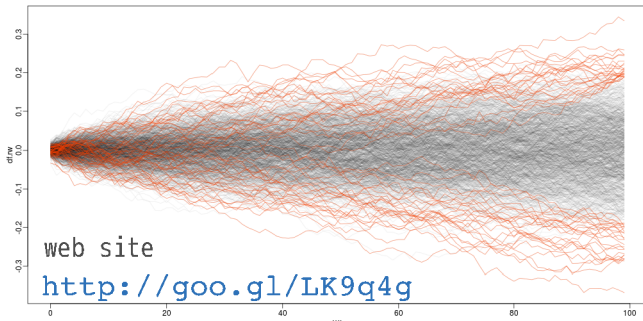


## 時系列データの統計モデリング入門

久保拓弥 (北海道大・環境科学)

kubo@ees.hokudai.ac.jp @KuboBook



2015-03-19

時系列データの統計モデル

1/42

## 今日のハナシ

### 「あぶない」時系列データ解析

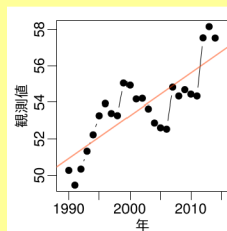
はやめましょう!

(危1) 時系列データの GLM あてはめ

(危2) 時系列  $Y_t \sim$  時系列  $X_t$

各時刻の個体数 ~ 気温 とか

### (危1) 時系列データを GLM で

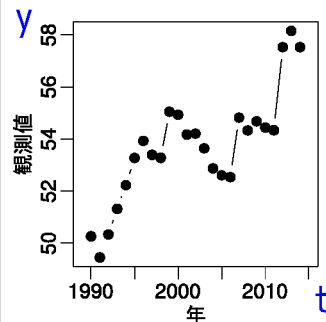


「ゆーいな傾き」を  
ねっぞうする原因

傾きの検定やめて  
AIC モデル選択  
しても同様になる

検定とかモデル選択とかそういう問題ではない  
統計モデルがおかしい?

### このような時系列データがあったとしましょう



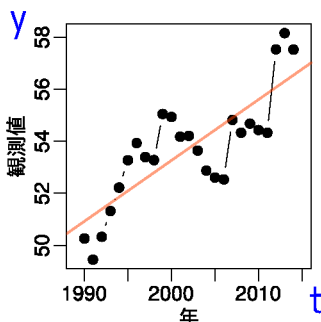
y は何か連続値と  
しましょう  
(今日でてくる y は  
連続値ばかり, と  
いうことで)

2015-03-19

時系列データの統計モデル

4/42

## 時系列データの統計モデリング入門



$glm(y \sim t)$

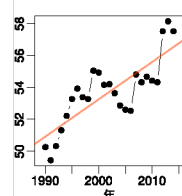
…とモデル  
をあてはめてみた

2015-03-19

時系列データの統計モデル

5/42

### 「やったーゆーいだ!!」……??



> summary(glm(formula = y ~ t))

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1295	-1.0583	-0.0817	0.9860	2.0188

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-414.5655	71.4761	-5.80	6.6e-06
t	0.2339	0.0357	6.55	1.1e-06

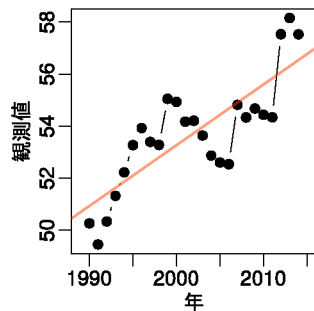
これはまちがい →  $glm(\text{時系列}Y \sim \text{時間 } t)$

2015-03-19

時系列データの統計モデル

6/42

## 時系列の各点は独立ではない



「ゆーいな傾き」を  
ねっぞうする原因

傾きの検定やめて  
AIC モデル選択  
しても同様になる

検定とかモデル選択とかそういう問題ではない

統計モデルがおかしい?

