スパース部分線形加法回帰モデルにおける変数重要度の算出に関する検討

宮澤 歩夢† 金子 奈々恵‡ 藤本 悠* 林 泰弘†

† 早稲田大学先進理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
‡ 早稲田大学先進理工学部 〒169-8555 東京都新宿区大久保 3-4-1
*早稲田大学スマート社会技術融合研究機構 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: † a-miyasawa@akane.waseda.jp, hayashi@waseda.jp, ‡ nananaco326.k@toki.waseda.jp, *y.fujimoto@aoni.waseda.jp

あらまし 説明変数のスパース性を想定した線形回帰は，一般に推定過程で変数を選択する機構を有し，回帰係数の比較によって説明変数の重要度の直感的な解釈を行うことが容易であることから，高次元多変量解析のための新たな統計的モデリング手法として定着しつつある．一方，目的変数と一部の説明変数との間の非線形な関係を考慮するべく拡張された加法モデルは同様に柔軟で解釈性の高いモデリングを可能にする一方で，変数重要度の定量化に関する議論があまりなされておらず，データの解析の文脈では利用しにくい側面がある．本稿では部分線形加法モデルにおける変数重要度指標としてブートストラップ標本の下での選択頻度に着目し，導出される重要度の性質について議論を行う．

キーワード 回帰，加法モデル，非線形変換，変数重要度

1. 概要

対象となる目的変数と，複数の説明変数との間の関係性を定量的に分析するための手段として，回帰分析が広く用いられている．回帰分析においては，予測精度やモデルの解釈性の向上を目的として，説明変数の選択が重要な意味を持つ．説明変数のスパース性を想定した線形回帰は，推定の過程において変数を選択する機構[1]を持ち，高次元多変量解析のための統計的モデリング手法として定着しつつある．変数選択の過程を経ることにより，目的変数との関係性が低い変数がモデルから除外され，予測精度が向上するだけでなく，得られたモデルを基にした説明変数と目的変数の関係性についての議論が容易になる．例えば，回帰係数は説明変数が目的変数に与える影響を直感的に表すため，変数の重要度を解釈する上で広く用いられるが，このような定量的な変数重要度は，モデルの解釈を行う上で重要な観点を提供する．

一方，目的変数に対して非線形の関係を持つ説明変数が混在する状況において，線形性を仮定した線形回帰モデルは，説明変数を適切に考慮することができない．このような問題に対し，非線形な関係を考慮するモデリング手法の一つとして加法モデル[2]がある．特に部分線形加法モデル[3]は，線形回帰の解釈性の高さを残しつつ，一部の説明変数が持つ非線形性を考慮できるよう線形回帰モデルを拡張したもので，各変数の加法的な貢献の度合いを直接的に表現する意味で場合によっては魅力的な解析アプローチとなり得る．しかし，非線形性を仮定した説明変数が存在する状況下においては，線形回帰モデルと同様に回帰係数を変数重要度とみなして議論を行うことができず，線形性を仮定した説明変数との重要度の相対的な大小を比較することが困難となる．非線形性を考慮した加法モデルにおける変数重要度の定量化については，解析において重要であるにも関わらず十分に議論が行われていない．本論文では，説明変数の非線形性を考慮したスパース部分線形加法回帰モデルにおける変数の重要度を測る指標として，ブートストラップ標本に基づくモデル作成における，変数選択頻度に着目した変数重要度の定量化を提案する．

以下，第 2 章では，高次元多変量解析において広く用いられる，スパース性を持つ線形回帰モデルについて説明し，また，非線形を持つ説明変数を考慮することの意義，及び部分線形加法回帰モデルの説明を行う．第 3 章では，説明変数の重要度について，従来手法を踏まえて議論を行った上で，ブートストラップ標本に基づく新たな変数重要度を提案する．第 4 章と第 5 章では，提案指標の妥当性を示すための数値実験の結果について報告する．第 4 章では人工的に作成されたデータについて，第 5 章では電力需要を目的変数とした実データについて数値実験を行う．最後に，第 6 章において，本論文のまとめを行う．

2. 加法モデルに基づく回帰

2.1 スパース性を想定した線形回帰モデル

p個の説明変数からなる入力$\boldsymbol{x}=[x^1,\dots,x^p]$に対して，線形回帰モデルは以下の式で表される．

$$f(\boldsymbol{x})=\beta_0+\sum_{j=1}^{p}\beta_j x^j+\varepsilon \qquad (1)$$

このとき，$\boldsymbol{\beta}=[\beta_0,\beta_1,\dots,\beta_p]$は回帰係数パラメータ，$\varepsilon$は誤差項である．回帰係数$\boldsymbol{\beta}$を推定する手法の一つとして広く用いられている最小二乗法では，N個の入力と出力の組$(\boldsymbol{x}_i,y_i)$が与えられたとき，二乗誤差が最小となるような$\boldsymbol{\beta}$を以下の式に基づいて決定する．

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_i^j\right)^2\right\} \qquad (2)$$

最小二乗法に基づく推定手法ではモデルの解釈性や予測精度の観点から，目的変数と関係性の低い説明変数をモデルから除外する変数選択の手続きが重要となる．一方，説明変数のスパース性を考慮するべく拡張された Lasso[1]は，回帰係数の大きさに対する罰則を与えることにより推定の過程で変数選択を行う機構を有しており，推定量$\boldsymbol{\beta}$は最小二乗法に制約を加えた以下の式に基づいて決定される．

