \qquad v_n \sim N(0,\tau^2)$$
$$y_n = t_n + w_n \qquad w_n \sim N(0,\sigma^2)$$

状態空間モデル

$$x_n = Fx_{n-1} + Gv_n$$
$$y_n = Hx_n + w_n$$

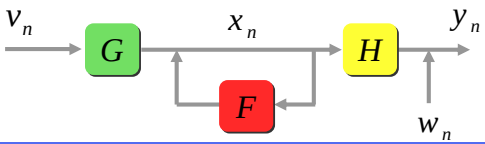
$$\lambda^2 = \frac{\sigma^2}{\tau^2}$$

状態空間モデル

$$x_n = Fx_{n-1} + Gv_n \quad \text{状態モデル}$$
$$y_n = Hx_n + w_n \quad \text{観測モデル}$$

y_n　時系列　　v_n　システムノイズ

x_n　状態ベクトル　　w_n　観測ノイズ



状態空間モデルの応用

- ARMA モデルの最尤推定 Akaike (1974)
- 非定常性のモデリング
  - トレンド推定　Harrison-Stevens (1976)
  - 季節調整　Kitagawa(1981), Harvey (1985)
  - 確率的ボラティリティ　Kitagawa- Gersch (1985), Harvey-Shepard(1996)
  - 時変係数 AR モデル　Kitagawa- Gersch (1985), Godsill-West(2000)
  - 信号抽出　Kitagawa- Gersch (1996)

状態空間モデルの拡張

線形・ガウス型

$$x_n = Fx_{n-1} + Gv_n$$
$$y_n = Hx_n + w_n$$

非線形・非ガウス型

$$x_n = f(x_{n-1}, v_n)$$
$$y_n = h(x_n, w_n)$$

関数：非線形
分布：非ガウス型

一般型

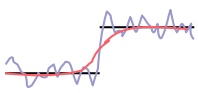
$$x_n \sim F(\cdot \mid x_{n-1})$$
$$y_n \sim H(\cdot \mid x_n)$$

条件付分布
離散状態・離散観測値

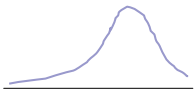
非ガウス型モデリングの必要性

- Non-standard Time Series -

Abrupt Structural Change



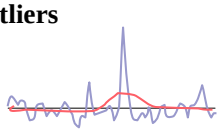
Asymmetry of Dist.



Nonlinearity

$$x_n = f(x_{n-1}) + v_n$$

Outliers



Discrete Process

Poisson Process

Binary Process

非ガウス型フィルタ・平滑化

一期先予測

$$p(x_n|Y_{n-1}) = \int_{-\infty}^{\infty} p(x_n|x_{n-1})p(x_{n-1}|Y_{n-1})dx_{n-1}$$

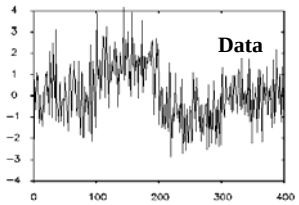
フィルタ

$$p(x_n|Y_n) = \frac{p(y_n|x_n)p(x_n|Y_{n-1})}{p(y_n|Y_{n-1})}$$

平滑化

$$p(x_n|Y_N) = p(x_n|Y_n)\int_{-\infty}^{\infty} \frac{p(x_{n+1}|x_n)p(x_{n+1}|Y_N)}{p(x_{n+1}|Y_n)}dx_{n+1}$$

