みどりぼん$10~$11

Yusuke Matsuyama@2016/6/30

おしながき

$10

- GLMMのベイズモデル化と推定
- Stanについてちょっと説明
- 事前分布の選び方
- 個体差+場所差の階層ベイズ

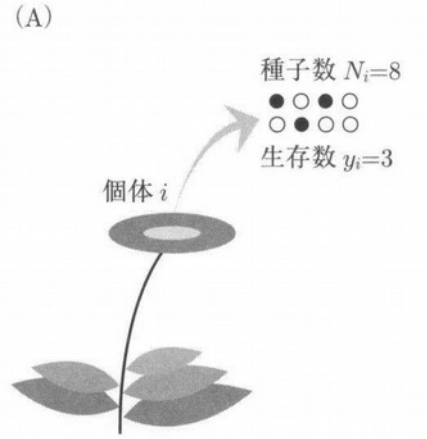
$11

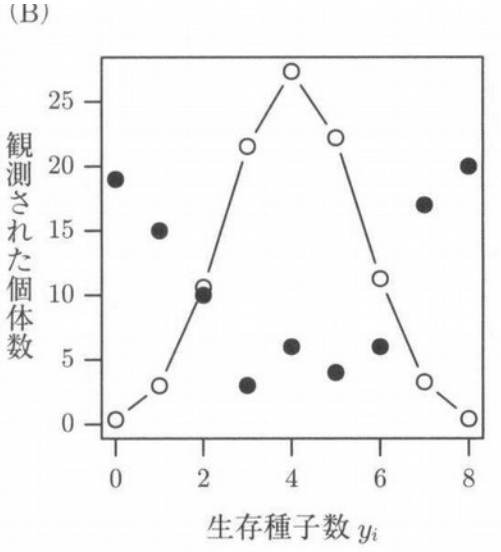
- 一次空間上の個体数分布
- 空間構造を組み込んだ階層ベイズモデル
- 空間相関モデルを使った欠損データへの対応

10章の問題設定

基本的には7章と同じ

- 種子の生存確率を調べる
- 各個体(100個体)から8個の種子を採取
- 生存数をyとする。
- (例によって)過分散なデータなので、
  単純な二項分布はNG





7章の復習(“個体差”を組み込んだGLMM)

[やったこと]

リンク関数と線形予測子 $\mathrm{logit}(q_i)=\beta+r_i$

尤度 $p(\boldsymbol{Y} \mid \beta, \{r_i\}) = \prod_i \binom{8}{y_i} q_i^{y_i}(1-q_i)^{8-y_i}$

β の事前分布 $p(\beta) = \frac{1}{\sqrt{2\pi\times 100^2}} \exp\left(\frac{-\beta^2}{2\times 100^2}\right)$

r_i の事前分布 $p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(\frac{-r_i^2}{2s^2}\right)$

階層事前分布の設定

・パラメータsに対しても事後分布を設定

事前分布...

...とりあえず無情報事前分布にしとく

[用語]

階層事前分布 $p(r_i \mid s)$

超パラメーター s

超事前分布 $p(s)$

階層ベイズモデル

階層事前分布を使っているベイズモデル

階層ベイズモデル

階層ベイズモデルの事後分布:

$$p(\beta, s, \{r_i\} \mid \boldsymbol{Y}) \propto p(\boldsymbol{Y} \mid \beta, \{r_i\})\ p(\beta)\ p(s) \prod_i p(r_i \mid s)$$

例によってMCMC使って事後分布を求める

階層ベイズモデルのMCMC

...のまえに

StanとBUGS

いずれも、ベイズモデリングをするためのDSL言語

それぞれのメリット

データサイエンティスト養成読本の人(http://d.hatena.ne.jp/EulerDijkstra/20130930/1380547174)より

STAN：インストールが簡単、Rから簡単に使える、
実行が早い、収束も早い

BUGS: 書くコードが少ない、パラメータの同時分布が得られる

Stanの文法

4つのブロックに別れる

data：入力するデータを指定

parameters：推定するパラメータを指定

transformed parameters:

(ざっくり)リンク関数と予測子の部分と記述

model：モデルの記述

階層ベイズモデルの記述

```
data {
  int<lower=0> N;
  int<lower=0> Y[N];
}
parameters {
  real beta;
  real r[N];
  real<lower=0> sigma;
}
```