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_i^j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \qquad (3)$$

ここで，λは非負のパラメータである．上式の解は，回帰係数の大きさに対する正則化の制約により複数の係数が 0 となる傾向を持ち，Lasso が推定の過程において変数を選択する機構を持つことを示している．

本論文では，Lasso に変数選択の一致性と，非 0 の推定量の漸近正規性[4]を持たせる目的で拡張された Adaptive Lasso[5]を前提とした議論を行う．Adaptive Lasso は，上記の性質を持つ最小二乗法や Ridge[6]によって得られた推定量$\tilde{\boldsymbol{\beta}}$を利用して，以下の式に基づいて決定される．

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_i^j\right)^2+\lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j|^{\nu}}\right\} \qquad (4)$$

ここで，νは非負のパラメータである．推定量$\tilde{\beta}_j$が 0 に近い値を取る場合にβ_jに対して大きな罰則が掛かることによって，変数選択の一致性が保証される．

2.2 加法モデルへの拡張

線形回帰モデルは一般に，説明変数群と目的変数が線形の関係を持つことを仮定している．一方で，推定に用いる説明変数が目的変数に対して必ずしも線形性を持つとは限らない場合，非線形性を考慮したモデリング手法による，より説明性の高いモデルの実現が望まれることになる．

図 1 は，気象庁が公開している東京都心部の 1 時間平均気温[7]と，電力広域的運営推進機関が公開している東京電力管内の 1 時間需要電力量[8]の関係性をプロットしたものである．図 1 では，ある一定の気温以上において気温と電力需要が正の相関を持つのに対して，ある一定の気温以下においては負の相関を持つことが確認できる．説明変数と目的関数が非線形の関係を持つこのような場合において，線形回帰モデルは必ずしも適さないことになる．

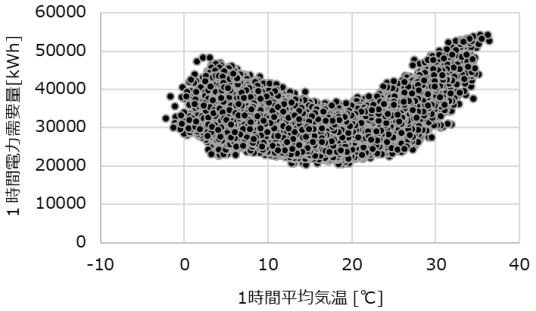


図 1 平均気温と需要電力量の関係

加法モデルは，目的変数と説明変数との間に非線形な関係の考慮が可能となるように，線形回帰モデルを拡張したモデルであり，以下の式で表される．

$$f(\boldsymbol{x})=\beta_0+\sum_{j=1}^{p}g_j(x^j)+\varepsilon \qquad (5)$$

線形回帰モデルと比較して，回帰係数$\boldsymbol{\beta}$の代わりに任意の関数$g_1,\dots,g_p$に置き換えることによって非線形の関係が表現可能となっていることが分かる．関数$g_1,\dots,g_p$が線形の関数に限られる場合，式(5)は線形回帰モデルと同等になる．本稿では，目的変数に対し線形性を持つ説明変数と，非線形性を持つ説明変数が混在する状況を想定する．以降，線形性を仮定するp個の説明変数$x^1,\dots,x^p$と，非線形な関数$\phi_1,\dots,\phi_q$に基づく非線形性を仮定するq個の説明変数$x^{p+1},\dots,x^{p+q}$を区別し，式(5)を変形した部分線形加法回帰モデル

$$f(\boldsymbol{x})=\beta_0+\sum_{j=1}^{p}\beta_j x^j+\sum_{j=1}^{q}\phi_j(x^{p+j})+\varepsilon \qquad (6)$$

を基に議論を進める．

部分線形回帰モデルの推定では，回帰係数$\boldsymbol{\beta}$と非線形関数$\phi_1,\dots,\phi_q$の両方を推定する必要があり，いくつかのスキームが提案されている[9,10]が，本稿では，$\phi_j(x^{p+j})=\beta_{p+j}\psi(x^{p+j})$として式(6)を

$$f(\boldsymbol{x})=\beta_0+\sum_{j=1}^{p}\beta_j x^j+\sum_{j=1}^{q}\beta_{p+j}\psi_j(x^{p+j})+\varepsilon \qquad (7)$$

と書き換え，関数$\psi_1,\dots,\psi_q$を対応する変数が目的変数に対して高い相関を持つような非線形変換によって得る Alternating Conditional Expectation algorithm (ACE)[11] により事前に導出し，推定結果を用いて，係数$\boldsymbol{\beta}$を Lasso により推定をするという二段階推定を考える．非線形変換により，変換後の変数に目的変数に対して線形の関係性を仮定することが可能となる．一方で，非線形変換の過程におけるスケーリングなど

