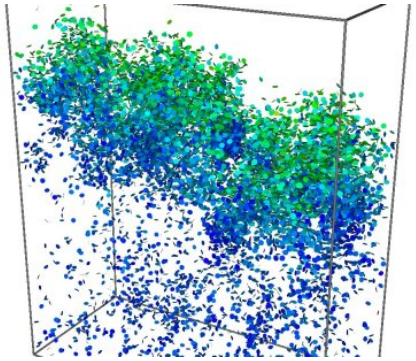
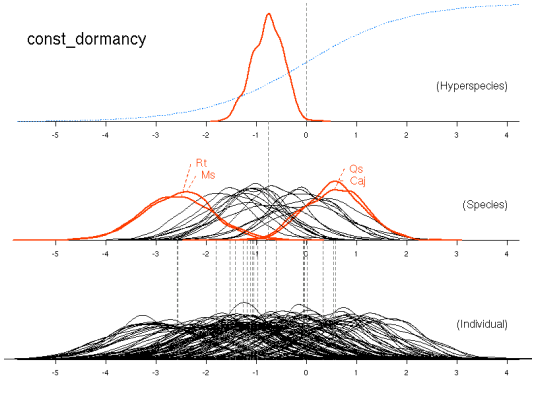
統計数理研究所共同研究「R の整備と利用」研究会

MCMC 計算まわりでさまよう R ユーザー





1. **エンドユーザーと階層ベイズモデル**
2. **R で MCMC 計算できるの?**
3. **階層ベイズモデルと Gibbs sampler**

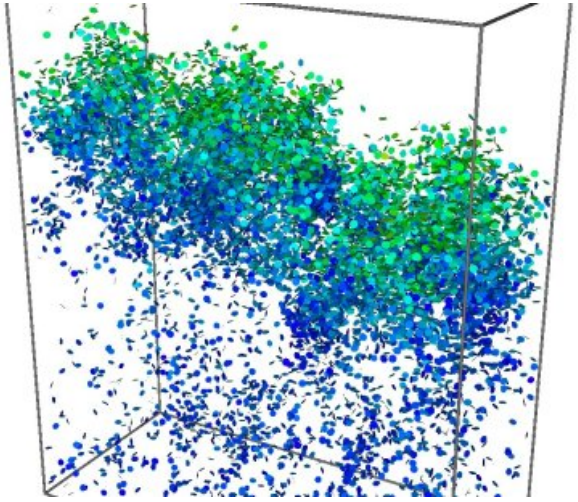
かなり「ソフトウェア紹介」方向のハナシです

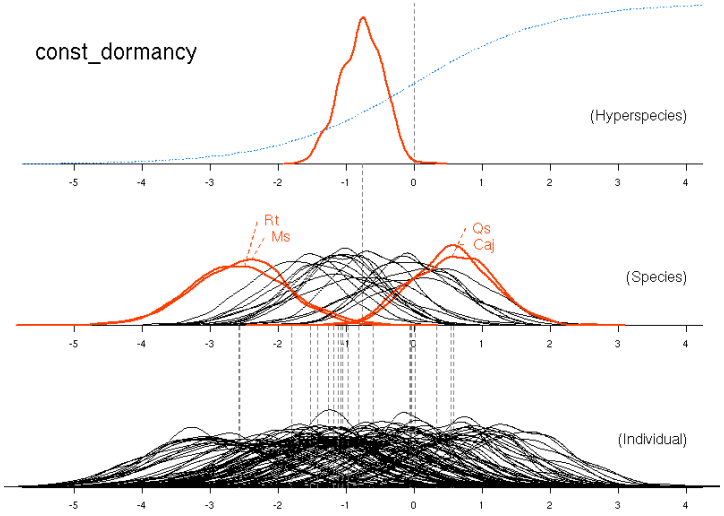
報告: 久保拓弥 kubo@ees.hokudai.ac.jp

http://hosho.ees.hokudai.ac.jp/~kubo/

今日の報告者: 久保拓弥 (自己紹介)

- 分野: 生態学，とくに植物生態学の問題をあつかうことが多い
- やってないこと: 野外調査 (やってないというよりできない)
- やってること: 他人がとってきたデータの解析 (寄生者?)