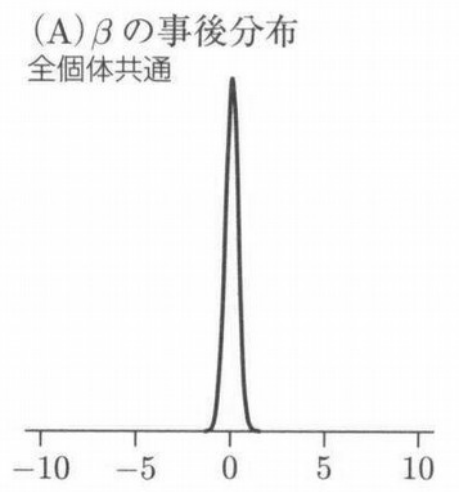
階層ベイズモデルの記述

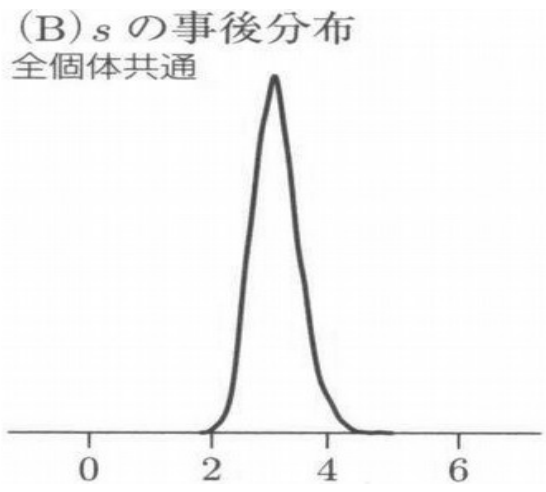
```
transformed parameters {
  real q[N];
  for (i in 1:N) {
    q[i] <- inv_logit(beta + r[i]);
  }
}
```

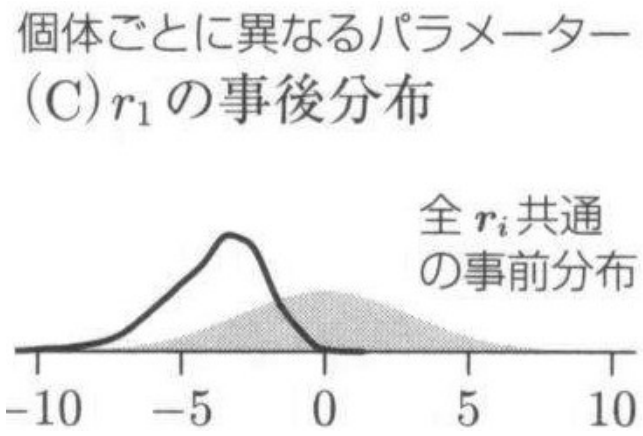
```
model {
  for (i in 1:N) {
      Y[i] ~ binomial(8, q[i]);
  }
  beta ~ normal(0, 100);
  r ~ normal(0, sigma);
  sigma ~ uniform(0, 1.0e+4);
}
```

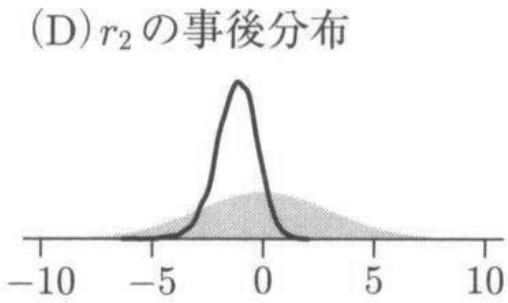
サンプリングの実行(Pystan)

