ハプロタイプ関連解析への漸近理論

東京大 数理科学研究科　　鎌谷研吾

Introduction

本講演ではハプロタイプ関連解析の漸近的手法の正当性を考える.

ハプロタイプ関連解析において, Excoffier and Slatkin [1] は次のモデルを考えた. パラメータ空間は $\Theta = \{\theta \in \mathbf{R^m}; \sum_{\mathbf{i=1}}^{\mathbf{m}} \theta_\mathbf{i} = \mathbf{1}, \theta_\mathbf{i} > \mathbf{0}\}$ とし, 観測される確率変数列 $(X_n; n = 1, 2, \ldots, N)$ は以下を満たすとする.

$$X_n = Y_{1,n} + Y_{2,n} \tag{1}$$

ここで, $Y_{i,j}$ は独立同分布で, $P_\theta(Y_{i,j} = a') = \theta_a$ $a \in \{\mathbf{0}, \mathbf{1}\}^\mathbf{m}$ と分布が定められている. ただし, 表記の簡単のため, $\{0,1\}^m$ と $\{1, 2, \ldots, 2^m\}$ を同一視している. $(Y_{i,j})$ は観測されず, (X_i) のみが観測される. パラメータ θ の真値 θ_0 を推定する事が目的である. このモデルの特徴は, $(X_i, Y_{1,i}, Y_{2,i})$ の分布は指数型分布型なので扱いやすいものの, X_i の分布は複雑であることである.
以降, $X^{(N)} = \{X_i; i = 1, \ldots, N\}, Y^{(N)} = \{X_{i,j}; i = 1, 2, j = 1, \ldots, N\}$, および $(X^{(N)}, Y^{(N)})$ の分布を $p_N(dX^{(N)}, dY^{(N)}|\theta)$ および, $Y^{(N)}$ の $X^{(N)}$ が得られたもとでの条件付き確率を $p_N(dY^{(N)}|\theta, X^{(N)})$ と書く.

サンプルサイズを固定したアプローチ

まずはサンプルサイズを固定した場合の推定量の構成を考える. 導出可能な方法としてモーメント推定量がある. しかしこのモデルでは X の分布が複雑なため, モーメント推定量の属の中から, 適当な基準のもと, 有効な推定量を導出するのは難しい. 一方, ベイズ推定量 $\tilde{\theta}_N$ や最尤推定量 $\hat{\theta}_N$ は漸近的に有効な推定量であるが, 直接求めるのは難しく, Markov chain Monte Carlo (MCMC) 法や Expectation Maximization (EM) アルゴリズムによる近似計算が必要である. これらの近似計算は (X,Y) が指数分布族であることから, 簡単にアルゴリズムを適用できて, 計算機によって実際に近似計算が実行できる.

これらの近似計算の正当性は大きな問題である. MCMC 法は事前分布の作用で容易に Uniformly Ergodic であることを示せるので, 近似計算は正当化される. 一方, EM アルゴリズムは, サンプルサイズを固定した状況では, 最尤推定量への収束を示すのは簡単ではない. EM アルゴリズムによる数列が収束しても, その収束先が local maxima や saddle point である可能性がある. このため, 数値計算においては得られた収束列が本当に最尤推定量を近似できているかチェックをする必要がある. この手順は簡単ではなく, パラメータの次元が大きくなるといっそう難しくなる. 様々な初期値を試してみたり, 数値実験を繰り返す事も有効だが, 今回はこれら収束の得られた後の操作をせずに収束を正当化することを考える.

EM アルゴリズムの収束の漸近的なアプローチ

漸近的な議論では、一致性を持つ推定量を見つけることができれば, EM アルゴリズムの最尤推定量への収束は示すことができる. これは真値の近傍での, 中間量とよばれる関数の展開により得られる. 定理は次のように表現される：

定理 初期推定量の $\hat{\theta}_{N,0}$ は $(\hat{\theta}_{N,0} - \theta_0) = o_{P_{\theta_0}}(1)$ とする. $(\hat{\theta}_{N,i}; i = 1, 2, \ldots)$ は EM アルゴリズムにより得られる推定量の系列で, $\hat{\theta}_{N,i-1}$ が得られたもと, $\hat{\theta}_{N,i}$ は

$$\int \partial_\theta \log p_N(X^{(N)}, Y^{(N)}|\hat{\theta}_{N,i}) p_N(dY^{(N)}|\hat{\theta}_{N,i-1}, X^{(N)}) = 0 \tag{2}$$

の解であるとする. このとき, ある $0 < r < 1$ があって,

$$\Omega_N := \{\omega \in \Omega; |\hat{\theta}_{N,i} - \hat{\theta}_N| \leq r|\hat{\theta}_{N,i-1} - \hat{\theta}_N| \text{ for all i}\} \tag{3}$$

とおくと, $P^N_{\theta_0}(\Omega_N) \to 1$ である.

この定理はサンプルサイズを固定した下での最尤推定量への収束を言うわけではないが, 漸近的に収束することを示している. この定理が示された利点は, local maxima ではないことを煩雑な計算で示す必要がなく, いくつかの正則条件を示せば, 漸近的収束が正当化されることである. またいわゆるレート行列と呼ばれる, EM アルゴリズムの収束速度をあらわすと考えられている行列に意味を与える. これにより, EM アルゴリズムの高速化の議論が意味を持つ.
ただし, この帰結は EM アルゴリズムの最初の論文である Dempster et al [2] により, 既に直感的に知られていた事実であることに注意したい.

EM アルゴリズムと MCMC 法との比較

次に, EM アルゴリズムの収束と MCMC 法の収束の速さの違いは興味のある問題である. 先ほど Uniformly Ergodic を示したような, 事前分布の作用は意味を持たないが, 異なった方法で MCMC 法について局所的な漸近的収束を示すことができる. これにより EM アルゴリズムと MCMC 法の比較が可能である. 実際に EM アルゴリズムと同様の議論のもと, 収束性を示す事ができて, この収束レートは EM アルゴリズムと同等であることが示せる.
これらの結果から, ハプロタイプのモデルにおいて, EM アルゴリズムの漸近的な収束は正当化されるが, その収束は MCMC 法と同等であることがわかった. これらによる帰結, および詳細は当日発表する.

参考文献

[1] L. Excoffier, M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution. 12. 5. 921-927, 1995

[2] A. Dempster, N. Laird, D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977