2010-11-10

生態学の統計モデリング 第 10 回

(第 9 章 – 第 11 章あたりの内容)

便利な道具: 階層ベイズモデル

久保拓弥 kubo@ees.hokudai.ac.jp

http://goo.gl/MNbr

今日の話: 階層ベイズモデル + WinBUGS

1. **階層ベイズモデル: GLMM のベイズモデル化**

   事前分布の設計について

2. **空間構造のある階層ベイズモデル**

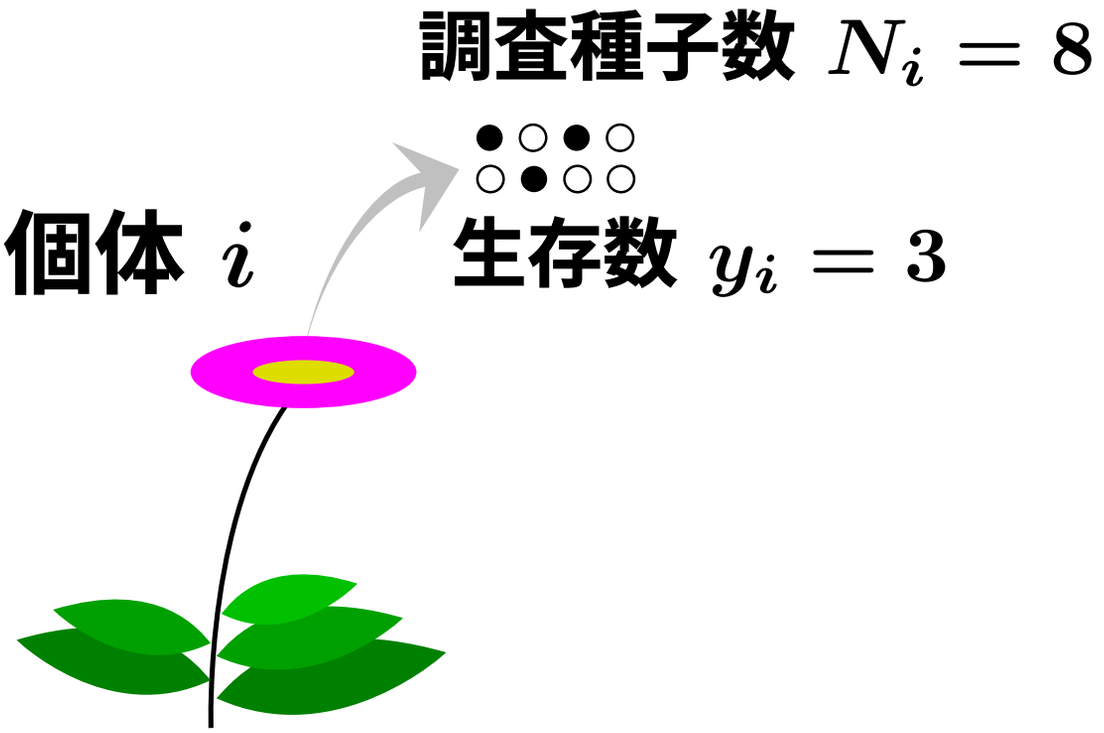
   空間的自己相関をくみこむ

1. 階層ベイズモデル: GLMM のベイズモデル化

事前分布の設計について

例題: 架空植物の種子の生存確率

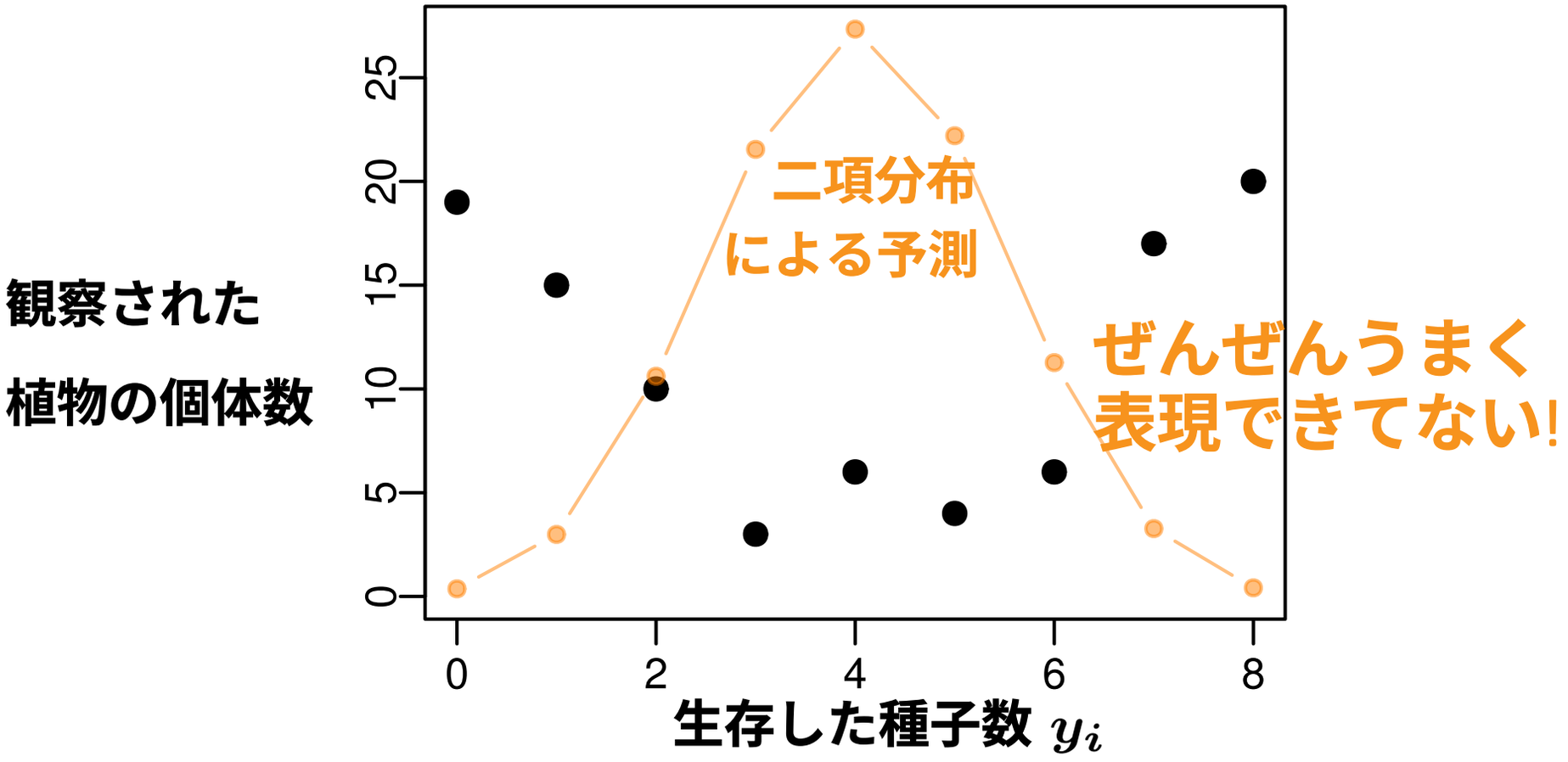
- 架空植物の種子の生存を調べた
  - この植物ではどの個体でも 8 個 調べたとする
  - 種子の中には発芽能力があるもの (生存)，ないもの (死亡) がある
  - 生存確率: ある種子が生存している確率



- データ: 植物 100 個体，合計 800 種子の生死を調べた
- 問: 種子の生存確率はどのように統計モデル化できるか?

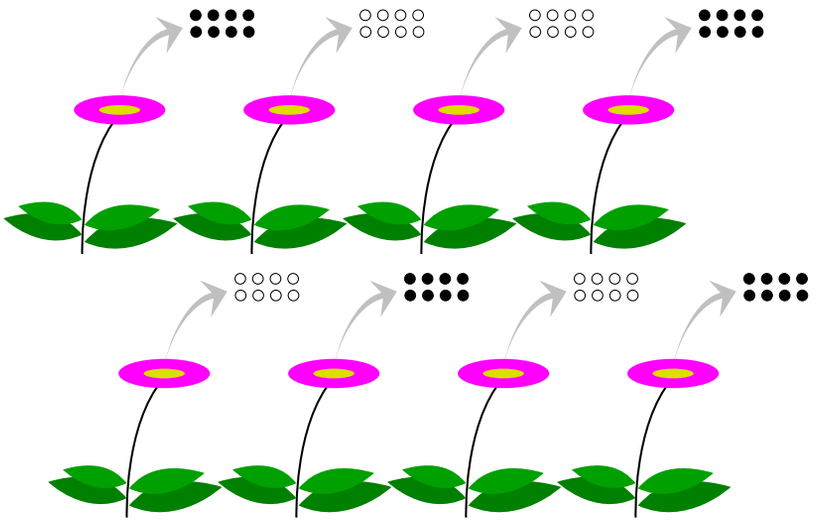
現実的な観測データ: 二項分布だめだめ?!

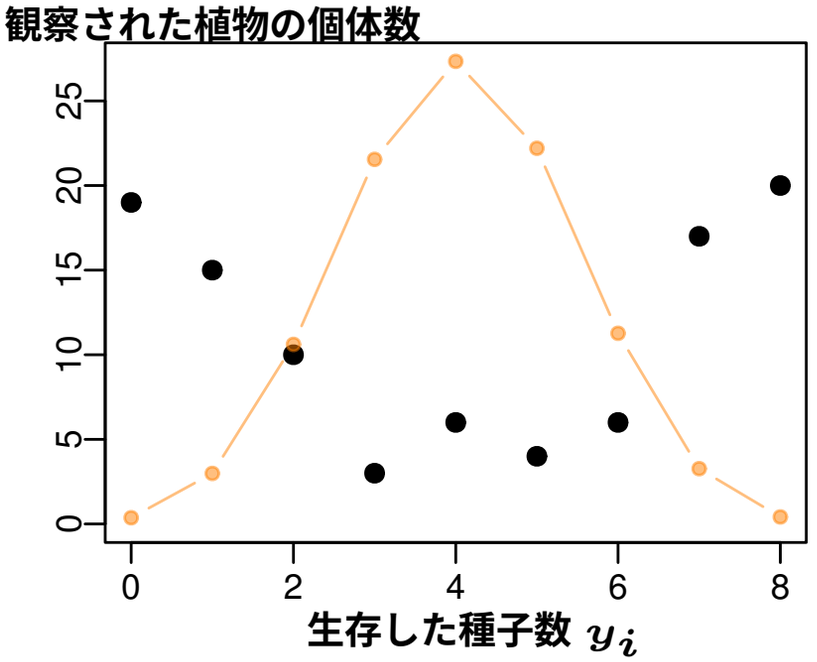
100 個体の植物の合計 800 種子中 403 個の生存が見られたので，平均生存確率は 0.50 と推定されたが……



「個体差」 $\longrightarrow$ 過分散 (overdispersion)

極端な過分散の例





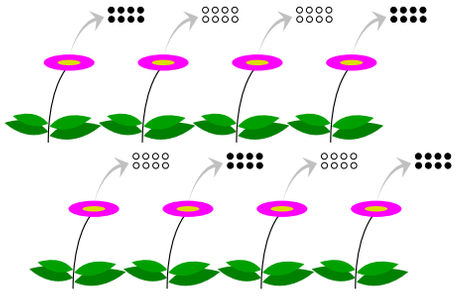
- 種子全体の平均生存確率は 0.5 ぐらいかもしれないが……
- 植物個体ごとに種子の生存確率が異なる: 「個体差」
- 「個体差」があると overdispersion が生じる
- 「個体差」の原因: ?

あのー …… 「個体差」とは?

- 生物学的には明確な定義はない
- しかしデータ解析においては人間が主観的に「これは個体差由来の効果であり，観察されたパターンに影響している」と定義，そして以下の二種類を区別する:
  1. fixed effects 的な効果
  2. random effects 的な効果
- 同様に，ブロック差・場所差・時間ごとに異なる差，などが統計モデルの中で定義される

「個体差」の fixed だの random だの …… って何?

- 「個体ごとに異なる何かに由来する効果」を fixed/random effects にわけて統計モデリングする:
  1. fixed effects 的な効果: 「この要因は生存確率を上下するだろう」と観測者が設定・測定した要因 (実験処理，植物のサイズなど)
     - この例題では fixed effects 的な個体差はない
  2. random effects 的な効果: fixed effects 的ではない要因 (観測対象個体に関連する，人間が設定・測定していないすべて)
     - 平均生存確率を変えずにばらつきだけを変えると考える



今回の例題では random effects 的な「個体差」の統計モデリングに専念

モデリングやりなおし: まず二項分布の再検討

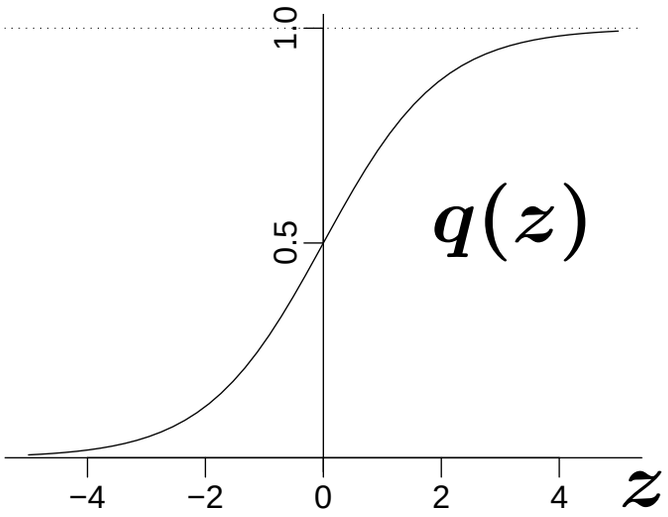
- 生存確率を推定するために **二項分布**という確率分布を使う
- 個体 i の N_i 種子中 y_i 個が生存する確率は二項分布

$$p(y_i \mid q_i) = \binom{N_i}{y_i} q_i^{y_i} (1-q_i)^{N_i - y_i},$$

- ここで仮定していること
  - **個体差がある**
  - **個体ごとに異なる生存確率 q_i**

ロジスティック関数で表現する生存確率

- そこで生存する確率 $q_i = q(z_i)$ をロジスティック (logistic) 関数 $q(z) = 1/\{1 + \exp(-z)\}$ で表現



- 線形予測子 $z_i = a + b_i$ とする
  - パラメーター a: 全体の平均
  - パラメーター b_i: 個体 i の個体差 (ずれ)

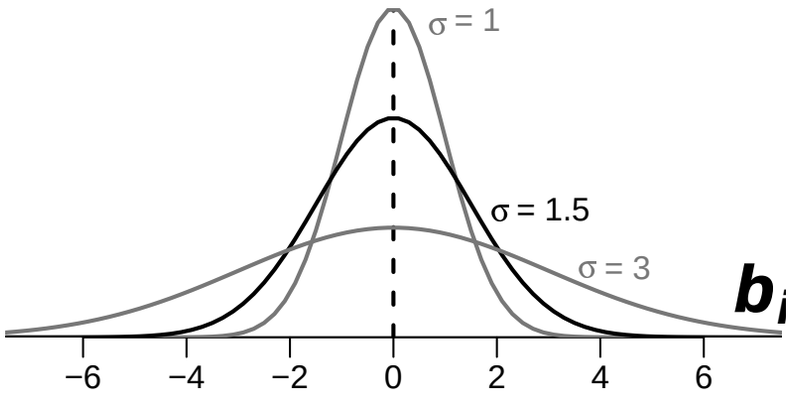
個々の個体差 b_i を最尤推定するのはまずい

- 100 個体の生存確率を推定するためにパラメーター 101 個(a と $\{b_1, b_2, \cdots, b_{100}\}$) を推定すると……
- 個体ごとに生存数 / 種子数を計算していることと同じ!
  (「データのよみあげ」と同じ)
- こう仮定すると問題がうまくあつかえないだろうか?
  - 個体間の生存確率はばらつくけど，そんなにすごく異ならない?
  - 観測データを使って,「個体差」にみられるパターンを抽出したい (統計モデル化)

階層ベイズモデル化: b_i の事前分布の設計

平均ゼロで標準偏差 s の正規分布

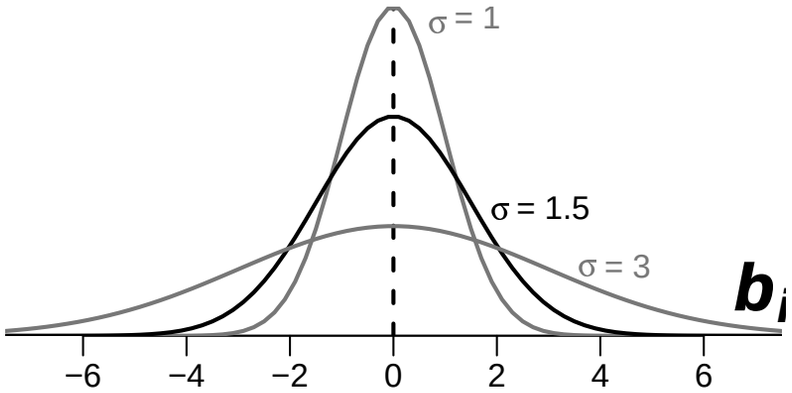
$$p(b_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp \frac{-b_i^2}{2s^2},$$



