4.3 マルコフ連鎖モンテカルロ法

MCMC（MARKOV CHAIN MONTE CARLO）は確率分布を明らかにするために用いられるシミュレーションに基づいた手法である。1950 年のはじめに開発されて以来（Metropolis et al., 1953）、Geman and Geman (1984)や Besag (1986) によって画像解析に対する有効な手法として発表される 1980 年代まで、MCMC を統計学的に応用しようという試みはほとんど存在しなかった。この後、このトピックへの関心は爆発的に高まることとなるが、この貢献に対し"statistical analysis of dirty pictures"という魅力的なタイトルをつけた Besag の役割を推測したくなるだろう。しかし、真の理由はほかにある。MCMC はその使いやすさと比較的容易な実装性により一世を風靡したのだ。

4.3.1 マルコフ連鎖と定常分布

k 次のマルコフ連鎖とはランダムな変数の連なり X_1、X_2、X_3、…である。これらランダムな変数は以下の確率に依る。過去の値すべてを得たとき、次の値を決定する確率分布は最後の k のときの値のみに依存する。それはつまり、

$$[X_t | X_{t-1}, X_{t-2},\ldots, X_1] = [X_t | X_{t-1}, X_{t-2},\ldots, X_{t-k}]$$

である。多数連なったマルコフ連鎖は時間を含む万物のプロセスを記述するために用いられる。インデックス t はしばしば時間を言及することがあるが、X_t の値は（万物の）"状態"を表している。

賭金 X_1 ドルを持ったギャンブラーがカジノに入り、ルーレットに賭けているとしよう。ルーレット盤は同程度の確率で 38 態の結果に帰する。それらのうち 18 態の結果ではギャンブラーが 1 ドル勝ち、のこり 20 態では 1 ドル負ける。t 回賭けた後のギャンブラーの賭金は X_t ドルとなる。もし、それぞれの賭けが独立であるならば、X_t,（t = 1,2,…）は 1 次のマルコフ連鎖になる。X_t はこれまでの経緯（$X_1, X_2, X_3,\ldots, X_{t-2}$）に関係なく、$X_{t-1}$ ＋1 ドル（確率：p = 18/38）または X_{t-1}－1 ドル（確率：1－p = 20/38）になる。

そのギャンブラーの資本金が X_1 = 20 であり、彼女が賭金を 2 倍に増やすか無一文になるかまで賭け続けると仮定すると、彼女の賭金 X_t は常に S = {0, 1, 2, 3,…, 40}のいずれかになり、これらは状態空間 (*state space*) と称される。状態 0 と状態 40 は吸収状態 (*absorbing states*)と呼ばれる。つまり、もし X_t = 0 なら X_{t+k} = 0 となり、X_t = 40 ならば X_{t+k} = 40 となる（k = 1,2,…）。

t－1 回賭けた後、ギャンブラーの賭金が X_{t-1} = \$5 であるとき、X_t = \$4 もしくは

$6（それぞれの確率は 20/38、18/38 となり、このとき X_{t-1} になるまでの一連の結果には左右されない）になる。彼女が続けざまに 20 回負け、彼女の賭金が$25 から次第に少なくなっていったという情報や彼女が 4 連勝して辛くも無一文から免れたという情報に対し、我々は興味を持つかもしれない。しかし、このような付加的な情報からは次の賭けで彼女が勝つ確率に対して何の洞察も得られない。また、もし $X_{t-1} = 0$ であるなら、$X_t = 0$ であるし、$X_{t-1} = 40$ であるなら $X_t = 40$ になる。これらはこれまでの経緯にかかわらず絶対である。時間 t における彼女の賭金は X_{t-1} のみを通じて $X_1, X_2, X_3, \ldots, X_{t-1}$ に依存する。この一連の X_t が一次のマルコフ連鎖である。

様々なマルコフ連鎖のうちいくつかは定常分布となる。定常分布とは標本空間の部分集合 A それぞれ（＊1）について

$$\pi(A) = \Pr(X_t \in A) \qquad (＊2)$$

であるような確率分布のことである。この確率分布の重要な特徴はこれが時間に対して定常(stationary）であるということである。つまり、X_t がある特定の状態になる、もしくは複数の状態がある特定の組み合わせ（＊3）になる確率は時間 t には依存しないということである。

＊1 標本空間の部分集合 A…ギャンブラーの例では部分集合 A は 0 から 40 の正の整数という状態空間。標本空間はお金がとりうる範囲（？）

＊2 $\Pr(X_t \in A)$：マルコフ連鎖 X_t に属するある値 A になる確率

＊3 複数の状態がある特定の組み合わせ…推定パラメータが複数ある場合に言及している

すべてのマルコフ連鎖が定常分布しているわけではない。例えば、ここで挙げたギャンブラーの例では成り立ってはいない。$\Pr(X_t = 19)$の場合を考えてみよう。時間 $t = 2$ の時、可能な状態は $X_2 = X_1 \pm 1 = 19$ or 21 となり、これは彼女が最初の賭けで勝つかどうかにかかっている。それ故、$\Pr(X_2 = 19) = 20/38$ になる。最終的に、彼女の連鎖は 0 か 40 に達し、そこにとどまるだろう。ゆえに、t が大きくなるにつれ $\Pr(X_t = 19)$は 0 に近づいてゆく。定常分布の存在条件としては $\Pr(X_t = 19)$が時間とともに変化しない事が要求される。

マルコフ連鎖が定常分布であるか否かは MCMC の根幹にかかわる問題である。ここで、分布 f をサンプリングしたいが、独立な標本を作る標準的な手法ではふさわしくない（＊1）としよう（標準的手法の例：cdf inversion (Section 4.2.1) 、rejection sampling (Section 4.2.2)）。このような場合でも、定常分布 f であるマルコフ連鎖 X_t は作ることができるだろう。これら抽出した値はおのおので独立ではないが、目標の分布に従うサンプルとなるだろう。