の影響により，線形性を仮定した説明変数の係数との相対的な大小を比較することが困難となる．ACE のアルゴリズムは Appendix に記載する．

3. 説明変数の重要度に関する検討

3.1 回帰における変数重要度

回帰モデルを分析に用いる際には，高い予測精度のほかに，説明変数と目的変数の関係の解釈性がしばしば求められる．各説明変数の目的変数への影響度合いを何らかの形で定量化することは，モデルの解釈を行う意味では重要となる．

線形回帰モデルにおいては，説明変数が単位量変化した時の目的変数の変化量を示す回帰係数の比較によって説明変数の重要度の直感的な解釈を行うことが容易である．予め入力データを標準化した上で得られるモデルの回帰係数は，変数間の相対的な変数重要度を表し，目的変数への影響度合いの観点から説明変数の順位付けを行うことも可能になる．

一方，部分線形加法回帰モデルを含む加法モデル全般においては，非線形性を仮定する説明変数について回帰係数に相当する貢献量の定量化ができないため，線形回帰モデルのように変数重要度の議論を行うことが難しい．そのため，非線形性を仮定する説明変数が，他の変数と比べて相対的にどの程度重要なのかを直感的に理解できず，実データ解析の文脈で定性的な解釈に窮することが出てくる．そのため，非線形性を考慮した部分線形加法モデルについて，モデルの解釈性を高めデータの解析の文脈で有効に活用するために，変数重要度の定量化に関する議論が必要となる．

3.2 変数重要度に関するこれまでの議論

線形回帰モデルにおいては単純に回帰係数の大小で議論する以外にも，説明変数の重要度を定量化するいくつかの取り組みが存在する．例えば，推定で得られた線形回帰モデルから，特に重要度の高い変数を除外することで汎化誤差が大きく悪化することに着目し，変数の回帰係数を 0 にすることによる交差検証誤差の改悪度合いとして定義された変数重要度について，Y. Gan らがその妥当性を検証している[12]．交差検証に基づいて変数重要度を算出するアルゴリズムを以下に示す．

＜汎化誤差の改悪度合いに基づく重要度算出＞

1) データセットをK個の部分集合に分割する．
2) $k\ (k=1,\dots,K)$番目の部分集合をテスト集合とし，残りのデータを用いてモデルを推定し，テスト集合に対する実測値と推定値の差の平均err_kを算出する．
3) 得られたモデルの，$j\ (j=1,\dots,p)$番目の変数の係数β_jを 0 としたモデルについても同様に差の平均$err_k^{(j)}$を推定する．
4) x^jの変数重要度を$I_j=\sum_{k=1}^{K}\left(err_k^{(j)}-err_k\right)^2$で算出する．

このアプローチでは，データのサンプル数に応じて変数重要度の大小関係が変化してしまうことが文献中でも指摘されており，また，正則化パラメータの観点からチューニングされたモデルは，変数の有無の観点から陽にチューニングした結果と一般に異なる．そのため，スパース正則化により得られたモデルにおいて同様の考え方を適用した場合，係数を 0 とみなすことが必ずしも交差検証誤差を改悪せず，重要度の直感的な解釈に適さない場合がある．

一方，S.Wang らは，サンプル数の影響を受けにくい変数重要度の算出とそれに基づく変数選択を行う手法として，Random Lasso を提案している[13]．Random Lasso では，元の標本からランダムにサンプリングされたブートストラップ標本に対し，Lasso に基づくモデリングを行う過程を繰り返すことで変数重要度を算出し，再抽出したブートストラップ標本で得られた重要度を基にした推定を行う．Random Lasso のアルゴリズムの中で，変数重要度は次のように算出されている．

＜Random Lasso における重要度算出＞

1) 元の標本からN個の標本をランダムに抽出したブートストラップ標本をB個作成する．
2) $q\ (\leq p)$個の説明変数をランダムに選択し，Adaptive Lasso により回帰係数$\hat{\beta}_j^{(b)}\ (b=1,\dots,B)$を決定する．$q$は交差検証により決定される．
3) x^jの変数重要度は$I_j=\left|\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^{(b)}\right|$で算出される．

Random Lasso は，ブートストラップ標本を用いることにより$N<p$の問題設定において変数重要度の算出が可能である点で，素朴な回帰係数の比較よりも解釈性の高い結果が得られる可能性がある．一方で，線形性の仮定の下で説明変数に掛かる係数に着眼した量のため，係数の平均として得られる変数重要度の比較は部分線形加法回帰モデルにおいては解釈に誤解を生む恐れがある．

これらの先行研究は，いずれも線形回帰モデルを前提としており，部分線形加法回帰モデルをはじめとする非線形性を考慮したモデルにおける変数重要度の議論は十分に行われていない．

3.3 ブートストラップ法に基づく変数重要度の提案指標

本稿では，ブートストラップ標本を用いる変数重要度の算出手法について検討する．Random Lasso ではブートストラップ標本に対して得られる回帰係数に着目していたが，(7)式で与えられる部分線形加法回帰モデ

ルを考えると，非線形性を仮定する変数に掛かる係数そのものは線形性を仮定したほかの変数の係数と単純比較ができないため，各ブートストラップ標本においてスパース推定を行った結果として変数が選択された頻度を変数重要度として定義する.変数の選択頻度は，直感的には重要な変数である程高く，目的変数に関係のない変数であれば 0 に近付くと考えられる.以下に，提案手法において変数重要度を算出するアルゴリズムを以下に示す.

