Bayesian Sparse Regression Modeling

数学専攻　保科 架風
HOSHINA Ibuki

1　はじめに

近年の計算機関連技術の発展により, 生命科学, 金融工学, 情報科学や情報通信関連産業などの諸科学・産業界において大規模・複雑な構造を持つデータが獲得・蓄積されている. 統計科学には膨大な情報から有用な知見を効率的に抽出し, 新たな価値創造へと結びつく分析手法開発が強く期待されている. 多くの情報を内包するデータに基づく現象のモデル化を可能とするスパース回帰モデリングは, 現在最も注目を集めている統計的モデリング手法の 1 つである.

スパース回帰モデリングには, 現象の結果と多数の要因を結びつける回帰モデルにおいて, 要因変数に対応する回帰係数を厳密にゼロと推定する特徴がある. これにより, 一部の要因から結果への影響を完全に遮断することが可能となる. この問題はこれまで最適な変数の組合せを探索する「変数選択」に則って行われてきたが, 大規模モデルへの従来の変数選択法の適用には理論的な問題と大きな計算コストの存在が知られており, これらの問題の打開策としてスパース回帰モデリングには大きな期待が集まっている.

しかし, スパース回帰モデリングはモデルの推定と選択を同時に実行できる優れた手法ではあるが, 多くのスパース回帰モデリングでは回帰係数の L_1 ノルムなどの微分不可能な正則化項を含む最適化問題によって回帰係数の推定を行うため, 推定値の解析解が得られず, また, 正則化項の強さを適切に決定することが回帰係数の推定や変数選択において本質的となる. これらの問題に対し, 推定では LARS (Efron *et al.*, 2004) などの有効な推定アルゴリズムが提案され, また, モデル選択ではモデルの有効自由度などのモデル選択基準の研究が行われている.

Park and Casella (2008) は, 代表的なスパース回帰モデリングである lasso (Tibshirani, 1996) に対し, ベイズアプローチに基づく Bayesian lasso を提案した. これは, 正則化推定をベイズモデルとして捉えたもので, lasso の推定やモデル選択の問題をベイズモデルのパラメータの推定や選択の問題として取り扱うことを可能にした. ベイズ統計学では非ベイズな手法とは異なり, モデルのパラメータの値ではなく非ベイズな手法が考慮しない事後分布 (事前情報とデータの情報を組合わせることで得られるパラメータの分布) を推定する. これにより, モデルパラメータの推定の信頼性についても評価することができ, より柔軟かつ定量的な判断を可能にする. ベイズスパース回帰モデリング (Bayesian Sparse Regression Modeling) も同様であり, Bayesian lasso は lasso の事後分布を獲得することを可能にした.

しかしながら, Bayesian lasso にはいくつかの問題点がある. まず, Bayesian lasso では事後分布を解析的に陽に推定できず, 数値的な手法による推定が必要となる. これにより, Bayesian lasso によって得られる回帰係数の推定値は厳密に 0 とならず, スパース回帰モデリングの最大の特徴を有さない (Bayesian lasso の sparse 性の欠如). また一般に, 統計モデルの精度はパラメータの推定精度や予測精度によって評価されるが, Bayesian lasso では事後確率や周辺尤度の最大化を基準にモデルの推定・選択を行っており, これらはモデルの予測精度を直接的に評価しない. さらに, Bayesian lasso の事後分布は調整パラメータの値が小さくなるにつれ分散が大きくなり, 回帰係数の推定値の推定変動を大きくしてしまう.

これらの問題点に対し, 本論文では “Sparse Algorithm” の提唱によって Bayesian lasso に sparse 性を与えることができた. また, ベイズ予測分布に基づく Bayesian lasso のモデル選択基準を提案し, 予測精度に基づく Bayesian lasso のモデル選択を可能にした. さらに, 事後分布の積分計算に対して離散近似とマルコフ連鎖モンテカルロ法を融合し, より安定的に Bayesian lasso における回帰係数の推定値を求める手法を提案した.

2 Bayesian sparse regression modeling

目的変数 y と p 次元説明変数ベクトル $\boldsymbol{x}$ に対するサンプルサイズ n の線形回帰モデル $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ を考える. ただし, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T : n \times p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T : p \times 1$ であり, 誤差ベクトル $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ に対しては n 次元ガウス分布 $N_n(\boldsymbol{0}, \sigma^2 I_n)$ を仮定する. また, 一般性を失うことなく目的変数と説明変数は $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = n$ $(j = 1, \ldots, p)$ と基準化されているものとする. このとき, $\boldsymbol{\beta}$ に対する lasso 推定値は

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} := \underset{\boldsymbol{\beta}}{\text{argmax}} \left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{1}$$

で与えられる. ただし, λ は正値の調整パラメータであり, λ を適切に選ぶことで $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ の幾つかの成分を厳密にゼロに推定することが可能となる.

