研究論文

Group fused lasso による料率区分の自動セグメンテーション

野村　俊一 *

2017 年 1 月 24 日投稿
2017 年 10 月 24 日受理

概要

本稿では，スパース回帰手法である group fused lasso を利用して，リスクファクター内の料率区分を同等のリスク水準をもつグループへと自動的にクラスタリングする新たな料率算定手法を提案する．タリフ分析では，信頼できる最良推計を得るために，多数のクラスをもつ料率ファクターを少数のカテゴリーへと再グループ化することが多い．しかしながら，料率区分の分割方法の組合せは膨大になることが多く，そのため可能な全ての分割方法を検討することは計算量的に困難である．このような状況下で，fused lasso と呼ばれる L1 正則化法は，推定の過程の中でリスク水準に有意差のない隣接区分を自動的に統合してくれるため非常に適している．ここでは特に，料率ファクター間の交互作用がある場合や，クレーム頻度と規模を同時にモデル化する場合の料率算定のための一般化線形モデルに対するグループ正則化項を導入する．正則化項付き対数尤度からのモデルパラメータ推定には交互方向乗数法を利用し，また，正則化パラメータは交差確認法により選択する．解析例として，提案手法を日本の自動車損害賠償責任保険のクレームデータへと適用し，都道府県を同水準のクレーム頻度または規模をもつクラスターへとグループ化を行う．

キーワード： タリフ分析，一般化線形モデル，Group fused lasso，交互方向乗数法，自動車損害賠償責任保険

1　はじめに

保険業界では古くから，保険契約ごとの保険料を決めるのにタリフ（tariff）と呼ばれる保険料率表が利用されてきた．タリフでは料率ファクターと呼ばれる保険リスクに関わる属性情報について，各ファクター内で属する料率区分から該当契約の保険料が参照できるようになっている．

タリフに定める保険料を過去のクレームデータから算定するための方法論は，現代の推測統計学が発展を遂げる以前から考案されており，その代表的なものとして Bailey(1963) による minimum bias 法がある．その後，Jung(1968) はクレーム頻度についてポアソン分布を仮定して最尤法により推定する手法を提案し，minimum bias 法とともに保険料率算定の実務的手法として利用されてきた．その後，90 年代に各国で保険の自由化が進むと，より柔軟なタリフ分析手法が求められるようになり，Nelder and Wedderburn(1972) による一般化線形モデルが利用されるようになった．Ohlsson and Johansson(2010) などの損害保険に特化した一

* 統計数理研究所 E-mail: nomura@ism.ac.jp

般化線形モデルの解説書籍も出版され，近年の料率算定では一般化線形モデルやその派生モデルが主流になっていると言える．

一方で，タリフ分析において障害になる要素として，被保険者年齢や居住地域，職業などの区分数の多いファクターの扱いが挙げられる．料率区分数が多くなる分だけ区分ごとの過去のクレーム実績のデータ量は減少するため，信頼のおける料率の最良推計が得られないという問題が生ずる．このような問題に対しては，リスク水準の近い料率区分を統合して，より少数からなるカテゴリーへとグループ化することが従来から実務的対処法として行われてきた．料率区分の統合は，タリフ適用の簡便性という面でも利点がある．

最適な料率区分のグルーピングを行うには，妥当と考えられる全てのグルーピングの組合せを試して，その中から最も当てはまりのよいものを選ぶべきである．しかし，ファクターの区分数が増えるほど，そのような組合せの数は指数関数的に増大するため，現実には全ての組合せを検討することは計算量的に不可能であることが多い．そのような事情から，従来より料率区分の統合のしかたは主に営業施策上の都合やアクチュアリーの経験に基づいて決められてきた．

しかし，近年ではビッグデータを扱うための統計的手法としてスパース統計モデルが発展し，膨大なモデル候補の中から最適化を通じて自動的にモデル選択を行えるようになった．スパース回帰手法としては，Tibshirani(1996) による lasso（least absolute shrinkage and selectioin operator）に端を発し，その様々な派生手法が提案されている．特に，ファクター内の区分統合を自動的に決めるための手法としては，Tibshirani et al.(2005) による fused lasso が有用であり，さらにグループ化されたパラメータに適用できるよう拡張したものとして，Bleakley and Vert(2011) による group fused lasso がある．そこで本研究では，一般化線形モデルにおいて group fused lasso を利用した料率区分の統合手法を提案する．提案手法は，ファクター間の交互作用を導入する場合や，クレーム頻度とクレーム規模について別々にモデル化した上で同時に区分統合を行いたい場合に適用することができる．

本稿の構成は次のとおりである．第 2 節では，本研究のベースとなるクラス料率算定法である一般化線形モデルについて解説し，クレーム頻度とクレーム規模のモデル化を行う．第 3 節では，料率区分の自動セグメンテーションを行うための L1 正則化手法を提案する．第 4 節では，前節で導入した L1 正則化項付き対数尤度を最適化してモデルパラメータを推定するアルゴリズムとして交互方向乗数法を紹介する．第 5 節では，提案手法を自動車損害賠償責任保険のクレームデータへと適用し，期待クレーム頻度および期待クレーム単価の推定結果を示す．そして第 6 節にて，本研究のまとめと今後の展望を述べて締めくくる．

2　一般化線形モデルによる純保険料の推定

本節では，代表的なクラス料率算定法である一般化線形モデルを用いた純保険料の推定法について解説する．以降では，保険料を定めるのに p 個の料率ファクターがあり，それぞれが $n_1, \ldots, n_p$ 個の区分に分類されているものとする．また，全体で T 件の契約あるいは同じファクター内区分をもつ契約集団があり，そのうち t 番目の契約あるいは契約集団が各ファクター内で属する区分を $x_{t1}, \ldots, x_{tp}$ と表す．一般化線形モデルとは，通常の線形回帰モデルに対して，正規分布以外の確率分布と非線形なリンク関数を扱うよう拡張されたモデルである．観測値 $y_1, \ldots, y_T$ は，その確率（密度）関数をそれぞれ次式の形で表すことができる指数分散モデル（exponential dispersion models）に従うものと仮定する．

$$f(y_t; \theta_t, w_t, \phi) = \exp\left\{ \frac{y_t\theta_t - b(\theta_t)}{\phi/w_t} + c(y_t, w_t, \phi) \right\}, \quad t = 1, \ldots, T. \tag{1}$$

