3-3-2-4-3 マルコフ連鎖モンテカルロ法による民族学的差の解析について

（ref Jonathan K. Prichard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959, 2000）

1. 全体像
   (1) 扱うパラメタは以下のとおり。
      ① 既知パラメタ
         a)個人別・遺伝マーカー別 genotype
      ② 未知(推定)パラメタ
         a)クラスタ数
         b)クラスタ別・遺伝マーカー別頻度
         c)個人別所属クラスタ
   (2) 解決しようとしているのは
      ① 既知パラメタから、未知パラメタの期待値(その他信頼区間など)を求める。
   (3) 上記の問題は以下のような理由により、算術的に得られないことがしばしばである。
      ① 複雑な積分を用いる必要がある。
      ② 高次元の解である。
      ③ そもそも解析解が得られえない分布である。
   (4) したがって、算術的方法以外の解法が必要である。
2. 算術的方法以外の方法として採用される「マルコフ連鎖モンテカルロ法(MCMC)」の概説
   (1) モンテカルロ法とマルコフ連鎖を組み合わせた推定法のことである。集団の遺伝的構造解析のように、多数の未知パラメタの推定をおこなう場合には、MCMC 法を実践するにあたって、さらに工夫が必要で、Metropolis-Hastings(M-H)アルゴリズムを用いて定常分布(推定解)に近づく必要があり、さらに、M-H アルゴリズムの実行時に使用する関数(分布)には、Gibbs サンプラと呼ばれる条件付き分布(関数)を用いるのが通例となっている。この Gibbs サンプラからの乱数発生には Adaptive Rejection sampling という方法を併用する必要がある。
   (2) モンテカルロ法とは
      ① 統計学の実験的手法の１つである。乱数を発生させてシミュレーションを行い、確率的に解を得る方法である。適切な解を得るためには、乱数の発生に条件を設定する必要がある。その条件設定方法の一つがマルコフ連鎖である。
   (3) マルコフ連鎖とは
      ① ある変数を順次、発生させるときに、現世代のパラメタの値(のセット)のみをもとに次世代のパラメタの値を発生させる方法のことである。
   (4) Metropolis-Hastings アルゴリズムとは
      ① マルコフ連鎖を用いて、世代を順次進めていくときに、次世代のパラメタの値を得るためには、次世代のパラメタ値をある分布(関数)(サンプラと呼ぶ)から作り出し、その作られた値を次世代のパラメタの値として妥当かどうかを判断する、という手続きを踏むと未知パラメタの値(のセット)が定常状態(推定値)に向かうことが知られている。このような新世代のパラメタ値の作り方を M-H アルゴリズムと呼ぶ。

(5) Gibbs サンプラとは

① M-H アルゴリズムを実行する(現世代パラメタ値(のセット)から次世代のパラメタ値(のセット)を得る)ときに用いるサンプラ(分布(関数))のタイプの一つである。このサンプラの特徴は、

「複数のパラメタがあり、その個々のパラメタを推定するときに、自身を除くその他の既知・未知パラメタの現時点での値によって推定することができる、言い換えれば、その推定作業には現時点の自身の値は影響を与えない」

というものである。このような分布のことを「フル条件付き分布」という。このような条件は、個々のパラメタの新世代の値を１つ１つ作ればよいのでアルゴリズムとして簡単である。しかしながら、アルゴリズムとして簡単であるがゆえに、乱数の発生が困難であることが多く、次の Adaptive Rejection sampling を導入することが必要である。

(6) Adaptive Rejection sampling とは

① フル条件付き分布から直接乱数を発生させることは容易ではないので(複雑な積分計算を伴うことが多いため)、乱数を発生させることが容易な代理の分布を作成し、その 2 分布の関係を用いて乱数を発生させるという方法である。