レベルシフトの自動検出



トレンドモデル

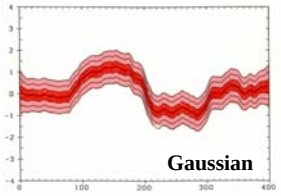
$t_n = t_{n-1} + v_n$

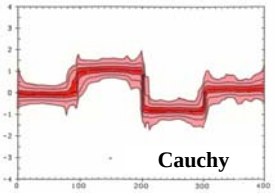
$y_n = t_n + w_n$

ノイズ分布

$v_n \sim C(0, \tau^2)$ コーシー分布

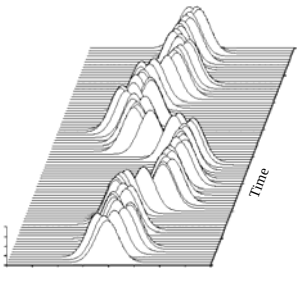
$w_n \sim N(0, \sigma^2)$ 正規分布

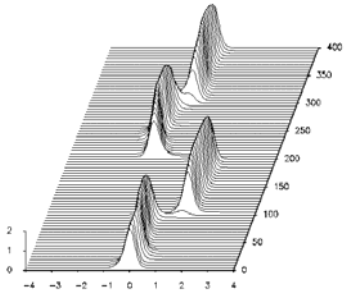




非ガウス型フィルタ・平滑化

Marginal Posterior Density





Gaussian Distribution **Cauchy Distribution**

自己組織型フィルタ・平滑化

状態空間モデル

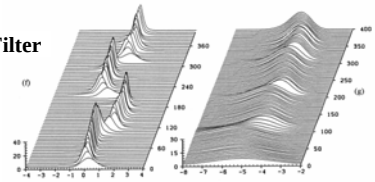
$x_n = x_{n-1} + v_n$

$y_n = x_n + w_n$

コーシー分布

$$p(v_n) = \frac{\tau}{\pi}\frac{1}{(v_n^2 + \tau^2)}$$

Filter

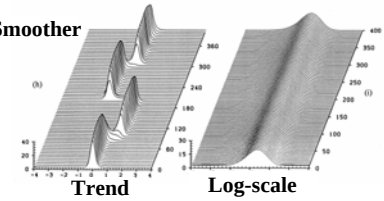


Augmented State Space Model

$$\begin{bmatrix} x_n \\ \log\tau_n^2 \end{bmatrix} = \begin{bmatrix} x_{n-1} \\ \log\tau_{n-1}^2 \end{bmatrix} + \begin{bmatrix} \tau_{n-1} \\ 0 \end{bmatrix} v_n$$

$$y_n = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ \log\tau_n^2 \end{bmatrix} + w_n$$

Smoother



Trend **Log-scale**

自己組織型フィルタ・平滑化

状態空間モデル $x_n = x_{n-1} + v_n$

$y_n = x_n + w_n$

$$p(v;\tau^2,b) = \frac{\Gamma(b)\tau^{2b-1}}{\Gamma(1/2)\Gamma(b-1/2)}\frac{1}{(v^2+\tau^2)^b}$$

Augmented State Space Model

$$\begin{bmatrix} x_n \\ \theta_{n,1} \\ \theta_{n,2} \end{bmatrix} = \begin{bmatrix} x_{n-1} \\ \theta_{n-1,1} \\ \theta_{n-1,2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} v_n$$

$$y_n = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ \theta_{n,1} \\ \theta_{n,2} \end{bmatrix} + w_n$$

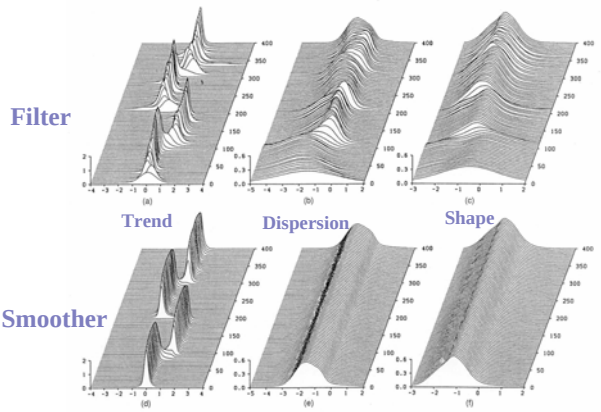
τ^2: Dispersion

b: Shape

$\theta_{n,1} = \log \tau_n^2 - 3\theta_{n,2}$

$\theta_{n,2} = \log(b_n - 1/2)$

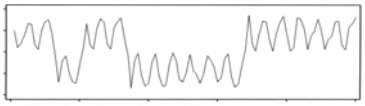
自己組織型フィルタ・平滑化

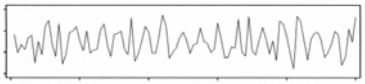


非線形モデル

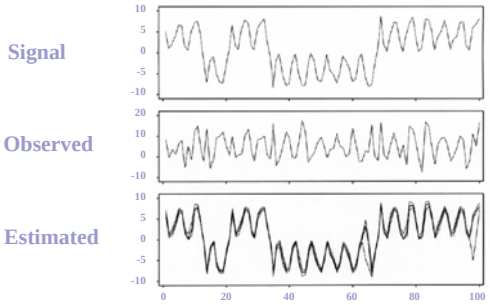
$$x_n = \frac{1}{2}x_{n-1} + \frac{25x_{n-1}}{1+x_{n-1}^2} + 8\cos(1.2n) + v_n$$

$$y_n = \frac{x_n^2}{20} + w_n \qquad v_n \sim N(0,0.1),\ w_n \sim N(0,1)$$

x_n  Signal

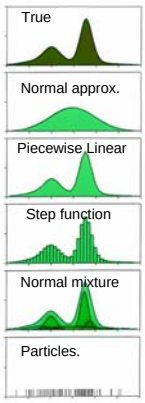
y_n  Observation

非線形フィルタ・平滑化



分布の近似

0. 正規分布近似
   （拡張）カルマンフィルタ・平滑化
1. 区分線形（階段）近似
   非ガウス型フィルタ・平滑化
2. 混合正規分布近似
   ガウス和フィルタ・平滑化
3. 粒子近似
   逐次モンテカルロフィルタ・平滑化