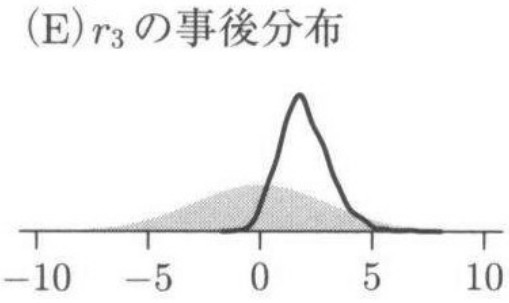
サンプリング結果

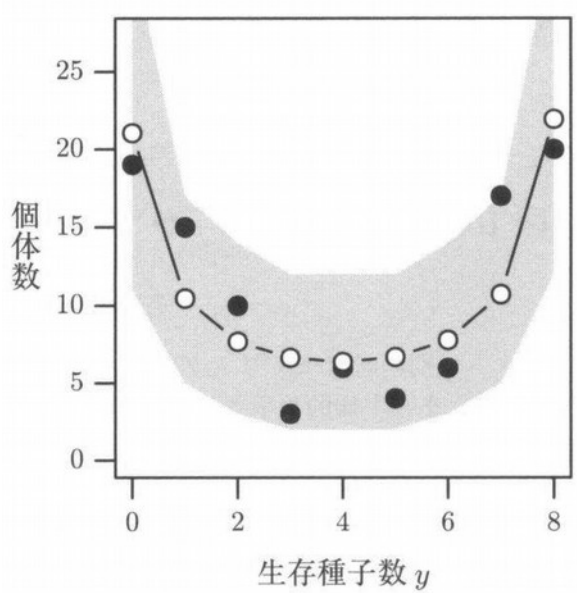

(A) β の事後分布
全個体共通
−10
−5
0
5
10


(B) s の事後分布
全個体共通
0
2
4
6


個体ごとに異なるパラメーター
(C) r_1 の事後分布
全 r_i 共通
の事前分布
−10
−5
0
5
10


(D) r_2 の事後分布
−10
−5
0
5
10


(E) r_3 の事後分布
−10
−5
0
5
10


個体数
25
20
15
10
5
0
0
2
4
6
8
生存種子数 y

ベイズモデルで使う事前分布

Q.ベイズモデルって、どんな事前分布を使えばいいの？

「個体差rの事前分布について、階層事前分布を使えばいいのはどんなとき?」

A.パラメータが説明する範囲、パラメータのばらつき具合を考慮して決めましょう

[前提知識]

統計モデルに含まれるパラメータには

・データを大域的に説明する少数パラメータ

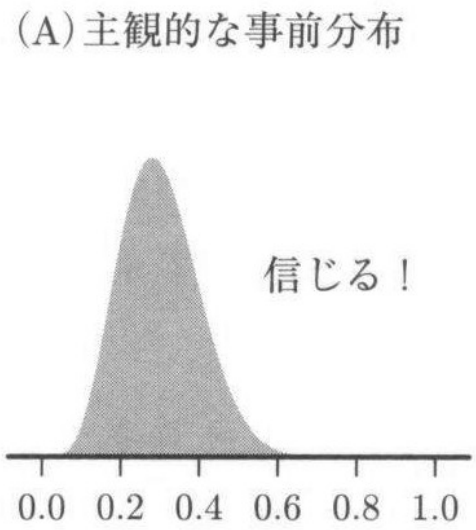
・データのごく一部を説明する多数の局所的パラメータ

がある。

主観的な事前分布

解析者

「たぶんrの分布はこうだろう(断言)」

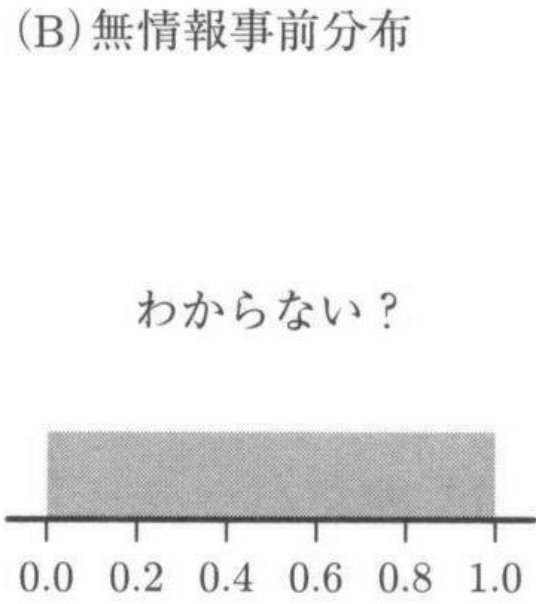


無情報事前分布や、階層事前分布で事足りるので、「主観的な事前分布」が出る幕はない

ただし...

例えば「測定誤差」の事前分布を考えるとき、測定機器によっては、カタログスペックから主観的な確率分布を設定することになることもある。

無情報事前分布



解析者

「rがどんな分布なんて知るか!」

今回の例において、切片 β は、データ全域を説明する大域的パラメータなので、無情報事前分布を用いる

個体差+場所差の階層BM

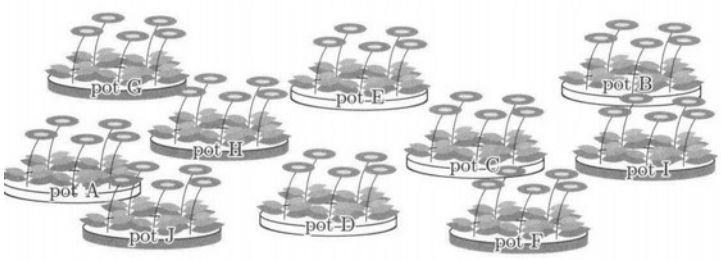
[問題設定]