＜ブートストラップ選択頻度に基づく重要度算出＞

1) データセット中の全てのデータを用いた交差検証により Adaptive Lasso のλを決定する.
2) 元の標本からN個の標本をランダムに抽出したブートストラップ標本をB個作成する.
3) $q\ (\leq p)$ 個の説明変数をランダムに選択し，Adaptive Lasso により$\hat{\beta}_j^{(b)}\ (b=1,\dots,B)$を決定する.$q$は交差検証により決定される.
4) x^jの変数重要度は$I_j=\left|\frac{1}{B}\sum_{b=1}^{B}\Omega\left(\hat{\beta}_j^{(b)}\right)\right|$で算出される．ただし，

$$\Omega\left(\hat{\beta}_j^{(b)}\right)=\begin{cases}1 & \left(\hat{\beta}_j^{(b)}\neq 0\right)\\ 0 & \left(\hat{\beta}_j^{(b)}=0\right)\end{cases} \tag{8}$$

とする.

4. 人工データに対するシミュレーション

4.1 問題設定

本章では，非線形性を持つ説明変数に対する変数重要度の提案指標の有効性を確認するため，人工的に作成されたデータに基づく数値実験を行う.

数値実験のため，正規分布に基づく$(p+q)$次元の入力ベクトル$\boldsymbol{x}_i$~$\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$と，式(7)のモデルにより得られる出力$y_i$の組$(\boldsymbol{x}_i,y_i)$を$N(=100)$個作成する．線形回帰モデルにおける従来の重要度算出手法との比較，及び部分線形加法モデルにおける提案手法の有効性を評価する目的で，以下の 3 つのケースを想定してそれぞれデータセットを作成する.

(ケース 1) $p=8,q=0$の低次元線形回帰モデルを考える．p個の説明変数$\boldsymbol{x}$に掛かる係数$\boldsymbol{\beta}=[\beta^0,\beta^1,\dots,\beta^p]$は

$$\boldsymbol{\beta}=[0,3,1.5,0,0,2,0,0,0]$$

とする．目的変数と関係性を持つ変数(係数が非 0)と，関係性を持たない変数(係数が 0)の存在を仮定している.また，入力ベクトルは平均$\boldsymbol{\mu}=[0,\dots,0]^{\mathsf{T}}$，$p\times p$次元の分散共分散行列$\boldsymbol{\Sigma}=\left[\sigma_{j_1j_2}\right]$が

$$\sigma_{j_1j_2}=0.5^{|j_1-j_2|}$$

で与えられる正規分布に従って生成されるものとする.

(ケース 2) $p=40,q=0$の高次元線形回帰モデルを考える．説明変数$\boldsymbol{x}$に掛かる係数$\boldsymbol{\beta}$は

$$\boldsymbol{\beta}=\left[\overbrace{3,\dots,3}^{5},\overbrace{-2,\dots,-2}^{5},\overbrace{-0.5,\dots,-0.5}^{5},\overbrace{0.2,\dots,0.2}^{5},\overbrace{0,\dots,0}^{20}\right]$$

とする．目的変数と高い関係性を持つ変数(係数が 3,-2)と，低い関係性を持つ変数(係数が-0.5, 0.2)と，関係性を持たない変数(係数が 0)の存在を仮定している.また，入力ベクトルは$\boldsymbol{\mu}=[0,\dots,0]^{\mathsf{T}}$，$\boldsymbol{\Sigma}=\left[\sigma_{j_1j_2}\right]$が

$$\sigma_{j_1j_2}=\begin{cases}0.9 & \left(\beta_{j_1}=\beta_{j_2}\text{ かつ }\beta_{j_1}\neq 0\right)\\ 0 & (\text{otherwise})\end{cases}$$

で与えられる正規分布に従って生成されるものとする.

(ケース 3) $p=40,q=10$の高次元部分線形加法モデルを考える．説明変数$\boldsymbol{x}$に掛かる係数$\boldsymbol{\beta}$は

$$\boldsymbol{\beta}=\left[\overbrace{3,\dots,3}^{5},\overbrace{-2,\dots,-2}^{5},\overbrace{-0.5,\dots,-0.5}^{5},\overbrace{0.2,\dots,0.2}^{5},\overbrace{0,\dots,0}^{20},\overbrace{-3,\dots,-3}^{5},\overbrace{0.5,\dots,0.5}^{5}\right]$$

とする．前述のケースと同様に，関係性の高い変数，関係性の低い変数,関係性を持たない変数を仮定する.また，入力ベクトルは$\boldsymbol{\mu}=[0,\dots,0]^{\mathsf{T}}$，$\boldsymbol{\Sigma}=\left[\sigma_{j_1j_2}\right]$が

$$\sigma_{j_1j_2}=\begin{cases}0.9 & \left(\beta_{j_1}=\beta_{j_2}\text{ かつ }\beta_{j_1}\neq 0\right)\\ 0 & (\text{otherwise})\end{cases}$$

で与えられた正規分布に従って生成されるものとする.また，$p+1$番目から$p+q$番目の説明変数については

$$\psi_j\left(x^{p+j}\right)=\left(x^{p+j}\right)^2$$

という変換を仮定する．その他の説明変数との整合を取るために，$\psi_j\left(x^{p+j}\right)$は平均 0，分散 1 となるよう正規化される.