＊1 例えば、多変量分布から確率変数をサンプリングするような場面を想定している？

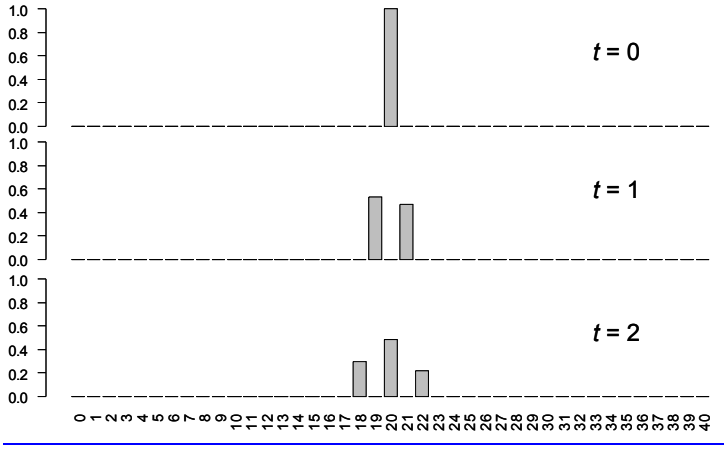
MCMC の使用者はエルゴード性の定理に気を留めておくべきであろう。エルゴード性の定理とは、正再帰的（*positive recurrent*）、かつ非周期的（*aperiodic*）なマルコフ連鎖は標本空間の部分集合 A に対して

$$\pi(A) = \lim_{n \to \mathrm{Inf}} \Pr(X_n \in A \mid X_1)$$

を満たす定常分布 $\pi(A)$をもつというものである。我々はまず、正再帰性と非周期性の性質によってもたらされる暗示を指摘し、つづいてそれらの用語を定義することにする。その暗示とは定常分布の存在を保証するだけではなく、初期値 X_1 が連鎖の非周期的な振る舞いに影響を及ぼさないということをも提示しているということである。ある連鎖の初期値にかかわらず、最終的には A を訪れるパターンは特定の確率 $\pi(A)$に対応するようになる。実際には、初期値を設定しなければならないため、MCMC の実装においてこの保証は有用である。この初期値をエルゴード性の定理はそれほど気にする必要がないということを示しているのである。一般的には、得られた連鎖から定常分布を上手く表現できていないかもしれない初期の観測結果のいくつかを放棄する”*burn-in values*”ことになる。

ギャンブラーの連鎖は定常分布ではないと注意を喚起した。そもそも、この例では非周期性を保っていない。つまり、周期が 2 で固定（＊1）されている。ギャンブラーはどんなに多数回賭けを行ってもある賭金（$0 < i < 40$）に戻ることができるだけだ（何故だろうか？）。（＊2）ある状態の周期というものは可能な再帰時間の最大公約数（＊3）として、いくらか不明瞭だが、決定される。もし、ある状態の周期が 1 であるとき、その状態は非周期的であると言える。非周期性がすべての状態で保証されている時に、マルコフ連鎖は非周期的であるといえる。

＊1　周期が 2 になるギャンブラーの例



＊2　意味がうまく取れない

＊3　可能な再帰時間の最大公約数…たとえば、状態 20 に戻ってくる可能性のある t は 2、

4、6、…という 2 の倍数であり、その最大公約数は 2 である

また、ギャンブラーのマルコフ連鎖の例は再帰的でもない。$X_1 = i$とし、$T_i = \min\{n > 1 : X_n = i\}$（$T_i$は状態 i へ戻るのにかかる時間）となるように仮定してみる。状態 i はもし、$\Pr(T_i < \infty) = 1$ であれば再帰的とよばれ、$E(T_i) < \infty$ならば正再帰的と呼ばれる。つまり、ある状態にふたたび戻るであろうということが確かであるとき再帰的であり、その時間がそれほど長くかからないであろう場合は正再帰的であるといえる。これらの描写がすべての状態に適用できるなら連鎖全体に適用される。大雑把に言うと、正再帰的連鎖ではすべての状態に適切な頻度で訪れることで、状態空間を効率良くよく探索できる。ギャンブラーの連鎖の喩えでは 0 と 40 の吸収状態が存在することから再帰的ではない。$0 < i < 40$ のいずれかの状態から、状態 i へと戻る前に吸収状態に達してしまう可能性が常に存在する。よって、$\Pr(T_i < \infty) < 1$ となる。

MCMC の適用に際し、周期的な振る舞いをするか否かが争点になることはまれだが、非再帰的な状態にはなりうる。以下からいくつかの例に戻る。

4.3.2 用例：標準正規定常分布

ここでは一様乱数を用いたマルコフ連鎖を使って標準正規分布のサンプルを作りだすという簡単な例を挙げる。この連鎖はひとつのチューニングパラメータ A（$A > 0$ でとりあえず機能するが、そのなかでもよりよく機能する値が存在する）（＊1）で定義される。

＊1　この A はこれまで出てきた「標本空間における部分集合 A」とは別物

アルゴリズム 1

$X_0 = 0$ とし、$t = 1, 2, \ldots,$の間、以下の規則に従って X_t を発生させる。

Step 1: 0 から 1 までの一様乱数 U(0,1)を二つ発生させる（二つの乱数は u_1、u_2 とする）

Step 2: 候補点(candidate value) X_{cand} を算出する。

$$X_{\text{cand}} = X_{t-1} + 2A(u_1 - 1/2)$$

Step 3: 以下を計算する

$$r = \exp(-(1/2)*X_{\text{cand}}^2) \,/\, \exp(-(1/2)*X_{t-1}^2)$$

Step 4: もし、$u_2 < r$ ならば、$X_t = X_{cand}$ とし、そうでないなら $X_t = X_{t-1}$ とする。

それぞれの時間ステップで連鎖は現在の値のままにとどまるか、もしくはランダムに発生された候補点に移動するかのどちらかである。Step 2 における増加は $U(-A, A)$の分布を持ち、候補点は現在の値を中心とした範囲で一様にサンプリングされる。つまり、$X_{cand} | X_{t-1} \sim U(X_{t-1} - A, X_{t-1} + A)$となる。$X_t$の値は明らかに一次のマルコフ連鎖である。つま