個体差 $\{b_1, b_2, \cdots, b_{100}\}$ がこの確率分布に従うとする

b_i の事前分布は無情報事前分布ではない

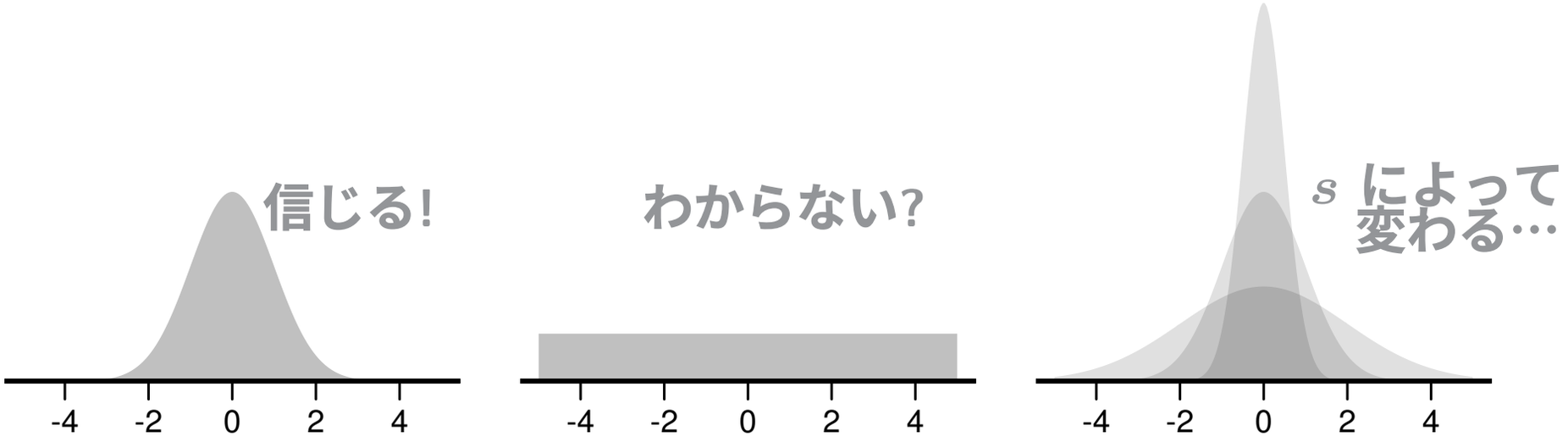
データにあわせて s が変化する階層的な事前分布



- s がとても小さければ個体差 b_i はどれもゼロちかくになる $\rightarrow$ 「どの個体もおたがい似ている」
- s がとても大きければ，b_i は各個体の生存数 y_i にあわせるような値をとる

図 10.3: 個体差 b_i の事前分布は?

(A) 主観的な事前分布(B) 無情報事前分布　(C) 階層的な事前分布



(A) 主観的な事前分布: 「自分の信じるところによれば，b_i たちはこんな分布になる」を表現している．

(B) 無情報事前分布: 「b_i たちがどんな値になるのかまったくわかりません」を表現しようとしている (しかし -5 から 5 ぐらい，という主観も表現している)．

(C) **階層的な事前分布**: b_i の事前分布のパラメーター s がいろいろな値をとる，そして s についての超事前分布を設定する．

図 10.4: 階層的な事前分布と $y_i = 2$ の個体の b_i

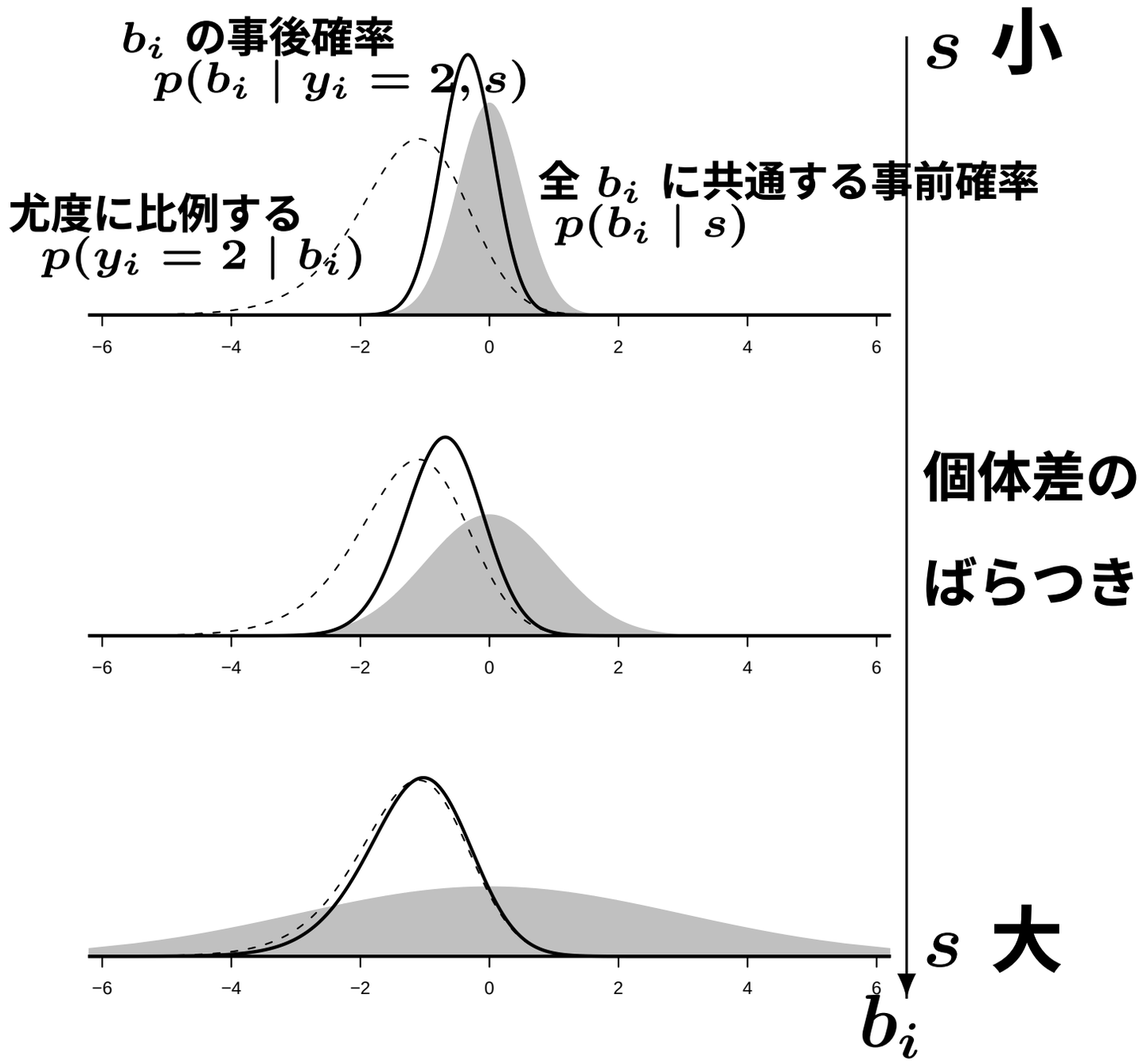
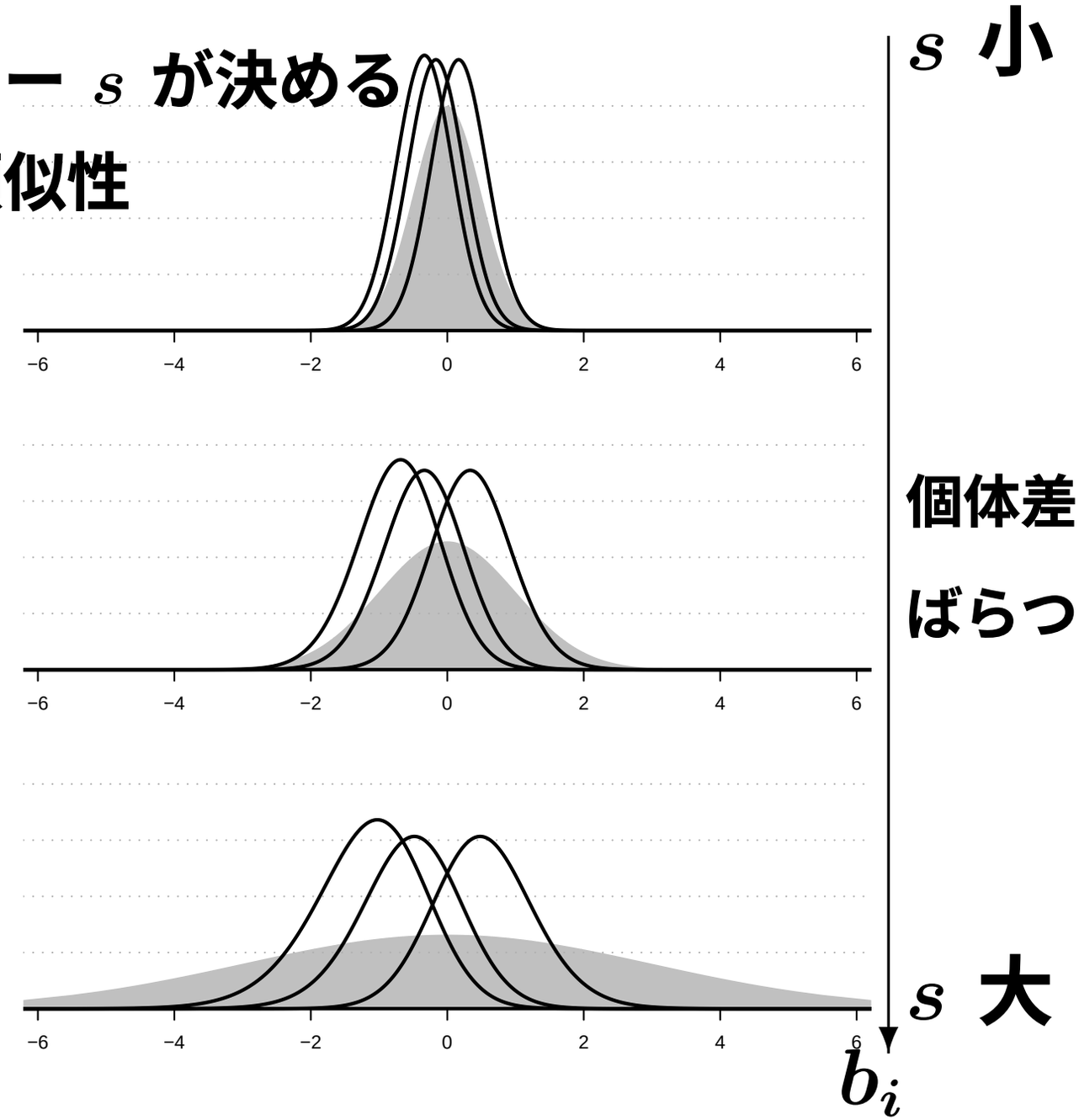
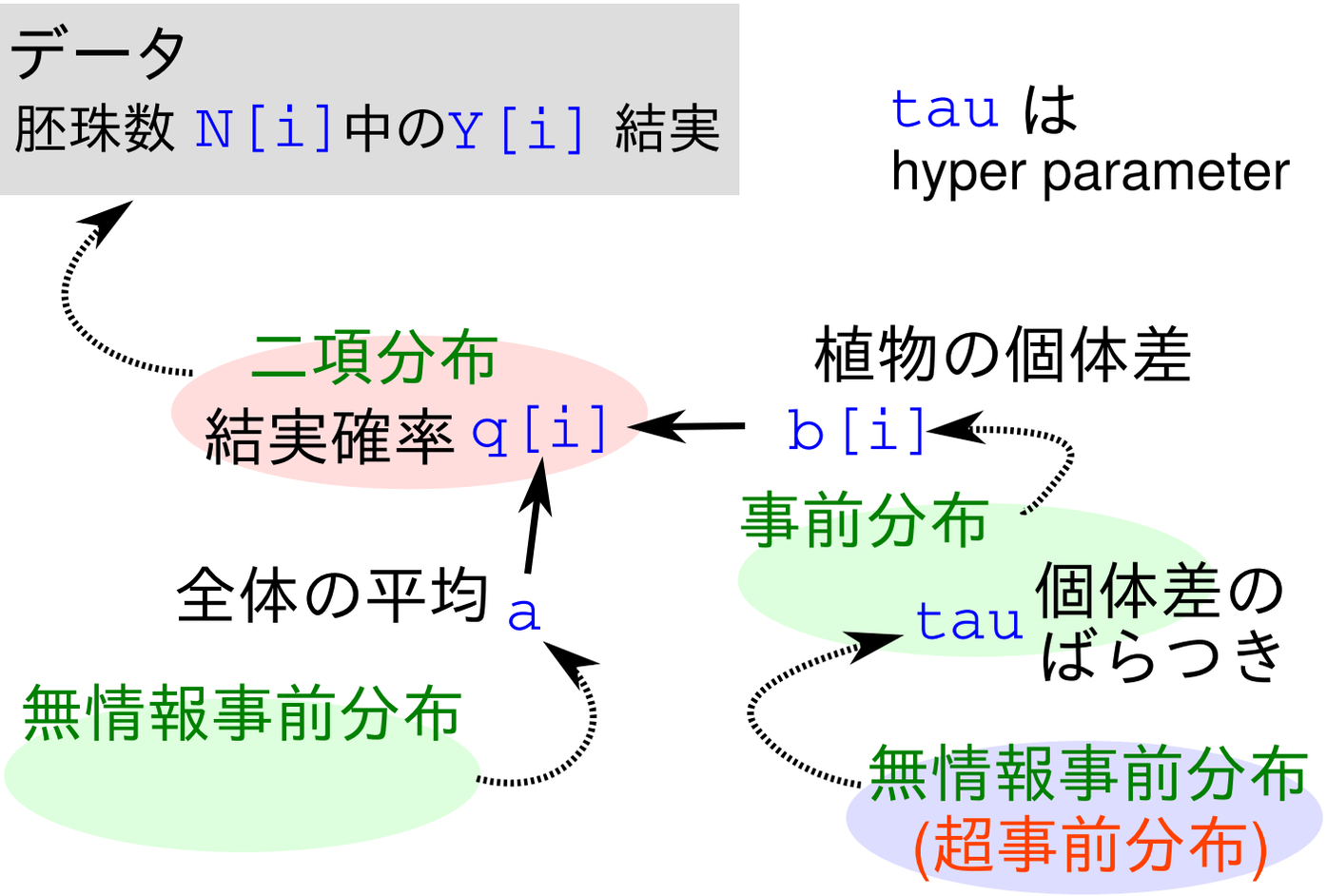


図 10.5: 階層的な事前分布と $y_i \in \{2, 3, 5\}$ の個体の b_i

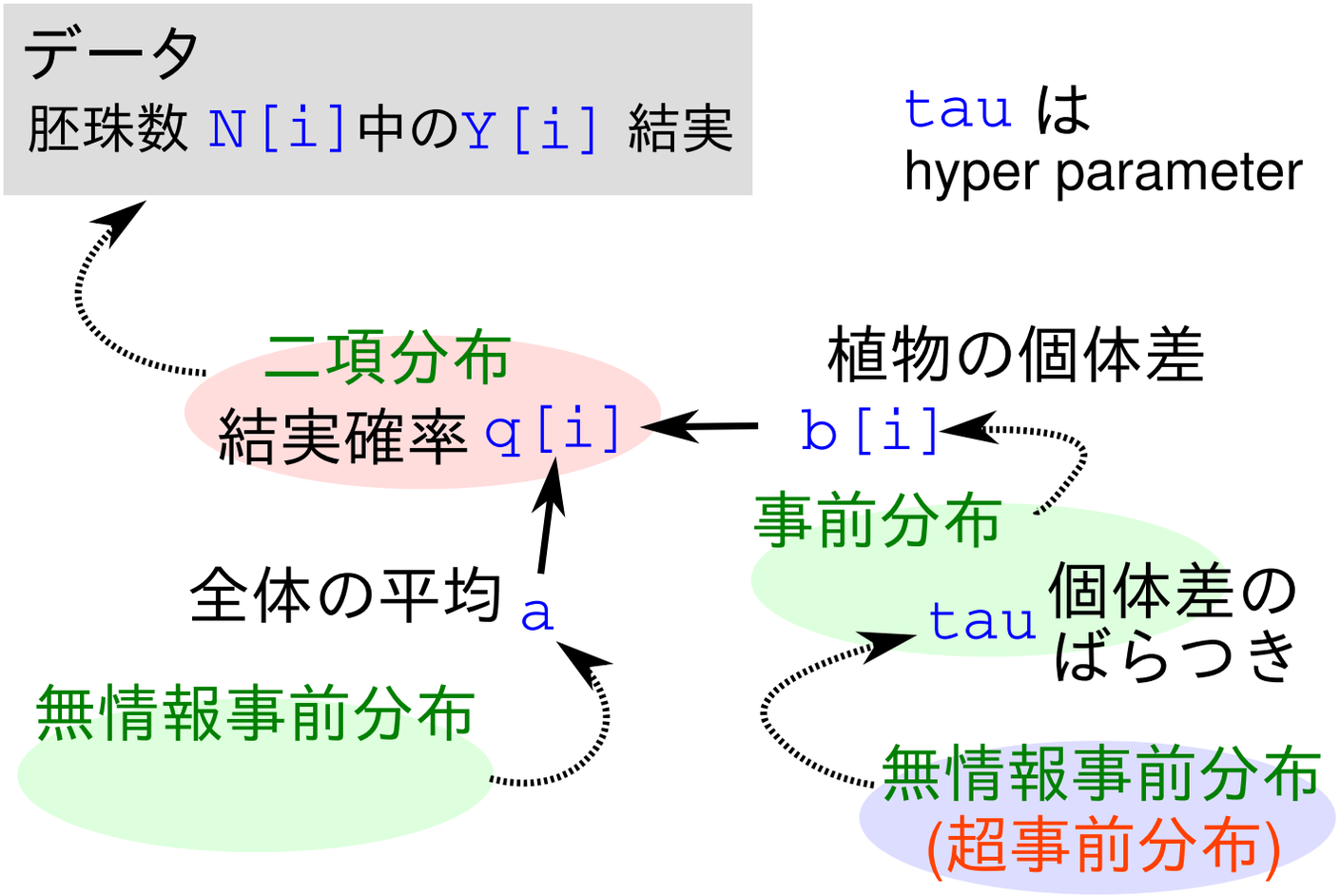


なぜ「階層」ベイズモデルと呼ばれるのか?