また，3 つのケースに共通して，誤差項をε~$\mathcal{N}(0,\ 1)$で与える．それぞれのケースにおいて，(a)回帰係数の絶対値$(I_j=|\hat{\beta}_j|)$，(b)交差検証の意味での汎化誤差の改悪度，(c)Random Lasso に基づく変数重要度，及び(d)提案手法に基づく変数重要度を比較し考察を行う．ただし，(c)及び(d)について，ブートストラップ標本の抽出回数$B=1000$とする.

4.2 数値実験結果と考察

はじめに，低次元の線形回帰モデルにおける検証を行うため,ケース 1 の条件設定の下で変数重要度$I_j\ (j=1,\dots,(p+q))$を算出した結果を図 2 に示す．変数重要度の並びは,4.1 項で示した問題設定と対応付いている.

全ての重要度指標に共通して，直感的にも目的変数との関係性が強い変数の重要度が高くなっている．ケース 1 のような低次元かつ，$N\gg p$の問題設定では，(a)回帰係数や(b)汎化誤差の改悪度は直感的な理解を得やすいものとなっている．(c)Random Lasso では，目的関数と関係を持たない変数について非 0 の重要度が得られているが，これは推定毎に説明変数の候補がランダムに選択されることに起因する．(d)提案指標は，全てのブートストラップ標本において選択される変数が

同様に重要であることを示している．

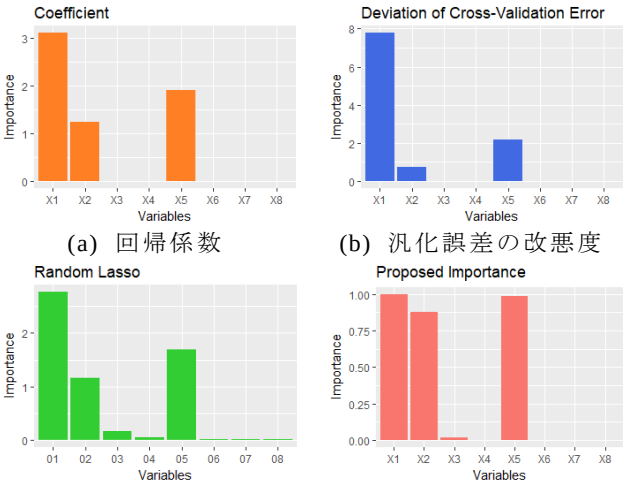
次に，ケース 2 の条件設定の下で変数重要度を算出した結果を図 3 に示す．目的変数と関係を持つが相対的に関係性が低い変数について，(a)や(b)の指標では重要度が 0 となるのに対して，(c)や(d)では非 0 の重要度が得られている点で，解釈性の高い指標となっているといえる．これは，ブートストラップ標本に基づきランダム性を持たせることに起因すると考えられる．また，(a)や(b)の指標において本来，目的変数と関係があるにも関わらず重要度が低いとみなされる説明変数が見られる．これは，説明変数同士が相関を持っていることで，一部の変数の係数が縮小された結果，重要ではないと推定されることによって生じる問題である．これは，高次元多変量解析において，互いに相関を持ついくつかの重要性の高い変数の存在により，回帰係数や改悪度に基づく重要度の妥当性が失われる可能性を示している．一方，ブートストラップ標本に基づく推定は，この問題設定においても，本来説明力を有する説明変数に対して重要度の値を算出していることを示している．

最後に，(7)式で与えられる部分線形加法モデルを仮定したケース 3 の条件設定の下で変数重要度を算出した結果を図 4 に示す．(b)汎化誤差の改悪度の観点では，線形性を持つ説明変数と非線形性を持つ説明変数が同等の重要度を持つと推定されているが，(a)の指標では，非線形性の関係を持つ説明変数の重要度が非常に高くみなされている．これは，非線形性の関係を持つ説明変数に関して，入力$\psi_j(x^j)$の導出に用いられる ACE による非線形変換の過程でのスケーリングなどの影響で，得られる回帰係数同士の相対的な比較が意味をなさなくなっていることを示唆している．回帰係数の平均を取る(c)Random Lasso においても，(a)の指標と近い傾向の結果が得られていることが確認できる．一方，(d)提案指標では，変数の選択頻度に着目することにより，係数そのものの相対比較を避け，より直感的に理解のしやすい重要度が算出できていることが分かる．各変数重要度はその算出方法によって様々な特徴を持ち，解析の用途によって使い分ける必要があるが，特に非線形性を持つ変数が考慮される加法モデルにおいては，提案指標の有効性が示されたと言える．