ここで，θ_t は契約 $t = 1, \ldots, T$ ごとの平均に関わるパラメータであり，ϕ は全契約に共通する分散に関わるパラメータとなっている．また，w_t は契約 $t = 1, \ldots, T$ ごとの分散に関わる所与の値であり重みと呼ばれる．$b(\theta_t)$ は 2 階微分可能な θ_t の関数であり，$c(y_t, w_t, \phi)$ は θ_t に依存しない関数で，確率の総和を 1 とするために用意された規格化定数である．このとき，指数分散モデル (1) の平均および分散は関数 b の 1 階微分 b' と 2

階微分 b'' を用いて次のように表される．

$$\mathrm{E}(y_t) = b'(\theta_t), \quad \mathrm{Var}(y_t) = \frac{\phi}{w_t} b''(\theta_t).$$

純保険料とは契約ごとの期待クレーム総額を指し，純保険料に手数料や管理費などを加えたものが販売価格である営業保険料となる．損害保険の料率算定では，契約毎の純保険料すなわち期待クレーム総額を直接推定するよりも，クレーム件数から期待クレーム頻度を，クレーム単価（= クレーム総額 ÷ クレーム件数）から期待クレーム単価を別個のモデルにより推定し，その積により期待クレーム総額が推定されることが多い．

クレーム件数の確率分布には，離散分布として例えば次のポアソン分布が当てはめられる．

$$f_1(z_t; \mu_t^{(1)}, w_t) = \frac{(w_t \mu_t^{(1)})^{z_t}}{z_t!} e^{-w_t \mu_t^{(1)}}, \quad z_t = 0, 1, \ldots. \tag{2}$$

ここで，w_t は t 番目の契約のエクスポージャーすなわち既経過保険期間であり，観測値 z_t は t 番目の契約におけるクレーム件数である．このとき $\mathrm{E}(z_t) = \mathrm{Var}(z_t) = w_t \mu_t^{(1)}$ となり，$\mu_t^{(1)}$ は年間あたりの期待クレーム頻度を表す．ポアソン分布は指数分散モデルに属さないが，指数分散モデルとほぼ同様に一般化線形モデルで扱うことができる．なお，クレーム件数 z_t を既経過保険期間 w_t で除したクレーム頻度 $y_t = z_t / w_t$ を観測値として置き換えたときの y_t の分布は相対ポアソン分布と呼ばれ，$\theta_t = \log \mu_t^{(1)}$, $\phi = 1$ とおくことで指数分散モデル (1) に属することが確かめられる．

続いてクレーム件数 z_t が与えられたとき，クレーム単価の確率分布には，例えば次のガンマ分布が当てはめられる．

$$f_2(y_t; \mu_t^{(2)}, z_t, \phi) = \frac{1}{y_t \Gamma(z_t/\phi)} \left(\frac{y_t z_t}{\mu_t^{(2)} \phi} \right)^{z_t/\phi} \exp\left(-\frac{y_t z_t}{\mu_t^{(2)} \phi} \right), \quad y_t > 0. \tag{3}$$

ここで，観測値 y_t は t 番目の契約における z_t 件のクレームの平均単価である．このとき $\mu_t^{(2)}$ は期待クレーム単価を表し，$\mathrm{E}(y_t) = \mu_t^{(2)}$, $\mathrm{Var}(y_t) = \phi \left(\mu_t^{(2)} \right)^2 / z_t$ となる．なお，ここでのクレーム件数 z_t は式 (1) の重み w_t に相当し，$\theta_t = -1/\mu_t^{(2)}$ とおくと式 (1) の形を導くことができる．以上のモデル (2)，(3) から推定された期待クレーム頻度 $\mu_t^{(1)}$ と期待クレーム単価 $\mu_t^{(2)}$ の積 $\mu_t^{(1)} \mu_t^{(2)}$ が契約 $t = 1, \ldots, T$ ごとに収受すべき純保険料となる．

一般化線形モデルでは，t 番目の契約あるいは契約集団の第 i ファクターが属する区分を x_{ti} $t = 1, \ldots, T$, $i = 1, \ldots, p$ とおいたとき，上記の確率分布に現れた平均パラメータ μ_t は次式によって与えられる．

$$g(\mu_t) = \beta_0 + \beta_{1x_{t1}} + \cdots + \beta_{px_{tp}}, \quad t = 1, \ldots, T. \tag{4}$$

ここで，β_0 は切片，β_{ij} は第 i ファクターにおける第 j 区分の料率較差を表す未知パラメータである．ただし，パラメータの識別性を確保するためファクターごとに 1 つだけ料率較差をゼロとおいた基準区分を設ける．また，関数 g はリンク関数と呼ばれる微分可能な単調関数である．リンク関数 g が恒等関数 $g(y) \equiv y$ であるとき，平均パラメータ μ_t は式 (4) 右辺のようにファクターごとの料率較差の総和で表される．一方，リンク関数 g を対数関数 $g(y) = \log y$ でおくと，平均パラメータ μ_t は

$$\mu_t = \exp(\beta_0) \times \exp(\beta_{1x_{t1}}) \times \cdots \times \exp(\beta_{px_{tp}}), \quad t = 1, \ldots, T. \tag{5}$$

のようにファクターごとの料率較差の積によって表されることとなり，このとき $\beta_{11}, \ldots, \beta_{pn_p}$ は料率区分間の対数料率較差を表す未知パラメータとなる．料率算定において前者は加法モデル，後者は乗法モデルと呼ばれている．

式 (4) では個々のファクターごとに料率較差を与えているが，現実には複数のファクターの組合せにより料率較差が複雑に変化する場合も考えられる．そのような場合には，複数のファクターを合成した交互作用項を

与えることができる．例えば，第 1 ファクターから第 i ファクターまでの交互作用を考慮したモデルは次式のように与えられる．

$$g(\mu_t) = \beta_0 + \beta_{1\cdots i,(x_{t1},\ldots,x_{ti})} + \beta_{i+1,x_{t,i+1}} + \cdots + \beta_{px_{tp}}, \quad t = 1,\ldots,T. \tag{6}$$

ここで，$\beta_{1\cdots i,(x_{t1},\ldots,x_{ti})}$ は第 1 ファクターから第 i ファクターまでがそれぞれ第 $x_{t1},\ldots,x_{ti}$ 番目の区分に属するときの料率較差を表す交互作用項であり，合計で $n_{1\cdots i} = n_1 \times \cdots \times n_i$ 個の未知パラメータをもつこととなる．ただし，前述のように識別性を保証するために，任意のある 1 つの区分 $(j_1,\ldots,j_i)$ を基準区分に定めて $\beta_{1\cdots i,(j_1,\ldots,j_i)} = 0$ とおく必要がある．