超事前分布 $\rightarrow$ 事前分布という階層があるから

全パラメーターを一斉に推定する

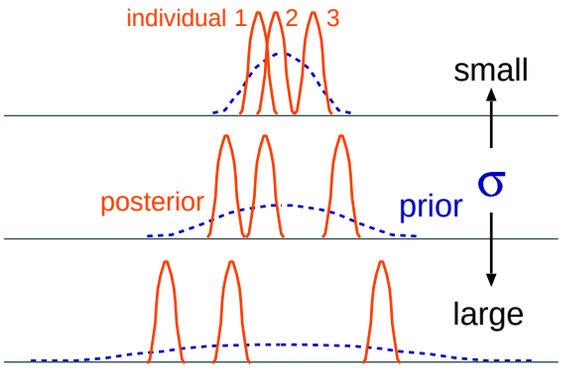


矢印は手順ではなく，依存関係をあらわしている

階層ベイズモデルではないベイズモデルって何でしょう?

個体差 b_i の事前分布の設定を例に検討してみる

- 事前分布を主観的に決める
  「自分は $s = 0.1$ と信じるので，それを使う」
- 以前のデータを使う?
  「これまでの経験から $s = 0.1$」
- 無情報事前分布ばかりにする
  「よくわからないので s をすごく大きくする」



(これらに対して)

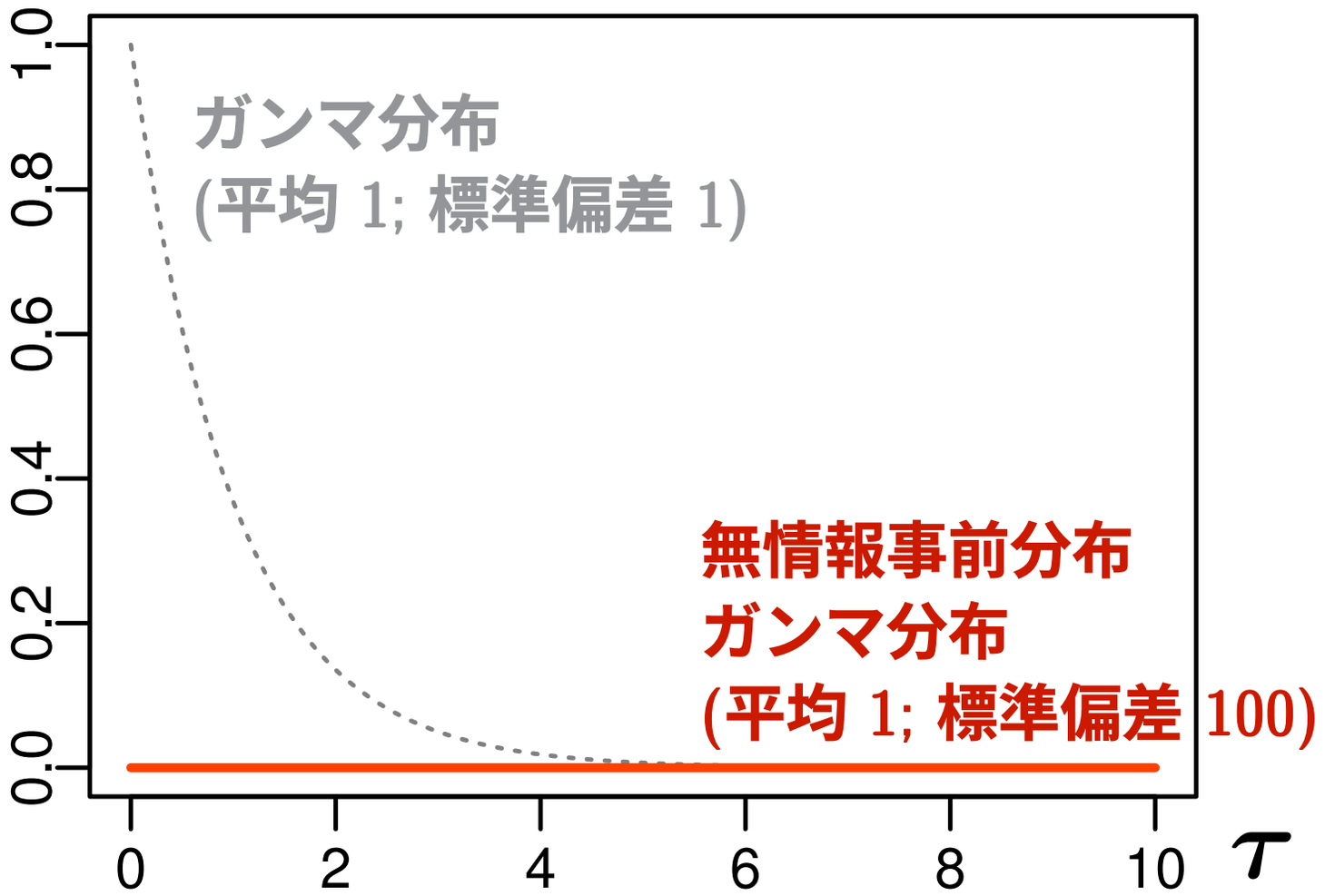
観測データにもとづいて s を決めようとするのが階層ベイズモデル

$\tau = 1/s^2$ の事前分布を無情報事前分布

- s はどのような値をとってもかまわない
- そこで τ の事前分布は **無情報事前分布**(non-informative prior) とする
- たとえば「ひらべったいガンマ分布」

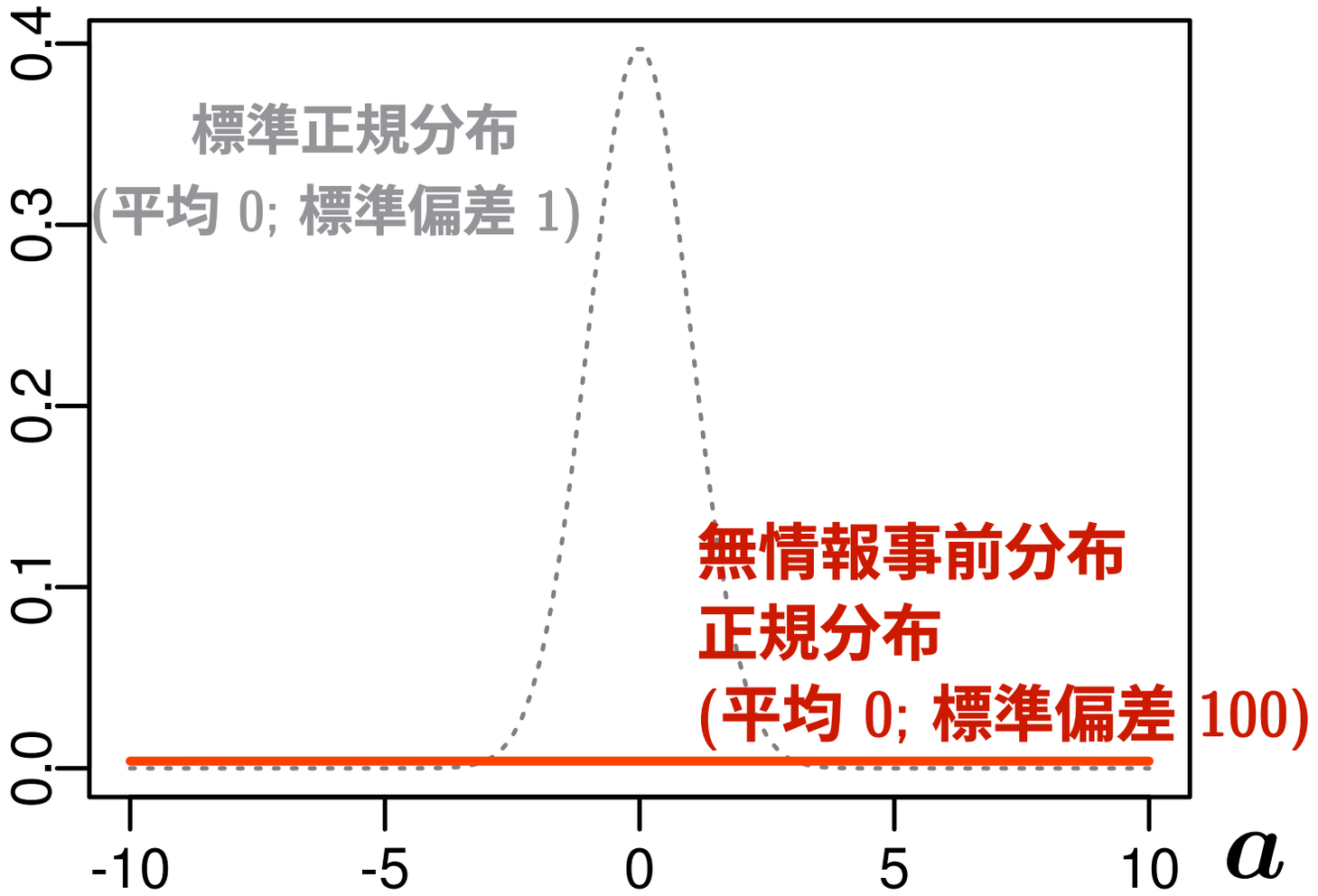
$$p(\tau) = \tau^{\alpha-1} \frac{e^{-\tau\beta}}{\Gamma(\alpha)\beta^{-\alpha}}, \qquad \alpha = \beta = 10^{-4}$$

無情報事前分布 (1) ばらつきパラメーター τ



「τ は正の値であれば何でもよい」と表現している

無情報事前分布 (2) 全個体の平均 a



「生存確率の (logit) 平均 a は何でもよい」と表現している

階層ベイズモデル全体の定式化

$$p(a, \{b_i\}, \tau \mid \text{データ}) = \frac{\prod_{i=1}^{100} p(y_i \mid q(a+b_i))\; p(a)\; p(b_i \mid \tau)\; h(\tau)}{\iint \cdots \int (\text{分子} \uparrow \quad \text{そのまま})\; db_i\; d\tau\; da}$$

分母は何か**定数**になるので

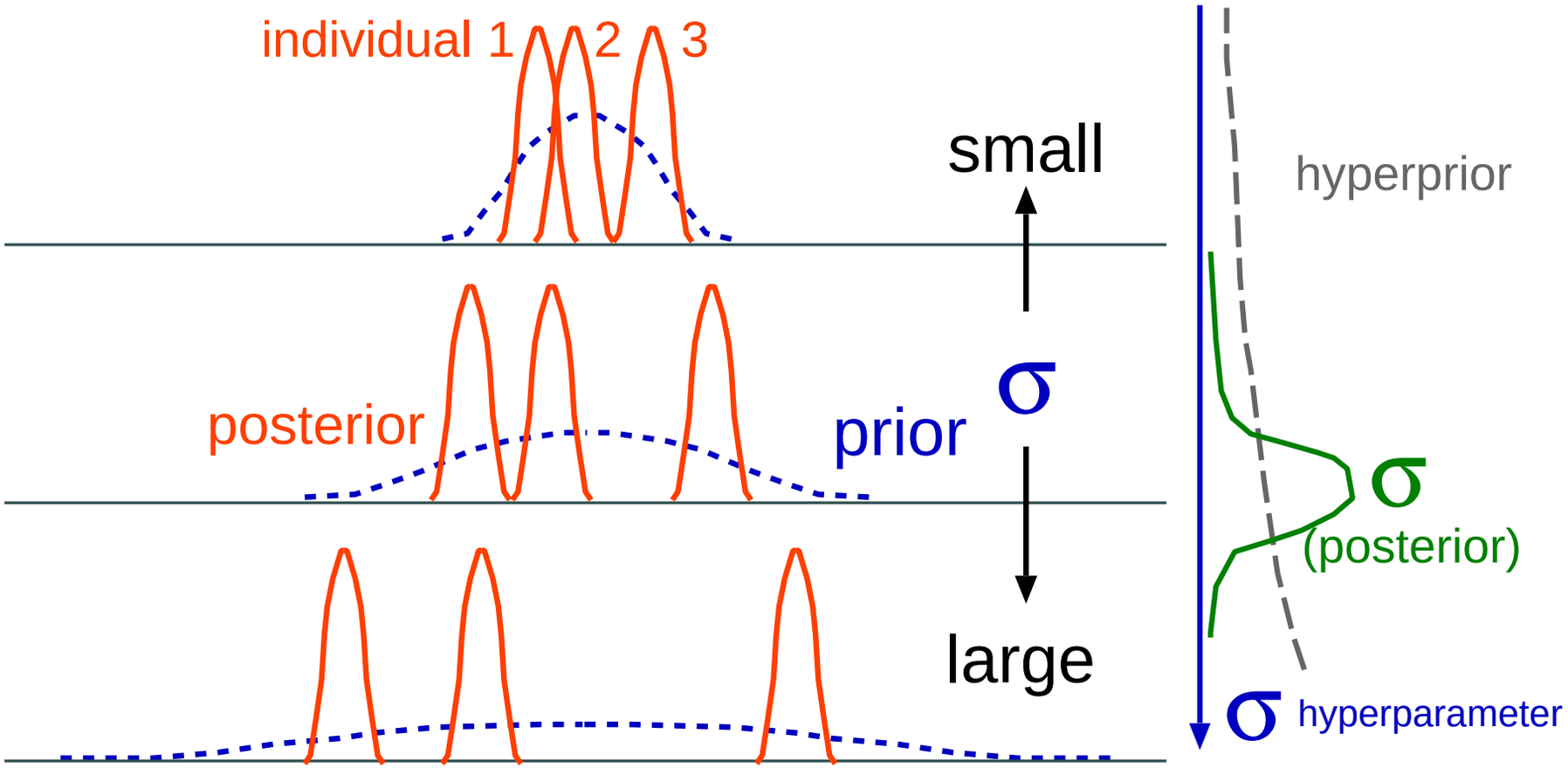
$$p(a, \{b_i\}, \tau \mid \text{データ}) \propto \prod_{i=1}^{100} p(y_i \mid q(a+b_i))\; p(a)\; p(b_i \mid \tau)\; h(\tau)$$

事後分布: $p(a, \{b_i\}, \tau \mid \text{データ})$

尤度: $\prod_{i=1}^{100} p(y_i \mid q(a+b_i))$

事前分布たち: $p(a)\; p(b_i \mid \tau)\; h(\tau)$

個体差 b_i とそのばらつき s の事前分布・事後分布



「ちょうどいいぐあい」の個体差のばらつきになる

あたりを s の事後分布となるようにしたい⇔ MCMC

どうやって事後分布を推定するの?

事後分布

$$p(a, \{b_i\}, \tau \mid \text{データ}) \propto \prod_{i=1}^{100} p(y_i \mid q(a + b_i))\, p(a)\, p(b_i \mid \tau)\, h(\tau)$$

- 観測データと事前分布を組みあわせれば 事後分布$p(a, \{b_i\}, \tau \mid \text{データ})$ を知ることができるはず
- しかし右辺をみてもよくわからない
- Markov chain Monte Carlo (MCMC) を使えば「よくわからない確率分布」から事後分布が得られる!

