Robust Estimation under Heavy Contamination using Unnormalized Models

金森 敬文 (名大)　　藤澤 洋徳 (統数研)

概要

外れ値が混入している状況で，統計モデルのパラメータと外れ値の割合 (外れ値比) を推定する方法を提案する．提案法では，定数倍のパラメータを導入して拡張した統計モデルの下で，proper score を最小化することで推定量を得る．回帰分析への応用についても考察する．詳細は [1] を参照のこと．

1. Proper Score によるロバスト推定

統計的推論においてデータから確率分布を推定するとき，最尤法など score (scoring rule) に基づく方法がよく用いられる．Score は損失の期待値として定義される．データ x を確率密度 q をもつ分布で予測するときの損失を $\ell(x,q)$ とし，分布 p の下での期待損失を

$$S(p,q)=\int \ell(x,q)p(x)dx$$

とする．$S(p,q)$ を score という．適当な関数空間に属する任意の p,q に対して $S(p,q)\geq S(p,p)$ が成り立つとき，$S(p,q)$ を proper score という．データ $x_1,\ldots,x_n\sim_{\text{i.i.d.}} p$ に対して，経験分布による proper score の近似式 $\frac{1}{n}\sum_{i=1}^n \ell(x_i,q)$ を小さくするような q を選択することで，分布 p を適切に推定することができる．

統計的推論の目的に合わせて，さまざまな score が提案されている．データに外れ値が含まれることが想定される場合，外れ値の影響を抑えるような score を用いることでロバストな推定を行うことができる．ロバスト推定のための代表的な proper score を以下に示す．関数 $f(x)$ に対して $\langle f\rangle=\int f(x)dx$ とする．

Density-power score： パラメータ $\gamma>0$ に対して $S_{\mathrm{DP}}(p,q)=\gamma\langle q^{1+\gamma}\rangle-(1+\gamma)\langle pq^{\gamma}\rangle$.

Pseudo-spherical score： パラメータ $\gamma>0$ に対して $S_{\mathrm{PS}}(p,q)=-\langle pq^{\gamma}\rangle/\langle q^{1+\gamma}\rangle^{\gamma/(1+\gamma)}$.

上記の例では, 確率密度だけでなく非負値関数 f,g に対して $S(f,g)\geq S(f,f)$ が成立する．また γ について適切に極限をとれば最尤推定量に対応する score が得られる．Density-power score では $S_{\mathrm{DP}}(f,g)=S_{\mathrm{DP}}(f,f)$ のとき $g=f$ となる．一方，pseudo-spherical score では $S_{\mathrm{PS}}(f,g)=S_{\mathrm{PS}}(f,f)$ のとき $g\propto f$ となる．

データの分布は $p(x)=c_0p_0(x)+(1-c_0)w(x)$ と表せるとする．ここで $p_0(x)$ が推定すべき分布であり，$w(x)$ が外れ値の分布である．また $1-c_0$ が外れ値比であり，必ずしも微小量であることは仮定しない．分布 $p_0(x)$ に対して統計モデルを設定することは可能である場合が多いが，$w(x)$ を適切にモデル化することは困難である．また一般に c_0 は未知である．このような状況で，統計モデル $p_\theta,\theta\in\Theta$ を用いて p_0 を推定することを考える．p_0 は統計モデルに含まれ，$p_0=p_{\theta_0}$ が成り立つとする．分布のパラメータ θ_0 に加えて外れ値比 $1-c_0$ を推定するために，統計モデルとして拡張モデル $\mathcal{P}=\{cp_\theta|0<c\leq 1,\theta\in\Theta\}$ を用いる．このとき

$$S_{\mathrm{DP}}(p,cp_\theta)=S_{\mathrm{DP}}(c_0p_0,cp_\theta)-(1+\gamma)(1-c_0)c^{\gamma}\varepsilon_\theta \tag{1}$$

となる．ここで $\varepsilon_\theta=\langle wp_\theta^{\gamma}\rangle$ であり，これは $\theta=\theta_0$ の近傍で十分小さいと仮定できる．データの経験分布を $\widetilde{p}$ とする．外れ値比が微小量でない場合でも (1) の右辺第 2 項が十分小さいなら，$S(\widetilde{p},cp_\theta)$ を c と θ について最小化することで c_0 と p_0 を精度よく推定することができる．推定量を $\widehat{c},\widehat{\theta}$ とすると，適当な条件下で $(\widehat{c},\widehat{\theta})=(c_0,\theta_0)+O(\varepsilon_{\theta_1}^{1/2})+O_p(n^{-1/2})$ が成り立つ．ここで $\theta=\theta_1$ は $\min_{cp_\theta\in\mathcal{P}}S(p,cp_\theta)$ の最適解である．推定値を求めるための最適化計算では，$\min_{c>0}S_{\mathrm{DP}}(p,cp_\theta)=-\{-S_{\mathrm{PS}}(p,p_\theta)\}^{1+\gamma}$ という関係式を用いて $S_{\mathrm{PS}}(p,p_\theta)$ と $S_{\mathrm{DP}}(p,p_\theta)$ の 2 段階最小化に帰着させることで，不等式制約 $c\leq 1$ を除くことができる．

2. 回帰分析への応用

データ $\{(x_i,y_i)|i=1,\ldots,n\}$ から回帰関数 $f(x)$ や条件付き密度 $p(y|x)$ を推定する問題を考える．外れ値は y に対して混入し，x には混入しないと仮定する．したがってデータの分布は

$$p(x,y)=p(x)\{c_0(x)p_0(y|x)+(1-c_0(x))w(y|x)\}$$

と表現でき，$p_0(y|x)$ を推定することが目標となる．外れ値比 $c_0(x)$ は x に依存する状況を考える．$p_0(y|x)$ に対して統計モデル $p_\theta(y|x)$ を設定することはできるが，$c_0(x)$ や $w(y|x)$ にパラメトリックモデルを仮定することは困難である．真の分布 $p_0(y|x)$ は，確率密度 $r(z)$ から定義される位置尺度モデル $p_\theta(y|x)=r((y-f(x;\beta))/\sigma)/\sigma,\theta=(\beta,\sigma)$ に含まれるとする．このとき，S_{DP} と S_{PS} に対応する条件付き score を用いることで，$p_0(y|x)$ と外れ値比の期待値 $1-\int c(x)p(x)dx$ を精度よく推定することができる．さらに，外れ値比の推定結果を用いてデータに含まれる外れ値を検出することでができる．

References

[1] T. Kanamori & H. Fujisawa, Robust Estimation under Heavy Contamination using Unnormalized Models, Biometrika, to appear.