ダイバージェンスを用いた有限混合分布モデルの均一性の検定

情報理工学系研究科数理情報学専攻　2年　48096224　中西 佑介
指導教員: 駒木 文保 教授

2011年2月2日

1　序論

均一性の検定は混合分布の研究の一分野であり，観測値が与えられた時に，それが均一な分布から生じたのか，それとも混合分布から生じたのかを検定する問題である．混合分布モデルとして

$$(1-w)f(x|\theta_1)+wf(x|\theta_2)$$

を考え，観測値 $X_1,\ldots,X_n$ はこの分布から生じたとする．w は混合比，θ_1,θ_2 は混合パラメータと呼ばれ，$0\leq w\leq 1$ であり，また $\{f(x|\theta):\theta\in\Theta\}$ はパラメータ θ で表される確率密度関数を持つ分布族である．均一性の検定とは, このモデルにおいて，帰無仮説

$$H_0: X_1,\ldots,X_n\sim f(x|\theta_0)$$

を，対立仮説

$$H_1: X_1,\ldots,X_n\sim(1-w)f(x|\theta_1)+wf(x|\theta_2)$$
$$(\theta_1\neq\theta_2, w\neq 0,1)$$

に対して検定することであり，遺伝学者達が，突然変異を引き起こす遺伝子の部分母集団の存在を証明したい時などに使われる．

既存の検定方法の一つとしては尤度比検定を用いた方法がある．均一性の検定は帰無仮説の元での識別不可能性のために正則条件を満たさず，漸近分布が良く知られた χ^2 型分布とはならないことが知られているが，それを改良した方法として，Chen らによる修正尤度比検定 (MLRT) がある．この方法では，$w\to 0$，$w\to 1$ で $P(w)\to-\infty$ となるようなペナルティー関数 $P(w)$ を従来の対数尤度に加えた

$$l_n(w,\theta_1,\theta_2)=\sum_{i=1}^{n}\log\{(1-w)f(x_i|\theta_1)+wf(x_i|\theta_2)\}+P(w)$$

を新しい対数尤度として考える．$\hat{\theta}_0$ を帰無仮説の元での θ_0 の最尤推定量，$\hat{\theta}_1$,$\hat{\theta}_2$ を対立仮説の元での θ_1,θ_2 の最尤推定量とすると，

$$M_n=2\{l_n(\hat{w},\hat{\theta}_1,\hat{\theta}_2)-l_n(1/2,\hat{\theta}_0,\hat{\theta}_0)\}$$

が MLRT の統計量となり，漸近分布は $1/2\chi_0^2+1/2\chi_1^2$ となる．

他にはシミュレーションを用いた方法があり，その一つに Charnigo らによる D 検定がある．この検定では

$$d(n)=\int\left\{\left((1-\hat{w})f(x|\hat{\theta}_1)+\hat{w}f(x|\hat{\theta}_2)\right)-f(x|\hat{\theta}_0)\right\}^2\mathrm{d}x$$

を統計量とする．式を見れば分かるように，帰無仮説のモデルに当てはめた場合と，対立仮説のモデルに当てはめた場合の差を L^2 距離を二乗したもので計算し，それを元に検定を行う方法であり，その棄却点は $n, f(x|\theta)$ に応じてシミュレーションにより定める．$f(x|\theta)$ が正規分布，ガンマ分布の場合，$d(n)$ を closed-form expression で表すことが可能である．しかし積分を解析的に解く事ができなくても，一次元の積分であるから数値的に時間をかけずに計算する事ができるので，他の尺度を用いても計算にかかる時間はほとんど変わらない．そこで，どのダイバージェンスを用いた時に最も精度良く検定を行うことができるのかについて調べた事が本研究のテーマとなる．

2　ダイバージェンスを用いた均一性の検定

D 検定では L^2 距離を二乗したものを用いていたが，今回の方法では，有名なダイバージェンスである α-ダイバージェンスと，L^2 距離を二乗したものも含まれる β-ダイバージェンスを用いて検定を行う．

2.1　α-ダイバージェンス

α-ダイバージェンスは，Csiszár の f ダイバージェンスからも得られ，同様に Bregman ダイバージェンスからも得ることができるダイバージェンスである．またその性質は Chernoff によって提議され，甘利 や他の研究者達によって幅広く調査，拡張が行われている．基本的な α-ダイバージェンスは，$\alpha\neq 0,1$ の時

$$\frac{1}{\alpha(\alpha-1)}\int\left\{p^\alpha q^{1-\alpha}-\alpha p+(\alpha-1)q\right\}\mathrm{d}x$$

となる．また $\alpha=1$，$\alpha=0$ の場合も，それぞれ上式の $\alpha\to 1$，$\alpha\to 0$ の極限値として定義され，

$$\int\left\{p\ln\frac{p}{q}-p+q\right\}\mathrm{d}x \quad (\alpha=1)$$
$$\int\left\{q\ln\frac{q}{p}-q+p\right\}\mathrm{d}x \quad (\alpha=0)$$

となり，α-ダイバージェンスは全ての α の値に対して定義される．

2.2 β-ダイバージェンス

β-ダイバージェンスは，江口らによって導入され，他の研究者達によって研究されているダイバージェンスである．基本的な β-ダイバージェンスは

$$\int\left\{p\frac{p^\beta-q^\beta}{\beta}-\frac{p^{\beta+1}-q^{\beta+1}}{\beta+1}\right\}\mathrm{d}x$$