  $\rightarrow$ ということで，WinBUGS のハナシに……

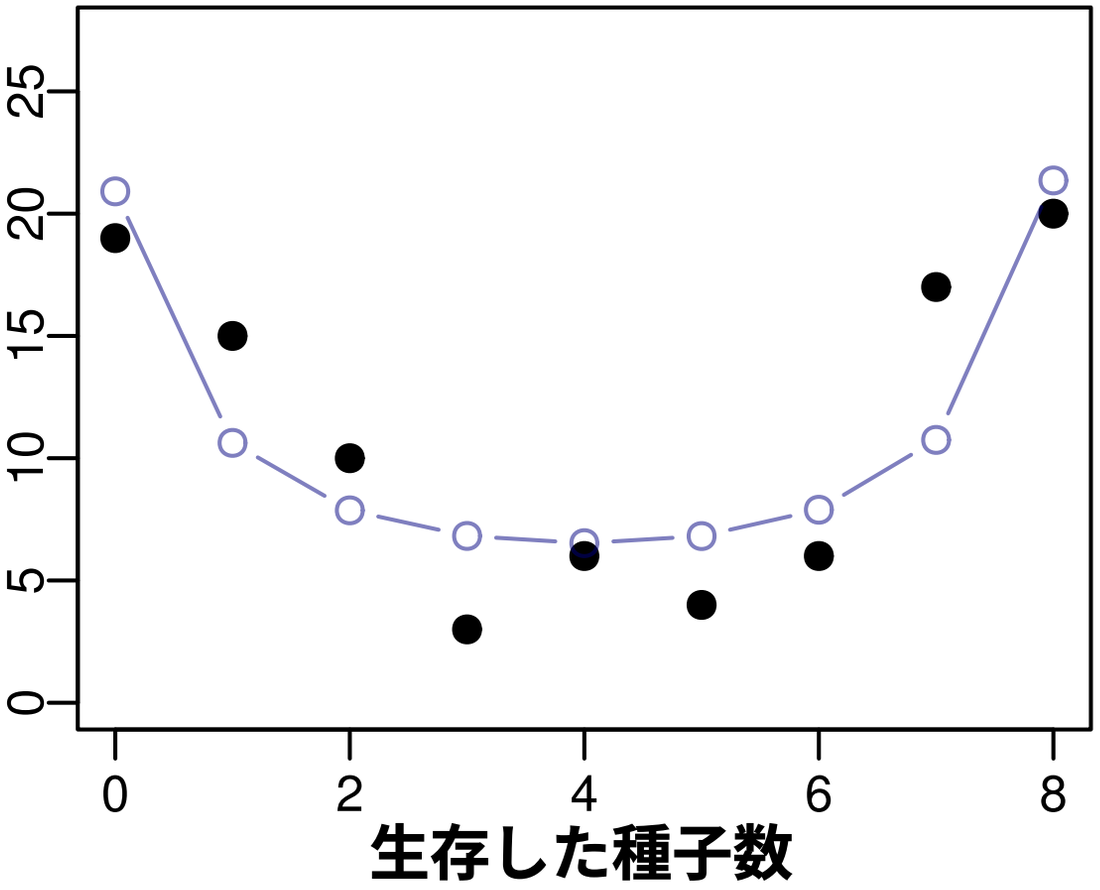
パラメーターの条件つき分布から Gibbs sampling

サンプリングの対象とするパラメーター以外は値を固定する

$$
\begin{aligned}
p(a \mid \cdots) &\propto \prod_{i=1}^{100} p(y_i \mid q(a + b_i))\, p(a) \\
p(\tau \mid \cdots) &\propto \prod_{i=1}^{100} p(b_i \mid \tau)\, h(\tau) \\
p(b_1 \mid \cdots) &\propto p(y_1 \mid q(a + b_1))\, p(b_1 \mid \tau) \\
p(b_2 \mid \cdots) &\propto p(y_2 \mid q(a + b_2))\, p(b_2 \mid \tau) \\
&\vdots \\
p(b_{100} \mid \cdots) &\propto p(y_{100} \mid q(a + b_{100}))\, p(b_{100} \mid \tau)
\end{aligned}
$$

推定された事後分布に基づく予測

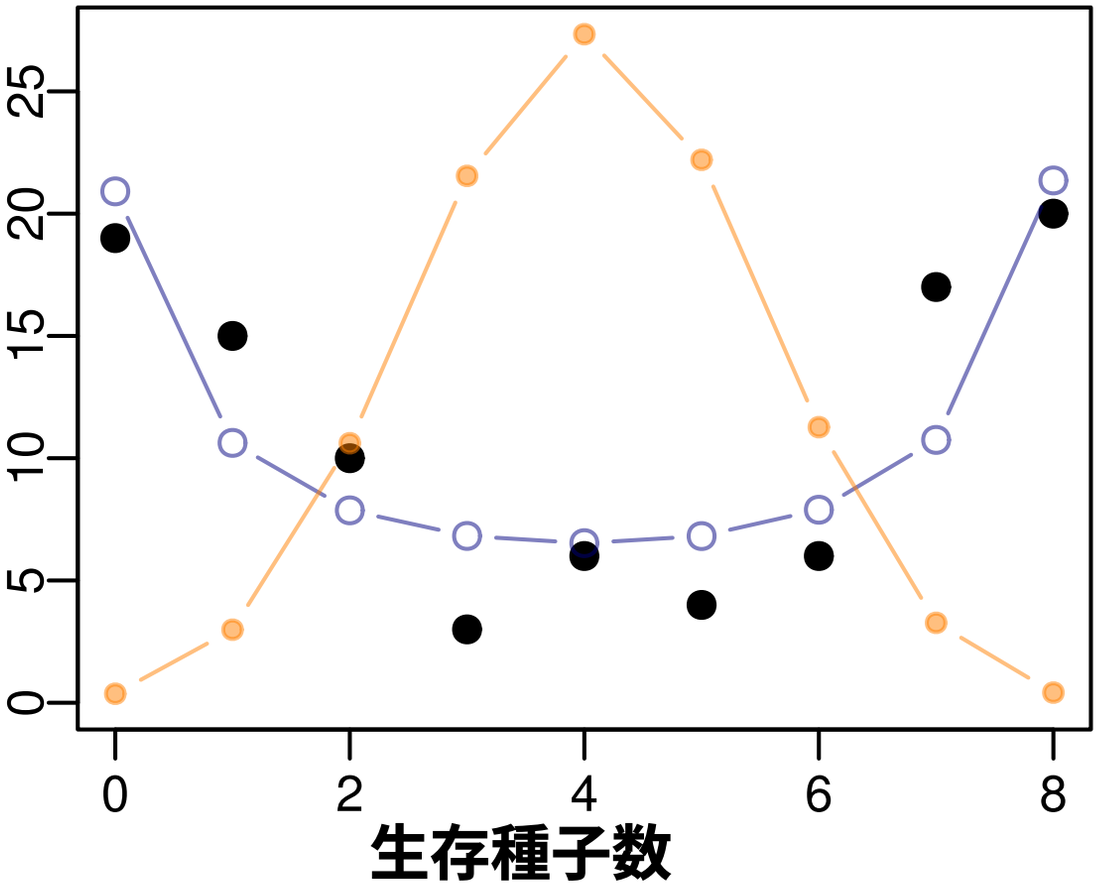
観察された
植物の個体数



「個体差」を考慮することで，
少しはマシな統計モデルが作れた

解決策: 二項分布と正規分布をまぜる

観察された
植物の個体数

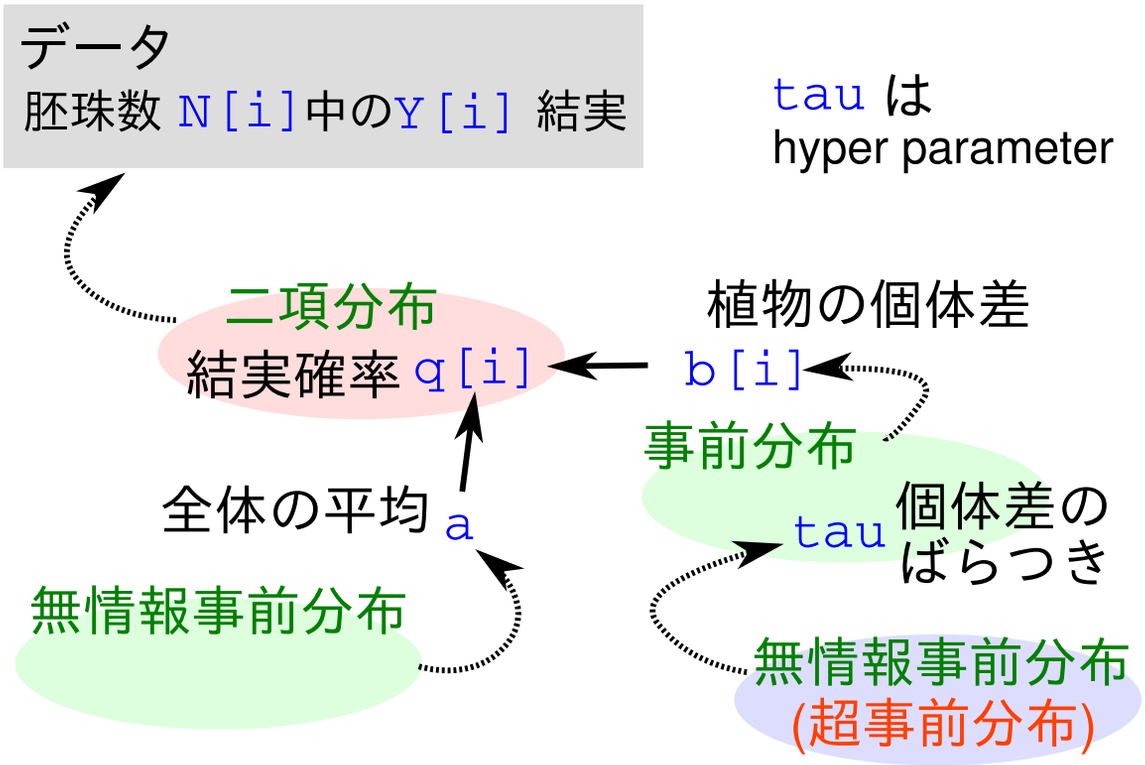


複雑な確率分布を新しく導入するのではなく
二項分布と正規分布をまぜることで現象を表現した

ここまでの用語の整理

- **階層ベイズモデル**

  (事後分布) $\propto$ (尤度) $\times$ (事前分布) $\times$ (超事前分布)



- **事後分布の推定計算方法**: **Markov Chain Monte Carlo (MCMC) 法**

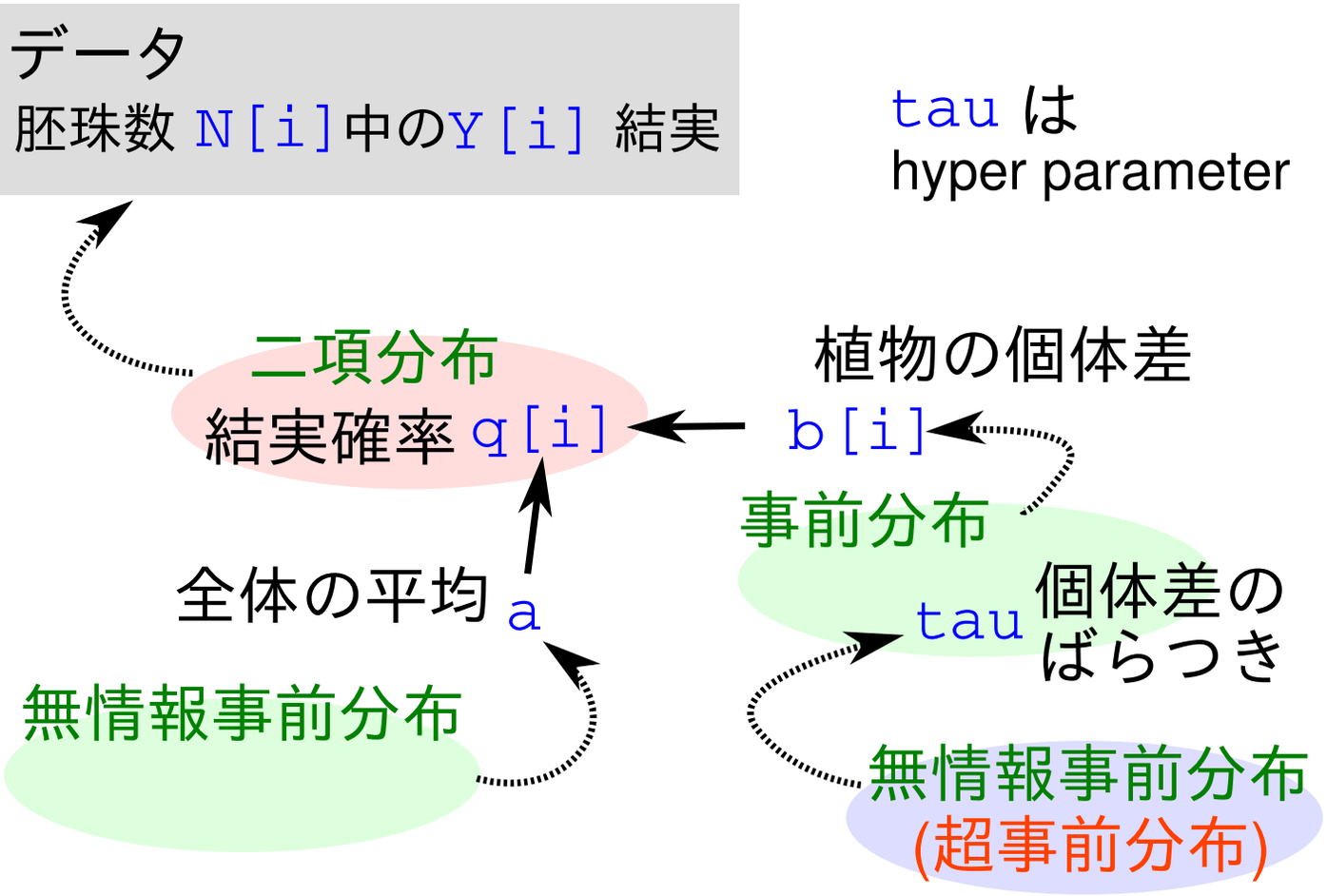
「生存確率の推定」例題を

WinBUGS で推定

「生存確率の推定」例題を WinBUGS に推定させる手順

1. 生存確率の階層ベイズモデルの構築する
2. それを BUGS 言語でかく (model.bug.txt)
3. R2WBwrapper 関数を使って R コードを書く (runbugs.R)
4. R 上で runbugs.R を実行 (source(runbugs.R) など)
5. 出力された結果が bugs オブジェクトで返される

生存確率の階層ベイズモデルってどんなでしたっけ?



$$p(a, \{b_i\}, \tau \mid \text{データ}) \propto \prod_{i=1}^{100} p(\text{データ} \mid q(a+b_i))\; p(a)\; p(b_i \mid \tau)\; h(\tau)$$

事前分布の設定方法

- 階層的な (hierarchical) 事前分布にする
  - random effects 的な個体差・場所差
- 無情報 (non-informative) 事前分布にする
  - 切片や説明変数の係数など fixed effects 的なパラメーター
- 主観的な (subjective) 事前分布にする
  - あまりおすすめできない
  - (反復測定していないときの) 測定時のエラーとか

生存確率の階層ベイズモデルを BUGS 言語で

ファイル model.bug.txt の内容 (一部簡略化)

```
model{
    for (i in 1:N.sample) {
        Y[i] ~ dbin(q[i], N[i])  # 観測値との対応
        logit(q[i]) <- a + b[i]  # 生存確率 q[i]
    }
    a ~ dnorm(0, 1.0E-4)         # 個体の平均
    for (i in 1:N.sample) {
        b[i] ~ dnorm(0, tau)     # 個体差
    }
    tau ~ dgamma(1.0E-4, 1.0E-4) # 個体差のばらつき
    sigma <- sqrt(1 / tau)       # tau から SD に変換
}
```

BUGS 言語について，いくつか

- BUGS 言語は普通の意味でのプログラミング言語ではない
  - 「式」を列挙しているだけ，と考える
  - 「式」の並び順を変えても計算結果は (ほぼ) 変わらない
- 各パラメーターは二種類の node それぞれで一度ずつ定義できる (二度以上は定義できない)
  1. ~ sthochastic node
  2. <- deterministic node

R2WBwrapper な R コード runbugs.R (前半部)

観測データの設定

```
source("R2WBwrapper.R")    # R2WBwrapper よみこみ
d <- read.csv("data.csv") # 観測データよみこみ

clear.data.param() # いろいろ初期化（まじない）
set.data("N.sample", nrow(d)) # データ数
set.data("N", d$N)                # 調査種子数
set.data("Y", d$Y)                # 生存
```