り、X_{t-1}, X_{t-2}, …, X_1 が得られたとき、X_t の分布は直前の値 X_{t-1} にのみ依存するということである。また、Step 4 はベルヌーイ試行であることを言及しておく。成功確率 = min(r, 1) であり、ここで r は候補と現在の値の連鎖による評価の目標となる標準正規密度の比率である（＊1）。この移動の成否を決定するパラメータ r は受容確率（acceptance probability）または移動確率（movement probability）と呼ばれている。

＊1　標準正規分布の確率密度関数は以下の式

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

アルゴリズム 1 はメトロポリス-ヘイスティングス法のひとつのかたちである。これについての詳細は後々述べる。ここまでの話は、MCMC を操作するうえで配慮すべきことへの洞察を深めるために示した簡単な実験例である。

チューニングパラメータと初期値の効果

t に対し X_t をプロットしたものがマルコフ連鎖 X_t の標本経路(history plot)である。図 4.6 はアルゴリズム 1 に従って作られた四つのマルコフ連鎖の最初の 1000 個を示した標本経路である。この内、三つは $X_0 = 0$ で始まり、ひとつは $X_0 = 20$ で始まっている。左上から時計回りにそれぞれ連鎖は $A = 0.5, 1.0, 3.7, 15$ となっている。

四つ全ての連鎖は標準正規定常分布であるが、それぞれの標本経路には明らかな違いが見られる。$A = 0.5, 1.0$ では、標準正規分布の範囲で連鎖が小さくたくさんのステップをとり、ゆっくりと動いている。$A = 1.0$、初期値 $X_0 = 20$ の連鎖はゆっくりと、だが標準正規乱数の典型的な範囲に着実に下がっている。一般的には、定常分布を現していないため”burn-in”として最初の 100 個の値は廃棄される。$A = 15$ の連鎖はときより大きなステップをとるが多くの場合はとどまっている。具体的には、700 から 750 の間の t ではほぼ 1 の値で止まっている。

これらの違いは X_t の値の間にあるつながりの強さを反映している。X_t と X_{t+h} のあいだの $\rho(X_t, X_{t+h})$はラグ h の自己相関と呼ぶ（$h = 1, 2,\ldots,$）。そして、$H(h) = p(X_t, X_{t+j})$は自己相関関数 ACF（autocorrelation function）と呼ばれる。図 4.7 は三つの連鎖の自己相関関数を表現している。

$A = 3.7$ のときでは、自己相関は急激に下がる。このときはラグ 9 で 0.01 より小さ

くなっている。これはつまり、9 番目の観測点ごとにほぼ独立だとみなすことができるということであろう。A = 15 ではラグ 35 のときにおいて同様の小さな値を示している。一方、A = 3.7 の連鎖がラグ 3 のときに示す値は A = 0.5 の連鎖がラグ 50 のときに示す値と同じ値であり、強い自己相関があることを示している。

高い自己相関を示すということは標的となる分布の特徴を推定する確度や精度を減少させてしまう。連鎖長 n が長くなるにつれ、エルゴード性の定理は三つのマルコフ連鎖のサンプルの特徴が近似的に正規分布の特徴と対応していくことを保証する。つまり、サンプルの平均はゼロに近づき、標準偏差は 1 に近づくだろう。そして、サンプルの 95 パーセンタイルは 1.960 に近づくだろう。しかしながら、n が限られている場合では、サンプルの特徴は定常分布の特性の推定値に過ぎず、偏っていたり、不正確であったりするかもしれない。連鎖の偏りと精度は A のチューニングにかかっている。

驚くべきことではないが、A = 3.7 である連鎖はこれらの量的推定をする際により効率的である。なぜなら、その抽出された値は他の二つの連鎖よりも強く相関していないからである（Fig. 4.7）。Table 4.1 では A = 0.5、3.7、15 であり、連鎖長が 1000 である連鎖を用いてアルゴリズム 1 を評価している。比較として、1000 回の独立抽出の類似性評価を盛り込んだ。（＊1）したがって、例えばその標準正規分布の 97.5 パーセンタイルは 1.960 である。A = 3.7、連鎖長 1000 の連鎖では平均すると 1.964（標準偏差は 0.158）の推定値が得られる。

＊1　正規乱数 rnorm(1000,0,1) を比較のため盛り込んだということ

自己相関が強い連鎖（A = 0.5, 15）は推定精度が実質的に劣る。さらに精度だけではなく、標準偏差と 97.5 パーセンタイルの推定値には標本数が少ないことによる偏りさえ生じている。（＊1）A = 0.5 または 15 では連鎖長 1000 でもその偏りに打ち勝つためには不十分である。注目すべきは独立抽出したときでさえ、n = 1000 では高い精度や確度を得るためには不十分であり、より多くの標本が必要とされることである。長い連鎖を使用することと同時に、自己相関ができるだけ小さくなる連鎖を作るような手法を用いることで高くなってしまう自己相関を補うことが望まれるだろう。Table 4.1 では、3 種のマルコフ連鎖 A = 0.5、3.7、15 と独立抽出した標本では推定パラメータの標準偏差はそれぞれ近似的に 6:2:4:1 の比率になる。標準誤差が標本数の平方根に比例する原則を考慮すると、A = 0.5、3.7、15 のそれぞれの連鎖の標本数は独立抽出した標本数の 1/36、1/4、1/16 であるとみなせるかもしれない。