で定義される．注目すべき点は $\beta=1$ の時，L^2 距離を二乗したものとなることである．また $\beta=0$，$\beta=-1$ の場合も，それぞれ上式の $\beta\to 0$，$\beta\to -1$ の極限値として定義され，

$$\int\left\{p\ln\frac{p}{q}-p+q\right\}\mathrm{d}x \quad (\beta=0)$$
$$\int\left\{\ln\frac{q}{p}+\frac{p}{q}-1\right\}\mathrm{d}x \quad (\beta=-1)$$

となる．

2.3 ダイバージェンスを用いた均一性の検定

具体的な検定方法は，D 検定と同じように行う．各 α，β に応じた $D_A^{(\alpha)}\Big((1-\hat{w})f(x|\hat{\theta}_1)+\hat{w}f(x|\hat{\theta}_2)||f(x|\hat{\theta}_0)\Big)$，$D_B^{(\beta)}\Big((1-\hat{w})f(x|\hat{\theta}_1)+\hat{w}f(x|\hat{\theta}_2)||f(x|\hat{\theta}_0)\Big)$ を統計量として定め，棄却点は，帰無分布における統計量の計算を一万回行い，そのシミュレーション結果から求める．α-ダイバージェンスの二次の近似は，g_{ij} をフィッシャー情報行列として

$$\begin{aligned}&D_A^{(\alpha)}\Big(q(x|\boldsymbol{\theta}+\mathrm{d}\boldsymbol{\theta})||q(x|\boldsymbol{\theta})\Big)\\ =&\frac{1}{2}\int q^{-1}(x|\boldsymbol{\theta})\Big(\frac{\partial}{\partial\theta_i}q(x|\boldsymbol{\theta})\Big)\Big(\frac{\partial}{\partial\theta_j}q(x|\boldsymbol{\theta})\Big)\mathrm{d}x\mathrm{d}\theta^i\mathrm{d}\theta^j\\ =&\frac{1}{2}g_{ij}\mathrm{d}\theta^i\mathrm{d}\theta^j\end{aligned}$$

となる．またまた $\beta>0$ の時の β-ダイバージェンスの二次の近似を求めると

$$\begin{aligned}&D_B^{(\beta)}\Big(q(x|\boldsymbol{\theta}+\mathrm{d}\boldsymbol{\theta})||q(x|\boldsymbol{\theta})\Big)\\ =&\int q^{\beta-1}(x|\boldsymbol{\theta})\Big(\frac{\partial}{\partial\theta_i}q(x|\boldsymbol{\theta})\Big)\Big(\frac{\partial}{\partial\theta_j}q(x|\boldsymbol{\theta})\Big)\mathrm{d}x\mathrm{d}\theta^i\mathrm{d}\theta^j\end{aligned}$$

となる．

3 数値実験

この章では，観測値を正規分布，指数分布，コーシー分布の混合分布から生じさせて，ダイバージェンスを用いた均一性の検定を行った．ダイバージェンスは $\alpha=2$, 1, 0, -1,$\beta=3$, 2, 1 の場合を用いた．ただし $\alpha=1$ の場合は $\beta=0$ の場合と一致する．また $\beta=1$ の場合は D 検定と同じである．観測値の個数は 50 個，100 個の 2 通りとし，各パラメータ w,θ の変化による違いを調べるために，混合比や混合パラメータを変えた場合でも数値実験を行った．棄却点は，帰無分布による 10000 回のダイバージェンスの計算から求める．結果は，各混合分布で 10000 回検定を行い，それを元に p 値の確率分布を求めたものである．見方としては，例えば棄却点を 5 パーセント点とすると，p 値が 0.05 以下となれば棄却される、つまり混合分布であると判定されることになる．そのため p 値が低い場所で確率密度が高くなっている方が良いということになる．

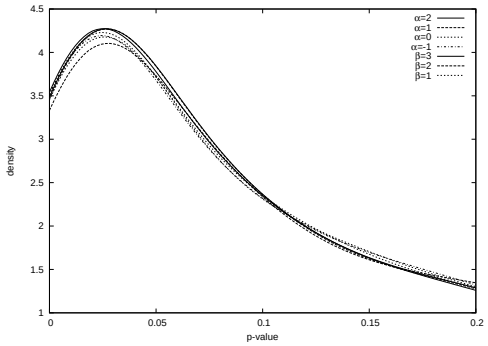

図 1: 数値実験の一例：$n=50$，正規分布

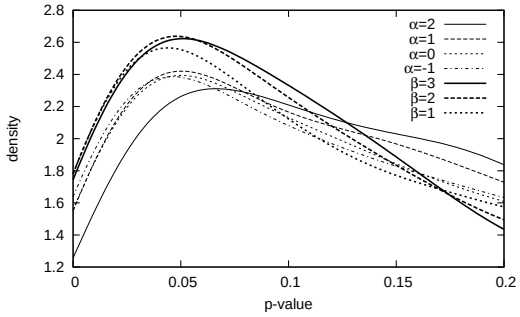

図 2: 数値実験の一例：$n=50$，コーシー分布

4 まとめ

数値実験の結果，正規分布，指数分布の場合はダイバージェンスによる違いは見られなかったが，コーシー分布の場合は明確な違いが見られた．この結果を元にダイバージェンスの二次の近似から考察を行った．