一般化線形モデルの未知パラメータは，一般に最尤法により推定される．以降では，式 (4) あるいは (6) の右辺に現れる切片および料率較差のパラメータをまとめて $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \ldots, \beta_{pn_p})$ と表すと，ここでの未知パラメータとは $(\boldsymbol{\beta}, \phi)$ である．契約 $t = 1,\ldots,T$ ごとの観測値 y_t の確率分布が指数分散モデル (1) の形で与えられ，その平均が式 (4) から $\mu_t = \mu(\theta_t) = g^{-1}(\beta_0 + \beta_{1x_{t1}} + \cdots + \beta_{px_{tp}})$ と表されるとき，一般化線形モデルの対数尤度は

$$\log L(\boldsymbol{\beta}, \phi; y_1,\ldots,y_T, w_1,\ldots,w_T) = \sum_{t=1}^{T} \log f(y_t; \theta_t(\boldsymbol{\beta}), w_t, \phi) \tag{7}$$

となる．ただし $\theta_t(\boldsymbol{\beta}) = \mu^{-1} \circ g^{-1}(\beta_0 + \beta_{1x_{t1}} + \cdots + \beta_{px_{tp}})$ と表した．この対数尤度を最適化することにより，未知パラメータの推定値が次のように得られる．

$$\begin{aligned}(\hat{\boldsymbol{\beta}}, \hat{\phi}) &= \mathrm{argmin}_{(\boldsymbol{\beta},\phi)} - \log L(\boldsymbol{\beta}, \phi; y_1,\ldots,y_T, w_1,\ldots,w_T) \\ &= \mathrm{argmin}_{(\boldsymbol{\beta},\phi)} - \sum_{t=1}^{T} \log f(y_t; \theta_t(\boldsymbol{\beta}), w_t, \phi).\end{aligned} \tag{8}$$

ここで，料率算定のために平均に関するパラメータ $\boldsymbol{\beta}$ のみ推定すればよい場合には，指数分散モデルの形 (1) を利用して

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{t=1}^{T} \{-y_t \theta_t(\boldsymbol{\beta}) + b(\theta_t(\boldsymbol{\beta}))\} \tag{9}$$

のようにより単純な形に表すことができる．例えばポアソン分布 (2) の場合は元々ばらつきのパラメータ ϕ を持たないため

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(1)} &= \mathrm{argmin}_{\boldsymbol{\beta}^{(1)}} \sum_{t=1}^{T} \{w_t \mu_t^{(1)} - z_t \log \mu_t^{(1)}\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta}^{(1)}} \sum_{t=1}^{T} \{w_t g^{-1}(\beta_0^{(1)} + \beta_{1x_{t1}}^{(1)} + \cdots + \beta_{px_{tp}}^{(1)}) - z_t \log g^{-1}(\beta_0^{(1)} + \beta_{1x_{t1}}^{(1)} + \cdots + \beta_{px_{tp}}^{(1)})\}\end{aligned} \tag{10}$$

により全ての未知パラメータが推定される．また，ガンマ分布 (3) の場合は

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(2)} &= \mathrm{argmin}_{\boldsymbol{\beta}^{(2)}} \sum_{t=1}^{T} \left\{ \frac{z_t y_t}{\mu_t^{(2)}} + z_t \log \mu_t^{(2)} \right\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta}^{(2)}} \sum_{t=1}^{T} \left\{ \frac{z_t y_t}{g^{-1}(\beta_0^{(2)} + \beta_{1x_{t1}}^{(2)} + \cdots + \beta_{px_{tp}}^{(2)})} + z_t \log g^{-1}(\beta_0^{(2)} + \beta_{1x_{t1}}^{(2)} + \cdots + \beta_{px_{tp}}^{(2)}) \right\}\end{aligned} \tag{11}$$

となり最適化の目的関数がより単純になる．この式 (11) から先に $\boldsymbol{\beta}^{(2)}$ を推定し，その後 ϕ に関して式 (8) を最適化して推定するのもよい．また，ポアソン分布これらの最適化問題は，通常はニュートン法などの勾配情報を利用した最適化手法を用いることで高速に解くことができる．

3　Group fused lasso による料率区分のグルーピング

前節の一般化線形モデルでは，料率ファクターの区分数に比して標本サイズが十分に大きければ信頼できる料率較差が得られる．しかし，たとえば車両種類や被保険者年齢のように区分数が多い料率ファクターがある場合や，さらに別のファクターとの交互作用を考慮する場合には，料率較差の未知パラメータの数が膨大になるために推定精度が落ちて料率の信頼性を確保できない恐れがある．そのような場合の対処法として，性質やリスクの近しい区分同士を統合し，満足する信頼性が確保できるまで区分数を小さくすることが考えられる．そこで本節では，料率較差の推定に際して，較差の小さい区分同士を自動的に統合するための (group) fused lasso と呼ばれる手法を導入する．

ここでは，p 個のファクターのうち第 1 ファクターが多数の料率区分 $V = \{1, \ldots, n_1\}$ からなるものとする．この料率区分を少数グループへと統合するために，区分間の隣接関係を定義して隣接する区分のペアの集合を $E = \{e_1, \ldots, e_m\} \subseteq V \times V$ とおいた無向グラフ構造 (V, E) を導入する．このとき，第 1 ファクターの料率較差を表すパラメータ $\beta_{11}, \ldots, \beta_{1n_1}$ に関する L1 正則化項を設けて，次式のように料率較差のパラメータ $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \ldots, \beta_{pn_p})$ を推定することをグラフ上の fused lasso という．

$$(\hat{\boldsymbol{\beta}}, \hat{\phi}) = \mathrm{argmin}_{(\boldsymbol{\beta}, \phi)} \sum_{t=1}^{T} q(\boldsymbol{\beta}, \phi; y_t, w_t) + \kappa \sum_{(u,v) \in E} |\beta_{1u} - \beta_{1v}|. \tag{12}$$

ここで，右辺第 1 項の $q(\boldsymbol{\beta}, \phi; y_t, w_t)$ は一般の損失関数であり，一般化線形モデルにおいては式 (8) の負の対数尤度や，式 (9)，(11) あるいは (10) の目的関数が入る．上式の右辺第 2 項が L1 正則化項であり，これは隣接する区分間の料率較差を縮小あるいはゼロにする効果をもつ．L2 正則化項をもつリッジ回帰では料率較差が縮小されるもののゼロとなることはないが，L1 正則化を用いることで小さい料率較差はゼロにもっていかれ，料率較差がゼロとなった料率区分は実質的に統合され，自動的に料率区分の統合が達成されることとなる．右辺第 2 項の係数 κ は正則化パラメータと呼ばれ，正則化項の効果の大きさを調整する役割を担う．