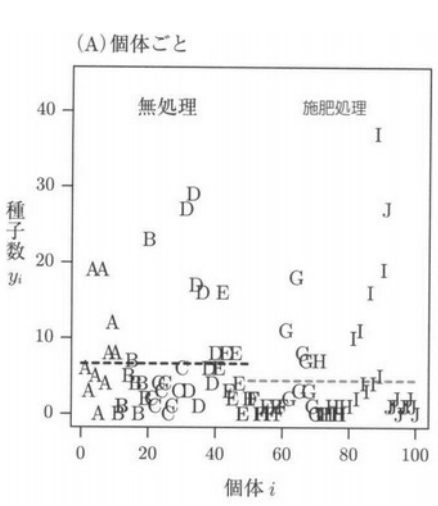
10個の植木鉢とそれぞれに10の個体

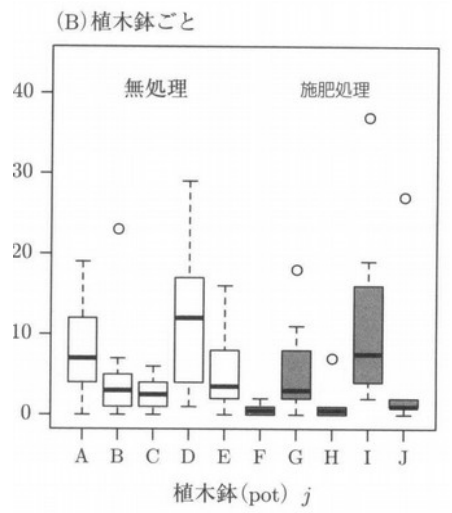
肥料を与える時の種子数の変化をモデリングしたい。

例によ(ry)過分散なデータ

　　個体差によるものと植木鉢の差によるもの

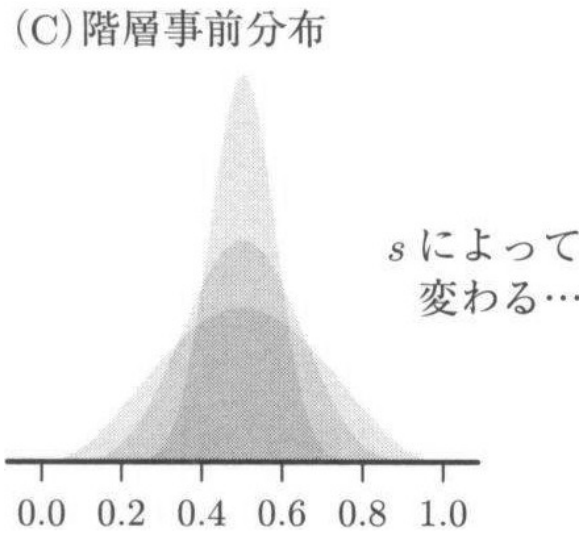






階層事前分布

解析者「超パラメータsによって、
rが説明する範囲が変わる...」



例題におけるパラメータrは、

・個々のrはデータ全体のごく一部を説明するだけ

・r_i は、ある範囲に存在するパラメータの集まり

→r_i は局所的パラメータ → 階層事前分布を選択

階層事前分布により、局所パラメータを「拘束」

個体差+場所差の階層BM

種子数のばらつきをポアソン分布で表現

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

平均種子数

$$\log \lambda_i = \beta_1 + \beta_2 f_i + r_i + r_{j(i)}$$

植木鉢差(標準偏差sp)

切片　施肥の有無　個体差(標準偏差s)

β_1, β_2: 大域的な平均パラメータ

→ 無情報事前分布(平らな正規分布)

s, s_p: 大域的なばらつきパラメータ

→ 無情報事前分布(一様分布)

r_i, r_j 局所パラメータ → 階層事前分布

個体差+場所差の階層BM

β_2の事後分布95%信頼区間：-2.47~0.70

→ 肥料はあんまり意味なさそう

→ そら (肥料効果が0な架空データを
生成したんだから)そうよ

→ (でもね)“手抜き”な統計モデリングだと、
肥料の効果が「推定」されてしまうこともある

例)個体差・植木鉢差を無視したGLMだと、「肥料によって平均種子数が低下」がAIC最良

10章まとめ(コピペ)

