spike and slab 事前分布を用いた罰則付き回帰

大阪大学大学院基礎工学研究科　田辺 竜ノ介

観測 $\boldsymbol{y} = (y_1, \ldots, y_n)$ が回帰モデル $y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i$, $(i = 1, \ldots, n)$ から観測される．ただし，$X_1, \ldots, X_p$ は説明変数，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ は回帰係数ベクトル，$\varepsilon_i (i = 1, \ldots, n)$ は互いに独立で中央値が 0 の未知分布である．

回帰係数の選択と推定を行う場合，スパース推定が利用できる．スパース推定は最小自乗法に罰則項を加えることで過適合を防いでいる．特に罰則項に $\boldsymbol{\beta}$ の L_1 ノルムを採用すると解がスパースになりやすい．言い換えると不要な変数を 0 にして必要な変数だけを残すことが出来る．しかし，罰則項を用いたスパース推定では推定量の分散が複雑になり，信頼区間の構成が困難であるという問題点を持つ (Knight and Fu, 2000).

Thibshirani(1996) により Lasso 推定量は尤度に正規分布，事前分布にラプラス分布を置いた事後分布の最頻値と一致することが示されている．この性質を用いることで，Park and Casella(2008) はベイズ的な解釈に基づいた Bayesian Lasso を提案した．これにより容易に信用区間を構成できる．さらに分散を考慮にいれた推定量が得られる．頻度論のスパース推定の方法に基づき，group lasso や fused lasso, elastic net に対してベイズ的な再構築が行われた．

しかし，ベイズ的な再構築を行った罰則付き回帰モデルは解のスパース性が失われているという問題点がある．この問題は Bayesian Lasso に始まるベイズ的なスパース推定量が，推定量に事後期待値を使用してることに起因する．同様な問題は罰則付き分位点回帰のにおいても議論されている．Li(2010) では L_1 罰則付き分位点回帰をベイズ的に解釈を行った手法が提案されているが，推定量が厳密に 0 を返さない問題が生じている．

本講演ではベイズ的手法に基づき線形回帰，分位点回帰ともに真に解がスパースな解を返せる推定量の構築を目指す.

spike and slab 事前分布はパラメータに対して点確率と連続分布の混合分布を設定するもので以下のように書き表される．

$$\pi(\beta|w) = (1 - w)\delta_0(\beta) + w\gamma(\beta).$$

$\delta_0(\cdot)$ はデルタ関数であり $\gamma(\cdot)$ は連続関数である．w $(0 \leq w \leq 1)$ は混合度を表す.

spike and slab 事前分布と事後中央値を用いて回帰係数の推定量に厳密に 0 を返すことが可能となる．これにより頻度論での罰則付き回帰モデルと同様に変数選択と係数の推定の両立が可能となる．加えて，spike and slab 事前分布を設定することで事後中央値が threshold 性を持つこと，モデルの一致性と漸近正規性があることを示す.

参考文献

[1] Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, **28** 1356-1378.

[2] Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, **5**(3), 533-556.

[3] Park, T., and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681-686.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, **58**(1): 267-288