3. SNP を用いた MCMC 法による遺伝的構造解析の実際

(1) パラメタ

① 既知パラメタ

a) 個人別・遺伝マーカー別 genotype X

- Xは要素数＝(人数)×(SNP 数)の行列である

② 未知パラメタ

a) クラスタ数（固定し、クラスタ数別にシミュレーションする）

b) クラスタ別・遺伝マーカー別頻度 P

- Pは要素数＝(クラスタ数)×(ＳＮＰ数)の行列である

c) 個人別所属クラスタ Z

- Zは要素数＝(人数)の１次元行列である

(2) 推定の流れ

① Zの初期値設定(以下の２方法で行える) Z_0

② 以下、マルコフ連鎖による推定の繰り返し(m=1,2,3,…は繰り返し回数)

a) X及びZ_{m-1}より、P_mを推定する

b) X及びP_mよりZ_mを推定する

③ Zの初期設定の影響がなくなった後の推定P_m、Z_mの分布が得られる。これがある観測データXが与えられたときのP及びZの推定値の分布である

(3) 推定の各段階の詳細

① Zの初期値設定(以下の２方法で行える) Z_0

a) ランダムに設定(指定クラスタに均等な確率で割り当て)

b) サンプリング地域などにより個人別に特定

② マルコフ連鎖

a) X及びZ_{m-1}より、P_mを推定する

- クラスタ c・SNP s のアレル頻度が t(0≦t≦1)の確率分布 Pr(t)は

  $Pr(t)=t^{n1}x(1-t)^{n2}/(Constant)$

  で与えられる。但し Constant = $\int_0^1 Pr(t)dt$ を満たす。

  これはβ分布であり、Dirichlet 分布の要素が２の場合に補正係数を(1,1)とした場合(Uniform distribution を仮定している)である。つまり、Prichard らのプログラム “structure”と同様の設定となっている。
- クラスタ別・SNP 別にアレルの本数を数える。
- アレル１の本数を n1、アレル２の本数を n2 とすると

  n1=n2=0 の場合と

  n1=1 または n2=1 の場合は

  分布 Pr(t)から直接乱数が発生できるので、その値を p の次世代の値として採用する

  それ以外の場合は

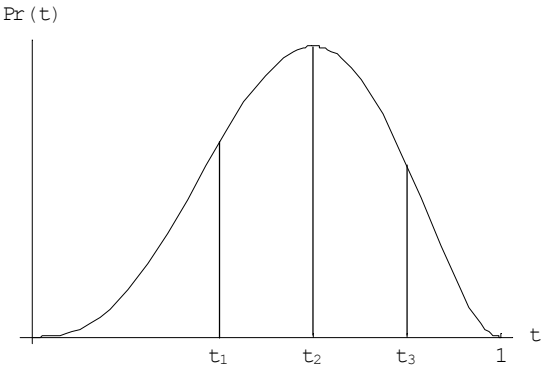
  Adaptive Rejection sampling を行って p の次世代の値を得る