1)GLMMをベイズモデル化すると階層ベイズモデルになる

2)階層ベイズモデルとは、事前分布となる確率分布のパラメータにも事前分布が指定されている統計モデル

3)無情報事前分布と階層事前分布を使うことで、ベイズ統計モデルから主観的な事前情報分布を排除できる

4)個体差+場所差といった複雑な構造のあるデータの統計モデリングでは、階層ベイズとMCMCサンプリングによるパラメータ推定の組み合わせで対処すれば良い

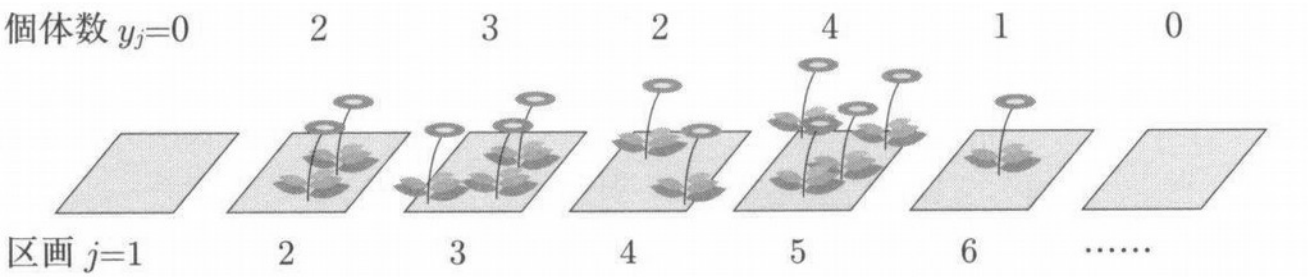
11章のabst

階層ベイズの応用例 -空間構造のあるベイズモデル-

これまで：植木鉢ごとの差は植木鉢ごとに独立に決まる

げんじつ：データをとる空間配置の影響を無視できない

れい：



場所差の空間(的自己)相関を考慮した統計モデリングの話

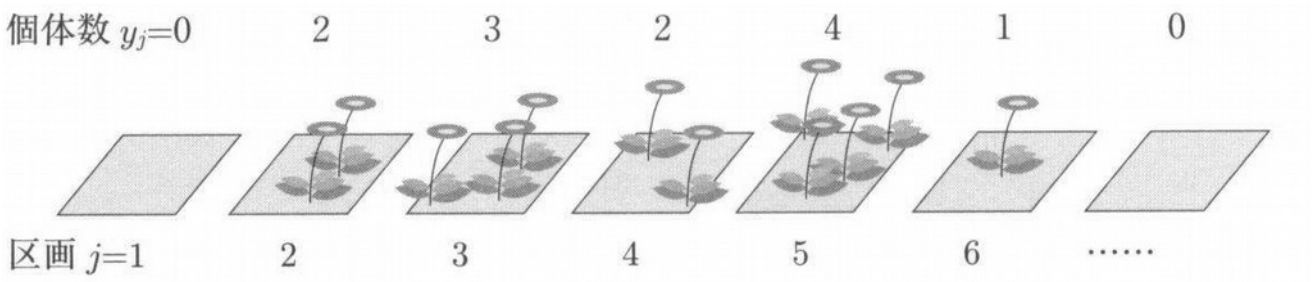
11章の問題設定

[目的]観測データに基づいて、位置によって変化する
局所密度を構成できるような統計モデルを作る

[データ]

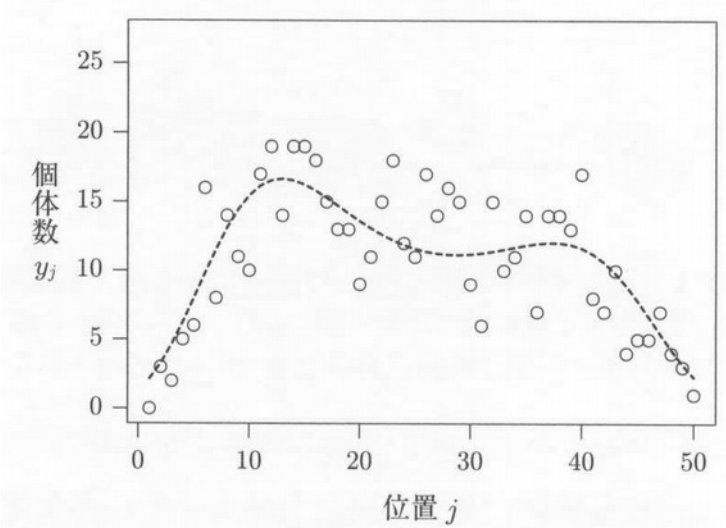
どこかの草原あたりで何かの生物の個体数をカウント

50個の直線的に並んだ”調査区域”を設定



11章の問題設定

とりあえずデータを見てみよう



[わかること]

この生物の局所密度は場所ごとに異なる

局所密度はなだらかに変化しているらしい

場所差を組み込まないと...

とりあえず、全ての区画が共通の平均値を持つ
ポアソン分布に従うと考える

$$p(y_j \mid \lambda) = \frac{\lambda^{y_j} \exp(-\lambda)}{y_j!}$$

標本分散は標本平均と同じ10.9くらいになるはず
→ 実際は27.4

→ なんらかの方法で
“場所差”を組み込んだ統計モデリングが必要

空間構造のある階層事前分布

調査区画ごとに平均パラメータを設定して、

50個のパラメータを推定するのは頭が悪い

「全体に共通する大域的な密度」と「局所的な差異」を同時に組み込みたい

→ 次のように平均個体数を表す

$$\log \lambda_j = \beta + r_j$$

β には無情報事前分布、r_i には階層事前分布を使用

空間構造のある階層事前分布

場所差 r_iの事前分布に、r_i が位置によって少しずつ変化する様子を入れるにはどうすれば...