- 統計学はいわゆる独学，統計学ツールエンドユーザーの一人
- 生態学研究者集団という community (つまりエンドユーザー集団) の中ではデータ解析に関してなんだかエラそうにしている

今日のハナシ: 「エンドユーザー」とベイズ統計学

統計学ツールのエンドユーザーとは?

1. **統計学「ツール」指向** (ふりまわされる，の同義語)
2. **統計学の基礎的なことほど勉強してない** (基礎ほど難しいから)
3. **保守的** or **「誤った成功体験」によるフィードバック**?

エンドユーザー的な「ベイズ? MCMC?」疑問

1. **なぜ階層ベイズモデルなんかが必要なの**?
2. **なぜ** MCMC **計算**?
3. **どういうソフトウェアで計算できるの?** (今日はここに偏る)

R **のせいで，エンドユーザーが「ベイズに手を染め」ねばと追い詰められる時代・状況になった** ?!

エンドユーザーからみた統計学ツール「含有関係」

(一般化) 線形モデル的に現象を表現する場合

[尤度をあつかうモデル]

「すべてのパラメーターは確率分布」とする Bayes 統計学

階層 Bayes モデルなどなど

[最尤推定法 であつかうモデル]

fixed effects **パラメーターは最尤推定値**

経験 Bayes 法や一般化線形混合モデル (GLMM) などなど

[一般化線形モデル (GLM)]

指数関数族の確率分布 + 線形モデル, fixed effects **のみ**

[最小二乗法 であつかうモデル]

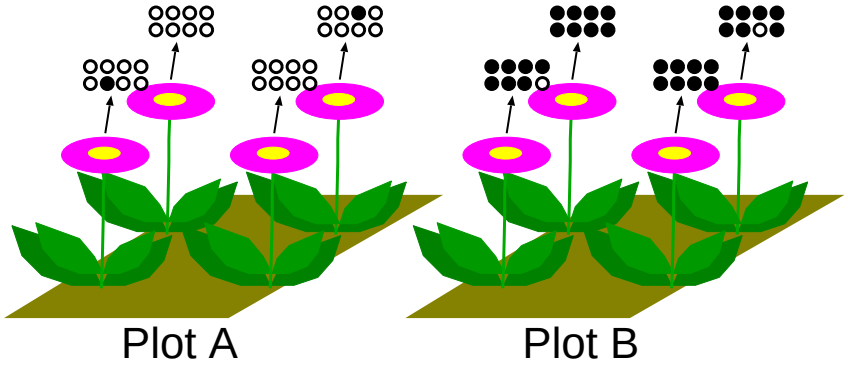
等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」などなど

階層ベイズモデルのご利益とは?

階層ベイズモデルでないとうまく表現できない現象がある

- **複数の** random effects (**個体差・ブロック差・縦断的データ・……**)
- **多重 nest した** random effects **の導入**