R2WBwrapper な R コード runbugs.R (後半部)

パラメーターの初期値の設定など

```
set.param("a", 0)        # 個体の平均
set.param("sigma", NA) # 個体差のばらつき
set.param("b", rep(0, N.sample))  # 個体差
set.param("tau", 1, save = FALSE) # ばらつきの逆数
set.param("p", NA)       # 生存確率

post.bugs <- call.bugs(      # WinBUGS よびだし
    file = "model.bug.txt",
    n.iter = 2000, n.burnin = 1000, n.thin = 5
)
```

WinBUGS に指示した事後分布のサンプリング

```
post.bugs <- call.bugs(      # WinBUGS よびだし
    file = "model.bug.txt",
    n.iter = 2000, n.burnin = 1000, n.thin = 5
)
```

- じつは default では独立に (並列に) 3 回(`n.chains = 3`) MCMC sampling せよと指定されている (収束性をチェックするため)
  - cf. 伊庭さんのたくさんの PC で MCMC する話
- ひとつの chain の長さは 2000 step (`n.iter = 2000`)
- 最初の 1000 step は捨てる(`n.burnin = 1000`)
- 1001 から 2000 step まで 5 step おきに値を記録する (`n.thin = 5`)

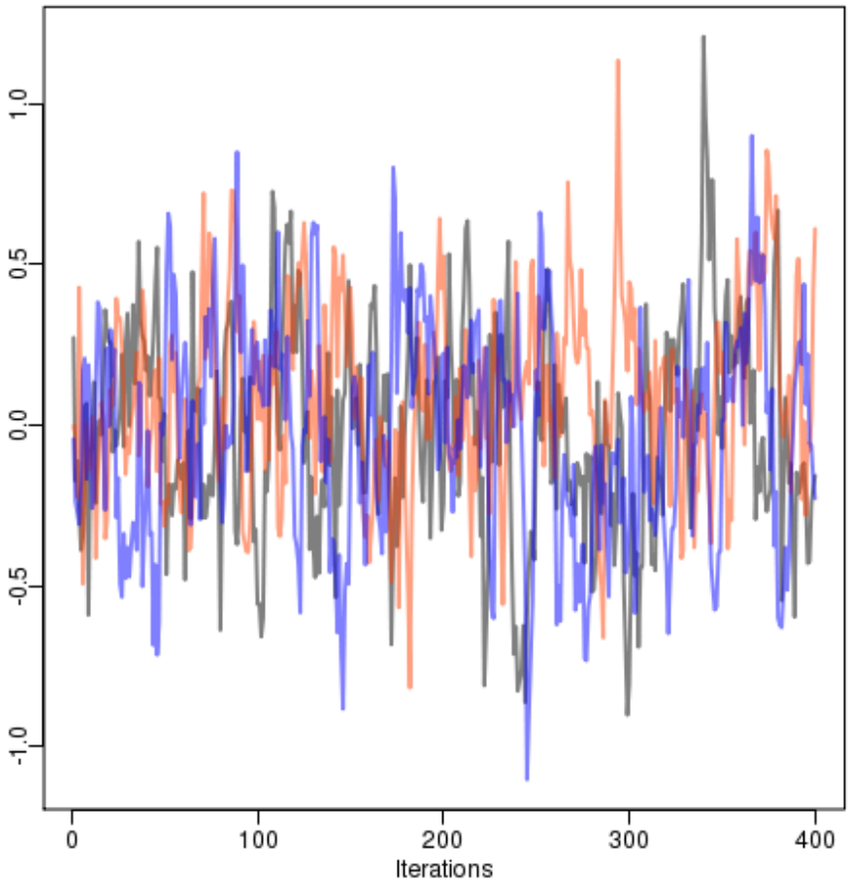
このあたりの設定はデータ・統計モデルによって変わる

で，実際に動かすには?

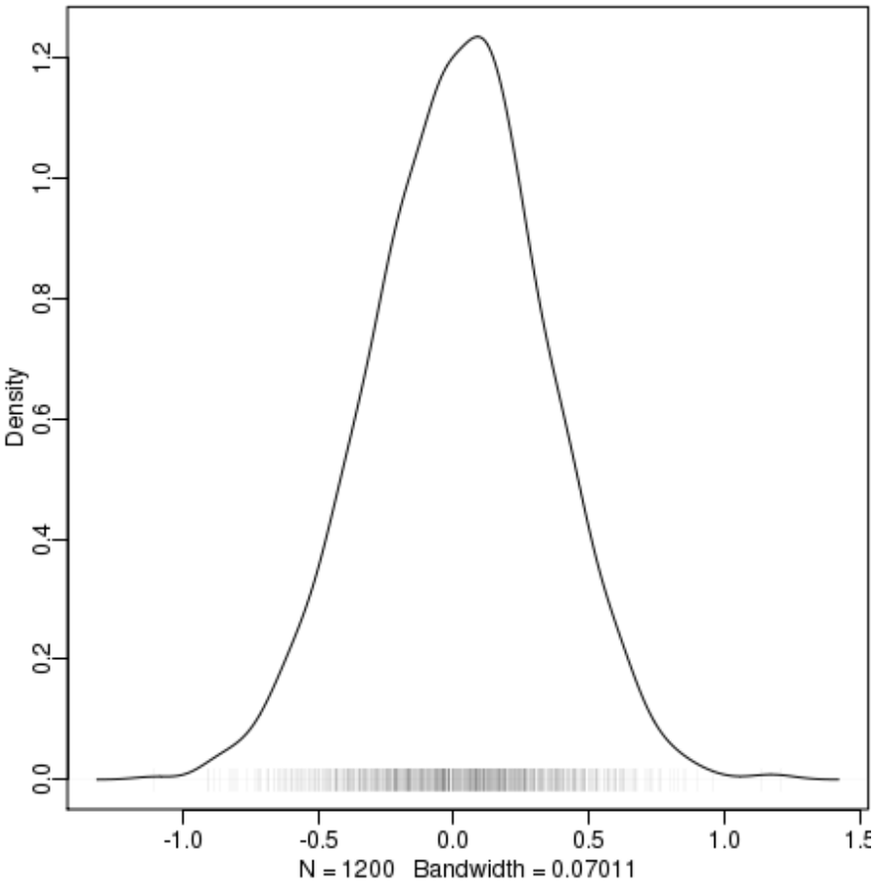
- たとえば，R 上で `source("runbugs.R")` とか
- すると WinBUGS が起動して MCMC sampling をはじめる
- この例題は簡単なのですぐに計算が終了する (WinBUGS 内で図などが表示される)
- 手動で WinBUGS を終了する
- すると WinBUGS が得た結果が R にわたされ，`post.bugs` というオブジェクトにそれが格納される

事後分布のサンプルを R で調べる

a のサンプリングの様子



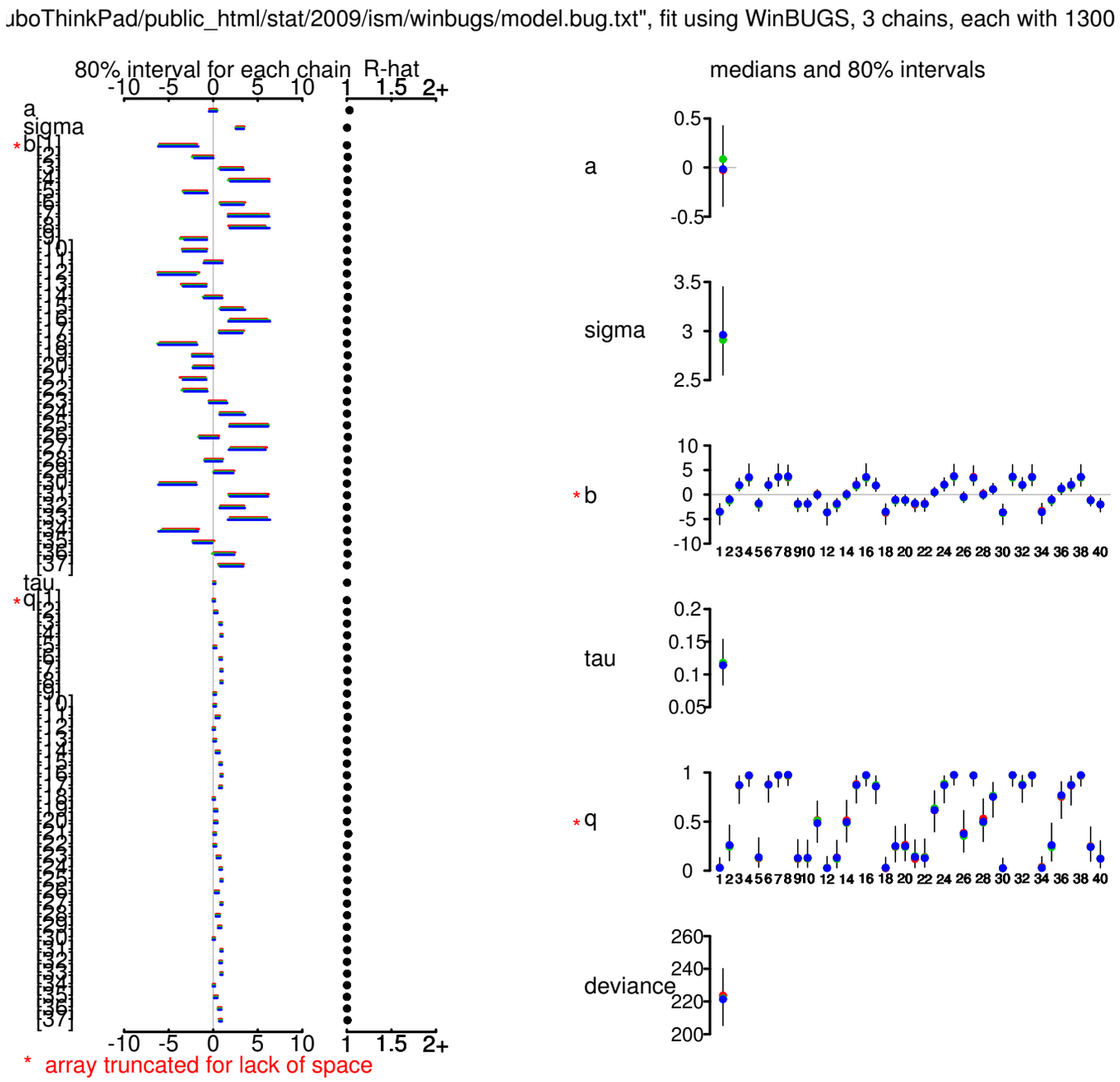
a の事後確率密度の推定



収束?

bugs オブジェクトの post.bugs を調べる (1)

- plot(post.bugs)　　→ 次のページ, 実演表示
- R-hat は Gelman-Rubin の収束判定用の指数
  - $\hat{R} = \sqrt{\dfrac{\hat{\mathrm{var}}^{+}(\psi|y)}{W}}$
  - $\hat{\mathrm{var}}^{+}(\psi|y) = \dfrac{n-1}{n}W + \dfrac{1}{n}B$
  - W : chain 内の variance
  - B : chain 間の variance
  - Gelman et al. 2004. Bayesian Data Analysis. Chapman & Hall/CRC


ubo ThinkPad/public_html/stat/2009/ism/winbugs/model.bug.txt", fit using WinBUGS, 3 chains, each with 1300
80% interval for each chain
R-hat
medians and 80% intervals
a
sigma
*b
tau
*q
deviance
-10 -5 0 5 10
1 1.5 2+
0.5
0
-0.5
3.5
3
2.5
10
5
0
-5
-10
0.2
0.15
0.1
0.05
1
0.5
260
240
220
200
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
* array truncated for lack of space

bugs オブジェクトの post.bugs を調べる (2)

- `print(post.bugs, digits.summary = 3)`
- **事後分布の 95% 信頼区間などが表示される**

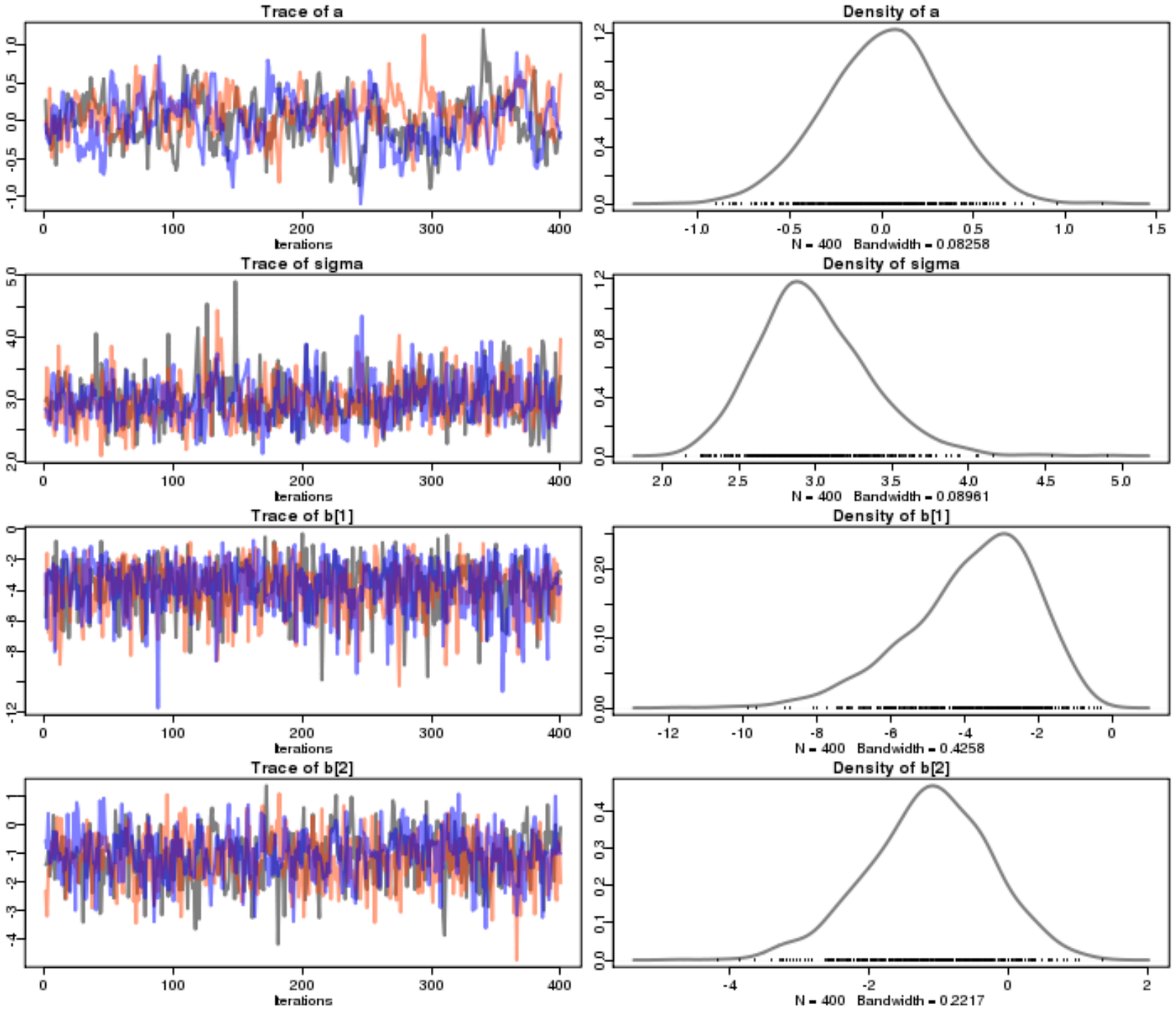
	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
a	0.018	0.322	-0.621	-0.202	0.025	0.233	0.628	1.030	75
sigma	2.980	0.361	2.346	2.738	2.948	3.205	3.752	1.003	590
b[1]	-3.800	1.711	-7.652	-4.776	-3.503	-2.554	-1.193	1.002	1100
b[2]	-1.142	0.874	-3.003	-1.688	-1.111	-0.530	0.464	1.010	200
b[3]	1.992	1.047	0.169	1.251	1.889	2.665	4.346	1.005	390
b[4]	3.745	1.781	0.975	2.503	3.408	4.751	7.926	1.008	520
b[5]	-2.005	1.066	-4.257	-2.719	-1.909	-1.257	-0.131	1.005	370
b[6]	2.047	1.077	0.147	1.310	1.933	2.716	4.456	1.002	1100
b[7]	3.765	1.763	1.023	2.482	3.593	4.811	7.515	1.000	1200
b[8]	3.782	1.661	1.133	2.591	3.570	4.703	7.621	1.003	640
b[9]	-2.049	1.106	-4.439	-2.745	-1.948	-1.255	-0.218	1.004	470
b[10]	-2.028	1.066	-4.340	-2.655	-1.902	-1.314	-0.175	1.002	750

...