[仮定]

区画の場所差は、近傍区画の場所差からのみ影響される

“近傍”の個数は有限であり、モデル設計者が指定する

“近傍”の直接の影響は等しく1/区画数

空間構造のある階層事前分布

さらに、「ある区画はそれと接してる区間のみと相互作用 しうる」と問題を簡単にする

[設定]r_jの近傍である r_{j-1}と r_{j+1}の値を固定したときのr_jの条件付き分布を正規分布とする

$$p(r_j \mid \mu_j, s) = \sqrt{\frac{n_j}{2\pi s^2}} \exp\left\{ -\frac{(r_j - \mu_j)^2}{2s^2/n_j} \right\} \qquad \mu_j = \frac{r_{j-1} + r_{j+1}}{2}$$

端の観測区画については、

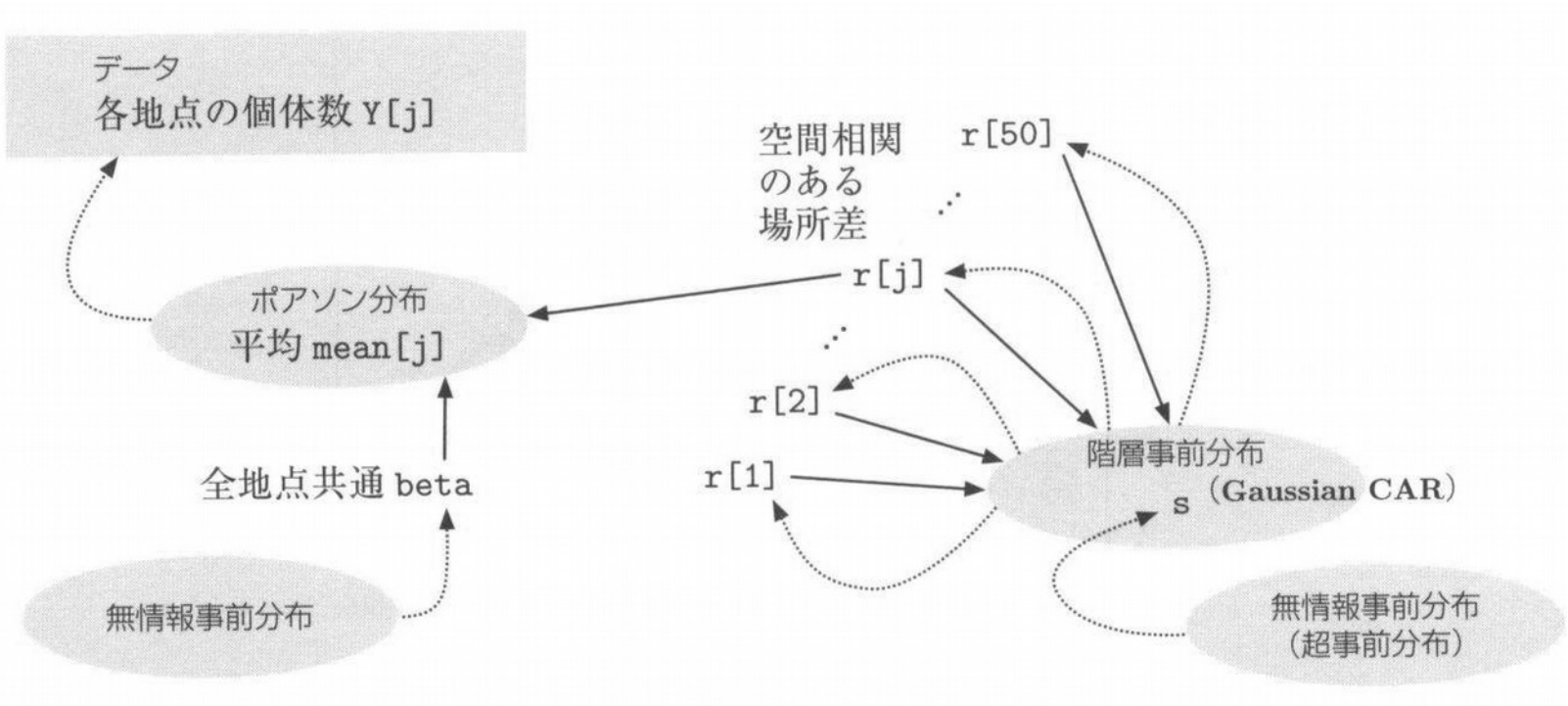
$$\mu_1 = r_2, \mu_{50} = r_{49}$$

空間構造のある階層事前分布

同時分布：

$$p(\{r_j\} \mid s) \propto \exp\left\{ -\frac{1}{2s^2} \sum_{j \sim j'} (r_j - r_{j'})^2 \right\}$$