5. 実データに対するシミュレーション

5.1 問題設定

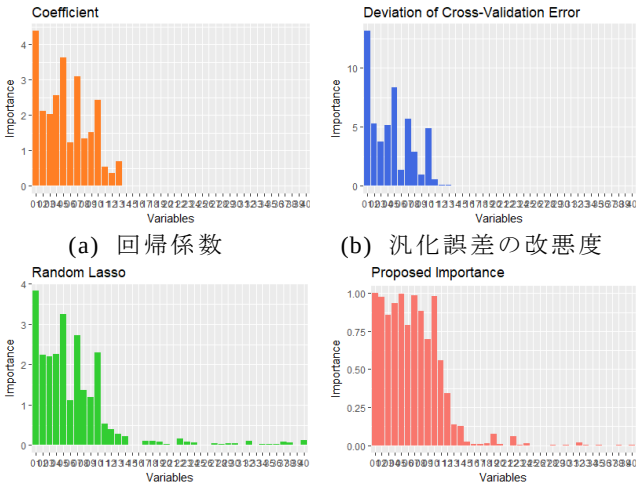
本章では，提案指標の妥当性についてより深い考察を行うため，実世界で観測されているデータを基にした推定を行う．目的変数に対し線形性を持つ説明変数と，非線形性を持つ説明変数が混在する例として，2.2 項で述べたような平均気温と需要電力量の関係に着目し，日毎の需要電力量を対象としたモデルの推定を行



(a) 回帰係数　(b) 汎化誤差の改悪度

(c) Random Lasso　(d) 提案手法

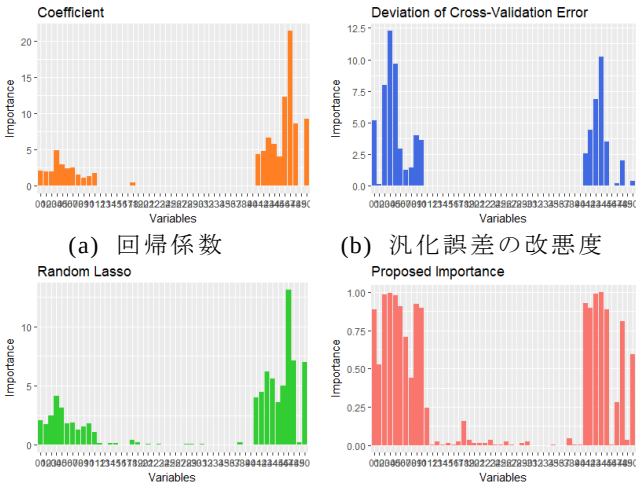
図 2 ケース 1 における変数重要度



(a) 回帰係数　(b) 汎化誤差の改悪度

(c) Random Lasso　(d) 提案手法

図 3 ケース 2 における変数重要度



(a) 回帰係数　(b) 汎化誤差の改悪度

(c) Random Lasso　(d) 提案手法

図 4 ケース 3 における変数重要度

う．説明変数には，気象庁が公開している各種気象データや，前日の需要電力量のほか，モデルの説明性を高める目的で月毎に観測されるデータを加える．ただし月別データは，同月中は常に同じ値が観測されると仮定し，日毎データに変換してモデルの推定に利用する．数値実験に用いる説明変数の一覧を表 1 に示す．また，モデルの学習には 2016 年 4 月 1 日～2017 年 11 月 30 日のデータを用いることとし，平均 0，分散 1 となるよう正規化される．需要電力量と非線形の関係を持つ説明変数が存在する場合を想定するため，説明変数群に対して事前に ACE に基づく非線形変換を行う．

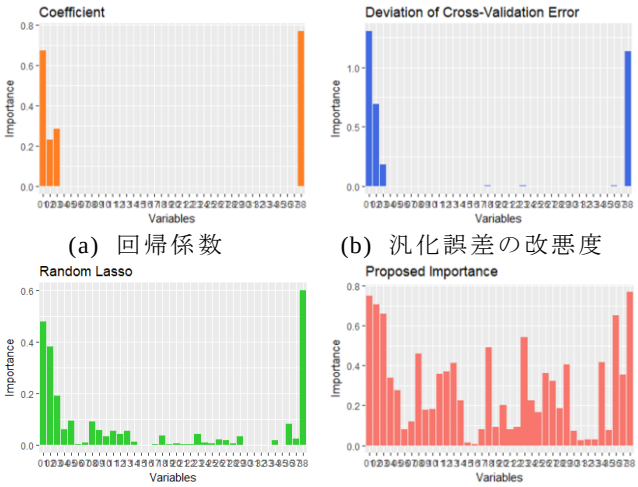
5.2 数値実験結果と考察

前項で述べた問題設定の下，変数重要度を算出した結果を図 5 に示す．図 5 より，実データに対するシミュレーションにおいて，各重要度指標の傾向は人工データのものと概ね一致していることが分かる．各指標の結果の違いとして，回帰係数そのものの値や汎化誤差の改悪度に基づく重要度では一律 0 と表現される説明変数群が，ブートストラップ標本に基づく重要度では相対的な重要度の順位付けがなされている点，及びブートストラップ標本に基づく指標(c)と(d)において，比較的重要とされる変数の傾向が異なっている点が挙げられる．