mcmc.list クラスに変換して作図

- `post.list <- to.list(post.bugs)`
- `plot(post.list[,1:4,], smooth = F)`

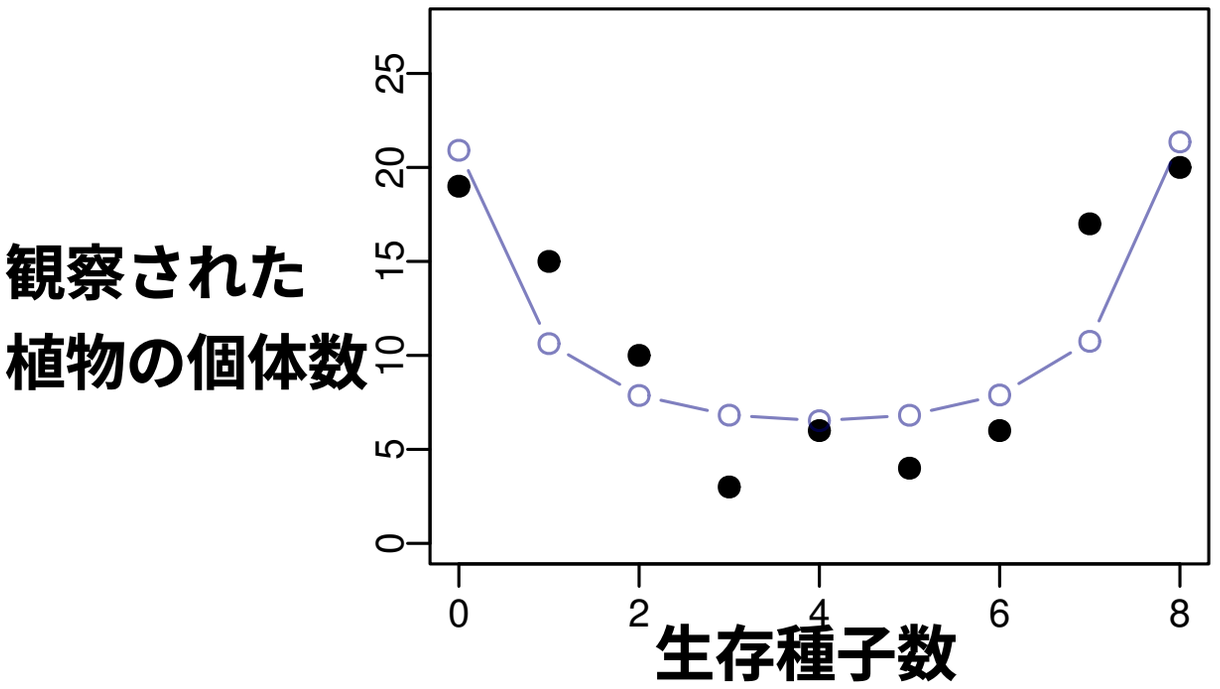
→ 次のページ, 実演表示


Trace of a
Density of a
N = 400 Bandwidth = 0.08258
Trace of sigma
Density of sigma
N = 400 Bandwidth = 0.08961
Trace of b[1]
Density of b[1]
N = 400 Bandwidth = 0.4258
Trace of b[2]
Density of b[2]
N = 400 Bandwidth = 0.2217
Iterations

mcmc クラスに変換して作図

- `post.mcmc <- to.mcmc(post.bugs)`
- これは `matrix` と同じようにあつかえるので，作図に便利

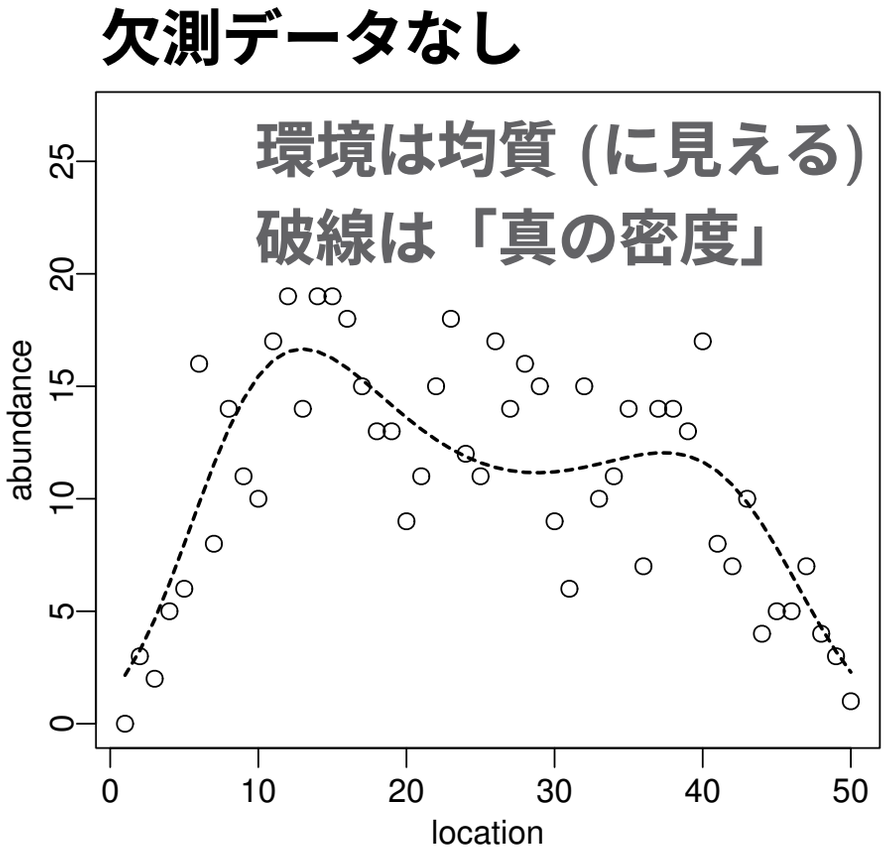
例: 推定された事後分布に基づく予測



2. 空間構造のある階層ベイズモデル

空間的自己相関をくみこむ

架空の例題: 個体数データ，一次元空間データ

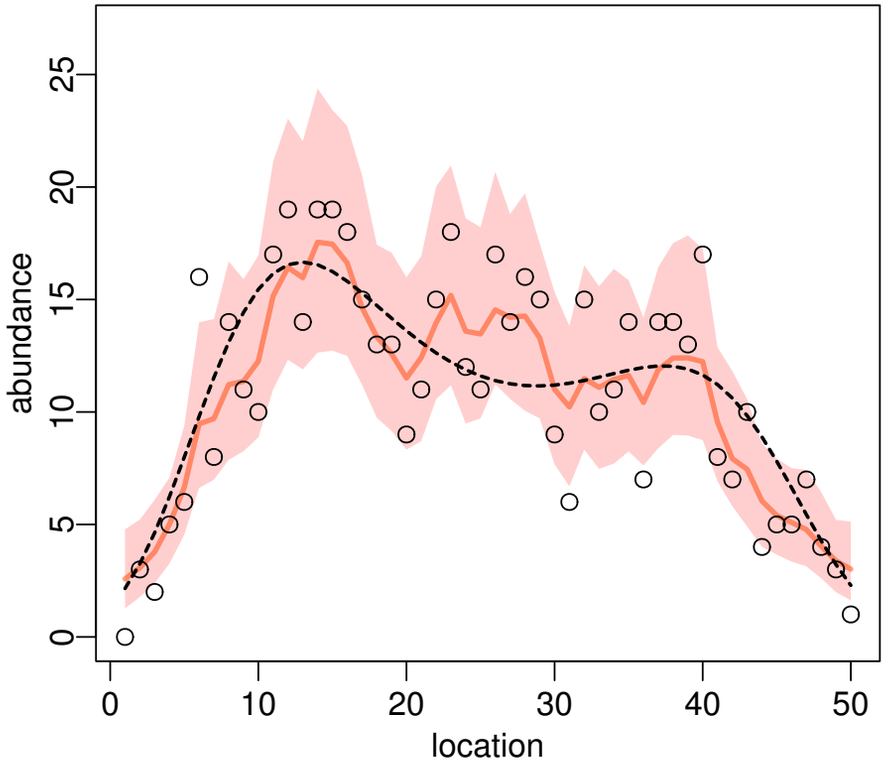


問: 空間自己相関を考慮して生物個体の密度推定

解析の目的: まずはこんな推定をしてみたい

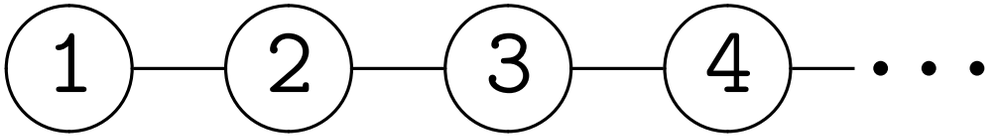
空間相関を考慮するモデル

欠測データなし



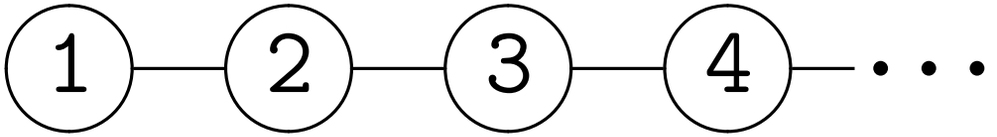
(彩色された領域は平均値の事後分布の 95% 区間，曲線は中央値)

空間相関のある「場所差」階層ベイズモデル



- 地点 i の観測個体数は平均 λ_i のポアソン分布にしたがう:
  $y_i \sim \text{Poisson}(\lambda_i)$
- 平均 λ_i の対数は (全体の平均) + (場所差) と分割する:
  $\log \lambda_i = \beta + r_i$
- ベイズモデルとしてあつかいたいので，推定したいパラメーターの**事前分布** を決めてやらなければならない
  - **事前分布** についてはあとで説明
- 全体の平均 β は無情報事前分布にしたがう:
  $\beta \sim \text{Normal}(0, 10^2)$,

空間相関のある「場所差」階層ベイズモデル (続)



- Conditional Autoregressive (CAR) モデルにおける場所差 r_i の**条件つき事前分布**(N_i は i の近傍場所数, J_i は i の近傍場所):

$$r_i \sim \mathrm{Normal}\left(\frac{\sum_{j \in \mathrm{J}_i} r_j}{N_i}, \frac{\sigma}{N_i}\right)$$

σについては次の次のスライドで

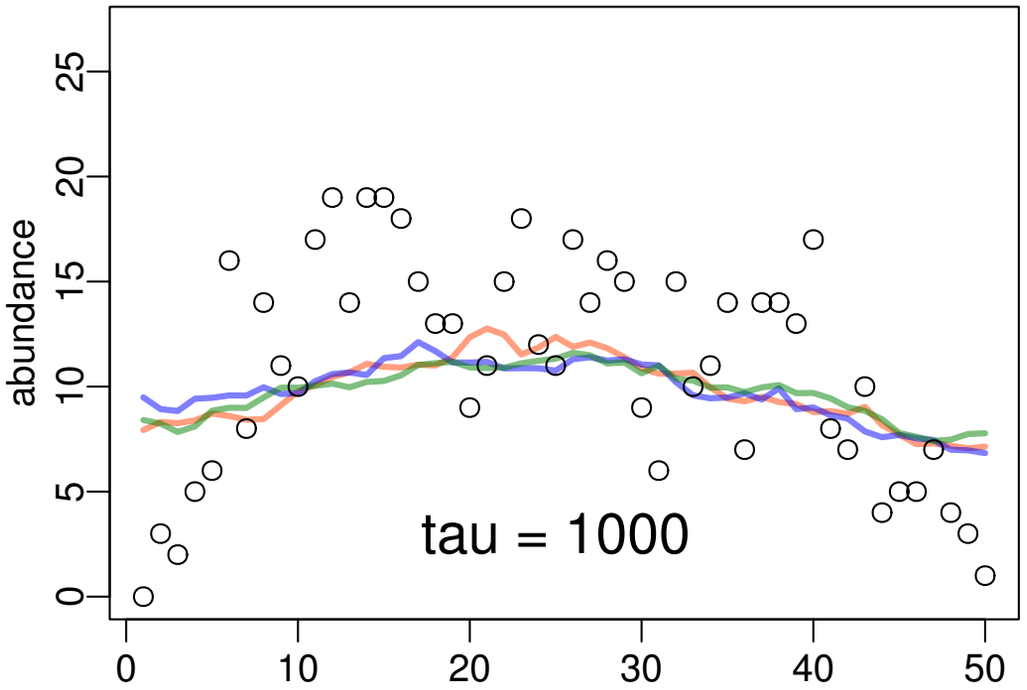
- σ は無情報事前分布にしたがう:
  $\tau = 1/\sigma$ $\mathrm{Gamma}(1.0^{-2}, 1.0^{-2})$
- ベイズの定理 $\rightarrow$ **事後分布**の導出

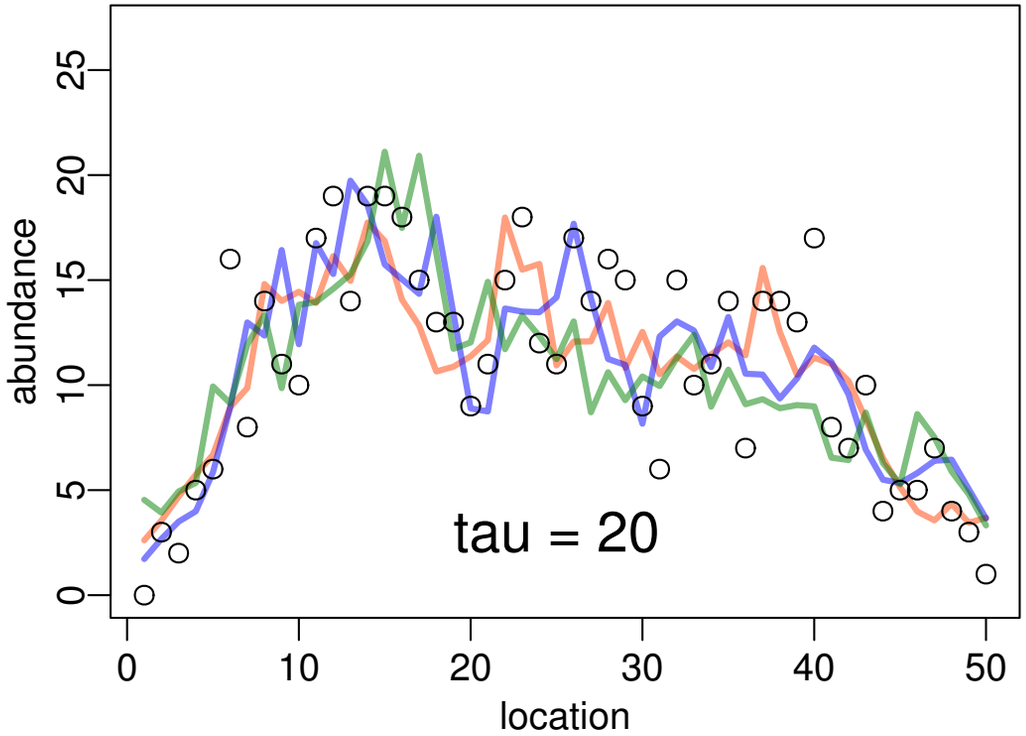
$$p(\beta, \{r_i\}, \tau \mid \{y_i\}) = \frac{p(\{y_i\} \mid \beta, \{r_i\}, \tau) \times (\text{事前分布あれこれ}}{\int\int \cdots \int (\uparrow \text{分子}) d\beta dr_1 \cdots dr_{50} d\tau}$$

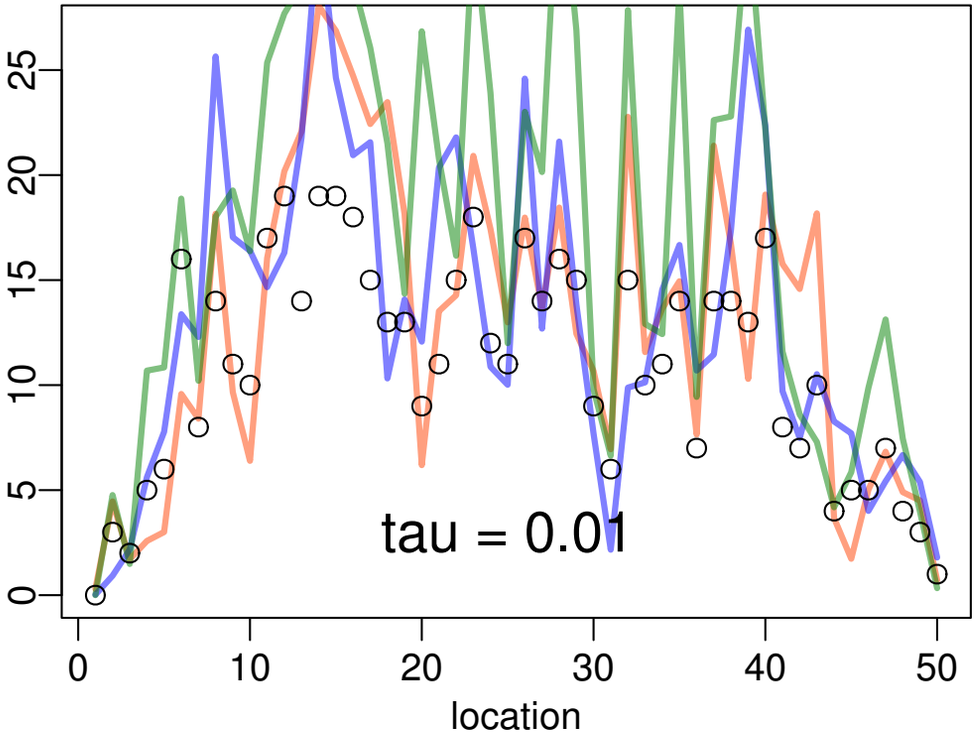
超パラメーター τ が決める

隣との類似度

- τ が大 (σ が小) だと隣と似ている
- τ が小 (σ が大) だと隣と似てない
- ベイズ推定によって適切な τ の範囲 (**事後分布**) が得られる