さらに，前節のように第 1 ファクターから第 i ファクターまでの交互作用を考慮する場合には，交互作用項を含む料率較差のパラメータ $\boldsymbol{\beta} = (\beta_0, \beta_{1\cdots i,(1,\ldots,1)}, \ldots, \beta_{1\cdots i,(n_1,\ldots,n_i)}, \beta_{i+1,1}, \ldots, \beta_{pn_p})$ を推定するのに，次式のようなグループ正則化項を設けた group fused lasso が適用できる．

$$(\hat{\boldsymbol{\beta}}, \hat{\phi}) = \mathrm{argmin}_{(\boldsymbol{\beta}, \phi)} \sum_{t=1}^{T} q(\boldsymbol{\beta}, \phi; y_t, w_t) + \kappa \sum_{(u,v) \in E} \|\boldsymbol{\beta}_{1u} - \boldsymbol{\beta}_{1v}\|_2. \tag{13}$$

ここで，$\boldsymbol{\beta}_{1u} = (\beta_{1\cdots i,(u,1,\ldots,1)}, \ldots, \beta_{1\cdots i,(u,n_2,\ldots,n_i)})$ は交互作用項のパラメータを第 1 ファクターが u 番目の区分であるものに関してグループ化した $n_{2\cdots i} = n_2 \times \cdots \times n_i$ 次元ベクトルであり，また $\|\boldsymbol{\beta}_{1u} - \boldsymbol{\beta}_{1v}\|_2$ はベクトル間のユークリッドノルム

$$\|\boldsymbol{\beta}_{1u} - \boldsymbol{\beta}_{1v}\|_2 = \sqrt{\sum_{j_2=1}^{n_2} \cdots \sum_{j_i=1}^{n_i} (\beta_{1\cdots i,(u,j_2,\ldots,j_i)} - \beta_{1\cdots i,(v,j_2,\ldots,j_i)})^2}$$

である．右辺第 2 項のグループ正則化項は，式 (12) の L1 正則化項と同様に第 1 ファクターに関して隣接した区分間の料率較差を縮小あるいはゼロにする効果をもつ．特に，第 1 ファクターの隣接区分 u と v に関して料率較差がゼロすなわち $\|\boldsymbol{\beta}_{1u} - \boldsymbol{\beta}_{1v}\|_2 = 0$ となるとき，第 2 ファクターから第 i ファクターまでの全ての区分 $j_2, \ldots, j_i$ に対して第 1 ファクターの隣接区分 u と v の間の料率較差が全てゼロすなわち $\beta_{1\cdots i,(u,j_2,\ldots,j_i)} = \beta_{1\cdots i,(v,j_2,\ldots,j_i)}$ となることがわかる．ゆえに，料率較差がゼロとなった第 1 ファクターの区分同士は，完全に同一料率が適用されるためタリフ上統合することができる．

以上で述べた (group) fused lasso では，正則化パラメータ κ の値を与えておく必要がある．正則化パラメータの大きさは料率較差の縮小度合いや区分の統合のしかたに影響するため，所与のデータに対して適切な値を定める必要がある．適切な正則化パラメータ κ の値の決め方としては，次の N 分割交差確認法と呼ばれる手法を用いるのが一般的である．まず，標本サイズ T の全データ（契約）を N 個のグループ $T_1, \ldots, T_N \subset \{1, \ldots, T\}$ に分割し，そのうち k 番目のグループを除いた標本を用いて式 (12) あるいは (13) により推定されたパラメータを $\hat{\boldsymbol{\beta}}_{-k}$, $\hat{\phi}_{-k}$ とおく．そして，推定の際に除かれた k 番目のグループのデータに推定値を当てはめた次式の交差確認損失が最小となるように正則化パラメータ κ を定めるのが N 分割交差確認法である．

$$\hat{\kappa} = \mathrm{argmin}_{\kappa} \sum_{k=1}^{N} \sum_{t \in T_k} -\log f(y_t; \hat{\boldsymbol{\beta}}_{-k}, w_t, \hat{\phi}_{-k}) \tag{14}$$

ここで，κ の値は連続的に最適化されるよりも，十分な候補値の集合 Λ を与えてその中から選択されるのが一般的である．標本の分割数 N には，5 から 10 程度が設定されることが多い．

以上の議論では，期待クレーム頻度と期待クレーム単価を別々に推定してきたが，最終的なタリフの料率区分を統合するためには，期待クレーム頻度と期待クレーム単価とで統合する区分を揃える必要がある．そこで，第 1 ファクターに関して，期待クレーム頻度の料率較差と期待クレーム単価の料率較差を同時にグループ化した正則化項を与えて，次のように期待クレーム頻度と期待クレーム単価を同時推定することを考える．

$$(\hat{\boldsymbol{\beta}}, \hat{\phi}) = \mathrm{argmin}_{(\boldsymbol{\beta}, \phi)} -\sum_{t=1}^{T} \{\log f_1(z_t; \mu_t^{(1)}(\boldsymbol{\beta}^{(1)}), w_t) + \log f_2(y_t; \mu_t^{(2)}(\boldsymbol{\beta}^{(2)}), z_t, \phi)\} + \kappa \sum_{(u,v) \in E} \|\boldsymbol{\beta}_{1u} - \boldsymbol{\beta}_{1v}\|_2. \tag{15}$$

ここで，f_1 と f_2 はポアソン分布の確率関数 (2) およびガンマ分布の確率密度関数 (3) であり，式 (15) の損失関数はクレーム件数 z_t とクレーム単価 y_t の同時分布に関する負の対数尤度となっている．また，料率較差のパラメータに関しては，期待クレーム頻度のパラメータを $\boldsymbol{\beta}^{(1)}$，期待クレーム単価のパラメータを $\boldsymbol{\beta}^{(2)}$ と表し，それらをまとめて $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$, $\boldsymbol{\beta}_{1u} = (\boldsymbol{\beta}_{1u}^{(1)}, \boldsymbol{\beta}_{1u}^{(2)})$ と表記している．このように，期待クレーム頻度のパラメータと期待クレーム単価のパラメータを一緒にグループ化した正則化項を設けることで，式 (13) のときと同様に第 1 ファクターに関して期待クレーム頻度の較差と期待クレーム単価の較差が同時にゼロとなるため，純保険料すなわち期待クレーム総額の料率区分を統一することができる．