逐次モンテカルロ・フィルタ

・システムノイズ
$$v_n^{(j)} \sim p(v) \qquad j = 1, \ldots, m$$
・予測分布
$$p_n^{(j)} = F(f_{n-1}^{(j)}, v_n^{(j)})$$
・重要度（ベイズ係数）
$$\alpha_n^{(j)} = p(y_n | p_n^{(j)})$$
・フィルタ分布のリサンプリング
$$\{p_n^{(j)}\} \Rightarrow \{f_n^{(j)}\}$$

Gordon et al. (1993), Kitagawa (1996)
Doucet, de Freitas and Gordon (2001) "Sequential Monte Carlo Methods in Practice"

一期先予測

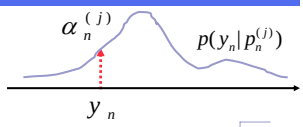
$$x_n = F(x_{n-1}, v_n)$$ システムモデル

$$v_n^{(j)} \sim p(v)$$
$$f_{n-1}^{(j)} \sim p(x_{n-1}|Y_{n-1})$$

$$p_n^{(j)} = F(f_{n-1}^{(j)}, v_n^{(j)})$$

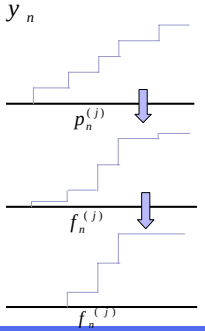
フィルタ（リサンプリング）

$\alpha_n^{(j)}$：粒子 $p_n^{(j)}$ のベイズ係数

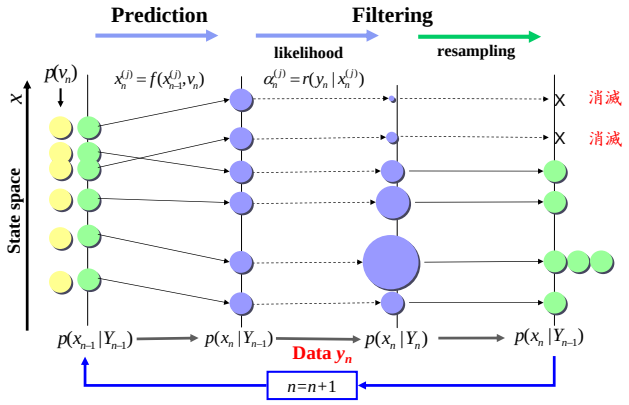


$$\alpha_n^{(j)} = p(y_n | X_n = p_n^{(j)})$$

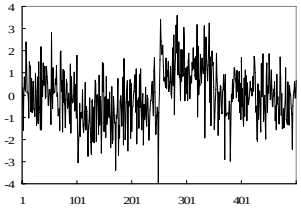
$$\Pr(X_n = p_n^{(j)}|Y_n) = \Pr(X_n = p_n^{(j)}|Y_{n-1}, y_n)$$
$$= \frac{\Pr(y_n|X_n = p_n^{(j)})\Pr(X_n = p_n^{(j)}|Y_{n-1})}{\sum_{i=1}^{m}\Pr(y_n|X_n = p_n^{(i)})\Pr(X_n = p_n^{(i)}|Y_{n-1})}$$
$$= \frac{\alpha_n^{(j)}\frac{1}{m}}{\sum_{i=1}^{m}\alpha_n^{(j)}\frac{1}{m}} = \frac{\alpha_n^{(j)}}{\sum_{i=1}^{m}\alpha_n^{(j)}}$$