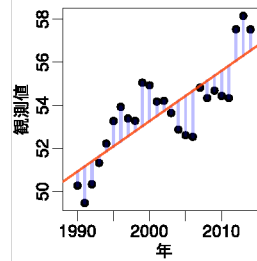
2015-03-19

時系列データの統計モデル

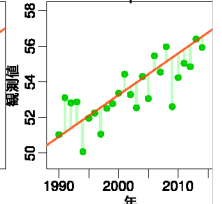
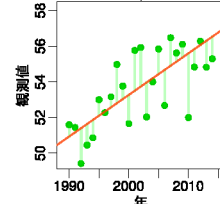
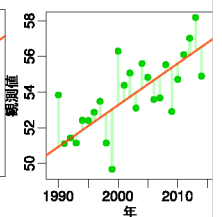
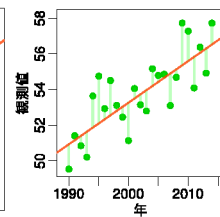
7/42

## 時系列の「ずれ」

## GLM のずれ



ずれかたが  
ちがってる?

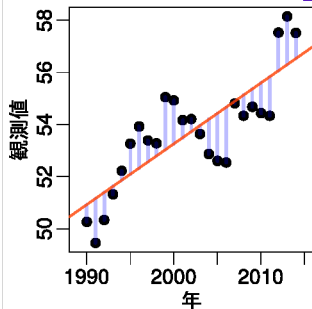


2015-03-19

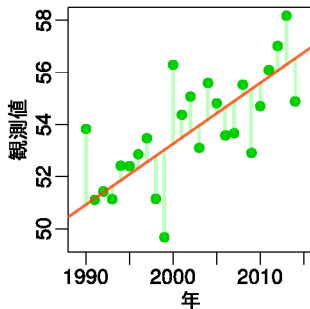
時系列データの統計モデル

8/42

## 時系列の「ずれ」



## GLM のずれ



直線からのずれがちがう!