条件付き自己回帰(CAR)



今回のように制約をつけて簡単にしたものは

intrinstic　Gaussian CARとよぶ。

Let's sampling

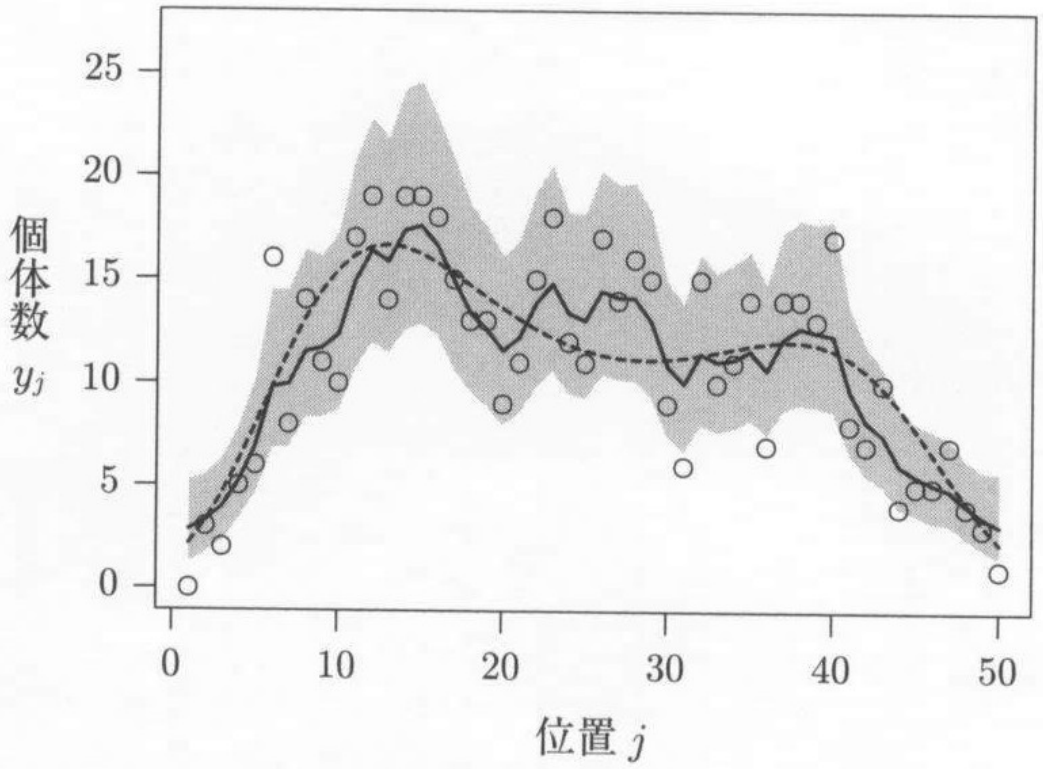
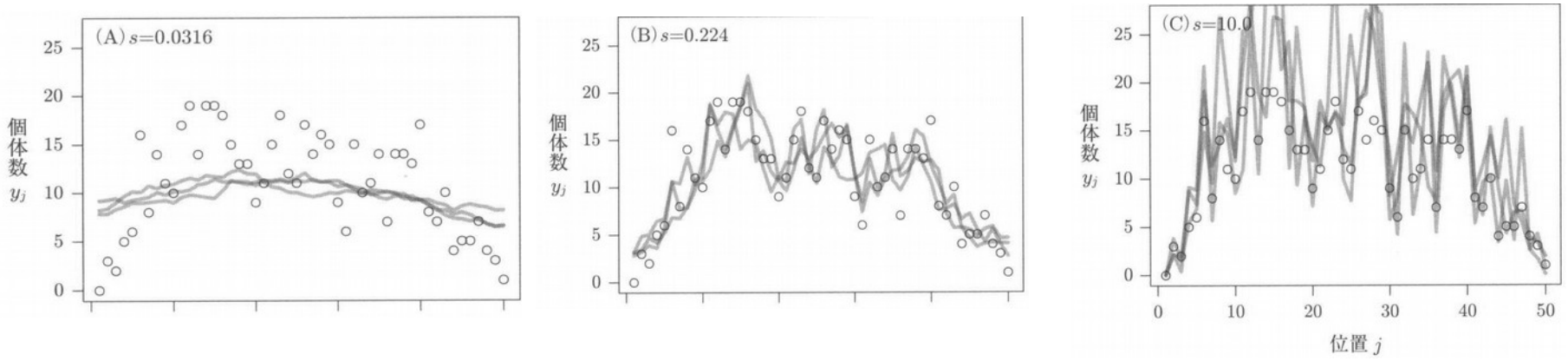