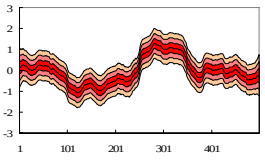
One Cycle of Monte Carlo Filtering



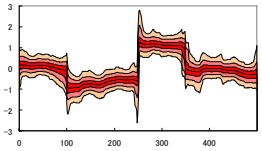
非ガウス型平滑化



Kalman Smoother



Exact Non-Gaussian Smoother



Trend Model

$$t_n = t_{n-1} + v_n$$

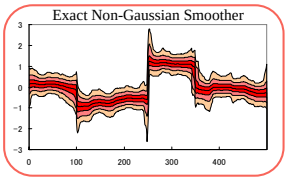
$$y_n = t_n + w_n$$

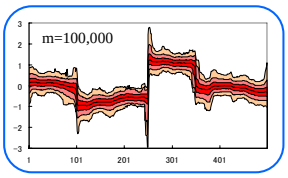
Noise Distribution

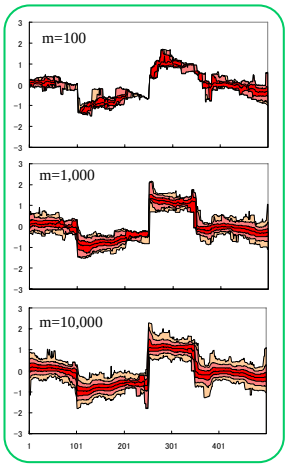
$$v_n \sim N(0,\tau^2) \quad \text{or} \quad C(0,\tau^2)$$

$$w_n \sim N(0,\sigma^2)$$

Single MCF







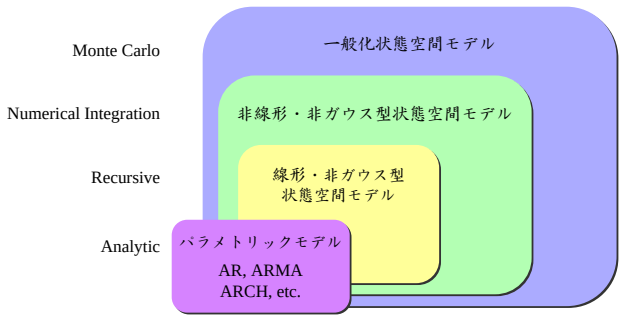
Applications of MCF

1. **Non-Gaussian smoothing**
   - Level shift
   - Non-Gaussian seasonal adjustment
   - Stochastic volatility models
2. **Nonlinear smoothing**
   - Tracking
   - Phase-unwrapping
3. **Signal extraction problems**
4. **Modeling count data**
5. **Self-organizing state space model**
6. **High-dimensional filtering/smoothing**

Gordon et al. (1993), Kitagawa (1996)
Doucet, de Freitas and Gordon (2001) "Sequential Monte Carlo Methods in Practice"

モデリングの問題点

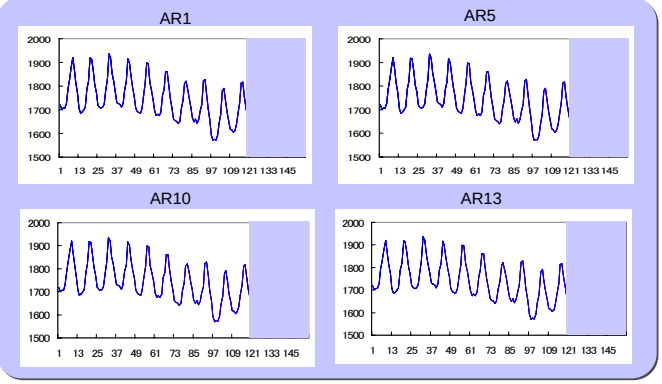
よいモデル族を導出する方法論がない　　モデリングは職人芸か



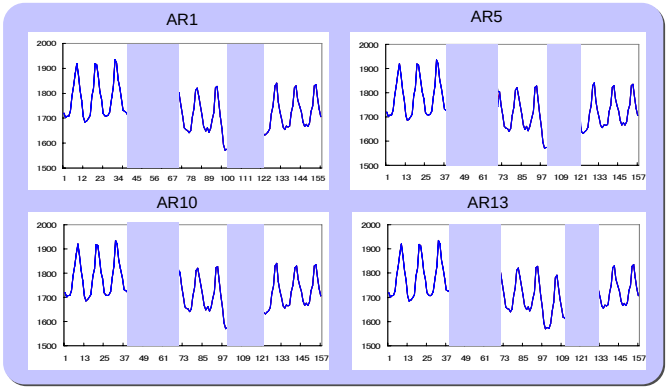
例：モデリングによる情報抽出

1. 時間軸モデリング
   - 時間構造が安定な場合（季節成分）
   - ランダム性が強い場合
2. 多変量時系列モデリング
   - 多変量ARの利用
   - 地下水位データ
3. 時空間モデリング
   - 海底地震計データ