時間的自己相関がある

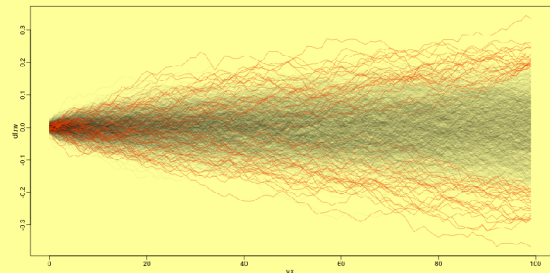
時間的自己相関がない

2015-03-19

時系列データの統計モデル

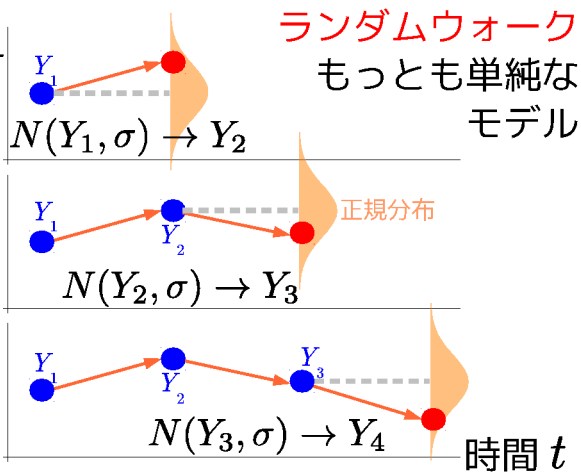
9/42

## 時系列の基本モデルのひとつ ランダムウォーク (乱歩)



変数

$Y$



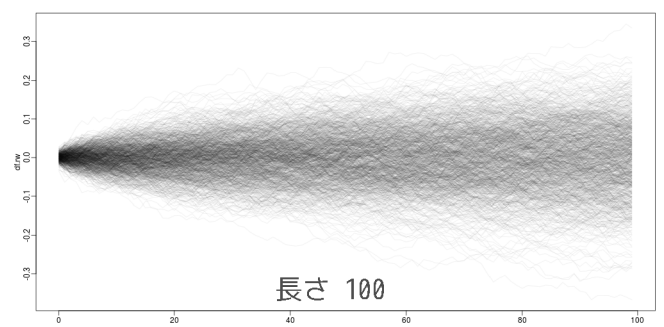
2015-03-19

時系列データの統計モデル

11/42

## ランダムウォークなサンプル時系列

とりあえず 1000 本ほど生成してみました



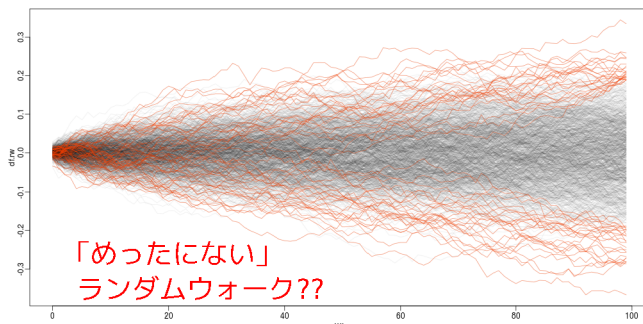
2015-03-19

時系列データの統計モデル

12/42

例外的な時系列というのはいりえる

たとえば  $t = 100$  でかなり外れている 50 本



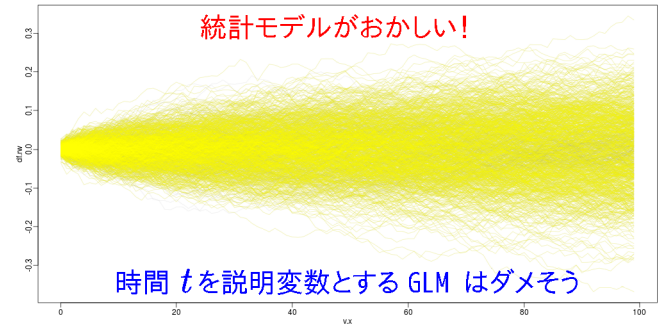
2015-03-19

時系列データの統計モデル

13/42

しかし直線回帰 GLM あてはめると…

ほとんどすべての場合で「ゆーい」!



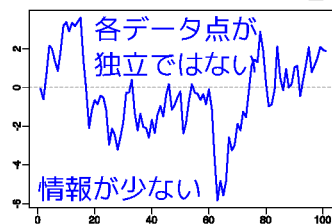
2015-03-19

時系列データの統計モデル

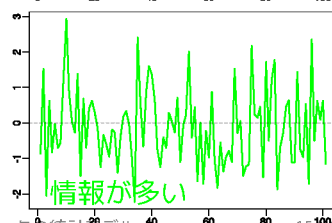
14/42

ちょっとでも傾いてたら「ゆーい」

実際には  
こんなデータ  
なのに



R の glm() は  
こんなデータ  
だとみなしている



2015-03-19

時系列データの統計モデル

15/42

時間的自己相関

(略称: 自己相関, 時間相関)

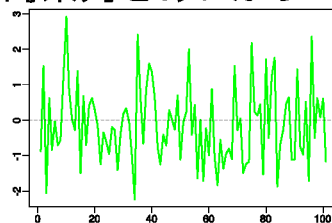
を調べたらいいの?