このような違いについて考察を行うため，どの重要

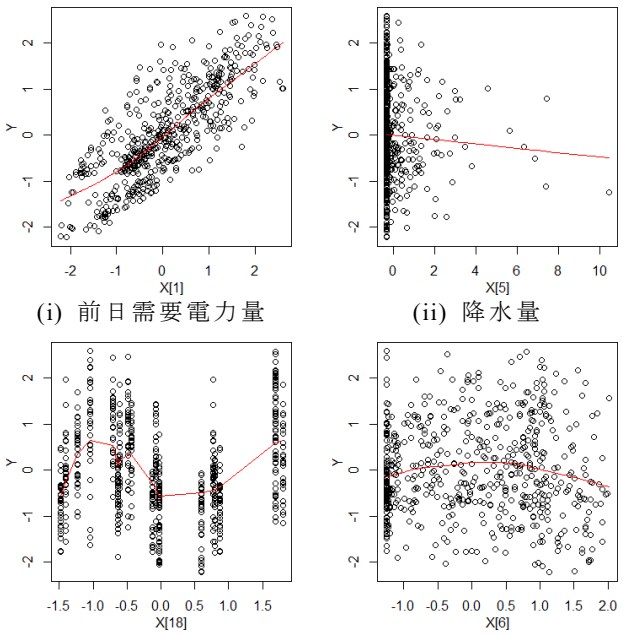
表 1 数値実験に利用する説明変数
表中の項目に付記された番号は，図 5 の変数重要度の並びと対応付けられている．

	分類	項目	引用元
日別	電力	東京電力管内の前日需要電力量 [1]	[8]
	気象	平均/最高/最低気温 [2-4]，降水量 [5]，日照時間 [6]，全天日射量 [7]，平均/最大/最大瞬間風速 [8-10]，平均雲量 [11]，平均湿度 [12]，平均現地気圧 [13]（東京都心部）	[7]
	暦	平・休日情報 [38]	
月別	IIP	IIP [14]	[14]
	企業物価指数	工業製品 [15]，農林水産物 [16]，鉱産物 [17]，電力都市ガス水道 [18]，スクラップ類 [19]	[15]
	サービス価格指数	金融保険 [20]，不動産 [21]，運輸郵便 [22]，情報通信 [23]，リースレンタル [24]，広告 [25]，諸サービス [26]	[15]
	消費者物価指数	食料 [27]，住居 [28]，光熱水道 [29]，家具家事用品 [30]，被服及び履物 [31]，保健医療 [32]，交通通信 [33]，教育 [34]，教養娯楽 [35]，諸雑費 [36]	[16]
	人口	全国大人口 [37]	[17]



(a) 回帰係数　(b) 汎化誤差の改悪度
(c) Random Lasso　(d) 提案手法
図 5 実データに対する変数重要度

度指標においても重要な変数とされた前日需要電力量[1]，(c) Random Lasso に基づく変数重要度において比較的重要な変数とされた降水量[5]，(d) 提案手法において比較的重要な変数とされた企業物価指数（電力都市ガス水道）[18]，どの重要度指標においても重要な変数とされなかった日照時間[6]の 4 つの説明変数に着目する．図 6 に，各説明変数と目的変数である需要電力量の関係性をプロットしたもの，及び ACE により学習された非線形関数を併せて示す．図 6 より，どの指標においても重要とされる(i)前日需要電力量には，目的変数に対して強い相関が見られる．一方で，(ii)降水量や(iii)企業物価指数(電力)は，(iv)日照時間よりも比較的目的変数に対する関係性が見られる．また図 6 からは，ACE による非線形関数の学習の過程に起こるスケーリングによって，変数同士の回帰係数の相対的な関係が失わ



(i) 前日需要電力量　(ii) 降水量
(iii) 企業物価指数(電力)　(iv) 日照時間
図 6 説明変数と目的変数の関係

れてしまう可能性があることが見て取れる．そのため，目的変数に対し非線形な関係を持つ説明変数の存在を想定する状況では，回帰係数への着目が必ずしも直感的な理解と結びつかないことに留意すべきである．

Random Lasso に基づく重要度と選択頻度に基づく重要度において比較的重要とされる変数の傾向が異なっている点について考察を行うため，ブートストラップ標本の抽出を繰り返す過程で作成されるB個のモデルの回帰係数を比較する．図 7 に，(ii)降水量と(iii)企業物価指数(電力)について，B個の回帰係数の絶対値を昇順に並び替えプロットしたものを示す．図 7 より，(ii)降水量の回帰係数は標本抽出毎のばらつきが大きく，(iii)企業物価指数(電力)は回帰係数が非 0 の値を持つ頻度が高くなっており，着目する説明変数の性質によって各指標で計算される重要度が異なることを示しており，解析の用途によって使い分ける必要があることを示している．

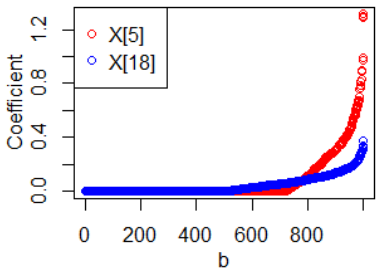


図 7 各ブートストラップにおける回帰係数の比較

6. まとめ

本稿では，目的変数と一部の説明変数との間の非線形な関係を考慮するべく拡張された部分線形加法回帰モデルに着目し，これまでにこのようなモデルを対象とした議論が十分になされていない変数重要度を定量化する指標の提案を行った．線形回帰モデルを前提に検討されていた従来の算出指標では，解釈性の高い結果が得られない可能性があることと，提案指標によって非線形性に起因する問題を緩和した，より妥当性の高い変数重要度が得られたことを数値実験により示した．非線形性を考慮したモデルにおける変数重要度の定量化をこのように行うことで，今後データの解析において変数重要度の直感的な解釈が容易になると考えられる．