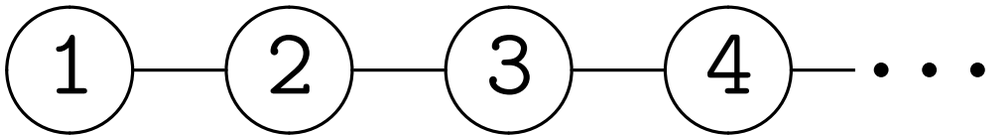






この例題の BUGS コード

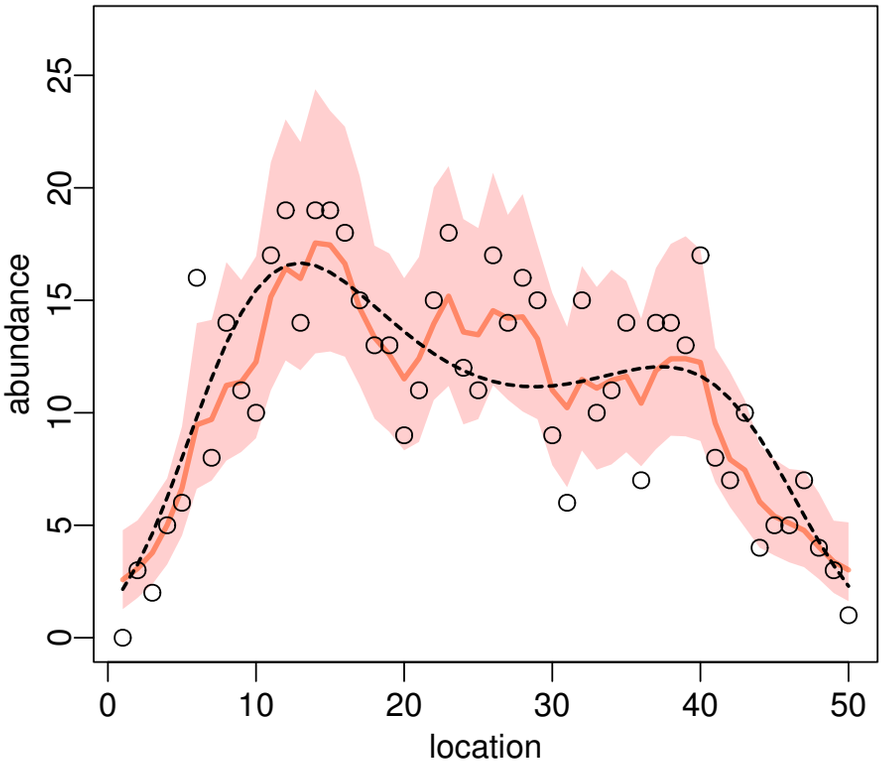
```
model { # BUGS コードで定義された階層ベイズモデルの例
  for (i in 1:N.site) {
    Y[i] ~ dpois(mean[i])           # 観測データと密度の関係
    log(mean[i]) <- beta + re[i]    # (全体の平均) + (場所差)
  }
  # 場所差 re[i] を CAR model で生成
  re[1:N.site] ~ car.normal(Adj[], Weights[], Num[], tau)
  beta ~ dnorm(0, 1.0E-2)           # 全体の平均は無情報事前分布
  tau ~ dgamma(1.0E-2, 1.0E-2) # 場所差のばらつきは無情報事前分布
}
```

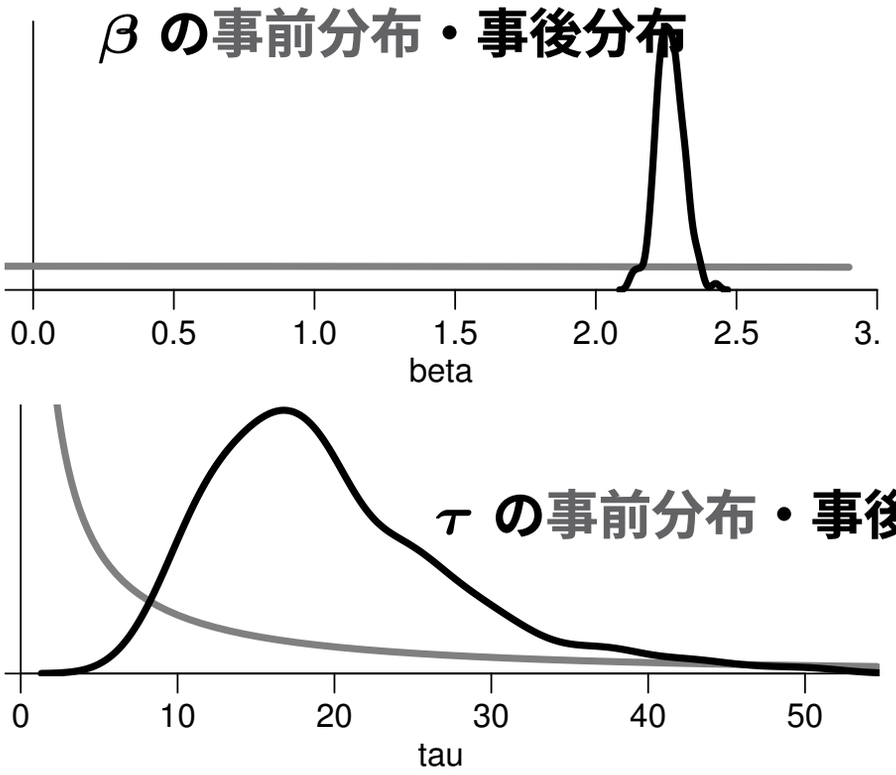


空間相関のある「場所差」モデルの推定結果

空間相関を考慮するモデル

欠測データなし



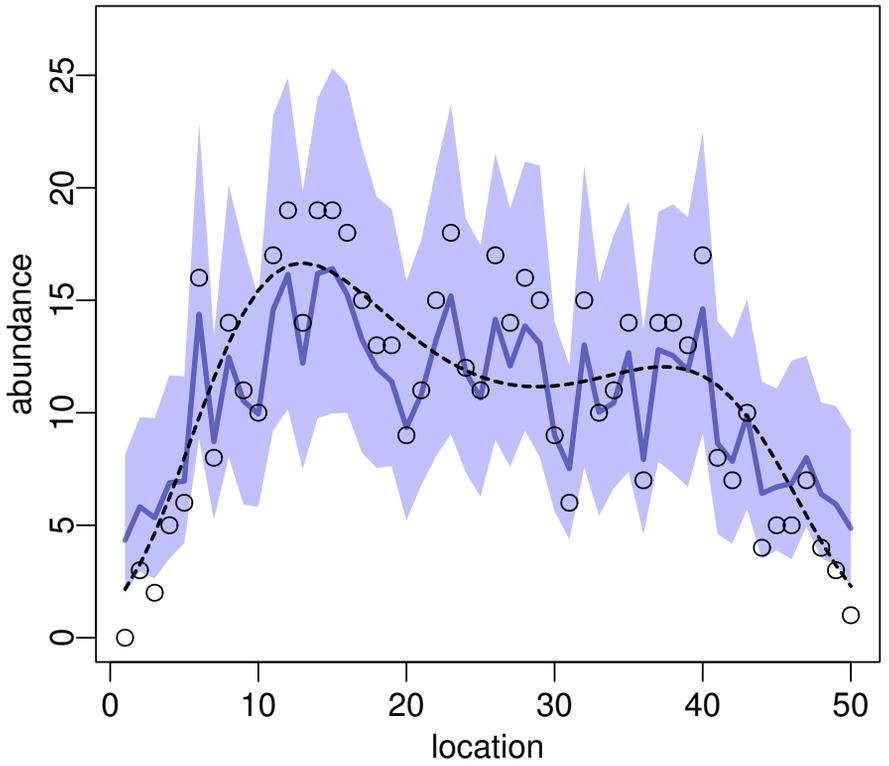


(彩色された領域は平均値の事後分布の 95% 区間，曲線は中央値)

空間相関を考慮しないベイズモデルの推定結果

空間相関を考慮しないモデル

欠測データなし

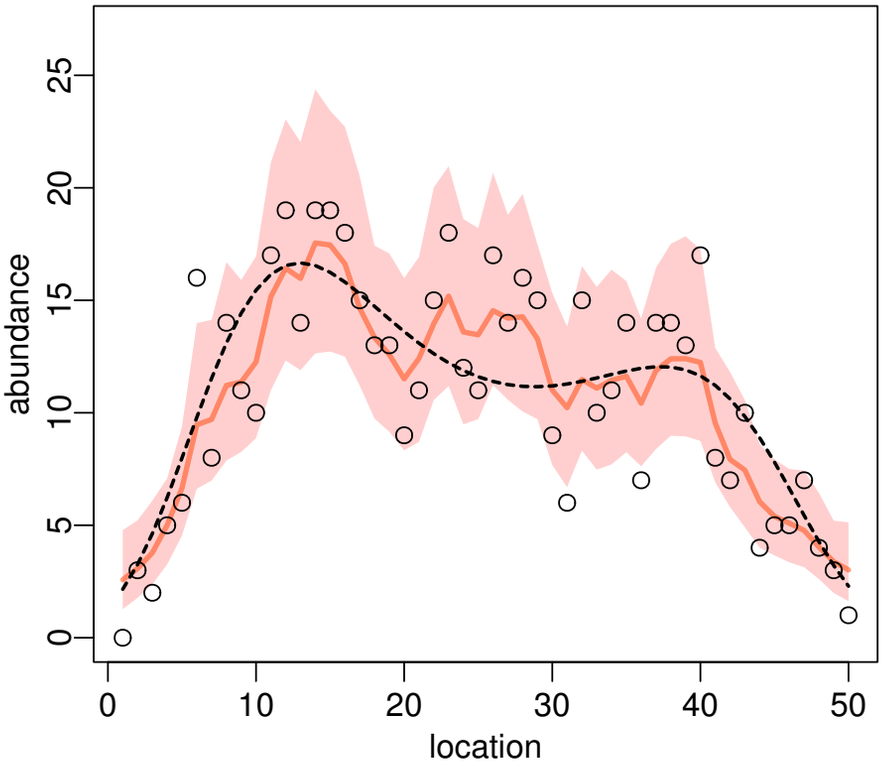


空間相関とか考えない GLMM 的なモデルでも OK?

空間相関を考慮する vs しないモデル

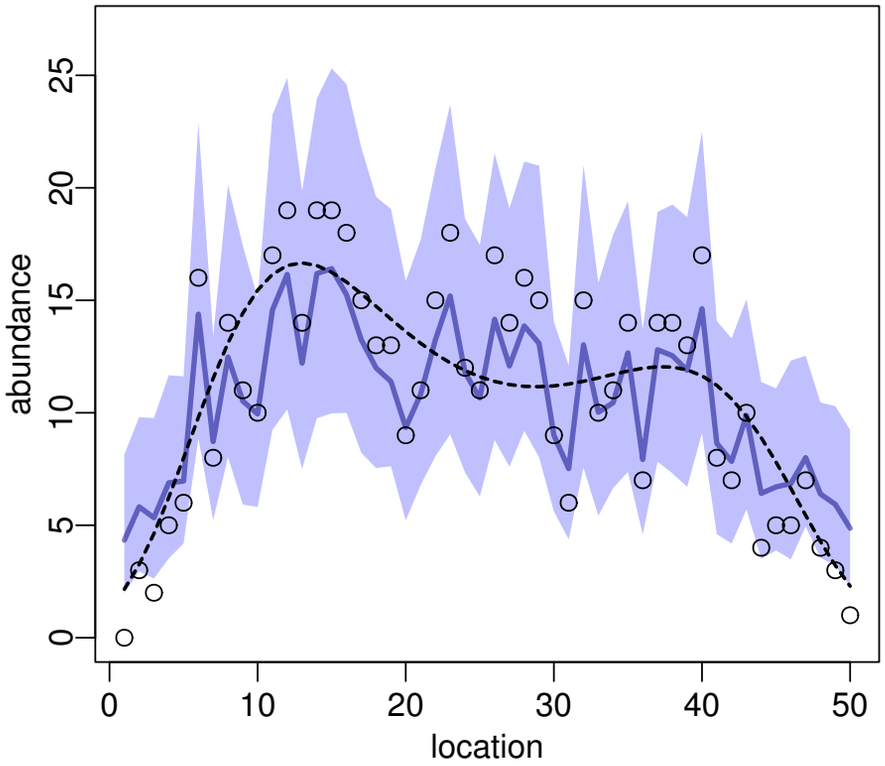
空間相関を考慮するモデル

欠測データなし



空間相関を考慮しないモデル

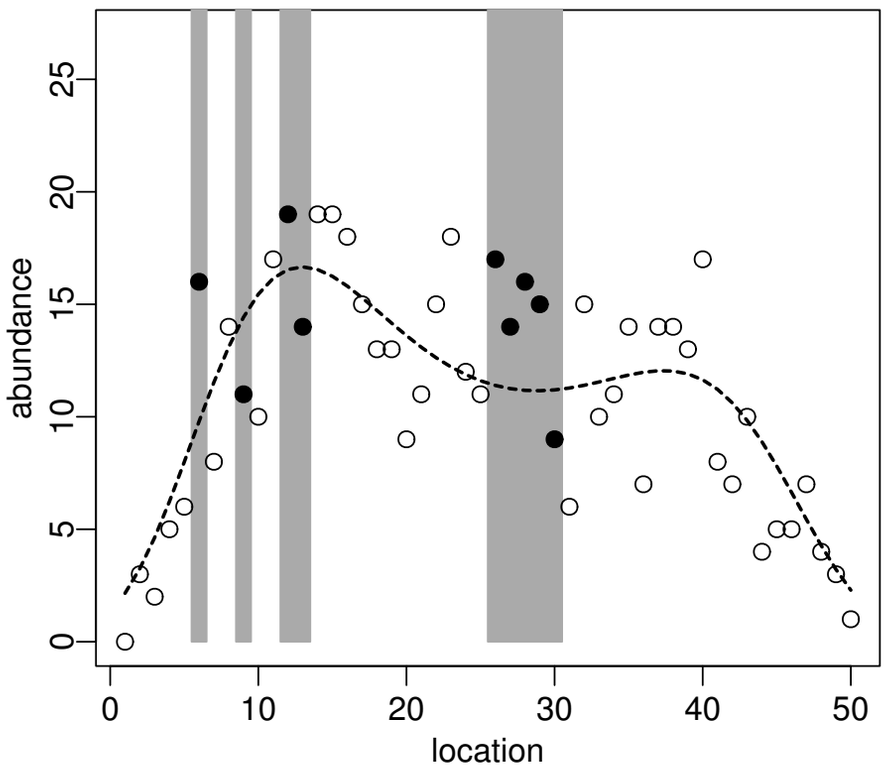
欠測データなし



空間相関を考慮する必要があるのだろうか?

架空の例題 (続): 欠測がある場合は?!