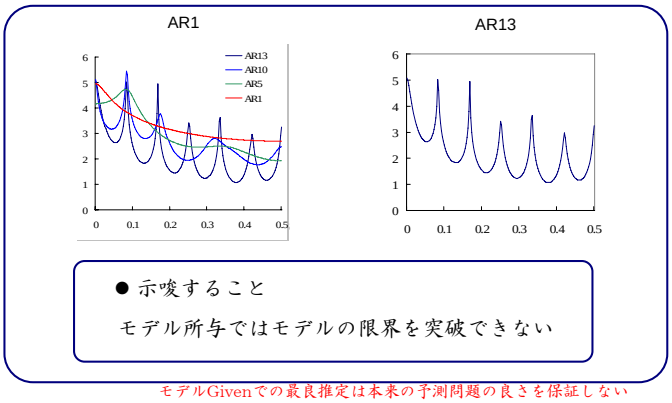
1．時系列構造のモデリングと予測



時系列構造のモデリングと補間



モデルの重要性



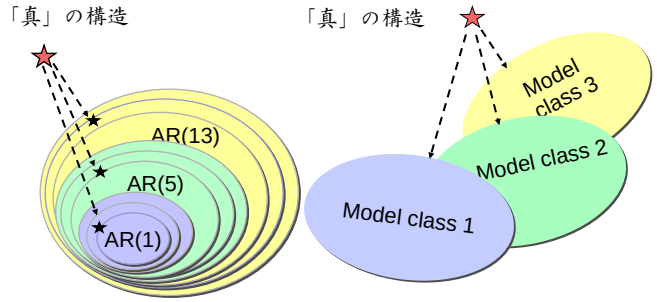
● 示唆すること

モデル所与ではモデルの限界を突破できない

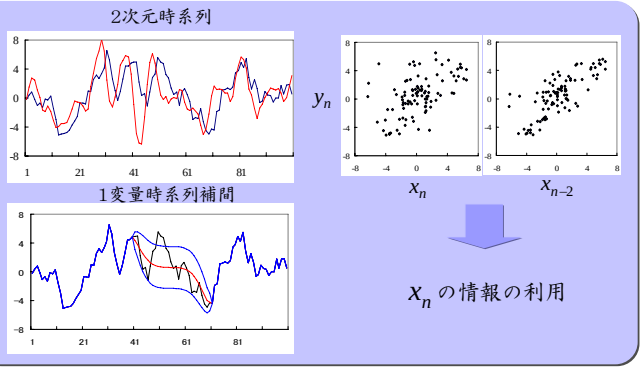
モデルGivenでの最良推定は本来の予測問題の良さを保証しない

These suggest …

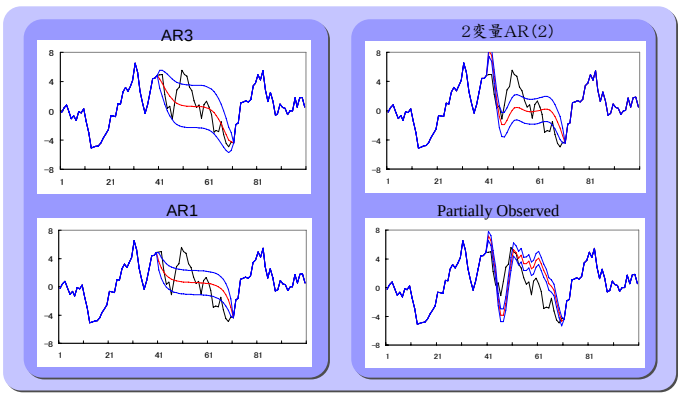
適切なモデル（仮説）提示の重要性



2．多変量構造のモデリング



1変量補間と2変量補間

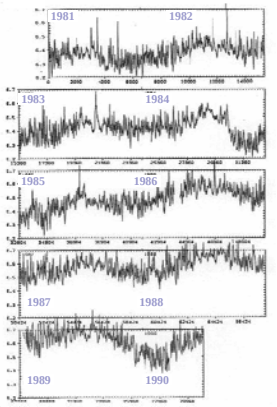


地下水位データ（地震の影響の抽出）

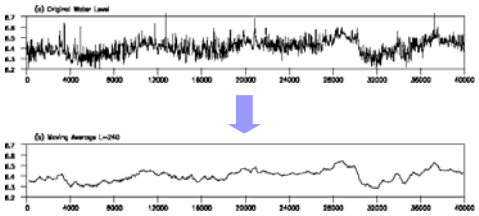
産総研観測井

気圧，地球潮汐，降雨などの影響を受ける

地震の影響の検出が困難



移動平均フィルタ



原データ

移動平均

何か知見が得られるか？

成分構造モデル

$$y_n = t_n + P_n + E_n + \varepsilon_n$$

y_n 観測値
t_n トレンド
P_n 気圧効果
E_n 地球潮汐効果
ε_n 観測ノイズ

成分モデル

$$\Delta^k t_n = w_n$$
$$P_n = \sum_{i=0}^{m} a_i p_{n-i}$$
$$E_n = \sum_{i=0}^{l} b_i et_{n-i}$$