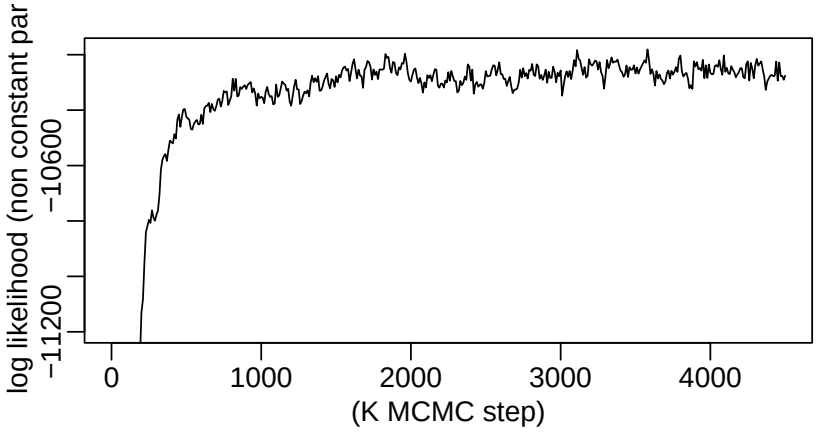


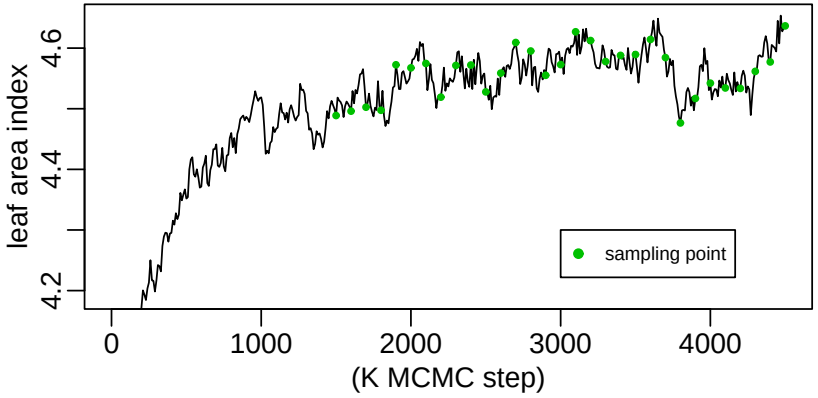
- **「隠れた」状態をあつかうモデル**
  - **例: 「欠側値を補う」処理**
- **空間構造ある問題も** MCMC **計算で**
  - **例: 「隣は似てるよ」効果** – Gaussian Random Field

MCMC 計算が何で必要? …… 今日は説明省略!

じゃあ，Markov Chain Monte Carlo って何?

- Gibbs 分布から逐次的に標本抽出 (sampling) する方法
  - 意訳: 「あてはまりの良さそうなところ」を「さまよう」
- どんな初期値から出発しても ……
- 「定常状態」に収束していく (sampling 開始!)
- 得られた sample は事後分布からの random sample set と考える
- 定常状態になるまでの step は捨てる (**burn-in**)





今日でてくるベイズ用語の整理

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

- **階層ベイズモデル** $p(\beta, \alpha|y) \propto p(y|\beta)p(\beta|\alpha)p(\alpha)$
  - 推定計算方法: Markov Chain Monte Carlo (MCMC) 法
    - MCMC 計算わざ 1: Metropolis-Hastings 法
    - MCMC 計算わざ 2: Gibbs sampler

    (上のふたつは本質的には同じもの)

- **経験ベイズ法** $\text{Likelihood}(\alpha|y) \propto \int p(y|\beta)p(\beta|\alpha)\mathrm{d}\beta$
  - 推定計算方法: α の点推定 (最尤推定)
    - 例: 一般化線形混合モデル (GLMM)
  - 簡単化した階層ベイズモデル，と考えるべきか?

(参照: 石黒ほか. 2004. 階層ベイズモデルとその周辺)

R まわりの MCMC 計算 / Gibbs sampler

- MCMC 計算はどういうソフトウェアで?
  - 自作する (問題によっては現実的)
  - R package: `library(MCMCpack)` など (いまいち)
  - Gibbs sampler ソフトウェア (R ではない世界)
    * `WinBUGS`
    * `OpenBUGS`
    * `JAGS`
  - `WinBUGS` と `OpenBUGS` の関係
    * `WinBUGS` , 2004 年ごろ開発停止，ソース非公開
    * `OpenBUGS` は `WinBUGS` の後継 project, GPL

君臨しつづける老雄: WinBUGS 1.4.1

- おそらく世界でもっともよく使われている Gibbs sampler
- **BUGS 言語の実装**
- adaptive rejection sampler
- 2004-09-13 に最新版 (ここで開発停止 → OpenBUGS)
- ソースなど非公開，無料，ユーザー登録必要
- Windows バイナリーとして配布されている
  - Linux 上では WINE 上で動作
  - MacOS X 上でも Darwine など駆使すると動くらしい
- ヘンな GUI (Linux ユーザーの偏見)
- R ユーザーにとっては R2WinBUGS が快適 (後述)

BUGS 言語で階層ベイズモデルを記述すると……

- Spiegelhalter et al. 1995. BUGS: Bayesian Using Gibbs Sampling version 0.50.

```
model {
    mu ~ dnorm(0, 1.0E-2)
    tau ~ dgamma(1.0E-3, 1.0E-3)
    for (i in 1:n.samples) {
        re[i] ~ dnorm(0.0, tau)
        p[i] <- 1.0 / (1.0 + exp(-(mu + re[i])))
        n.seeds[i] ~ dbin(p[i], n.ovules[i])
    }
}
```

- JAGS **だと行末に ; が必要，といった方言がある**

BUGS 言語雑感

- 複雑な尤度方程式をもつような統計モデルは直接的には記述できない (「書かせない」ポリシー?)
- しかし「隠れ変数」などをうまく使うことで等価なモデルに書き直せる場合がある
  - 例: zero-inflated Poisson (ZIP) model など
- だからといって BUGS で全てのベイズモデルが記述できるわけでもなさそう
  - 理由: 使えるデータ構造などが貧弱だから
- 汎用だけど，一般化線形モデル的なモデルの拡張に最良，か?