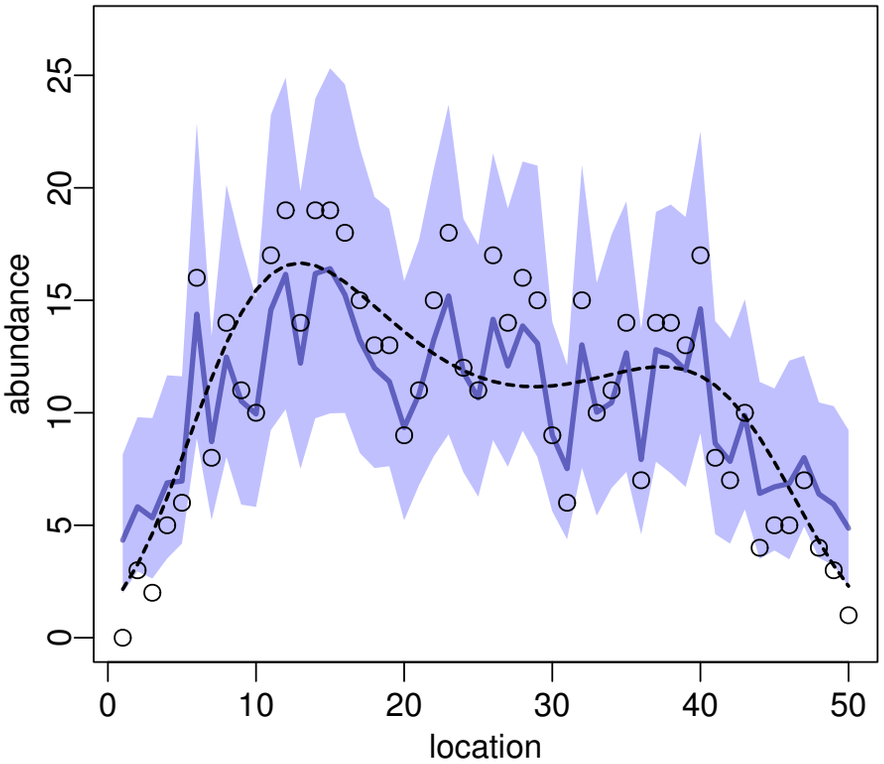
欠測あり　欠測値の予測!



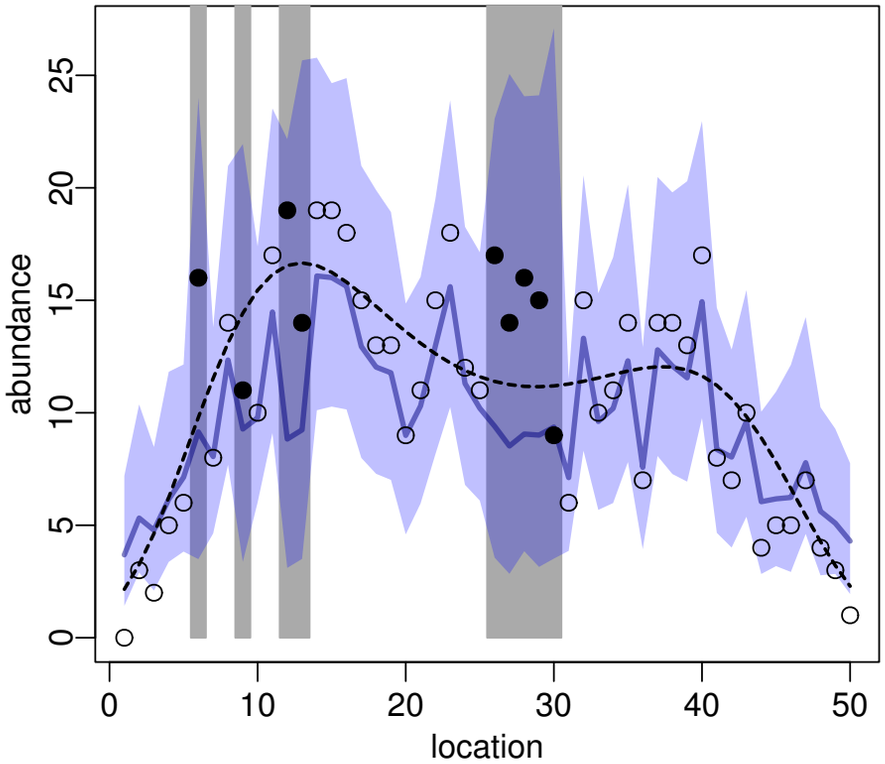
灰色の領域で観測できなかった (●は観測できなかった点

空間相関を考慮しないベイズモデルは欠測にヨワい

空間相関を考慮しないモデル
欠測データなし



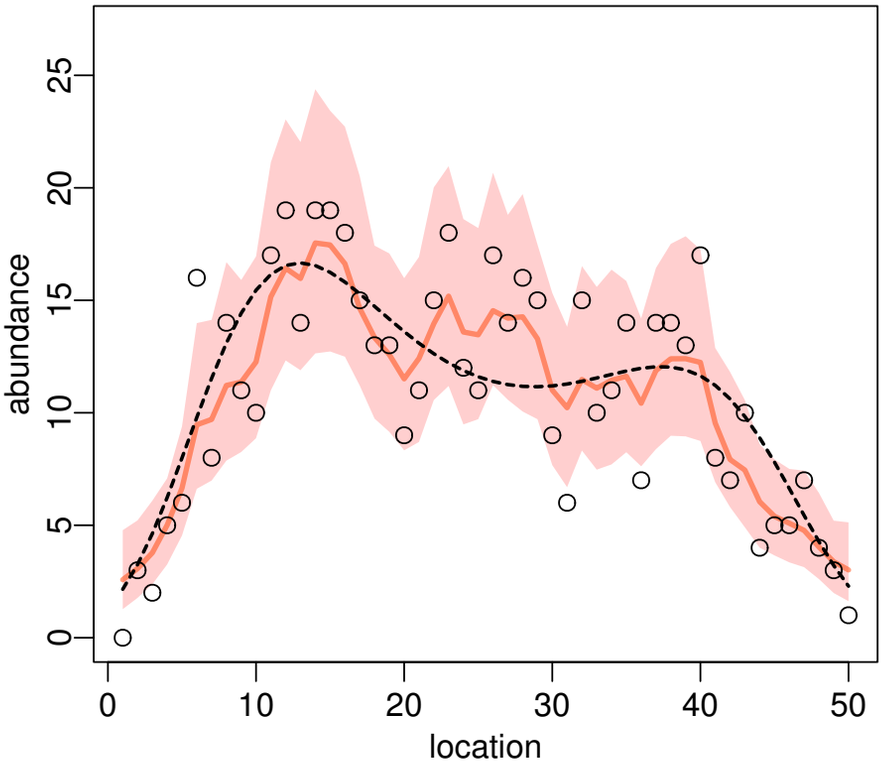
空間相関を考慮しないモデル
欠測あり



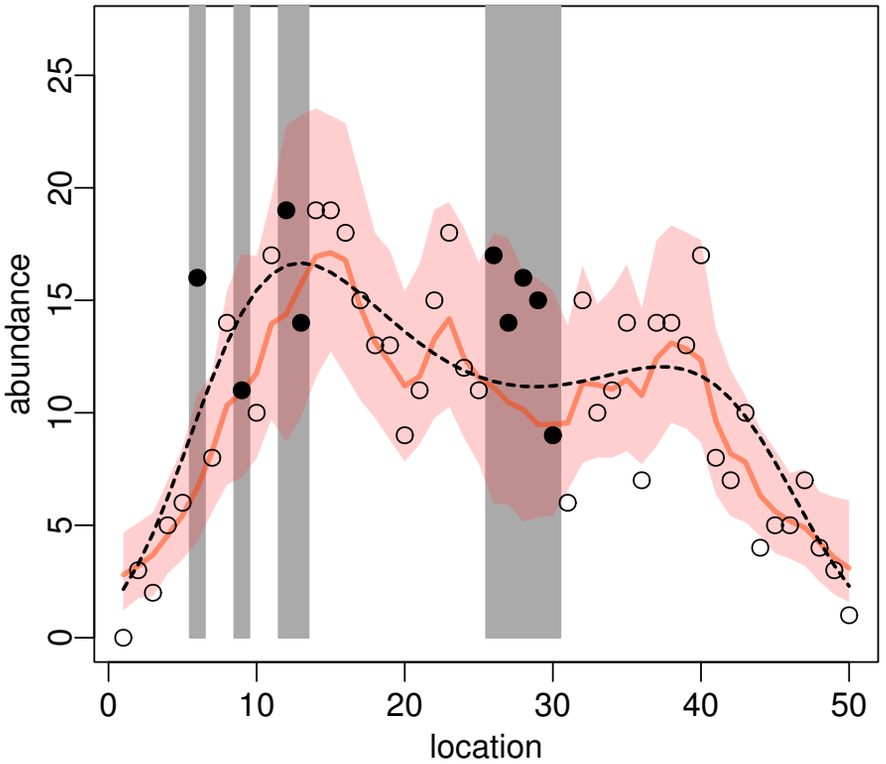
欠測領域で事後分布がひろがる!

空間相関を考慮するモデルは欠測に頑健

空間相関を考慮するモデル
欠測データなし



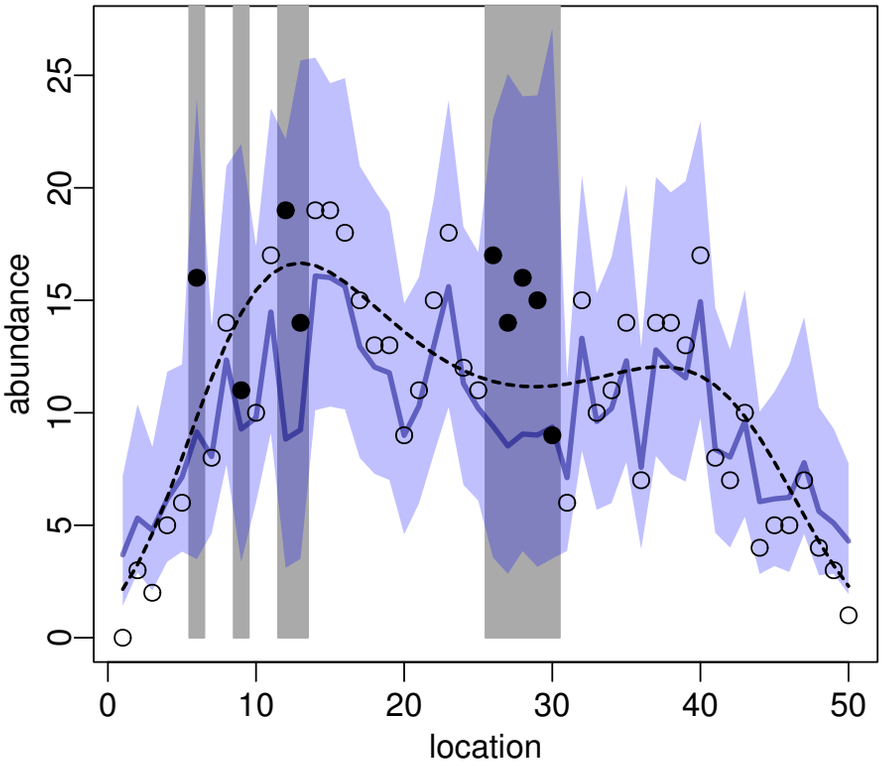
空間相関を考慮するモデル
欠測あり



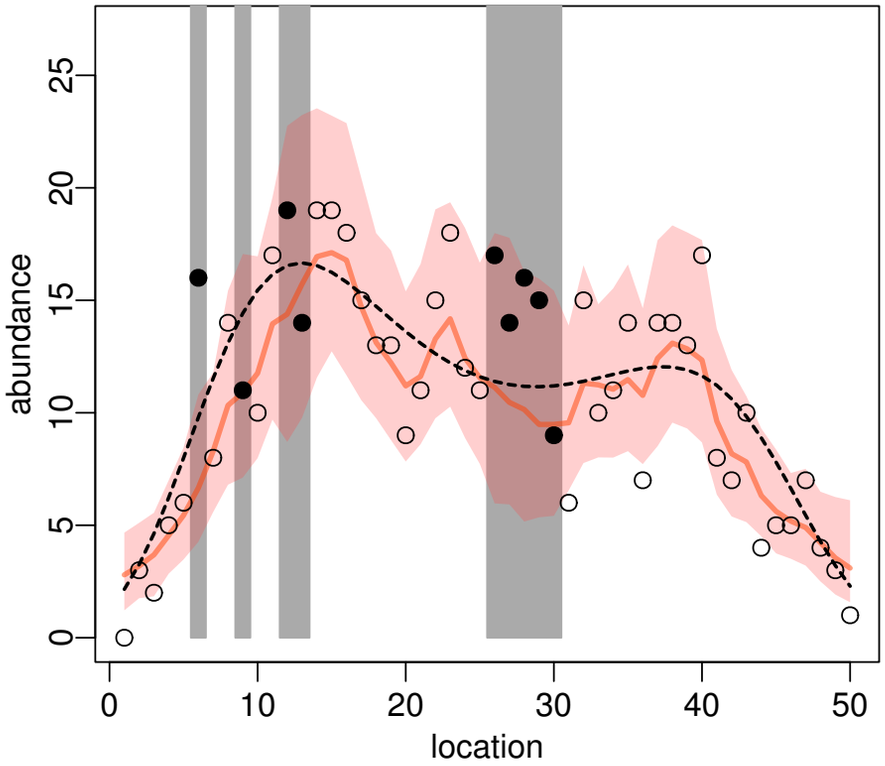
CAR 階層ベイズモデルで「隣は似てるよ」効果を表現

ベイズモデルの御利益: 空間的・時間的な欠測にも対処可能

空間相関を考慮しないモデル
欠測あり



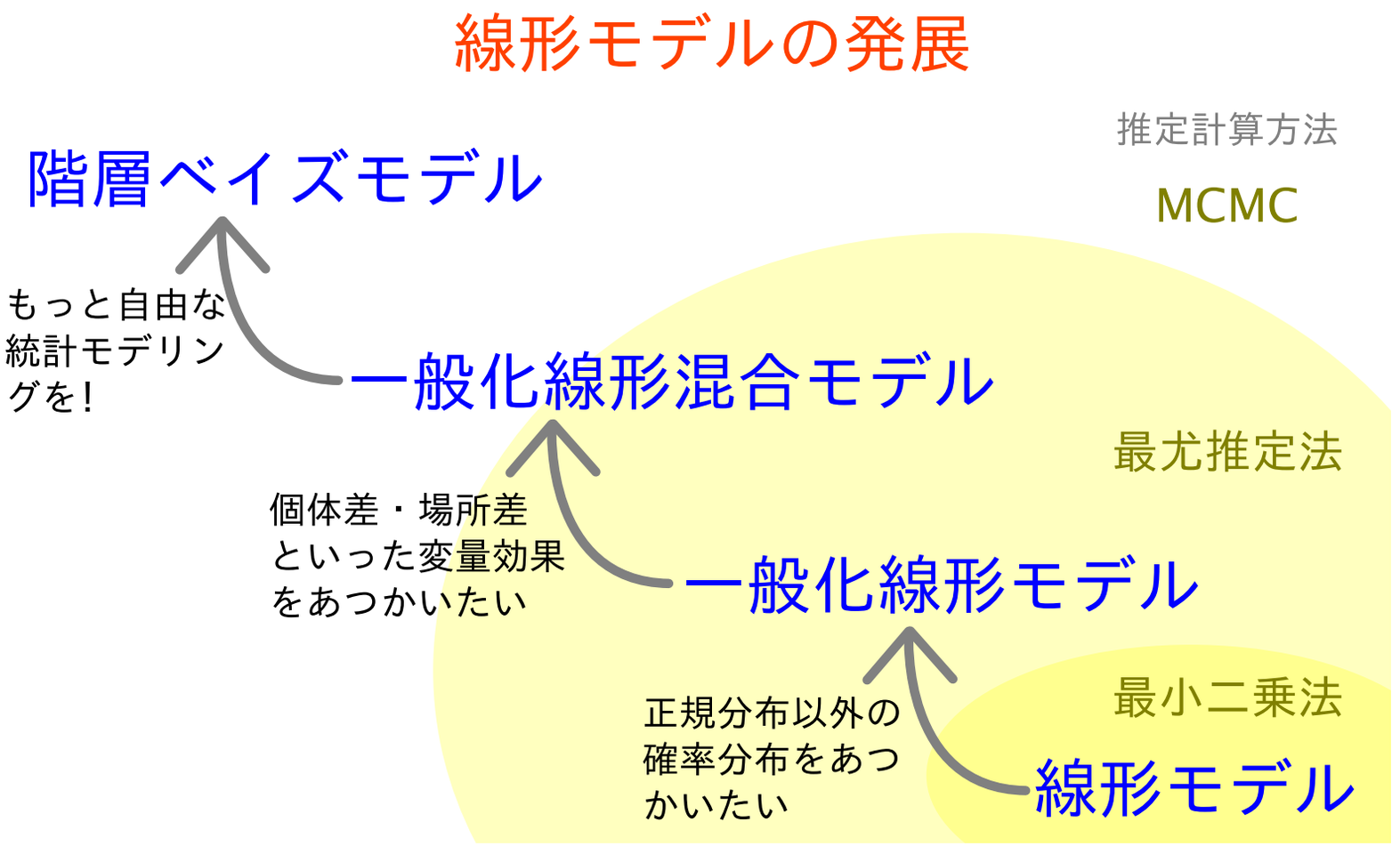
空間相関を考慮するモデル
欠測あり



この単純な例題を拡張して環境要因などをくみこめる

便利な道具: 階層ベイズモデル

1. **階層ベイズモデル: GLMM のベイズモデル化**
2. **空間構造のある階層ベイズモデル**



統計モデリング授業，終了!

皆さん，ありがとうございました．