Appendix

Alternating Conditional Expectation algorithm (ACE)は，出力Yに対し非線形な関係を持つ入力$\psi_1(X^1),\dots,\psi_q(X^q)$について，その関係を推定する非線形変換手法の一つで，以下のようなモデルを仮定する．

$$\theta(Y)=\sum_{j=1}^{q}\psi_j(X^j)+\varepsilon \tag{9}$$

θおよびψ_jが平均 0 となるような変換を行う任意の関数であるならば，式(9)は一般化加法モデルを表す．今回の検討では，θが線形変換により得られ，$E[\theta^2(Y)]$=1となるような関数であると仮定を置いた，加法モデルについて議論する．このとき，誤差関数e^2は

$$\begin{aligned} e^2(\theta,\psi_1,\dots,\psi_q) &= \frac{E\left\{\left[\theta(Y)-\sum_{j=1}^{q}\psi_j(X^j)\right]^2\right\}}{E\{\theta^2(Y)\}} \\ &= E\left[\left\{\theta(Y)-\sum_{j=1}^{q}\psi_j(X^j)\right\}^2\right] \end{aligned} \tag{10}$$

と定義され，e^2を最小化するため，θを固定した上で$\psi_j(X_j)$は以下の式

$$\psi_j(X^j)=E\left[\theta(Y)-\sum_{i\neq j}^{q}\psi_i(X^i)\,|X^j\right] \tag{11}$$

で更新され，ψ_jを固定した上で$\theta(Y)$は以下の式

$$\theta(Y)=\frac{E\left[\sum_{j=1}^{q}\psi_j(X^j)\,|Y\right]}{\left\|E\left[\sum_{j=1}^{q}\psi_j(X^j)\,|Y\right]\right\|} \tag{12}$$

で更新される．連続値を持つ有限個数のデータに対しては，誤差関数e^2を

$$e^2(\theta,\psi_1,\dots,\psi_q)=\frac{\sum_{i=1}^{N}\left[\theta(y_i)-\sum_{j=1}^{p}\psi_j(x_i^j)\right]^2}{N} \tag{13}$$

と置き換え，条件付き期待値の推定をデータの平滑化手法の一つである super smoother[18]に基づき行う．更新を繰り返すことによって最終的な変数変換の推定量$\hat{\theta}$および$\hat{\psi}_j$が得られる．なお，$\theta(.)$はyに対して線形な変換との制約を与えることによって，導出されるモデルは上記の加法モデルのクラスとなる．

参　考　文　献

[1] R. Tibshirani, “Regression Shrinkage and Selection via the lasso,” Journal of the Royal Statistical Society, 58, pp.267-288, 1996.

[2] J. H. Friedman and W. Stuetzle, “Projection Pursuit Regression,” Journal of the American Statistical Association, 76, pp.817-823, 1981.

[3] Y. Lou et al., “Sparse Partially Linear Additive Models,” Journal of Computational and Graphical Statistics, 25, pp.1126-1140, 2016.

[4] J. Fan and R. Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” pp. 1348-1360, 2001.

[5] H. Zou, “The Adaptive Lasso and Its Oracle Properties,” Journal of the American Statistical Association, 101, pp.1418-1429, 2006.

[6] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” Technometrics, 12, pp.55-67, 1970.

[7] 「気象庁｜過去の気象データ検索」, http://www.data.jma.go.jp/obd/stats/etrn/index.php (最終閲覧日: 2018 年 1 月 10 日)

[8] 「広域機関システム」, http://occtonet.occto.or.jp/public/dfw/RP11/OCCTO/SD/LOGIN_login# (最終閲覧日: 2018 年 1 月 10 日)

[9] A. Chouldechova and T. Hastie, “Generalized

Additive Model Selection," The Annals of Applied Statistics, arXiv preprint arXiv:1506.03850, 2015.

[10] H. H. Zhang, G. Cheng and Y. Liu, "Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models," Journal of the American Statistical Association, 106, pp.1099-1112, 2011.

[11] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," Journal of the American Statistical Association 80, pp.580-598, 1985.

[12] Y. Gan et al., "A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model," Environmental Modelling and Software, 51, pp.269-285, 2014.

[13] S. Wang, B. Nan, S. Rosset and J. Zhu, "Random Lasso," The Annals of Applied Statistics, 5, pp.468-485, 2011.

[14] 「鉱工業指数」, http://www.meti.go.jp/statistics/tyo/iip/index.html (最終閲覧日: 2018 年 1 月 22 日)

[15] 「物価、資金循環、短観、国際収支統計データの一括ダウンロード」, https://www.stat-search.boj.or.jp/info/dload.html (最終閲覧日: 2018 年 1 月 22 日)

[16] 「消費者物価指数」, https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&lid=000001196105 (最終閲覧日: 2018 年 1 月 22 日)

[17] 「統計局ホームページ/人口推計の結果の概要」, http://www.stat.go.jp/data/jinsui/2.htm (最終閲覧日: 2018 年 1 月 22 日)

[18] J. H. Friedman and W. Stuetzle, "Smoothing of Scatterplots," Technical Report ORION006, Stanford University, 1982.