＊1　精度 precision（ばらつきの少なさ）と確度 accuracy（偏りの少なさ）はそれぞれ Table4.1 の SD と Mean の列から評価

ラグ 1 の自己相関を比較することによって、A = 3.7 の連鎖は他の A の値をもつ連鎖よりも妥当であることを確かめた。Fig 4.8 は A が 0 から 15 の範囲で作られた連鎖の自己相関と移動確率をプロットしたものである。A の値が小さいとき、候補点の変化量は小さい。そして、移動確率 r が高くなる。なぜなら、隣接した値はほぼ同じ確率（＊1）になるからであり、それゆえ移動確率 r が 1 に近くなる。A の値が小さい連鎖はゆっくりと動き、高い自己相関を持つ。A が大きい連鎖では、より候補点の変化量が大きくなり、移動確率が低くなる傾向がある。A の増加に伴い、ラグ 1 の自己相関は頻繁に $X_t = X_{t-1}$ となるくらい移動確率が低くなる手前までは減少するが、そこから再び自己相関が高くなっていく。連鎖の自己相関は A がおよそ 3.7 で最も小さくなる。このとき、移動確率は約 0.42、ラグ 1 の自己相関は約 0.56 である。一般的には移動確率が 30～50%の時に最適になることが知られている。

＊1　このセクションの例では標準正規分布の確率密度関数から得られる確率

この例に取り組むことで、すこしだがその中身をのぞき込んだ（have a bit of a look under the hood）。多くの読者がこの車を走らせたいと望み、と同時にそれら詳細が手強いものだと感じたかもしれない、ということを我々筆者は想像する。我々はその恐怖を和らげたい。WinBUGS のような高い品質を持つソフトウエアが利用出来ることに読者は気づくべきである。このプログラムは様々なモデルに対し MCMC を実行する際に利用できる。しかしながら、このようなソフトウエアを使うときでさえ、いくつかの技能が要求される。それは自己相関関数とその他診断ツール（Section 4.3.5）に精通していることであり、結果を評価するためには必要である。

簡単なモデルを手始めに、自分自身の MCMC コードを書いてみることを勧める。自身のコードを書くことは MCMC のパフォーマンスに対する直感を養う最善の方法であり、かつ複雑なモデルを解析する際には折にふれて必要なことである。

MCMC を行う上で最も初歩的ツールはメトロポロスーヘイスティングスアルゴリズム（MH アルゴリズム）である。これについては Section 4.3.3.で詳細は述べる。MH アルゴリズムは多くを内包しており、その実行も比較的簡単である。しかし、この方法を適用する、特に多変量分布に適用する、際にはかなりのチューニング（これまでの議論の中の A の選択に似た調整）が要求される。

ギブスサンプリング（Section 4.3.4）は多変量分布に適用するために MH アルゴリズムを洗練したものである。

4.3.3 メトロポリス–ヘイスティングアルゴリズム

標準正規分布からサンプルを引くために用いたアルゴリズム 1 はメトロポリスーヘイスティングアルゴリズム（MH アルゴリズム）の特殊な例である。ここから、その MH アルゴリズムを紹介する。

メトロポリスーヘイスティングスアルゴリズム

ある標的分布 $f(x)$からサンプルを得たいとしよう。ここで、$j(x|y)$を候補生成分布(candidate generating distribution)とする。候補生成分布とは現在の値 y が既知のときに候補点 x になる確率を記述するものである。X_0 をある値に固定し、X_tを以下の規則に従って $t = 1, 2, \ldots,$で発生させる。

Step 1: $j(x|X_{t-1})$から抽出することで、候補点 X_{cand} を発生させる
Step 2: 以下を計算する

$$r = \frac{f(X_{cand})\, j(X_{t-1} \mid X_{cand})}{f(X_{t-1})\, j(X_{cand} \mid X_{t-1})} \tag{4.7}$$

Step 3: $U \sim U(0,1)$を発生させる
Step 4: $U < r$ ならば、$X_t = X_{cand}$、そうでなければ $X_t = X_{t-1}$ とする

MH アルゴリズムで最初に気づくことは、標的分布が r の計算にのみ関与し、そしてそれが分子と分母の両方に存在するということである。これがベイジアンにとって何を意味するのかを考えてみよう。: f が[γ]に呪われた事後分布（式 4.1）であるなら、ベイジアンの初期世代にとっては悩みのたねであったその規格化定数(normalizing constant)は消去することができる！

MH アルゴリズムは候補生成分布を選択する際に広い許容範囲を持つことができることに気づくべきである。定常分布であるためにマルコフ連鎖は再帰的でなければならないという必要条件（すべての状態は他のすべての状態に到達可能でなければならない）に、$j(x|y)$にかかる制約は関連している。実際には、この最小必要条件を満たす以上に、連鎖が無理なく低い自己相関を持つに十分なだけ自由に動けるという要求が存在する。

ここで、Section 4.3.2 で挙げたアルゴリズム 1 の二つの特徴に焦点を当てる。一つ目の特徴は、候補生成分布はその独立変数内で対称性を保っている(symmetric)ということである（＊１）。y を中心とし、範囲が $2A$ である一様分布は密度関数

$$j(x \mid y) = \frac{I(|x - y| < A)}{2A}$$

を持つ。ここで、$I(\cdot)$は指標関数である。$j(x \mid y) = j(y \mid x)$がなりたつため、Step 2 における r を

$$r = \frac{f(X_{cand})}{f(X_{t-1})}$$

と単純化できる。対照的な候補生成分布を利用することは計算の単純化や計算時間の節約になる。

＊１　別の表現ではこの状態を詳細釣合条件が保たれているともいう

同様に、アルゴリズム 1 のチューニングパラメータ A にも注意を喚起する。適切な値の選択は連鎖を生成すると同時に行える。まず $A = 1$ から始め、候補点が採択されたそれぞれの時間では（仮に）1.01 を A にかける。そして、候補地が棄却された各時間では（仮に）1.01 を割る。（仮に）5000 ステップのあと、A の値をその現在の値で固定する。そして、それ以前に抽出した値をすてる。このプロセスは採択率約 50%を目指している。より低い採択率を得るには、採択されたら 1.01 をかけ、棄却されたらより小さい値（仮に 1.007 とする）を割ることによって達成できる。このアプローチに従うと、$A = 3.48$、採択率 0.44 に決まり、最適値 3.7 に近づく。（＊１）[10]

