2019年6月5日　統計数理研究所　オープンハウス

principal variablesを用いたデータの次元における情報量損失の最小化手法

楠本　英子　　総合研究大学院大学　統計科学専攻　5年一貫制博士課程5年

【目的　及び　背景】

自動運転の技術やビッグデータ分析が目覚ましく発展を遂げる今日、逐次的な計算を素早く行う手法の開発の必要性が高まっている。その一つに観測変数自体を減らし、計算の効率化を図るということが考えられる。本研究では、特に条件つき確率分布に着目する。即ち、スパースなグラフィカルな構造を持つq次元部分ベクトルy_1（$q < p$）（以後、被説明変数とよぶ）とr(= p − q)次元部分ベクトルy_2（以後、説明変数とよぶ）が与えられ、幾つかの観測データから条件付き分布$p(y_2|y_1)$を推定することを考える。その際、情報量の損失を出来るだけ小さなものとするような形で、q′次元ベクトル y_1^*（y_1の部分ベクトル）、r′次元ベクトルy_2^*（y_2の部分ベクトル）を同時選択する手法について考察する。

【応用例】

本研究の具体例として、ロボットを使いサッカーを行なうロボカップが挙げられる。実際のサッカーにおいては、例えば選手がパスを行う際、選手はフィールド全体を見てパスをするのではなく、自身（ボール）の傍にいる選手たち、パスを渡したい味方の選手及びその周辺の選手たちの位置、動きを把握し、直感的にパスの出し方、即ち、方向、高さ、スピード等を判断しているものと思われる。一方、ボールを奪い取ろうとする敵方の選手はボールを持っている選手とその周辺の状況把握に集中し、次の行動を決めているものと思われる。前者の例においては、選手はフィールドにおける全てのプレイヤーの過去の位置情報からパスを行いたい味方の周辺の一部過去データを選択し、将来の一定期間の特定のプレイヤーの位置を推定し適切なパスを選択するものである。

【変数選択のための統計的手法】

今回、変数選択手法を提案するに当たり、従来から存在する主変数選択と共分散選択の2つの手法を組み合わせる。

主変数選択

Xを、p変量からなるベクトルとする。Xは、平均は0とし、分散共分散行列をΣと表わす。ここで、Xの中で残す変量をr個とする。残す変量からなるr次元ベクトルをX_1、消去する変量からなるp-r次元ベクトルをX_2で示す。これに伴い、Σも以下のように表示する。

①　$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Σ_{11}はX_1に対応する分散共分散行列である。ここで、行列式$|\Sigma_{11}|$を最大にすることを目的とする。即ち、選択基準は

②　$$\max|\Sigma_{11}|$$

となる。

共分散選択

この手法は、偏相関行列の一部をゼロとすることで、グラフィカル的な構造を考えた際、変数間の直接的なリンクを消していくものである。データの相関行列をR、偏相関行列をPとし、まず、相関行列から偏相関行列の算出を行うが、その手順は以下の通りである。

最初に、相関行列の逆行列R^{-1}を求める。ここでR^{-1}の(i,j)成分をr^{ij}とする。

次に、偏相関係数$r_{ij\cdot rest} = -r^{ij}/\sqrt{r^{ii}r^{jj}}$を計算する。これを(i,j)成分として持つ偏相関行列をPとする。

Dempsterは、一部の$r_{ij\cdot rest}$の値を零とすることで、偏相関行列の疎化を行う方法を述べている。その手法としては、RからPを上記手順で算出し、最も値の小さい成分$r_{i'j'\cdot rest}$を0した行列をP’とする。その後、新たに得られた行列P’を用いて相関行列Rを更新させる。この際、P’で$r_{i'j'\cdot rest} = 0$としたことによりRも変化するが、その際Rの(i', j')成分のみが大きさは変わるという条件を入れ、R’を更新する。更に、更新させたR’から再びP”を求めるという作業を行列式の値がある一定の基準を満たすまで繰り返す。

この2つの手法はサッカーの例では主変数選択がパスの話、共分散選択がボール追う側の考え方に対応している。

【情報量】

情報量としては、次のような形で定義されるエントロピーを使用する。

③　$$S = -\int_{-\infty}^{\infty} p(x)\, lnp(x)\, dx$$

例えば、r次元正規分布$N(\mu, \Sigma)$のエントロピーは

④　$$S = -\int_{-\infty}^{\infty} p(x)\, lnp(x)\, dx$$

⑤　$$= \frac{1}{2}\ln|\Sigma| + \frac{r}{2}\ln(2\pi e)$$

と計算される。

ある操作によりエントロピーがSからS’に変わったとき、エントロピーの差

⑤　S-S’

を情報量の減少分とする。Hoodaは主変数選択の基準として、この情報量基準を導入した。元のp次元ベクトルをV、次に残す変量からなるr次元ベクトルをWとする。ここである行列Aを選び、W=AVが満たされるとする。今、p次元ベクトルVはV~N(μ, Σ)であると仮定する。この場合、W~N(Aμ, A′ΣA)となる。このようにすれば、VからWの変換によって損失するエントロピーは、

⑥　$$\frac{1}{2}\ln(|\Sigma|) + \frac{p-r}{2}\ln(2\pi e) - \frac{1}{2}\ln(|A'\Sigma A|)$$