$$\rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-1})}}$$

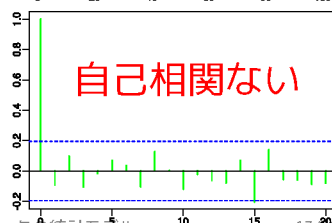


R の ts クラス: 時系列をあつかう

`plot(ts(Y))`



`plot(acf(ts(Y)))`



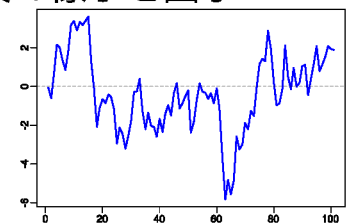
2015-03-19

時系列データの統計モデル

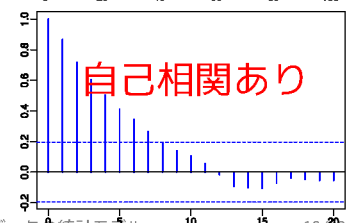
17/42

自己相関減衰の様子を図示

`plot(ts(Y))`



`plot(acf(ts(Y)))`



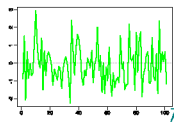
2015-03-19

時系列データの統計モデル

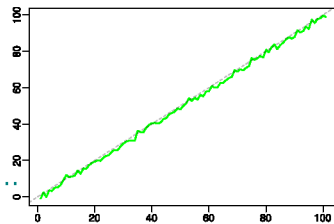
18/42

各点独立のデータをナナメにすると？

`plot(ts(Y))`

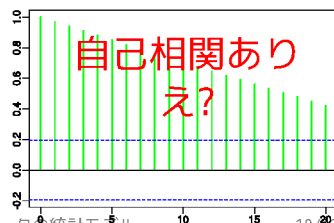


これを  
ナナメに  
したもの  
なんだけど..



`plot(acf(ts(Y)))`

自己相関あり  
え？



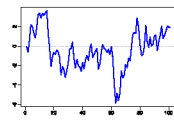
2015-03-19

時系列データの統計モデル

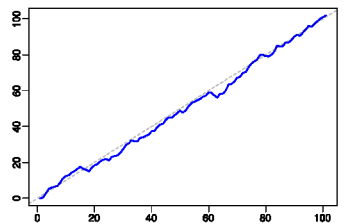
19/42

各点独立のデータをナナメにすると？

`plot(ts(Y))`

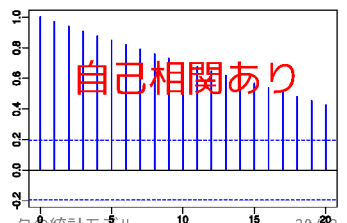


これを  
ナナメに  
したもの



`plot(acf(ts(Y)))`

自己相関あり

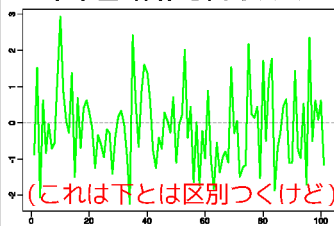


2015-03-19

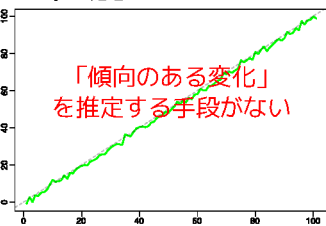
時系列データの統計モデル

20/42

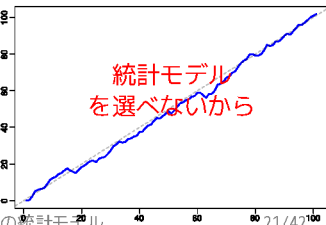
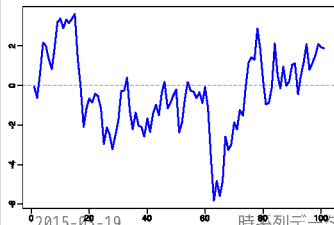
自己相関係数みても区別がつかない



(これは下とは区別つくけど)



「傾向のある変化」  
を推定する手段がない



統計モデル  
を選べないから

2015-03-19

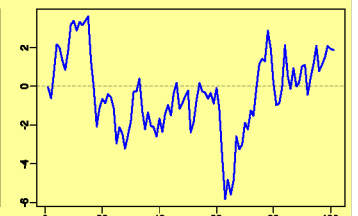
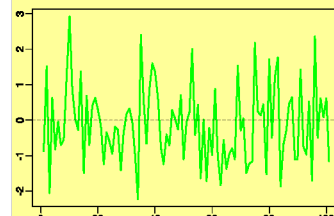
時系列データの統計モデル

21/42

状態空間モデルでたちむかう

時系列データ解析

いろいろな時系列データを  
統一的にあつかえないか？



「統計モデル」とは何か？

どんな統計解析においても  
統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり、これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手づき が準備されていて、モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