＊１　一変量正規分布では候補点の採択率が約 40%のときにサンプリング効率がよくなるという目安に過ぎない、らしい

10.　1.007 を割ることにより最適な採択率である 42%に近づく。そして、それは $p = 0.42$、$c = 1.01$ であるときの $p \times c + (1-p)/d = 1$ の解である。徐々に 1 に減少する c 値を使うことでこの調整プロセスはさらに精度を上げることができ、$d = (1-p)/(1-pc)$に対応する結果を返すようになる。

4.3.4 ギブスサンプリング

ギブスサンプリング (Gibbs sampling)は事後分布が多変量であったときのために設計されたものである。$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'$は未知の複数の(母)数とし、$\boldsymbol{X}$ をデータであるとする。ギブスサンプリングの目的は$[\boldsymbol{\theta} \mid \boldsymbol{X}]$からサンプルを抽出することである。

$\boldsymbol{\theta}_{(-j)}$は$\boldsymbol{\theta}$のすべての要素から$\theta_j$を除いた長さ$k$-1のベクトルを意味する。ここで、

$$[\theta_j \mid \boldsymbol{\theta}_{(-j)}, \boldsymbol{X}]$$

をθ_jの全条件付分布(*full conditional distribution*)(＊1)とよぶ。これは$\boldsymbol{\theta}$のj番目の要素の分布であり、このときθ_j以外の全要素は固定され、かつデータ$\boldsymbol{X}$によって情報を与えられた条件にある。事後分布$[\boldsymbol{\theta} \mid \boldsymbol{X}]$と同じように、この全条件付分布は$[\boldsymbol{X} \mid \boldsymbol{\theta}][\boldsymbol{\theta}]$に比例するが、その違いは規格化定数が$[\boldsymbol{X}]$でなく$[\boldsymbol{\theta}_{(-j)}, \boldsymbol{X}]$になることである。

＊1　　条件付提案分布ともいう

周辺事後分布(marginal posterior distribution)$[\theta_j \mid \boldsymbol{X}]$または結合事後分布 (joint posterior distribution)$[\boldsymbol{\theta} \mid \boldsymbol{X}]$がわからない場合でも、全条件付分布$[\theta_j \mid \boldsymbol{\theta}_{(-j)}, \boldsymbol{X}]$の特定は$[\boldsymbol{X} \mid \boldsymbol{\theta}][\boldsymbol{\theta}]$を調べればしばしば容易である。この利便性とパラメータベクトルのひとつの要素を同時に抽出できる能力がギブスサンプリングを魅力的なものにしている。

ギブスサンプリングアルゴリズム

結合事後分布$[\boldsymbol{\theta} \mid \mathrm{X}]$からサンプルを抽出しようとしている、としよう。まず、$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)})$を固定する。そして、$t = 1, 2, \ldots,$ で$\boldsymbol{\theta}^{(t)}$を以下の規則に従って発生させる。

Step 1: $\boldsymbol{\theta}_1^{(t)}$を$[\theta_1 \mid \boldsymbol{\theta}_{(-1)}^{(t-1)}, X]$から抽出する

Step 2: $\boldsymbol{\theta}_2^{(t)}$を$[\theta_2 \mid \boldsymbol{\theta}_{(-2)}^{(t-1)}, X]$から抽出する

・・・

Step k: $\boldsymbol{\theta}_k^{(t)}$を$[\theta_k \mid \boldsymbol{\theta}_{(-k)}^{(t-1)}, X]$から抽出する

Step k+1: $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)})$を決定する

注釈：このアルゴリズムにおいて、それぞれのステップの後で連続して$\boldsymbol{\theta}^{(t)}$を更新し、次の全条件付分布を抽出するときに部分的に更新された$\boldsymbol{\theta}^{(t)}$を用いる方法も可能である。たとえば、$\boldsymbol{\theta}_3^{(t)}$は全条件付分布

$[\theta_3 \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \theta_4^{(t-1)}, \ldots, \theta_k^{(t-1)}, X]$からではなく、

$[\theta_3 \mid \theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \ldots, \theta_k^{(t-1)}, X]$から抽出することもできるということである。

例

ここで、$n = 10$の二項乱数（指数$N = 25$、成功確率pは平均μ、精度τのロジット

正規分布からの無作為抽出）である $\boldsymbol{\gamma}$ = (6, 7, 9, 9, 12, 13, 14, 15, 17, 18)を観測したとしよう。φ_i = logit(p_i)とする。; 推定は事後分布[$\boldsymbol{\theta}$ | $\boldsymbol{\gamma}$]に基づいている。ただし、$\boldsymbol{\theta}$ = (μ, τ, φ_1, φ_2,…, φ_{10})’

モデルを特定するためには、超パラメータ μ と τ の事前分布が必要である。徐々に明らかになるであろう理由から、Section 4.1.3 で議論した正規-ガンマ事前分布を選択する。つまり、事前分布として[μ | τ] = N (η_0,1/($\kappa_0\tau$))、[τ] = Ga(a_0, β_0)を指定する。

事後分布[$\boldsymbol{\theta}$ | $\boldsymbol{\gamma}$] ∝ [$\boldsymbol{\gamma}$ | $\boldsymbol{\theta}$][$\boldsymbol{\theta}$]は

$$\propto \Pi_{i=1}^{n}\{\text{expit}(\varphi_i)^{y_i}[1-\text{expit}(\varphi_i)]^{25-y_i} \times \tau^{1/2}\exp[-(\tau/2)\,(\varphi_i-\mu)^2]\}$$
$$\times\ \tau^{1/2}\exp[(-\kappa_0\tau/2)(\mu-\eta_0)^2] \times \tau^{\alpha_0-1}\exp(-\beta_0\tau) \qquad (4.8)$$