式 (15) の正則化パラメータ κ の選択には，前述の N 分割交差確認法を同様に適用することができる．交差確認損失には式 (15) の損失関数に合わせクレーム件数 z_t とクレーム単価 y_t の同時分布を用いることもできるが，ここでは，クレーム総額が予測目標であるとの考えから，クレーム総額 $s_t = y_t z_t$ の周辺分布による負の対数尤度を採用する．特に，クレーム件数にポアソン分布 (2)，クレーム単価にガンマ分布 (3) を仮定したとき，クレーム総額は Tweedie(1984) により提案された Tweedie 分布と呼ばれる特殊な確率分布に従い，このとき N 分割交差確認法では次式の交差確認損失を最小にする κ を採用することとなる．

$$\hat{\kappa} = \mathrm{argmin}_{\kappa} \sum_{k=1}^{N} \sum_{t \in T_k} -\log \sum_{z=1}^{\infty} \left\{ f_1(z; \hat{\boldsymbol{\beta}}_{-k}^{(1)}, w_t) \frac{f_2(s_t/z; \hat{\boldsymbol{\beta}}_{-k}^{(2)}, z, \hat{\phi}_{-k})}{z} \right\}. \tag{16}$$

ただし，$\hat{\boldsymbol{\beta}}_{-k}^{(1)}$, $\hat{\boldsymbol{\beta}}_{-k}^{(2)}$, $\hat{\phi}_{-k}$ は式 (14) と同様に，標本サイズ T の全データ（契約）を N 個のグループ $T_1, \ldots, T_N \subset \{1, \ldots, T\}$ に分割したときの，k 番目のグループを除いた標本から推定された各パラメータを表す．この式 (16) は，クレーム件数 z とクレーム総額 s_t の同時分布について，z に関して周辺化したクレーム総額分布による負の対数尤度となっている．

4 Group fused lasso の最適化アルゴリズム

本節では，前節で導入したパラメータ推定のための正則化項付き最適化を実行するためのアルゴリズムを示す．第 2 節で紹介した正則化項のない対数尤度の最適化 (8) においては，ニュートン法などの勾配を用いた最適化手法を使うことで高速にパラメータの推定値を得ることができる．しかしながら，前節で導入した正則化項は，その零点において勾配が存在せず，勾配に基づく連続最適化アルゴリズムがそのままでは適用できない．Group fused lasso の最適化アルゴリズムとして，Bleakley and Vert(2011) ではブロック座標降下法，Wahlberg et al.(2012) では交互方向乗数法，Wytock et al.(2014) では Active Set Projected Newton 法が提案されている．ブロック座標降下法は，鎖状グラフ上の Group fused lasso に対して，通常の Grouped lasso にパラメータ変換することで適用できる手法であり，一般のグラフに対しては適用できない．また，Wytock et al.(2014) による Active Set Projected Newton 法は一般のグラフ上の Group fused lasso に対して高速に解が得られるものの，残差平方和を損失関数とした場合の解法に特化しており，一般化線形モデルの対数尤度を損失関数としての適用は難しい．一方，交互方向乗数法は汎用性が高く，一般のグラフおよび凸性を満たす一般の損失関数に適用できるため，本稿では交互方向乗数法を用いた最適化アルゴリズムを採用する．

交互方向乗数法は，ラグランジュの未定乗数法を拡張した最適化手法であり，元の目的関数に拡張ラグランジアンと呼ばれる項を加えた上で，損失関数と正則化項を分けて交互に最適化していく手法である．拡張ラグランジアンを導入する前に，式 (13) における最適化を次の同値な制約付き最適化問題へと書き換える．

$$\begin{aligned} \min_{(\boldsymbol{\beta},\phi,\boldsymbol{\xi})} \quad & \sum_{t=1}^{T} q(\boldsymbol{\beta},\phi;y_t,w_t) + \kappa \sum_{l=1}^{m} \|\boldsymbol{\xi}_l\|_2, \\ \text{s.t.} \quad & \boldsymbol{\xi}_l = \boldsymbol{\beta}_{1e_{l1}} - \boldsymbol{\beta}_{1e_{l2}}, \qquad l = 1,\ldots,m. \end{aligned} \tag{17}$$

ここで，$\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m)$ は $mn_{2\cdots i}$ 次元ベクトルであり，また $l = 1,\ldots,m$ について無向グラフ (V,E) の l 番目の辺を $e_l = (e_{l1},e_{l2}) \in E \subseteq V \times V$ と表記している．式 (17) の制約式は，さらに次のように 1 つにまとめることができる．

$$\begin{aligned} \min_{(\boldsymbol{\beta},\phi,\boldsymbol{\xi})} \quad & \sum_{t=1}^{T} q(\boldsymbol{\beta},\phi;y_t,w_t) + \kappa \sum_{l=1}^{m} \|\boldsymbol{\xi}_l\|_2, \\ \text{s.t.} \quad & \boldsymbol{\xi} = \boldsymbol{\beta}_1 A. \end{aligned} \tag{18}$$

ただし，$\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{11},\ldots,\boldsymbol{\beta}_{1n_1})$ は 1 番目から i 番目までの交互作用項のパラメータをまとめた $n_{1\cdots i}$ 次元ベクトルである．また，A は $n_{1\cdots i} = n_1 n_{2\cdots i}$ 行 $n_{2\cdots i}m$ 列の行列であり，次式のように $n_{2\cdots i}$ 次正方行列 A_{lv}, $l = 1,\ldots,m$, $v = 1,\ldots,n_1$ によるブロック行列で表される．

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n_1 1} & \cdots & A_{n_1 m} \end{pmatrix}, \qquad A_{lv} = \begin{cases} I, & \text{if } e_{l1} = v, \\ -I, & \text{if } e_{l2} = v, \\ O, & \text{otherwise.} \end{cases}$$

ただし，I は単位行列，O は零行列である．

ここで，制約付き最適化問題を解くための拡張ラグランジアンを導入する．通常のラグランジアンは目的関数に対して $-\boldsymbol{\lambda}(\boldsymbol{\beta}_1 A - \boldsymbol{\xi})$ という $mn_{2\cdots i}$ 次元の未定乗数 $\boldsymbol{\lambda}$ を含んだ新たな項 $-\boldsymbol{\lambda}(\boldsymbol{\beta}_1 A - \boldsymbol{\xi})$ を加えて最適化を行うものであるが，拡張ラグランジアンではさらに項を 1 つ追加した $-\boldsymbol{\lambda}(\boldsymbol{\beta}_1 A - \boldsymbol{\xi}) + \frac{\rho}{2}\|\boldsymbol{\beta}_1 A - \boldsymbol{\xi}\|_2^2$ を目的関数に加えることとなる．ただし，ρ は任意の定数である．通常のラグランジアンと同様に，拡張ラグランジアンを目的関数に加えることで，未定乗数 $\boldsymbol{\lambda}$ を変数に加える代わりに制約式を外した制約なし最適化問題を解くことにより式 (18) と同じ最適解を得ることができる．そして，交互方向乗数法は，拡張ラグランジアンを加