一方, Lasso をベイズモデルに拡張した Bayesian lasso (Park and Casella) は $\boldsymbol{\beta}$, σ^2 に事前分布

$$p(\boldsymbol{\beta}, \sigma^2) = \int \cdots \int \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j}\right) \cdot \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right) d\tau_1^2 \cdots d\tau_p^2 \cdot \text{IG}\left(\sigma^2 | \frac{\nu_0}{2}, \frac{\eta_0}{2}\right) \tag{2}$$

などを与えたときの事後分布を, 以下の全条件付モデルから反復抽出した確率サンプルによって推定する手法として提案された.

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &\sim N_p(\boldsymbol{\beta} | (A^{(k)})^{-1} X^T \boldsymbol{y}, \sigma^{2(k)} (A^{(k)})^{-1}), \\ \sigma^{2(k+1)} &\sim \text{IG}(\sigma^2 | \nu_1, \eta_1^{(k+1)}), \\ (1/\tau_1^{2(k+1)}, \ldots, 1/\tau_p^{2(k+1)}) &\sim \prod_{j=1}^{p} \text{IGauss}(1/\tau_j^2 | \mu_j'^{(k+1)}, \lambda^2), \quad k = 0, 1, \ldots, . \end{aligned} \tag{3}$$

ただし, $\text{IG}(x|\nu, \eta)$ は形状パラメータ ν, 比率パラメータ η を持つ逆ガンマ分布の確率密度関数, $A^{(k)} = X^T X + (D^{(k)})^{-1}$, $D^{(k)} = \text{diag}(\tau_1^{2(k)}, \ldots, \tau_p^{2(k)})$, $\nu_1 = (n + p + \nu_0)/2$, $\eta_1^{(k+1)} = \{\|\boldsymbol{y} - X\boldsymbol{\beta}^{(k+1)}\|^2 + (\boldsymbol{\beta}^{(k+1)})^T (D^{(k)})^{-1} \boldsymbol{\beta}^{(k+1)} + \eta_0\}/2$, $\mu_j'^{(k+1)} = \sqrt{(\lambda^2\sigma^{2(k+1)})/(\beta_j^{(k+1)})^2}$ であり, $\text{IGauss}(x|\mu, \lambda)$ は平均 μ, 形状パラメータ λ の逆ガウシアン分布の確率密度関数である.

2.1 Bayesian lasso 点推定値のスパース化

Bayesian lasso における $\boldsymbol{\beta}$ の点推定値 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ としては, カーネル密度推定によって求めた事後モードや事後中央値, 事後平均を用いることが可能である. しかし, そのどれも値が 0 となる成分を持たない. 理論的には Bayesian lasso の事後モードは lasso 推定値と同じではあるが, lasso が推定アルゴリズムの中で幾つかの成分が 0 になるような工夫を施している一方, Bayesian lasso の事後モードは数値的にモードを推定したものであり, 成分が厳密に 0 になる性質を有さない. また, 事後中央値や事後平均はそもそも lasso 推定値とは異なるものであり, スパースになる根拠を持たない.

そこで本論文では, 事後密度の大きさに基づいて $\hat{\beta}_j$ を 0 に修正するアルゴリズム (Sparse Algorithm, 表 1) を提案することでこの問題を解決した. このアルゴリズムでは, 事後モードの推定精度の良さは事後密度の大きさによって評価できるとの考えに則り, 回帰係数ベクトルの推定値の全ての成分に対し, それぞれに 0 の値を代入することで事後密度が上昇するか否かによってその成分の推定値を 0 とするかを決定している. これにより, Bayesian lasso の事後分布の推定に一切影響を与えることなく, 点推定値にスパース性を付与することを可能にした.

表 1　Sparse algorithm.