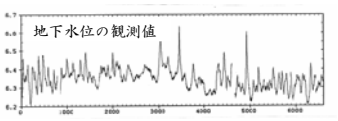
状態空間モデル

$$x_n = Fx_{n-1} + Gv_n$$
$$y_n = Hx_n + w_n$$

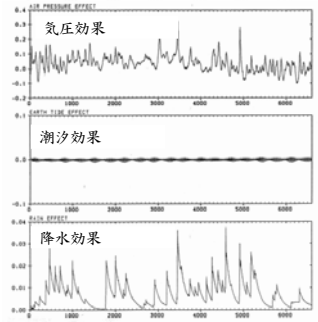
$$x_n = \begin{bmatrix} t_n \\ a_0 \\ \vdots \\ a_m \\ b_0 \\ \vdots \\ b_l \end{bmatrix}, \quad F = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

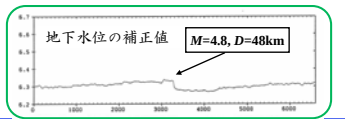
$$H = [1 \quad p_n \quad \cdots \quad p_{n-m} \quad et_n \quad \cdots \quad et_{n-l}],$$

地震の影響の抽出

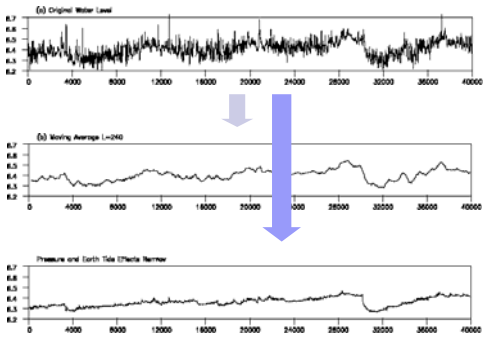


min AIC model
m=25, *l*=2, *k*=5





移動平均 vs. 統計的モデリング

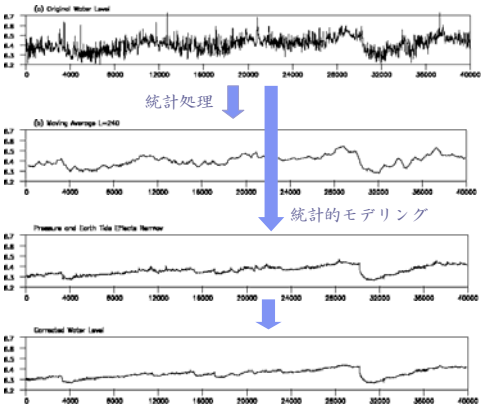


Original

MA filter

Modeling

降水効果項の導入



原データ

統計処理

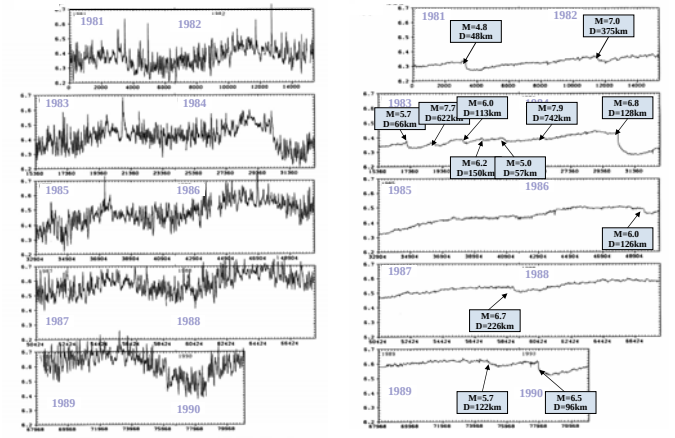
移動平均

統計的モデリング

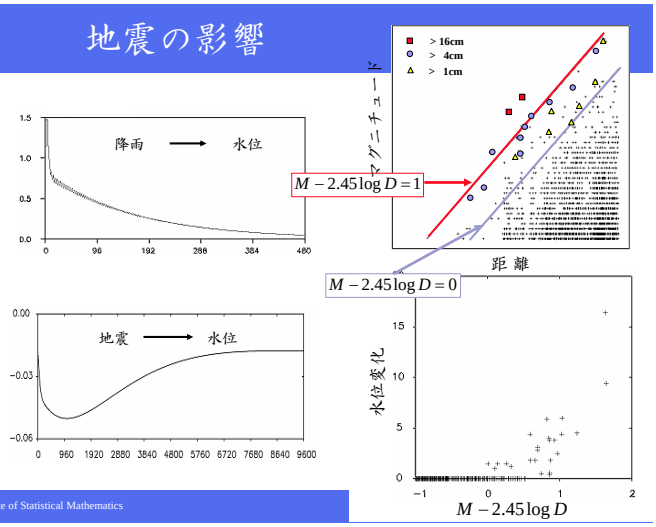
気圧・地球潮汐

気圧・地球潮汐・降雨

地震の影響の検出



地震の影響