えた目的関数について損失関数と正則化項を分離して，未定乗数 $\boldsymbol{\lambda}$ を更新しながら次のように交互に最適化を繰り返して最適解へと収束させる手法である．

$$(\boldsymbol{\beta}^{\mathrm{new}}, \phi^{\mathrm{new}}) = \mathrm{argmin}_{(\boldsymbol{\beta},\phi)} \sum_{t=1}^{T} q(\boldsymbol{\beta}, \phi; y_t, w_t) - \boldsymbol{\lambda}(\boldsymbol{\beta}_1 A - \boldsymbol{\xi}) + \frac{\rho}{2}\|\boldsymbol{\beta}_1 A - \boldsymbol{\xi}\|_2^2, \tag{19}$$

$$\boldsymbol{\xi}^{\mathrm{new}} = \mathrm{argmin}_{\boldsymbol{\xi}}\, \kappa \sum_{l=1}^{m} \|\boldsymbol{\xi}_l\|_2 - \boldsymbol{\lambda}(\boldsymbol{\beta}_1^{\mathrm{new}} A - \boldsymbol{\xi}) + \frac{\rho}{2}\|\boldsymbol{\beta}_1^{\mathrm{new}} A - \boldsymbol{\xi}\|_2^2, \tag{20}$$

$$\boldsymbol{\lambda}^{\mathrm{new}} = \boldsymbol{\lambda} - \rho(\boldsymbol{\beta}_1^{\mathrm{new}} A - \boldsymbol{\xi}^{\mathrm{new}}). \tag{21}$$

上記の 3 つの更新式を経て得たパラメータの組 $(\boldsymbol{\beta}^{\mathrm{new}}, \phi^{\mathrm{new}}, \boldsymbol{\xi}^{\mathrm{new}}, \boldsymbol{\lambda}^{\mathrm{new}})$ を新たに $(\boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \boldsymbol{\lambda})$ とおいて，再び上の更新式を 1 つ目から繰り返していくことで最適解へと収束させることができる．

1 つ目の更新式 (19) における最適化は，損失関数 q が微分可能であれば勾配を用いた通常の最適化手法により高速に解くことができる．また，2 つ目の更新式 (20) は，実際には次の式によって最適解が解析的に得られる．

$$\boldsymbol{\xi}_l^{\mathrm{new}} = \begin{cases} \left(1 - \dfrac{\kappa}{\|\rho\boldsymbol{\beta}^{\mathrm{new}} A_l - \boldsymbol{\lambda}_l\|_2}\right)(\boldsymbol{\beta}^{\mathrm{new}} A_l - \boldsymbol{\lambda}_l/\rho), & \text{if } \|\rho\boldsymbol{\beta}^{\mathrm{new}} A_l - \boldsymbol{\lambda}_l\|_2 > \kappa, \\ \mathbf{0}, & \text{if } \|\rho\boldsymbol{\beta}^{\mathrm{new}} A_l - \boldsymbol{\lambda}_l\|_2 \le \kappa, \end{cases} \qquad l = 1, \ldots, m.$$

ただし，A_l および $\boldsymbol{\lambda}_l$ は，行列 A とベクトル $\boldsymbol{\lambda}$ をそれぞれ $n_{2\cdots i}$ 列ずつに分割して $A = (A_1 \ \cdots \ A_m)$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ と表したときの l 番目の部分行列および部分ベクトルを指す．最後の更新式 (21) では，定数 ρ が変数 $\boldsymbol{\lambda}$ の更新幅を定めていることがわかる．定数 ρ は小さすぎると最適解への収束が遅くなるが，大きすぎるとパラメータが大きく振動して最適解に収束しないため，試行錯誤によって適切な値を設定する必要がある．

なお，式 (15) に示した期待クレーム頻度と期待クレーム単価の同時推定モデルにおいても上記のアルゴリズムは同様に適用できるが，特に更新式 (19) においてクレーム件数モデルのパラメータ $\boldsymbol{\beta}^{(1)}$ とクレーム単価モデルのパラメータ $(\boldsymbol{\beta}^{(2)}, \phi)$ の項が完全に分離されるため，それぞれのパラメータについて別々に最適化すればよいことになる．

5　自動車損害賠償責任保険の解析例

本節では，前節までに解説した手法について，損害保険料率算出機構がディスクロージャーとして公表している「損害保険料率算出機構統計集」における自動車損害賠償責任保険のクレームデータへと適用した解析例を示す．

自動車損害賠償責任保険の純保険料率（基準料率）は，普通乗用自動車・軽自動車などの車種，主に営業用・自家用の別による用途，本土・沖縄本島・本土離島・沖縄離島の 4 区分からなる地域の 3 つのファクターのみで構成されている．損害保険料率算出機構より年度ごとに発行されている損害保険料率算出機構統計集では，自動車損害賠償責任保険について車種別，用途別，そして都道府県別に契約および支払の集計がされており，基準料率の 4 地域区分よりも細かい都道府県ごとのクレーム発生頻度やクレーム単価を算出し比較することができる．そこで，本解析では幾つかの車種・用途を取り上げて，本土地域を 46 都道府県に分けて期待クレーム頻度と期待クレーム単価，ひいては期待クレーム総額すなわち純保険料の推定を試みる．

ただし，46 もある都道府県をそのまま料率区分とすると，区分あたりのデータ量が少なくなり料率算定の信頼性が損なわれることから，前節までで導入した group fused lasso の手法を用いて幾つかの都道府県を統合した新たな料率区分を構成することを考える．都道府県の隣接関係については，公道により接続された都道府県同士を辺で繋ぎ，図 1 に示した隣接構造をもつ無向グラフ上の group fused lasso を導入する．

本節では，普通貨物自動車の都道府県別・用途別・重量別での期待クレーム単価の推定と，自家用乗合自動車の都道府県別の期待クレーム頻度・期待クレーム単価・期待クレーム総額（純保険料）の推定の 2 つの解析例を扱う．

5.1　普通貨物自動車の期待クレーム単価の推定

自動車損害賠償責任保険では普通貨物自動車について，営業用または自家用の用途別，および，2 トン超または 2 トン以下の重量別に分けて純保険料を算定している．ここでは，さらに沖縄県および離島地域を除いた都道府県別に細分化したときの期待クレーム単価の推定を行う．損害保険料率算出機構統計集では，普通貨物自動車について年度ごとに都道府県別・用途別・重量別の支払件数および支払保険金総額が集計されており，さらにその内訳として死亡事故による支払件数および支払保険金総額が掲載されている．死亡事故による高額支払は期待クレーム単価の推定に外れ値として影響を及ぼしうるため，ここでは平成 22 年度から平成 26 年度までの死亡事故を除いた都道府県別・用途別・重量別の支払件数をクレーム件数，支払保険金をクレーム総額として，そこからクレーム単価をクレーム総額÷クレーム件数により算出し，期待クレーム単価の解析に用いる．