2015-03-19

時系列データの統計モデル

23/42

「統計モデル」のしくみを理解しよう！

もうすこし「わかった」ような気分？

種子数の平均値はサイズ  $x$  とともに増大する

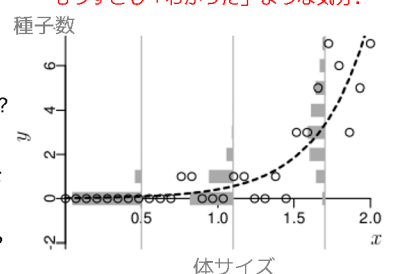
➡ どのように変化するか？  
数式で書くとどうなる？

平均値が増大するとばらつきが変化する

➡ どのようにばらつくのか？  
確率分布？

統計モデルをデータにうまくあてはめる

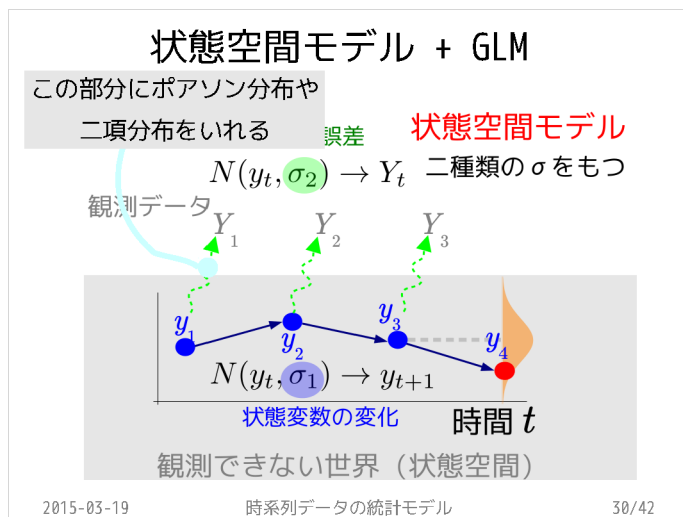