と表記できる。ここで、"expit"とはロジット変換の逆変換のことをいう。ゆえに、expit(φ_i) = p_iとなる。

この接合事後分布、もしくは周辺事後分布の計算は…論外である。つまり[$\boldsymbol{\theta}$ | $\boldsymbol{\gamma}$]はまったくひどいことになっているのである。しかしながら、ギブスサンプリングは公正で単純な方法である。我々は μ と τ と φ_i に関して全条件付分布の分布を特定する必要がある。これらのすべては式 4.8 と比例関係にある。

例えば、τ の全条件分布は式 4.8 の τ を含むすべての項に比例することだろう。簡便な表記で全条件分布を表すと、以下のようになる。

$$[\tau \mid \cdot] \propto \tau^{n/2+1/2+\alpha_0-1}\exp\{[-(1/2)\Sigma_{i=1}^{n}(\varphi_i-\mu)^2 - (\kappa_0/2)(\mu-\eta_0)^2-\beta_0]\tau\}$$

つまり言い換えれば、τ の全条件付分布は[τ | ・] = Ga(a_1, β_1)が成り立つということである。このとき、

$a_1 = (n+1)/2 + a_0$、かつ

$\beta_1 = \beta_0 + (1/2)\Sigma_{i=1}^{n}(\varphi_i - \mu)^2 + (\kappa_0/2)\,(\mu - \eta_0)^2$

である。

同様に、μ の全条件分布は式 4.8 における μ を含むすべての項に比例する。すると、以下のように表すことができる。

$$[\mu \mid \cdot] \propto \exp\{(-\tau/2)[\Sigma_{i=1}^{n}(\mu - \varphi_i)^2 + \kappa_0(\mu - \eta_0)^2]\}$$

これは、以下の形に変換できる。

$$[\mu \mid \cdot] \propto \exp\{(-\kappa_1\tau/2) \times (\mu-\eta_1)^2\}$$

このとき、

$$\kappa_1 = \kappa_0 + n$$

かつ、

$$\eta_1 = \frac{n}{n+\kappa_0}\bar{\phi} + \frac{\kappa_0}{n+\kappa_0}\eta_0$$

である。それゆえ、μ の全条件分布は $N(\eta_1, 1/(\kappa_1\tau))$である。

ここで留意すべきは、二組のパラメータ(μ,τ)'の事前分布と全条件付分布は両方共正規-ガンマ分布、$[\mu|\tau] = N(\eta, 1/(\kappa\tau))$、$[\tau] = Ga(a, \beta)$になるということである。この正規-ガンマ分布は共役ではない。共役性が成り立つためには事前分布と事後分布が同族分布になる必要がある。しかし、共役の様式に関する知識があれば全条件付分布を簡単に特定する事前分布選択法がおのずとわかる。原則として、ギブスサンプリングは共役分布族を利用することによって迅速に処理される。

あとは全条件付分布$[\varphi_i|\cdot]$を特定するのみである。もう一度、式 4.8 から関連する項を抜き出し、以下を得る。

$$[\varphi_i|\cdot] \propto \text{expit}(\varphi_i)^{\gamma_i}[1-\text{expit}(\varphi_i)]^{25-\gamma_i}\exp[(-\tau/2)(\varphi_i-\mu)^2] \qquad (4.9)$$

式 4.8 の φ_i を含まない他の部分は無視。; φ の関数としての式 4.8 に関して、それらは単なる定数であり、規格化定数に吸収される。残念なことに、式 4.9 残ったものはよく知られた分布ではなく簡単にサンプリングできない。μ と τ の全条件付分布は特定され、簡単に抽出されたが、$[\varphi_i|\cdot]$はできていない。

しかし、まだギブスサンプリングは実装することができる。これは式 4.9 の分布から抽出する方法を見つけるだけでいい。ひとつの可能性は棄却サンプリングを行うことである。変数変換の定理（Section 2.2.4）を用いることで、もし、$p \sim Be(\gamma, 25-\gamma)$ならば、そのときは $\varphi = \text{logit}(p)$が

$$j(\varphi) \propto \text{expit}(\varphi)^{\gamma}(1-\text{expt}(\varphi))^{25-\gamma} \qquad (4.10)$$

という密度関数をもつということがいえる。式 4.10 と式 4.9 の違いは項 $\exp[(-\tau/2)(\varphi_i-\mu)^2]$ のみであり、この項は 1 以上にならない。ゆえに、候補点 $p \sim Be(\gamma_i, 25-\gamma_i)$を抽出し、$\varphi = \text{logit}(p)$を計算し、そして確率 $r = \exp[(-\tau/2)(\varphi-\mu)^2]$に従い候補点を受容するか棄却するかを決定することによって棄却サンプリングを実装することができる。

$[\varphi_i|\cdot]$を抽出するためのこの棄却サンプリング法にはひとつの欠点がある。そこで、標本数 10 で解析を行った。: 時折、μ と τ の値は受容確率がとても低くなる。単一の候補点の時間の 34%は十分だった。; 第一と第二の候補点の時間の 49%は受容されたが、300 回に 1 回のケースで 1000 を超える候補点が必要だった。（＊１）この結果は平均して 25.6 の候補点が必要であり、長い計算時間がかかる。

＊１　よくわからない

もうひとつの有用な方法はギブスサンプリングの流れの中で[φ_i|・]を抽出するために M-H アルゴリズムを利用することである。式 4.10 を用いて候補点 φ_{cand} を発生させ、式 4.9 を標的分布として設定すると、MH 受容確率（式 4.7）は以下のように単純に表すことができる。

$$r = \exp[(-\tau/2)[(\varphi_{\text{cand}}-\mu)^2 - (\varphi_{\text{cuur}} - \mu)^2]]$$