期待クレーム単価の推定は，ガンマ分布 (3) と対数リンク関数 $g(\mu) = \log \mu$ による乗法モデル (5) を用いた一般化線形モデルを当てはめ，L1 正則化項付き対数尤度の最適化によって行う．

L1 正則化項は図 1 に示した隣接する都道府県間の対数期待クレーム単価の較差へと適用し，L1 正則化項の正則化パラメータ κ は，都道府県の較差が完全にゼロとなる境界値 κ_0 を最大値とした $\kappa_i = \kappa_0 10^{-\frac{3i}{99}}$, $(i = 99, 98, \ldots, 0)$ を候補値として，負の対数尤度を交差確認損失として年度ごとにデータをグループ化した 5 分割交差確認法により選択する．ただし，較差がゼロとなる境界値 κ_0 は解析的には求まらないため，まず適当な初期値 $\kappa = 1$ から較差がゼロになるまで κ を 10 倍ずつ倍増させ，はじめて較差がゼロとなった値とその前の値との区間を二分探索することにより境界値 $\kappa = \kappa_0$ を求めた．最適化アルゴリズムには前節で紹介した交互方向乗数法を用い，定数 ρ については試行錯誤の結果，正則化パラメータと同じ値 $\rho = \kappa$ に設定することで安定して早い収束が得られた．

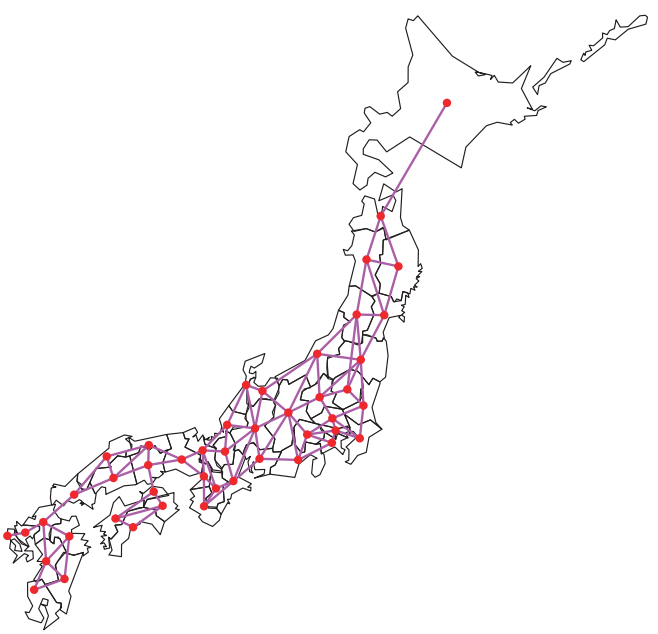

図 1　都道府県の隣接関係を表す無向グラフ

さらに，ファクター間の交互作用について，(a) 交互作用のない場合，(b) 都道府県×用途の交互作用，(c) 都道府県×重量の交互作用，(d) 都道府県×用途×重量の交互作用を導入する場合の 4 つの候補となるモデルを検討し，式 (14) に示された負の対数尤度による交差確認損失が最も小さくなるモデルを選択することとする．

以上の方法によって，交互作用の候補モデルごとに選択された正則化パラメータと，そのときの都道府県の料率区分数および負の対数尤度による交差確認損失を表 1 に示している．ただし都道府県の料率区分数とは，L1 正則化により較差ゼロとなった区分を統合した後に残った料率区分数を指す．表 1 より，交差確認損失を最小とするモデルは，都道府県と用途（営業用または自家用）の交互作用のみを取り入れたモデルであることがわかる．このことは，用途が営業用か自家用かによって，普通貨物自動車の期待クレーム単価の地域的傾向が異なることを意味している．また，選択されたモデルでは都道府県と重量（2 トン超または 2 トン以下）の間に交互作用がないため，重量が 2 トン超か 2 トン以下かによって期待クレーム単価の地域的傾向は変化しないこととなる．

選択された都道府県と用途の交互作用モデルにおける，正則化パラメータによる交差確認損失と期待クレーム単価較差の変化を図 2 に示した．図 2(a) から (c) にかけて描かれた緑の縦点線は，最小の交差確認損失（6042.7）をとる正則化パラメータの値 $\kappa = 86.6$ を表している．図 2(a) から見てとれるように，交差確認損失は正則化パラメータが小さすぎても大きすぎても悪くなる．なお，L1 正則化項を除外し，区分統合をせずに最尤法で推定した場合の交差確認損失は 6104.3 となるため，L1 正則化項の導入によって新規データに対する予測性能が大きく向上していることがわかる．また，ナイーブな区分統合の方法として，46 都道府県を八地方区分（北海道地方，東北地方，関東地方，中部地方，近畿地方，中国地方，四国地方，九州地方）に統合して通常の最尤法で一般化線形モデルを推定したところ，交差確認損失は 6045.0 となり区分統合しない場合よりも大きく改善したが，提案手法の方が予測精度をより向上させる結果となった．

図 2(b) と (c) では，北海道の自家用普通貨物自動車を基準とした対数期待クレーム単価較差について，正則化パラメータの値を変化させたときの推移を営業用と自家用にわけて示している．正則化パラメータの値が大きくなるにつれて 46 都道府県間の較差は縮まっていき，徐々に区分が統合されて最終的に 1 つの値に収束する様子が見てとれる．都道府県と用途の間に交互作用を導入しているため営業用と自家用とで較差は異なっているが，用途でグループ化された L1 正則化項によって，同じ正則化パラメータ値にて営業用と自家用で同時に区分統合が行われていることに注意されたい．緑縦線の交差確認損失を最小にする正則化パラメータ値では，表 1 にもある通り 46 都道府県のうち幾つかの区分が統合されて 25 個の都道府県区分へと再編成されており，タリフを作る上では 25 種類の都道府県区分を設けて純保険料率を定めればよいことになる．