Sparse algorithm	
1.	Estimate the coefficient vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$
2.	$\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T \leftarrow \hat{\boldsymbol{\beta}}$
3.	For $j = 1, \ldots, p$, set $\tilde{\beta}_j \leftarrow 0$
3.1	if $g(\tilde{\beta}, \hat{\boldsymbol{\xi}}, \boldsymbol{y}) \geq g(\hat{\beta}, \hat{\boldsymbol{\xi}}, \boldsymbol{y})$ then $\hat{\beta}_j \leftarrow \tilde{\beta}_j$
3.2	else $\hat{\beta}_j \leftarrow \hat{\beta}_j$
	where $g(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{y}) = \log f(\boldsymbol{y}\|\boldsymbol{\beta}, \boldsymbol{\xi}) + \log \pi(\boldsymbol{\beta}, \boldsymbol{\xi})$,
	$f(\boldsymbol{y}\|\boldsymbol{\beta}, \boldsymbol{\xi})$ is a likelihood, $\pi(\boldsymbol{\beta}, \boldsymbol{\xi})$ is a prior on $(\boldsymbol{\beta}, \boldsymbol{\xi})$,
	and $\hat{\boldsymbol{\xi}}$ is point estimates of the parameter vector, $\boldsymbol{\xi} = (\sigma^2, \tau_1^2, \ldots, \tau_p^2)^T$.

2.2　予測分布に基づく Bayesian lasso の調整パラメータの決定

ベイズモデリングにおける調整パラメータの決定は主にベイズモデルを階層化したときの事後確率の最大化 (maximum a posteriori; MAP) 推定, あるいは, 事後分布に含まれる回帰係数や誤差分散などのモデルパラメータを周辺化して得られる周辺尤度の最大化を通して行われる. MAP 推定はモデルのデータへの当てはまりを評価する尤度とモデルパラメータの事前確率の積を評価するため, データへの過度な適合 (過適合) を避けることが可能であり, また, 周辺尤度最大化は尤度と事前確率の積をモデルパラメータに関して積分したものを評価するため, MAP 推定よりもさらにデータへの過適合を避けることが可能であるといわれる (Murphy, 2012). この性質により, これらの手法は予測精度の高いモデルが導出可能な手法ともいわれるが, あくまでも結果的に予測精度が向上するのであって, 予測精度自体を評価しているものではない.

ベイズモデリングにおける予測精度はベイズ予測分布 $h(\boldsymbol{z}|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$ によって捉えることが可能である. ただし, $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ は将来のデータ, $f(\boldsymbol{z}|\boldsymbol{\theta})$ は $\boldsymbol{z}$ の尤度, $\boldsymbol{\theta}$ はモデルパラメータベクトル, $p(\boldsymbol{\theta}|\boldsymbol{y})$ は $\boldsymbol{\theta}$ の事後分布である. これに対し, Kitagawa (1997) はベイズ予測分布と $\boldsymbol{y}$ の真の分布 $q(\boldsymbol{y})$ との間の Kullback-Leibler 情報量 (Kullback and Leibler, 1951) を評価する PIC 基準,

$$\mathrm{PIC} = -2 \log h(\boldsymbol{y}|\boldsymbol{y}) + 2B_p \tag{4}$$

を提案した. ただし, B_p はバイアス項であり, $B_p = \mathrm{E}_{q(\boldsymbol{y})}[\log h(\boldsymbol{y}|\boldsymbol{y}) - \mathrm{E}_{q(\boldsymbol{z})}[\log h(\boldsymbol{z}|\boldsymbol{y})]]$ である.

Bayesian lasso においてもこの PIC 基準を適用すれば, ベイズモデリングにおける予測精度の観点で調整パラメータを選択することが可能となる. しかしながら, $\boldsymbol{\beta}$ の事前分布であるラプラス分布が正規尤度に対して共役ではないため, Bayesian lasso に対する PIC 基準の導出には困難な点が存在する. そこで本論文では Bayesian lasso のラプラス事前分布を Kullback-Leibler 情報量の意味で最も近い正規尤度と共役な事前分布 (正規分布) で近似し, 以下の Bayesian lasso に対する近似 PIC 基準を導出した (Kawano *et al.*, 2015):

$$\begin{aligned} \mathrm{aPIC}(\lambda) = &-2 \log \Gamma((n+\nu_n)/2) + 2 \log \Gamma(\nu_n/2) + n \log(\pi \nu_n) + \log |\hat{\Sigma}| \\ &+ (n+\nu_n) \log[1 + (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_n)^T \hat{\Sigma}^{-1} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_n)] + 2\mathrm{tr} X A_n X^T. \end{aligned} \tag{5}$$

ただし, $\nu_n = n + \nu_0$, $\hat{\Sigma} = (\hat{\eta}_n/\nu_n)(X A_n X^T + I_n)$, $\hat{\eta}_n = \eta_0 + \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_n^T A_n^{-1} \hat{\boldsymbol{\beta}}_n$, $A_n = (X^T X + (n^2 \lambda^2/2) I_p)^{-1}$, $\hat{\boldsymbol{\beta}}_n = A_n X^T \boldsymbol{y}$ である. これにより, Bayesian lasso においてもベイズモデリングにおける予測精度に則って調整パラメータの選択を行うことが可能となった.