図 11.4 図 11.2 にモデルによる予測を追加した．切片 β と場所差 r_j の事後分布から予測された場所ごとの平均 λ_j の分布の中央値(黒線)と 80% 区間(グレイの領域)．

確率場

$\{r_j\}$のように、相互作用する確率変数で埋め尽くされた空間(今回は一次元)のことを確率場という

空間統計モデルと確率場



小さい ← s → 大きい

(a) sが小さい → 両隣の平均と似ている

→ r_j のばらつきは小さい

(c) s が大きい → r_jは隣の値に左右されづらい

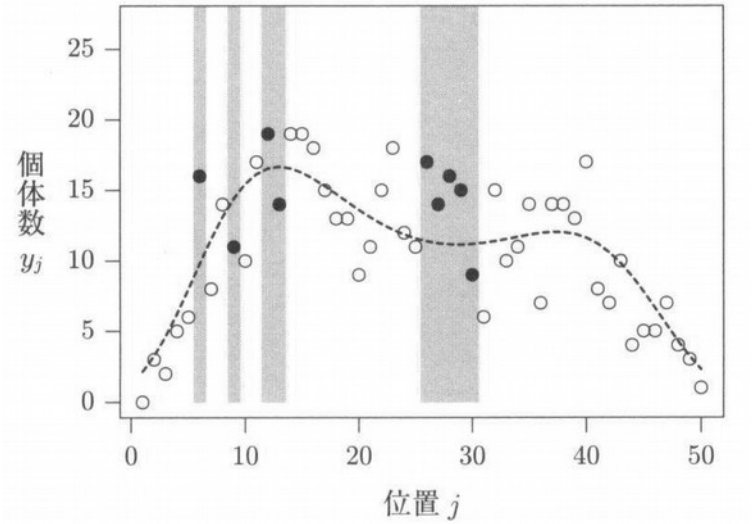
→ r_jのばらつきは大きい

欠測のある観測データ

空間相関と組み込んだ階層ベイズは、欠測のあるデータに対して良い予測を得られることがある

例)下グラフのような架空データ

(灰色の部分は欠測範囲、黒丸は欠測値)



欠測のある観測データ

図Aは空間相関を考慮したモデル

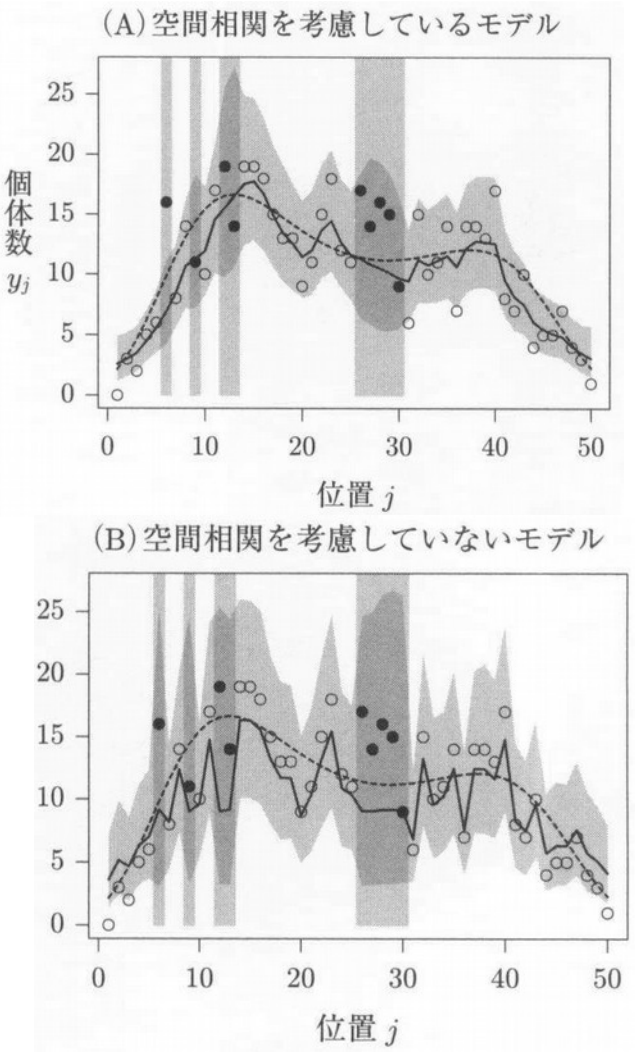
図Bは空間相関を考慮しないモデル

(個体差は全て独立)

図B r_j ではデータがない区間では、

を決められない

→ 予測区間の幅が大きい



11章まとめ(コピペ)

1)空間構造のあるデータを統計モデル化する場合、
空間相関を考慮しなければならない

2)空間相関のある場所差は確率場を使って表現できる

3)空間相関を考慮した階層ベイズモデルは観測データの欠測部分の予測(補完?)の用途にも利用できる