それは、ちょうど MH アルゴリズムによる標本が独立標本よりも劣っていたということと同じように、この Metropolis-within-Gibbs アルゴリズムは rejection-within-Gibbs よりも劣ると予想する人がいるかも知れない。これを Fig. 4.9 で確かめた。ここでは、$\sigma = 1/\sqrt{\tau}$ の自己相関関数 ACF がサンプルデータとして表されている。Metropolis-within-Gibbs アルゴリズムを用いた ACF（上方の破線）は rejection-within-Gibbs の ACF（中間の実線）よりもゆっくりと減少していることがわかる。しかし、計算時間は rejection-within-Gibbs のほうが 16 倍かかっている。そこで、より相対的な比較をするために Metropolis-within-Gibbs アルゴリズムを用いた一本の連鎖の長さを 16 倍にし、16 番目ごとの標本以外を捨てることで 16 倍に連鎖を“薄める”必要がある。その結果、薄めた連鎖の自己相関は実質的に rejection-within-Gibbs よりも低いことがわかった。[11]

11. ありがちな誤解として、MCMC の出力結果は決まって薄めなければならない、と考えてしまうことがある。せっかくシミュレーションした結果をなぜ捨てようとするのか？メモリや格納領域の制限によって、もしくはこのセクションのように手法の比較のために、それが余儀なくされているときにだけ、このようなことをすることを勧める。あくまで、荒く控えめな精度で標的分布の特徴を推定するために値がほぼ独立であるとみなすことができるというくらいに、薄めた連鎖については考えるとよいだろう。

このセクションを通じて、多くの読者はサンプルデータの解析に必要な BUGS コード（Panel 4.2）の単純さに気づき、喜ぶだろう。自身の MCMC コードを書く際に要求されることと比較して、すべての重労働は自動的にやってくれるため、これを“コード”というには図々しいくらいだ。BUGS“コード”は特定の MCMC の手法を考えることなしにあるモデルを書き留めるホワイトボードのようなものである。

4.3.5 MCMC の診断法

すべての数値的手法は誤りを犯すことから免れることができない。数的最大化の手法であるニュートン・ラプソン法について考えてみよう。この手法はしばしば最大尤度の推定値を発見するために用いられる。あるひとつの変数をもつ関数 f を最大化するために、

x_1から開始し、次いで$n = 1, 2,\ldots,$について、以下を計算する。

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

ここで、f'、f''はそれぞれfを一階微分、二階微分したものである。上手く行けば、配列$\{x_n\}$はfの最大値(maximizer)に収束する。例えば、この手法を密度関数$X \sim N(\mu,\sigma^2)$に適用したとしよう。このとき$\{x_n\}$は最頻値μに収束しなければならない。そして、それが実現したとしたら、μを中心とした$(\sqrt{2}/2)\sigma$の範囲に初期値x_1を運良く設定できたということである。もし、$x_1 = \mu + (\sqrt{2}/2)\sigma$であったなら、配列は$\mu \pm (\sqrt{2}/2)\sigma$の範囲であちらこちらに永遠に動き回っていたことだろう。また、もし$x_1 > \mu + (\sqrt{2}/2)\sigma$であったなら、配列は$\pm\infty$に発散してしまうだろう。

幸運にも、ニュートン・ラプソン法におけるこのような誤りは普通一目でわかる。[12] MCMC に関しても同様の誤りを犯しうるが、その恐ろしいところはそれに気づかないかもしれないということである。

12. しかし、常にわかるわけではない。ニュートン･ラプソン法は広域的な最大値を探そうとしているときに、局所的な最大値から抜け出せなくなることがある。同様に、このようなものが存在するときでも、最終的に極小値に行き着くことについてはまったく問題ない。（＊１）

＊１　意味がうまく取れない

最低限でも、連鎖の標本経路は一度見るべきである。Fig. 4.10 の最初の図に似て”草のよう”であったなら、確実にとまでは行かないが、十中八九は結果がうまく行っている。一般的に、最初にチェックするべきことは余分なパラメータがあるかどうかを見て、モデルを正しく指定できているかを確かめることである。ほかにも、マルコフ連鎖の生成元が正しく調整できていなかったり、もしくはあらかじめ準備されたパッケージを使っている場合、試みようとしている問題がその能力を超えていたりするかもしれない。もしくは、良くない初期値をその連鎖に選んでしまったかもしれない。それは特に、自動的に初期値を発生させてくれるパッケージを使っているときや、拡散した（diffuse）事前分布を使っているときにおきうる。時には、公正な情報の事前分布（fairly informative priors）を用いて予備運転し、その結果得られた連鎖から値を抽出し、無情報解析の適当な初期値として利用することも役に立つことがある。

たとえ Fig. 4.10 の左側のように（＊１）草のような標本経路でもすべてうまく行ったということを保証しているわけではない。かつて、筆者は実行に数時間かかるような複雑で長いコードを書いたことがある。その結果は素敵な草のようなプロットだった。ただひとつの問題は再実行したときに結果が僅かに違っていたことだった。その矛盾は大きいものではなかったが、問題点を洗い出すには多くの手間がかかった。コードを詳細に調

べたストレスに満ちた数日後、我々はその問題を発見した。それは、コードの打ち間違いでランダムな初期値を与えるためのパラメータの一つである迷惑パラメータ（nuisance parameter）の更新をコメントアウトしてしまっていたということであった。それゆえ、この迷惑パラメータを実行する度にある異なる値に固定されていたのだった。ただ、それは数ある迷惑パラメータの中のひとつであったため、事後分布からのサンプルを観測することは妨げられることはなかった。

＊１“上側のように”の間違え

しかしながら、我々筆者は知らず知らずのうちに MCMC のよい演習を行っていたといえるだろう。この方法ははっきりと異なり拡散した初期値を持った複数の連鎖を実行し、それらの結果を比較するための良いアイデアである。BUGS に実装されている Brooks-Gelman-Rubin 診断法（Brooks and Gelman, 1998）はこのような比較をする一つの方法である。中心の 100（1-α）%の信用区間は個々の連鎖のデータを使って計算され、複数の連鎖をプールしたデータを用いて作られた区間と比較される。結果間の一致の度合いはそれらの定常分布に連鎖が収束する尺度が用いられる。これは、続く評価で適正な burn-in の区間のガイドラインも提供する。（＊１）