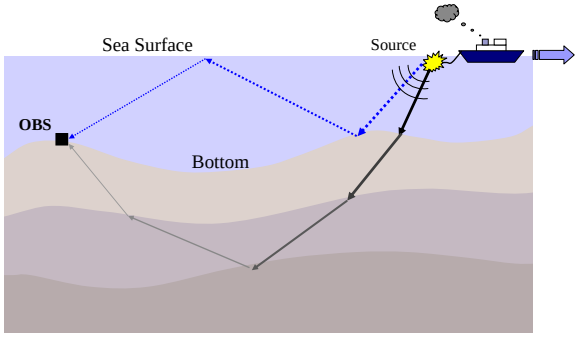
得られた知見

- $M > 2.45 \log D + 0.2$ の地震に水位変化が見られる
- 変化量は $M - 2.45 \log D$ の関数
- 地震がないとき 6cm/年の水位上昇

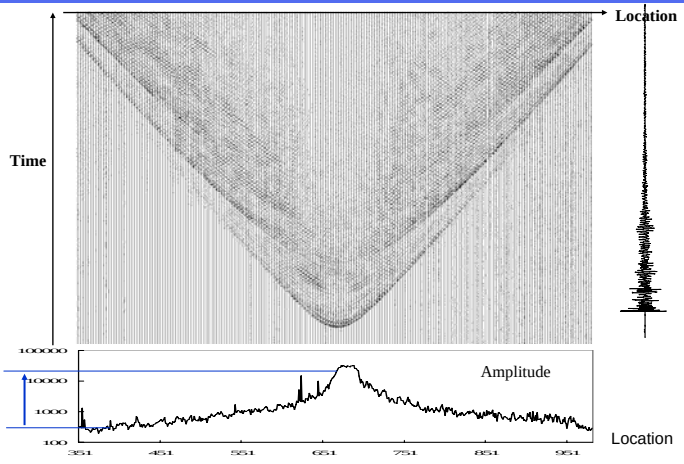
➡ ひずみの増加に対応か？

3. OBS（海底地震計）Data

― 地下構造の探索 ―



観測時系列（調整済，982ch）



時空間モデルによる分解

Basic observation model

$$y_{n,j} = r_{n,j} + s_{n,j} + w_{n,j}$$

$r_{n,j}$ Direct wave

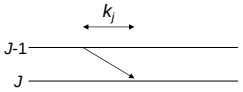
$s_{n,j}$ Reflection wave

Time series model

$$r_n = a_1 r_{n-1} + \cdots + a_\ell r_{n-\ell} + v_{n,r}$$
$$s_n = b_1 s_{n-1} + \cdots + b_m s_{n-m} + v_{n,s}$$

Spatial model

$$r_{n,j} = r_{n-k_j,j-1} + u^r_{n,j}$$
$$s_{n,j} = s_{n-h_j,j-1} + u^s_{n,j}$$



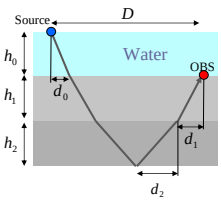
$k_j = \Delta T_j(\mathrm{W0}), \quad h_j = \Delta T_j(\mathrm{Wx})$