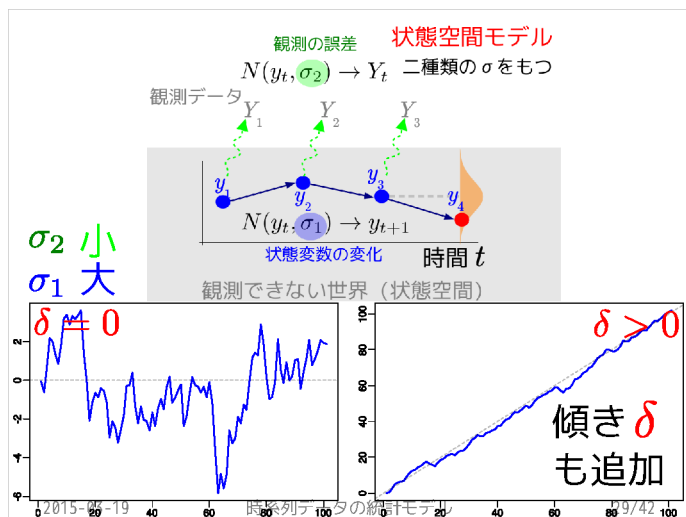
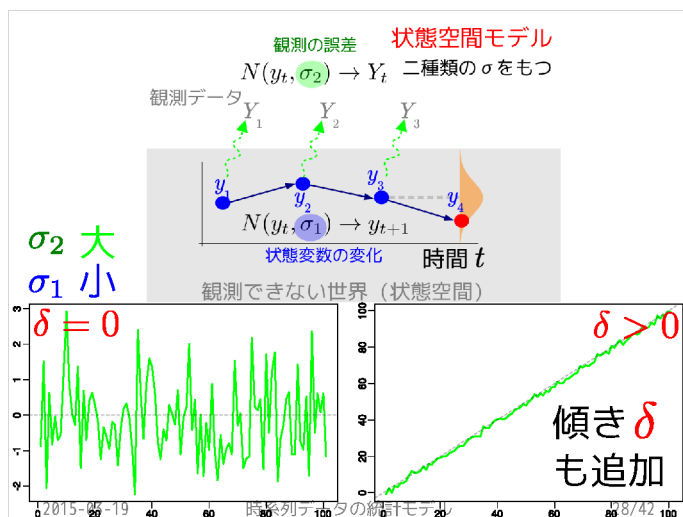
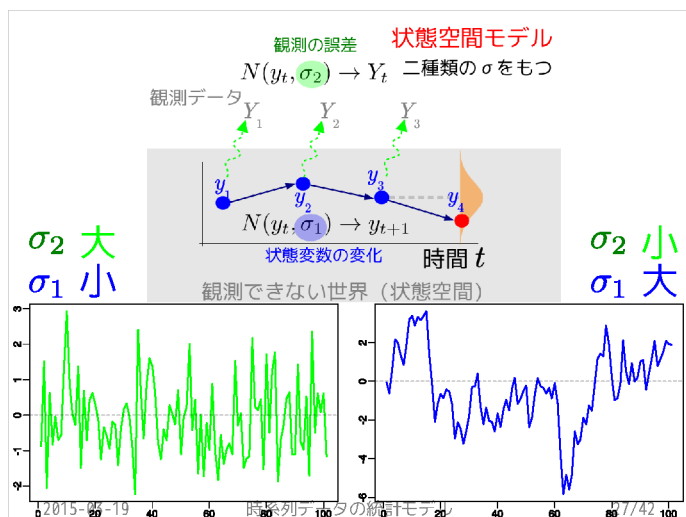
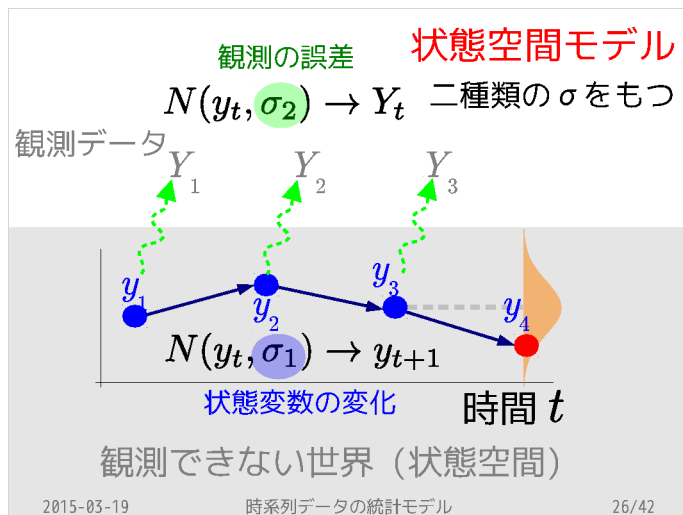
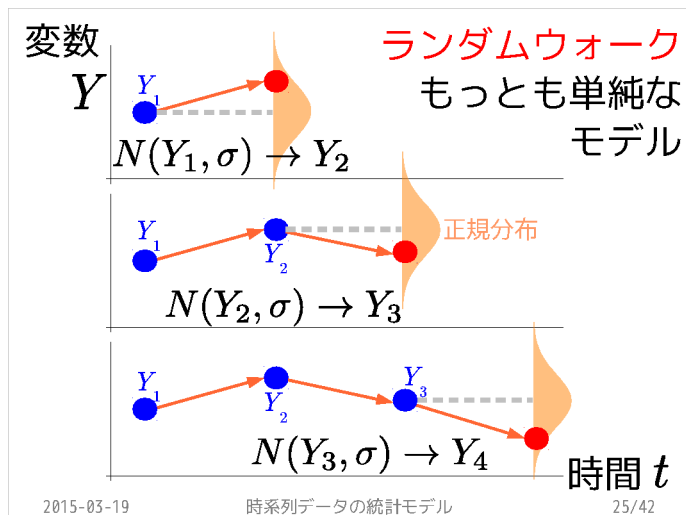
➡ どのようにあてはめるのが妥当なのか？ パラメーター推定法？



2015-03-19

時系列データの統計モデル

24/42



どうやってモデルをあてはめる?