GPL な WinBUGS めざして: OpenBUGS 2.2.0

- Thomas Andrew さん他が開発している
- WinBUGS の後継プロジェクト
- OpenBUGS is still in development and suffers frequent crashes.
- ソースは公開しているが ……
  - Component Pascal で実装
  - ソースを読んだりするには
    BlackBox Component Builder が必要
- Windows バイナリ配布，Linux でもなんとか使える
- R ユーザーにとっては library(BRugs) が快適そう
  - しかしこれは Windows 版 R でしか使えない!!

R 寄り (?) な Gibbs sampler: JAGS 0.97

- R core team のひとり Martyn Plummer さん作
  - Just Another Gibbs Sampler
- C++ で実装されている，誰でもコンパイルできる
  - R がインストールされていることが必要
- バイナリ版
  - 谷村晋さん Vine Linux 用 RPM package
  - Windows 版バイナリは JAGS サイトにある
  - MacOS X でもコンパイルできる
- R からの使う: `library(rjags)`
  - まだまだ未完成

JAGS のこれまでとこれから

- JAGS にできない計算
  - directed cycle (有向巡回閉路) を含むモデル

```
model {
    x ~ dnorm(y, tau)
    y ~ dnorm(x, tau)
}
```

  - 空間データ解析などに使えない!
  - 今後改良したいとのこと
- WinBUGS に計算速度・収束速度でちょっと負けてる?
- R ユーザーとしては今後の発展に期待したいソフトウェア

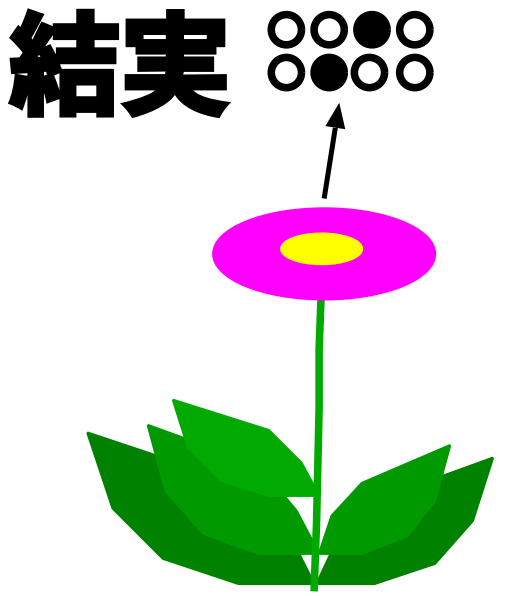
実演篇: JAGS と WinBUGS で

階層ベイズモデルをあつかう

今日の例題: 胚珠が種子になる確率

[架空植物の性質あれこれ]

- 花 $i \in \{1, \cdots, 100\}$
- 花の胚珠数 `n.ovules[i]` = 8
- 花の種子数 `n.seeds[i]`
- 今日は葉数 関係なし
- 結実確率 $p[i]$: ある胚珠が種子になる確率
  - ●: 結実成功胚珠，○ 結実失敗胚珠
- 結実確率: logit(`p[i]`) = 0 + `re[i]`
- 個体の random effects: `re[i]` $\sim N(0, 1/0.5)$
- 集団の平均結実確率 = 0.5, 期待結実数 = 4