となる。

【提案手法】

本研究で提案する手法を簡単に述べると、以下の2つのステップで構成されている。

Step1　被説明変数・説明変数の削減：McCabeの主変数選択を使用

Step2　グラフィカルモデルの簡素化：Dempsterの共分散選択を使用

更に、【Step1】を以下のように2つの段階にわける。

イ. 被説明変数の削減：被説明変数y_2(r次元)の分散共分散に主変数選択を適応させ、次元削減したr' 次元ベクトルy_2^*を選ぶ。

ロ. 説明変数の削減：説明変数y_1 を与えた場合のy_2^*の条件つき分布を考え、主変数選択を用い、q次元ベクトルy_1 からq′次元の部分ベクトルy_1^*を選ぶ。

【Step2】ではよりグラフ構造を疎にしていく中で、被説明変数とリンクを持たない説明変数及び、説明変数とリンクない被説明変数を除外する。

この、情報量の損失を最小限に抑えることを目指す。

【情報量の損失】

被説明変数を減らす場合

r個の被説明変数をr'個に減らす。

この際、分散共分散行列Σ_2 からΣ_2'に変わったとする。

⑥より元々エントロピーが

⑦　$$S = \frac{1}{2}ln|\Sigma_2| + \frac{r'}{2}\{ln(2\pi e)\}$$

であったが、変換後エントロピーは

⑧　$$S' = \frac{1}{2}ln|\Sigma_2'| + \frac{r'}{2}\{ln(2\pi e)\}$$

となるので、情報量の損失は以下のようになる。

⑨　$$S = \frac{1}{2}ln\left(\frac{|\Sigma_2|}{|\Sigma_2'|}\right) + \frac{r-r'}{2}\{ln(2\pi e)\}$$

説明変数を減らす場合

この場合、説明変数自体の分散共分散行列でなく、説明変数を与えた場合、前工程で選んだ被説明変数の分散共分散行列を考える。この場合は、説明変数を減らすことにより分散共分散行列の値は大きくなり、その値が小さいものを選ぶ。

元の分散共分散行列を$\Sigma_{22.1}$、主変数選択により得られた分散共分散行列を$\Sigma_{22\cdot 1}'$とすれば

変換前のエントロピー

⑩　$$S = \frac{1}{2}ln|\Sigma_{22.1}| + \frac{r'}{2}\{ln(2\pi e)\}$$

変換後のエントロピー

⑪　$$S' = \frac{1}{2}ln|\Sigma_{22.1}'| + \frac{r'}{2}\{ln(2\pi e)\}$$

である。目的変数の数は変わらない。説明変数の次元は$q \to q'$ に変わる。故に情報量の損失は次の通りになる。

⑫　$$S - S' = \frac{1}{2}ln\left(\frac{|\Sigma_{22.1}|}{|\Sigma_{22.1}'|}\right)$$

【シミュレーションについて】

本発表では、被説明変数を3つ(y_1, y_2, y_3)、説明変数を5つ$(x_1, x_2, x_3, x_4, x_5)$からなる仮想条件付き分布$p(y_1, y_2, y_3|x_1, x_2, x_3, x_4, x_5)$を用いる。

今回は以下の流れで計算を行う。

イ. 主変数選択を使い、被説明変数を1つにする。

ロ. 主変数選択を使い、説明変数を3つにする。

ハ. 共分散選択を使い、説明変数を1つ消去する。

【結果】

イ. について

被説明変数(y_1, y_2, y_3)の分散共分散行列は以下のように仮定する。

$$\Sigma = \begin{pmatrix} 400 & 300 & -420 \\ 150 & 625 & 300 \\ -420 & 300 & 900 \end{pmatrix}$$

よって、最大の分散900を持つy_3が選ばれる。ここでのエントロピー損失は、$|\Sigma| = 2.07 \times 10^7$なので⑨より、

$$S = \frac{1}{2}ln\left(\frac{2.07 \times 10^7}{900}\right) + \frac{3-1}{2}\{ln(2\pi e)\} = 7.17$$

である。

ロ. について

条件付き分布$p(y_3|x_1, x_2, x_3, x_4, x_5)$を考える。この時の分散共分散行列は

$$\begin{pmatrix} 64 & & & & & \\ 40 & 100 & & & & \\ 48 & 15 & 225 & & & \\ -86.4 & -162 & -81 & 324 & & \\ -20 & -100 & -112.5 & 315 & 625 & \\ -216 & -60 & -225 & 216 & 75 & 900 \end{pmatrix}$$

とする。この時のエントロピーは3.19である。この中から3つの変数を選んだ場合（x_1, x_2, x_4)が選ぶことができ、その時のエントロピーは3.21となり、その変化は0.06と小さい。

ハ. について

条件付き分布$p(y_3|x_1, x_2, x_4)$の相関行列R及び偏相関行列Sは

$$R = \begin{pmatrix} 1 & & & \\ 0.5 & 1 & & \\ -0.6 & -0.9 & 1 & \\ -0.9 & -0.2 & 0.4 & 1 \end{pmatrix}, S = \begin{pmatrix} - & & & \\ -0.616 & - & & \\ -0.309 & 0.875 & - & \\ 0.938 & -0.687 & -0.459 & - \end{pmatrix}$$

である。この時のエントロピーは7.56である。

なお、4行目が被説明変数に対応している。つまり偏相関行列4行目のいずれかの値が零になるまで共分散選択を数度適応させると以下のようになる。

$$R = \begin{pmatrix} 1 & & & \\ 0.324 & 1 & & \\ -0.36 & -0.9 & 1 & \\ -0.9 & -0.36 & 0.4 & 1 \end{pmatrix}, S = \begin{pmatrix} - & & & \\ 0 & - & & \\ 0 & 0.884 & - & \\ 0.884 & 0 & -0.476 & - \end{pmatrix}$$

このように2番目の説明変数を消すことが出来、エントロピーは9.66に変化した。今後はエントロピーの変化についての考察を行い、2つのステップの関連性を調べたい。