  このとき必要な

  Π(t)≧Pr(t)を常に満たすような分布Π(t)は以下のようにして作成する。

  <Π(t)の作成>

  Pr(t)は下図のように描かれて、それは１次導関数、２次導関数の正負によって４区分に分けられる。



  区分１：0－t_1は下に凸の増加関数

  区分２：t_1－t_2は上に凸の増加関数

  区分３：t_2－t_3は上に凸の減少関数

  区分４：t_3－t_4は下に凸の減少関数

  例図は n1=3，n2=2 の場合である。

  但し、

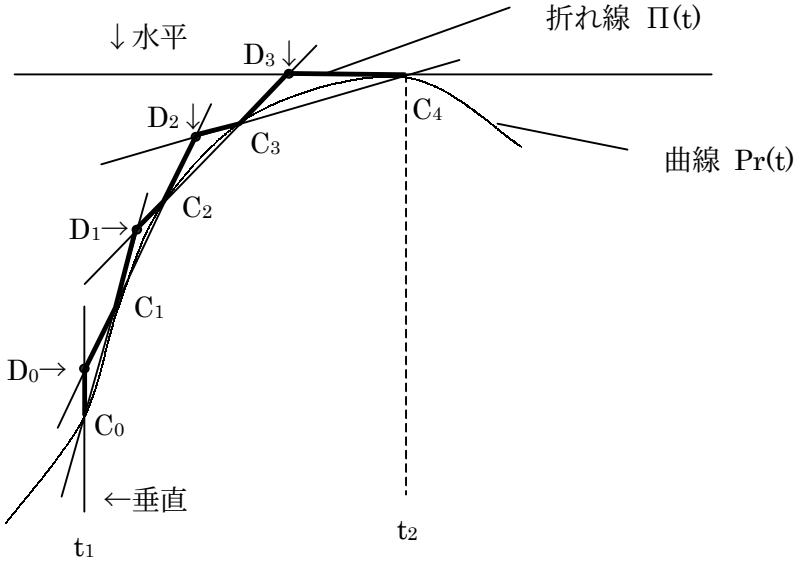
$$t_1 = \frac{n_1}{n_1 + n_2} - \frac{\sqrt{n_1 \times n_2 \times (n_1 + n_2 - 1)}}{(n_1 + n_2) \times (n_1 + n_2 - 1)}$$

$$t_2 = \frac{n_1}{n_1 + n_2}$$

$$t_3 = \frac{n_1}{n_1 + n_2} + \frac{\sqrt{n_1 \times n_2 \times (n_1 + n_2 - 1)}}{(n_1 + n_2) \times (n_1 + n_2 - 1)}$$

区分１と区分４とは下に凸なので、任意の点を結んだ直線がΠ(t)として採用できる。

一方、区分２と区分３とは、上に凸なので、Pr(t)上の 2 点(C_0,C_1…)を結ぶ直線は必ず Pr(t)そのものよりも小さい値をとる。したがって下図で→で指し示したように、Pr(t)上の隣り合う 2 点同士を結ぶ直線の交点として得られる点(D_0,D_1…) をΠ(t)の点として採用し、これと、Pr(t)上の点(C_0,C_1…)とを交互に結んだ折れ線をΠ(t)とする。



このようにして作成されたΠ(t)を用いて、乱数を発生させる。

Π(t)は有限個の点を結んだ折れ線なので、その積分は容易であり、したがって、このΠ(t)に比例する確率密度関数からは、乱数を発生させることは容易である。

今、Π(t)から発生させた乱数を r とする。

これとは別に[0-1]の一様乱数 q を発生させ、

q<Π(r)/Pr(r)

を満たした場合のみ r を採用すると

r は Pr(t)に比例する確率密度関数から発生させた乱数に一致する.

この採用された乱数が p の次世代の値である。

もし、

$q<\Pi(r)/Pr(r)$

が満たされなまった場合には、$\Pi(t)$の再作成をして、乱数を発生しなおす。そのときに、$Pr(t)$上の点$(t,Pr(t))$を新たに加えて、より細かい折れ線$\Pi(t)$を作成する。この新しい$\Pi(t)$は必ず、それよりも前の段階の$\Pi(t)$よりも$Pr(t)$に近く、乱数が採用される確率は高くなる。

b) X及びP_mよりZ_mを推定する

- 個人 i の genotype data から 1,…nc のそれぞれのクラスタの事後確率を求め、その比率に応じて、i の出身クラスタを推定する。
- 前の段階において、クラスタ別・SNP 別のアレル頻度は得られているからそれを $p_{(c,s)}$と表すと、個人 i がクラスタ c の出身である尤度 Pr(i from c)は

$$Pi(i \text{ from } c) = \prod_{s}^{\text{全てのSNP}} g_{(c,s)}$$ とする。

ただし、g(c,s)はクラスタ別・SNP 別の genotype 頻度であり、

個人 i の SNP s の genotype がアレル 1 のホモの場合は

$g_{(c,s)} = p_{(c,s)}^2$

genotype がヘテロの場合は

$g_{(c,s)} = 2 \times p_{(c,s)} \times (1 - p_{(c,s)})$

genotype がアレル 2 のホモの場合は

$g(c,s)=(1-p(c,s))^2$

で与えられる。

上記アルゴリズムを実行する上での詳細は、プログラムソース内の説明文を参照のこと。