R の状態空間モデルの  
package いろいろある

`library(dlm)`

`library(KFAS)`

しかしより一般化したモデルに  
ついての理解が必要かも

2015-03-19

時系列データの統計モデル

31/42

JAGS でいきましょう

BUGS 言語でこの単純な  
モデルを記述できる



R の「したっば」として  
動かすことができる

時間があれば demo

2015-03-19

時系列データの統計モデル

32/42

```
model
{
  Tau.Noninformative <- 0.0001
  Y[1] ~ dnorm(y[1], tau[2])
  y[1] ~ dnorm(0, Tau.Noninformative)
  for (t in 2:N.Y) {
    Y[t] ~ dnorm(y[t], tau[2])
    y[t] ~ dnorm(m[t], tau[1])
    m[t] <- delta + y[t - 1]
  }
  delta ~ dnorm(0, Tau.Noninformative)
  for (k in 1:2) {
    tau[k] <- 1 / (s[k] * s[k])
    s[k] ~ dunif(0, 10000)
  }
}
```

2015-03-19

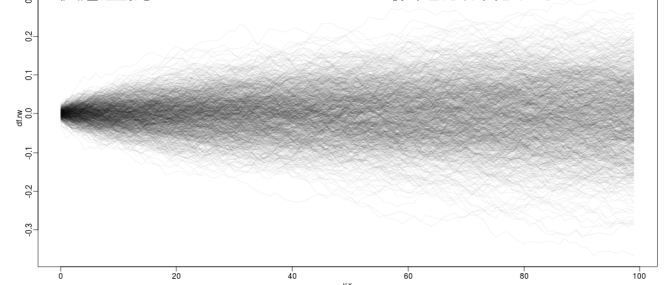
時系列データの統計モデル

33/42

1000 個の架空データを推定

いろいろなランダムウォークが生成される

状態空間モデルのパラメータ推定は成功するか?



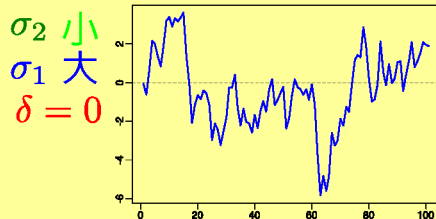
2015-03-19

時系列データの統計モデル

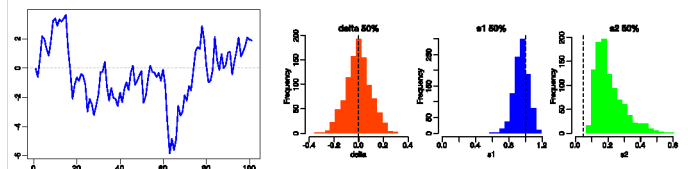
34/42

状態空間モデルを

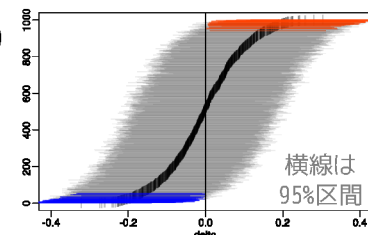
「かたむきゼロ」ランダムウォーク  
 $\delta = 0$   
な架空データにあてはめる