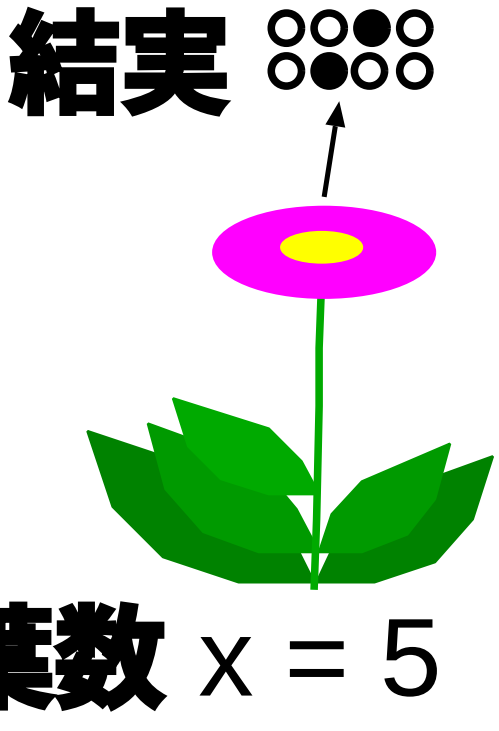


今日の例題: 架空植物の観測データ

結実数	0	1	2	3	4	5	6	7	8	(合計)
観測個体数	9	9	7	9	14	18	15	15	4	100
期待個体数	0.2	2.2	8.6	19.2	27.0	24.2	13.6	4.4	0.6	100

[観測者の判断]

- 集団全体の平均結実確率 = 0.529
- しかし $p = 0.529$ の二項分布からはあまりにも逸脱している
- overdispersion (過分散)!
- 個体差による random effects を組みこんだモデルでなければこの現象は説明できない



これを (あえて) 階層ベイズモデル化してみる

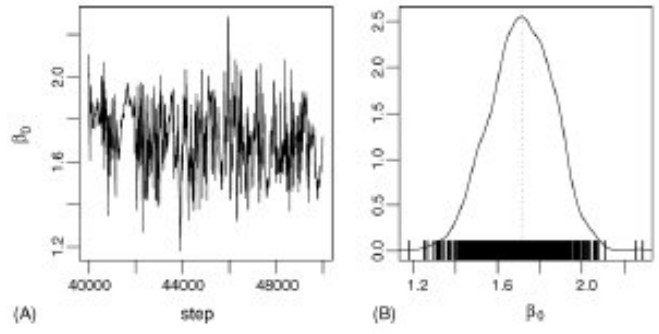
JAGS を使ってみる

1. BUGS 言語でかかれた model ファイルを準備する
2. R を使って以下のファイルを準備する
   - データファイル
   - パラメーター初期化ファイル
3. JAGS のコマンドファイルを書く (`foo.cmd`)
4. コマンドライン上で `jags foo.cmd`
5. JAGS 出力を R の `library(coda)` の `read.coda()` で読む (mcmc オブジェクト化)
6. mcmc オブジェクトを `plot()`, `summary()`, ……

`library(coda)`: MCMC 計算出力をあつかう

Convergence **D**iagnosis and **O**utput **A**nalysis for MCMC

- R の package (もともとは S-plus 用)
  - Martyn Plummer さん他の作
  - MCMC 計算出力の読みこみ → `mcmc`, `mcmc.list`
  - さまざまな収束診断
  - MCMC 遷移や事後分布の作図



mcmc オブジェクトの summary メソッド

```
> summary(r.mcmc)

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

     Mean    SD Naive SE Time-series SE
mu  0.254 0.187  0.00835          0.0254
tau 0.355 0.090  0.00402          0.0101

2. Quantiles for each variable:

      2.5%   25%   50%   75% 97.5%
mu  -0.108 0.125 0.251 0.372 0.622
tau  0.208 0.290 0.349 0.409 0.560
```

WinBUGS を使ってみる: R2WinBUGS 経由で

1. BUGS 言語でかかれた model ファイルを準備する
2. R2WinBUGS package を使う R コードを書く
3. R 上で 2. を実行
4. 出力された結果が bugs オブジェクトで返される
5. これを plot() したり summary() したり……
6. あるいは mcmc / mcmc.list オブジェクトに変換したり

今日のまとめ: 「エンドユーザー」とベイズ統計学

R エンドユーザーたちが選ぶ MCMC 計算の手段とは?

- (汎用 MCMC 計算では) 今後しばらくは WinBUGS が**支配的**であろう
  - R2WinBUGS 経由で使う人が多数派になる?
  - WinBUGS 使用には R が必要不可欠!
- (報告者の願望としては) JAGS に**がんばって**もらいたい
  - もっと R と連携できれば便利 (しかし Plummer さん多忙?)
- 「組みこみ」 MCMC 計算も普及するかも
  - 空間統計学の geoR 系: Gaussian Random Field
    - GRF の MCMC 計算: Langevin-Hastings 法
- 次は何? : OpenBUGS ? ... or ... Umacs + rv ⟶ "B"?