$T_j(\mathrm{W0})$: Arrival time of W0

$T_j(\mathrm{Wx})$: Arrival time of Wx

経路モデルと到着時刻

Parallel Layer Structure



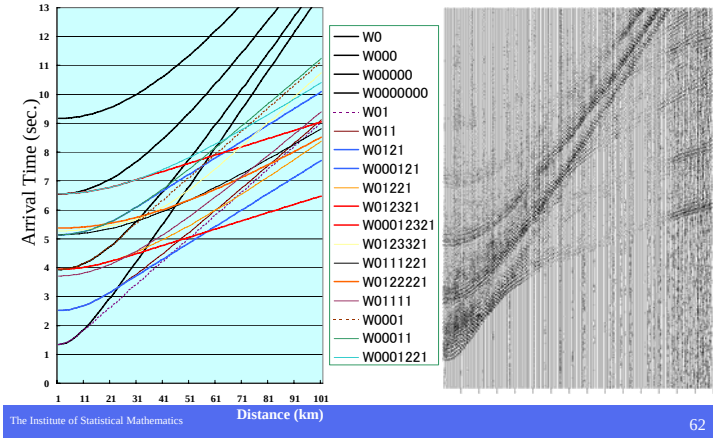
Width $h_0, h_1, h_2, h_3 \cdots$ km

Velocity $v_0, v_1, v_2, v_3 \cdots$ km/sec

経路モデル	到着時刻
Wave(0)	$v_0^{-1}\sqrt{h_0^2+D^2}$
Wave(000)	$v_0^{-1}\sqrt{9h_0^2+D^2}$
Wave(00000)	$v_0^{-1}\sqrt{25h_0^2+D^2}$
Wave(01)	$v_0^{-1}\sqrt{h_0^2+d_{01}^2}+v_1^{-1}(D-d_{01})$
Wave(0001)	$3v_0^{-1}\sqrt{h_0^2+d_{01}^2}+v_1^{-1}(D-3d_{01})$
Wave(0121)	$v_0^{-1}\sqrt{h_0^2+d_{02}^2}+2v_1^{-1}\sqrt{h_1^2+d_{12}^2}+v_2^{-1}(D-d_{02}-2d_{12})$
Wave(00012)	$3v_0^{-1}\sqrt{h_0^2+d_{02}^2}+2v_1^{-1}\sqrt{h_1^2+d_{12}^2}+v_2^{-1}(D-3d_{02}-2d_{12})$
Wave(01232)	$v_0^{-1}\sqrt{h_0^2+d_{03}^2}+2v_1^{-1}\sqrt{h_1^2+d_{13}^2}+2v_2^{-1}\sqrt{h_2^2+d_{23}^2}+v_3^{-1}d_3$

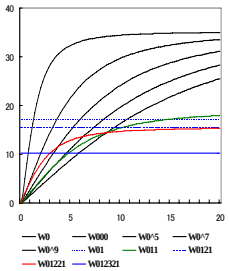
$d_{ij} = v_i h_i / \sqrt{v_j^2 - v_i^2}, \quad d_3 = D - d_{03} - 2d_{13} - 2d_{23}$

経路モデルと到着時刻（OBS4）

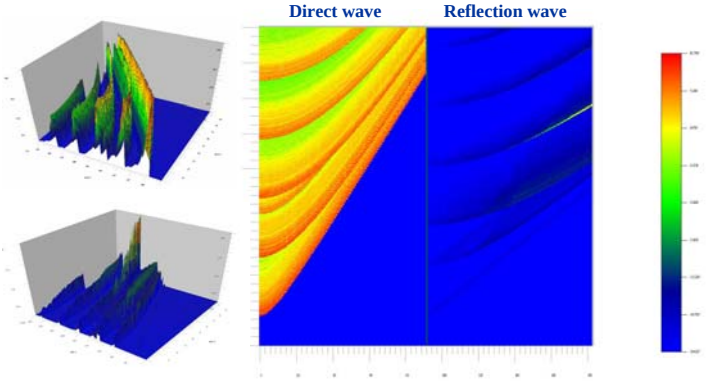


経路モデルと隣接する系列での到着時点差

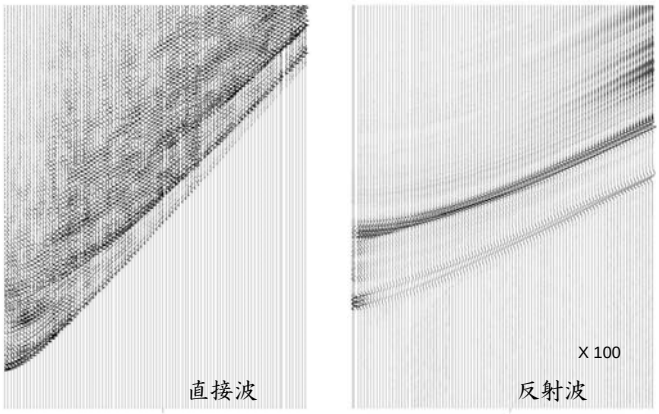
	Epicentral Distance (*km*)				
Path Model	0	5	10	15	20
Wave0	1.7	32.3	34.4	34.8	35.0
Wave000	0.6	21.6	29.7	32.4	33.5
Wave00000	0.3	14.9	24.1	28.8	31.1
Wave011	–	10.5	15.4	17.2	17.9
Wave01221	–	12.9	14.7	15.1	15.3
Wave012321	–	–	10.2	10.2	10.2



空間構造モデルによる分解



時空間モデルによる分解



まとめ

- 社会の変化、研究スタイルの変化
- 知識社会における統計科学の役割
- 状態空間モデリング
  - 非線形・非ガウス型フィルタ・平滑化
  - モンテカルロフィルタ
  - 情報統合・情報抽出