「傾き」 $\delta$ の事後分布を見る



真の $\delta$ は 0



1000回中  
63回ずれた

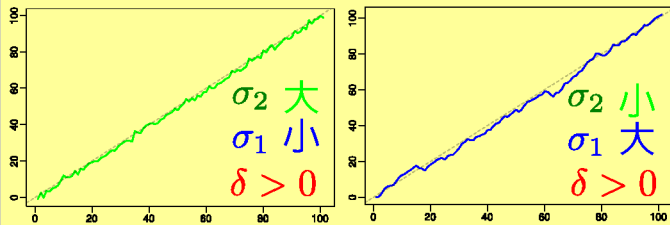
2015-03-19

時系列データの統計モデル

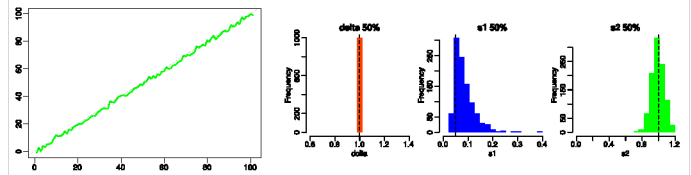
36/42

## 状態空間モデルを

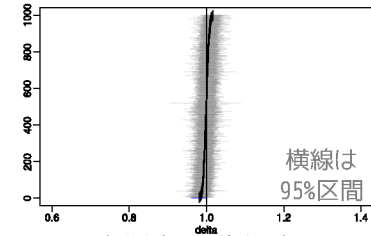
「かたむきあり」ランダムウォーク  
 $\delta > 0$   
 な架空データにあてはめる



## 「傾き」 $\delta$ の事後分布を見る



真の $\delta$ は 1



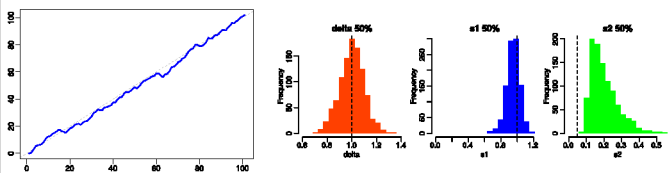
1000回中  
1回ずれた

2015-03-19

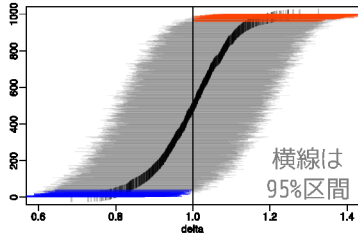
時系列データの統計モデル

38/42

## 「傾き」 $\delta$ の事後分布を見る

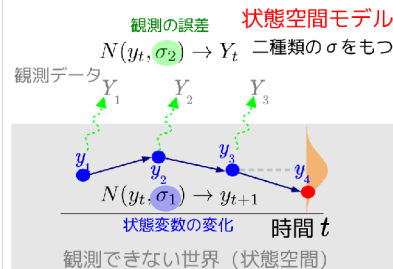


真の $\delta$ は 1



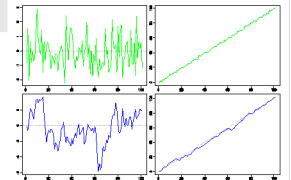
1000回中  
62回ずれた

## とりあえずの結論



ひとつの状態空間  
モデルを使って

右の4状態は  
区別可能でしょう



2015-03-19

時系列データの統計モデル

39/42

2015-03-19

時系列データの統計モデル

40/42

(危2) 時系列データ  $X_t$   
 と 時系列データ  $Y_t$   
 $Y_t \sim X_t$  なうたがわしい回帰  
 spurious regression

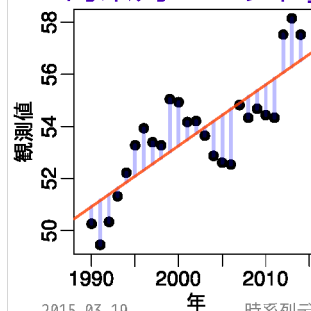
時間があればデモ  
 (すみません)

## 時間的な相関はデータの

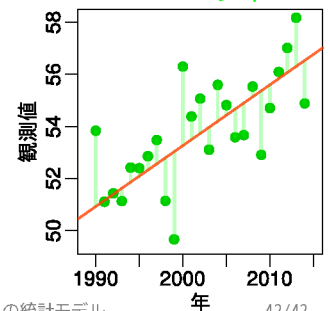
情報量を減少させる

空間相関も...

## 時系列の「ずれ」



## GLM のずれ



2015-03-19

時系列データの統計モデル

42/42