＊１　よくわからなかった

マルコフ連鎖を長くとることはどのような場合でも良い考えであり、長ければ長いほど良い。なぜなら、連鎖長が伸びるほど連鎖の特性が事後分布の特性により類似していくからである。[13]　近似の精度を評価する単一の基準というものはない。ただし、数個の簡単な選択肢から洞察を得ることができるかもしれない。一つ目は、観測値が実質独立とみなせるくらい ACF が小さくなったときのラグが k であるならば、長さ N の連鎖$\{X_i\}$は連鎖長を N/k まで薄めることができる。このため、標本平均の分散は $\sigma^2/(N/k) = k\sigma^2/N$ に近似される（$\sigma^2 = \mathrm{Var}(X_i)$）。よって、近似した標本平均の分散は連鎖全体の平均の精度についての控えめな尺度となる。

13　MCMC を事後分布の特徴を“推定(estimating)”すると表現することは表向きとしては正しいが、我々は“近似する(approximating)”という言葉を使うことを好む。なぜなら、“推定する”とは制限のある既知のデータセット、またはサンプルサイズのような資源を暗示するからである。しかしながら、この場合においては我々の時間や忍耐に制限を課すだけである。

精度を評価するもうひとつの手段は幾何曲線で ACF を近似する方法である。ラグ t の自己相関が ρ^t であるとき、標本平均の分散は以下のように近似できる。

$$\{(1+\rho)/(1-\rho)\}\sigma^2/N \qquad (4.11)$$

このとき σ^2 は連鎖の標本分散によって推定される。

さらにもうひとつの手段は BUGS に実装してある”batch means”(Roberts, 1996)を使う方法である。連鎖$\{X_i\}$を平均 n 個の連続した値でわけると、その平均が μ、分散が v/n であると仮定する。さらに長さ N の連鎖はほぼ独立なバッチ平均$\hat{\mu}_k$もつ長さ n の K 個の束(batch)に分割できるとしよう。そして、$K\to\infty$ となるとき、K 個の平均値$\hat{\mu}_k$間の標本標準偏差は一致推定量(consistent estimator) v/n となる。（＊１）ゆえに、v を以下のように推定できるだろう。

$$\hat{v} = n/(k\text{-}1)\,\Sigma_{k=1}^{\kappa}(\hat{\mu}_k - \bar{\mu})^2$$

このとき$\hat{\mu}$は連鎖全体の標本平均である。$\hat{V}/N$ によって Var($\bar{\mu}$)を推定できる。この量の平方根は“MC Error”と BUGS では説明されている。[14]

＊１　何をやっているかわからない

14.　ここで留意すべきは、バッチ平均を使う方法を説明するときに σ^2 よりも v を使って説明したことである。我々は連鎖の X_i の分散には σ^2 を使う。σ^2 は複数の X_i の標本分散によって自己相関にかかわらず一貫して推定される。v は式 4.11 異なり、$v = (1 + \rho) / (1 - \rho)\sigma^2$ とあらわせる。

実例

渡りをしてきた浜鳥に対しカブトガニ *Limulus polyphemus* の卵が食物としてどの程度重要かという研究の中で、Haramis *et al.* (2007)は捕獲した鳥類の体重量(x) に対する安定同位体（SI）比の関係 $y = \delta^{15}N$ を非線型モデルにあてはめた。測定値は以下の式に従うと仮定した。

$$y_i = A[1 - b\exp(-cx_i)] + \varepsilon_i$$

ここで、ε_i は分散 $\sigma_i^2 = a\exp(-\beta x_i)$、平均値がゼロである正規分布に従う独立な変数であるとした。A、b、c、β は 0 より大きいと仮定した。また、モデルは体重量の増加に従い、安定同位体比は増加して A に漸近し、漸近したときの個体間のバラつきは減少すると仮定した。

漸近線 A の事後分布をサンプリングするために、長さ 102.5 万のマルコフ連鎖を生成し、最初の 2.5 万個の値を burn-in として廃棄した。このプロセスを 2.4GHz のプロセッサ上で BUGS を走らせたとき、所要時間は 30 分程だった。連鎖が高い自己相関を持つことはこのモデルが複雑であるため、とくに驚くことではない。Fig. 4.11 にラグ 250 までのこの自己相関関数を示した（実線）。また、近似した幾何級数曲線 $f(t) = \rho^t$ (ρ = 0.9856)

も示した（破線）。この近似曲線のパラメータ ρ は対数変換した ACF をラグに回帰することで得た。式 4.11 を用い、$\mathrm{Var}(\bar{\mu}) \approx 137.6\ \sigma^2/N$ を得た。100 万個の標本分散 0.0431 代入すると、$\mathrm{Var}(\bar{\mu}) \approx 0.0024^2$ が得られ、推定値が小数点第二位までの精度があると結論づけられる。実際、A の事後分布の平均は 14.35 であった。

ここで、BUGS から MC error = 0.0029 であったことを指摘しておく。バッチ平均をもとにしたこの値はバッチのサイズに敏感に影響されるようだ。バッチサイズが 100、1000、5000 であるとき、それぞれ 0.0015、0.0029、0.0034 という値が得られた。重要なことは、これらの値は精度を示しているだけだと認識することである。サンプルを独立なものとみなす間違った結論 $s/\sqrt{N} = 0.0002$ よりも、すべての方法で明らかに低い精度が示めされた。 $(1+\rho)/(1-\rho) \approx 137.6$ という値は長さ 100 万の連鎖をより公正に評価している。つまり、精度が 1e6/137.6 ≈ 7300 の独立標本に匹敵するくらいであることをこの値は示している。こうして見ると、連鎖長 1e6 はそれほど無駄にはならなかったようだ。