2.3　Bayesian lasso に対する MAP 推定

Sparse Algorithm によって Bayesian lasso に sparse 性を与えることが可能となった. しかし, 数値的な $\hat{\boldsymbol{\beta}}$ の推定は変動が大きく, 推定値が不安定になるという問題が存在し, また, Bayesian lasso の事後分布を解析的

に陽な形で求めることや事後密度関数の理論的な最大化は困難であった.

これに対し本論文では, Bayesian lasso の事後分布が混合正規分布として表現可能であるという性質を用いて, そこから Bayesian lasso の事後分布をモンテカルロ積分によって近似し, ニュートン・ラフソン法によって安定的に MAP 推定値を求める MAP Bayesian lasso を提案した (Hoshina, 2015).

誤差分散 σ^2 と 超パラメータ λ の推定値 $\hat{\sigma}^2$, $\hat{\lambda}$ が与えられたとき, Bayesian lasso における $\boldsymbol{\beta}$ の事後分布は以下の積分を含む関数の定数倍で与えられる:

$$\prod_{j=1}^{m} \int N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \hat{\sigma}^2A^{-1})|D|^{-\frac{1}{2}}|A|^{-\frac{1}{2}} \exp\left\{-\frac{\boldsymbol{y}^T(I_n - XA^{-1}X^T)\boldsymbol{y}}{2\hat{\sigma}^2}\right\} \mathrm{Exp}\left(\tau_j^2 \middle| \frac{\hat{\lambda}^2}{2}\right) d\tau_j^2. \tag{6}$$

ただし, $A = X^TX + D^{-1}$, $D = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$ であり, $\mathrm{Exp}(x|\lambda) = \lambda\exp(-\lambda x)$ $(x, \lambda > 0)$ である. この積分を求めることは難しいが, モンテカルロ積分によって近似することが可能であり, 次の $\boldsymbol{\beta}$ の近似事後分布を得る.

$$\frac{1}{M}\sum_{m=1}^{M} N_p(\boldsymbol{\beta}|A_{(m)}^{-1}X^T\boldsymbol{y}, \hat{\sigma}^2A_{(m)}^{-1})|D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2} \cdot \exp\left\{-\frac{\boldsymbol{y}^T(I_n - XA_{(m)}^{-1}X^T)\boldsymbol{y}}{2\hat{\sigma}^2}\right\}. \tag{7}$$

ただし, $D_{(m)} = \mathrm{diag}(\tau_{1(m)}^2, \ldots, \tau_{p(m)}^2)$, $A_{(m)} = X^TX + D_{(m)}^{-1}$ であり, $\{\tau_{1(m)}^2, \ldots, \tau_{p(m)}^2 : \ m = 1, \ldots, M\}$ は $\prod_{j=1}^{p} \mathrm{Exp}(\tau_j^2|\hat{\lambda}^2/2)$ からのサイズ M の確率サンプルである.

この近似事後分布は微分可能な関数の和であり, それ自体も微分可能な関数である. よって, $\boldsymbol{\beta}$ の事後モードはこの近似事後分布にニュートン・ラフソン法を適用することで求めることが可能となる.

これにより, Bayesian lasso の点推定値が不安定になってしまう原因であった数値的手法による事後分布の推定を行う必要がなくなり, 安定的な Bayesian lasso の点推定が可能となった.

参考文献

[1] 保科 架風 (2012). Bayesian lasso によるスパース回帰モデリング (in Japanese). 計算機統計学, 25, 73–85.

[2] Efron, B., Hasite, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32, 407–499.

[3] Hoshina, I. (2015). Sparse regression modeling via the MAP Bayesian lasso. *Bulletin of the Informatics and Cybernetics*, 47, 37–58.

[4] Kawano, S., Hoshina, I., Shimamura, K., and Konishi, S. (2015). Predictive model selection criteria for Bayesian lasso regression. *Journal of the Japanese Society of Computational Statistics*, 28, 67–82.

[5] Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics – Theory and Methods*, 26, 2223–2246.

[6] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

[7] Murphy, K. (2012). *Machine Learning – a Probabilistic Perspective*, MIT Press.

[8] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

[10] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.