続いて表 2 に，25 個の都道府県区分の構成と，区分ごとの期待クレーム単価の推定値を示した．東北地方 6 県と四国地方 4 県はそれぞれ地方全体で 1 つの区分を構成しており，その他にも中部地方，中国地方，九州地方では大きくまとまった区分に統合されている．一方で，関東地方や近畿地方などの契約台数が多いためデータ量が豊富な地域では，統合されず単独の都道府県からなる区分が多い．また，同じ都道府県区分および重量区分で営業用と自家用の期待クレーム単価を比較すると，概して営業用普通貨物自動車の方が高いものの，東京都と埼玉県に限り大小関係が逆転していることがわかる．なお，交互作用のない重量（2 トン超または 2 ト

表 1　交互作用モデルの比較

交互作用をとるファクター	正則化パラメータ	都道府県の料率区分数	交差確認損失（負の対数尤度）
(a) 交互作用なし	147.3	12	6050.4
(b) 都道府県×用途	86.6	25	6042.7
(c) 都道府県×重量	75.3	27	6053.6
(d) 都道府県×用途×重量	108.7	24	6050.1

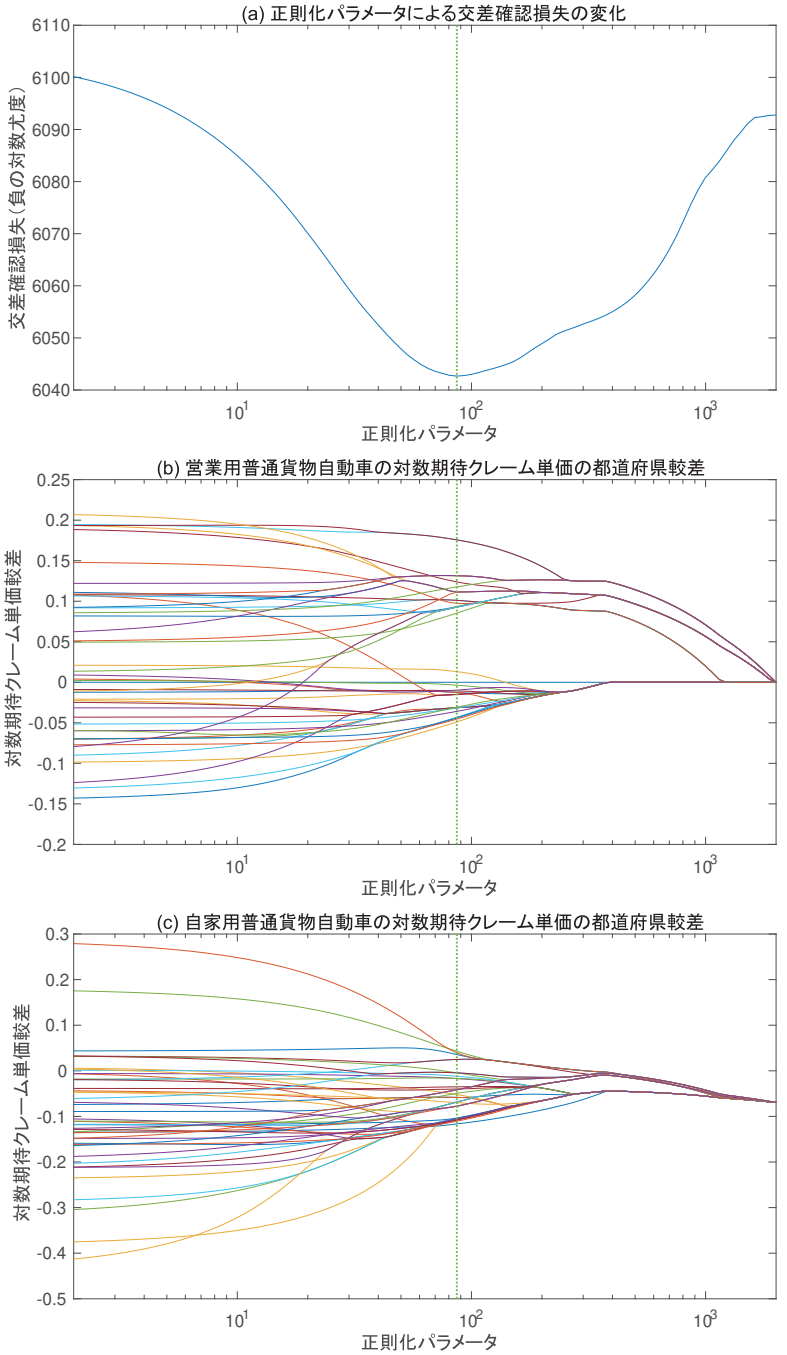

図 2　正則化パラメータによる交差確認損失と期待クレーム単価較差の変化

ン以下）による対数期待クレーム単価較差は 0.138 と推定されており，すべての都道府県と用途において，2 トン超の普通貨物自動車は 2 トン以下のものに比べて期待クレーム単価がおよそ $\exp(0.138) = 1.15$ 倍に推定されている．

最後に図 3 には，表 3 に示した期待クレーム単価の推定値を地図上にカラースケールで示している．図 3(a) および (b) の営業用普通貨物自動車から単価の地域的傾向を読み取ると，東日本の中では都道府県間に大きな差はなく，そこから西へ進むにつれて徐々に期待クレーム単価が上昇する傾向となっている．一方，図 3(c) および (d) の自家用普通貨物自動車については，同じように東日本より西日本の方が単価が高いものの，その差

表 2　普通貨物自動車の期待クレーム単価の推定値（千円）

都道府県の区分	営業用 2 トン超	営業用 2 トン以下	自家用 2 トン超	自家用 2 トン以下
北海道	832	724	740	645
青森，岩手，宮城，秋田，山形，福島	806	702	748	651
茨城，栃木	821	715	785	683
群馬	792	690	789	688
埼玉	802	699	818	712
千葉	829	722	818	712
東京	806	702	822	716
神奈川	823	717	803	699
新潟，長野	796	694	749	652
富山，石川，福井，岐阜	819	714	747	650
山梨	807	702	792	689
静岡	796	693	750	653
愛知	795	692	745	649
三重	843	734	776	676
滋賀	824	717	749	652
京都	906	789	868	756
大阪，兵庫	920	801	853	743
奈良	913	795	862	751
和歌山	917	799	865	754
鳥取，島根，岡山，広島	913	795	777	676
山口	941	820	799	696
徳島，香川，愛媛，高知	930	810	770	670
福岡	936	815	828	721
佐賀，長崎	992	864	828	721
熊本，大分，宮崎，鹿児島	949	826	798	695

は図 3(a)(b) の営業用に比べると縮小しており，その他に関東や関西などの大都市圏で単価が上昇する傾向が現れている．これらの観察結果から，同じ車種であってもその用途によってクレーム単価の地域的傾向に違いがあるものと結論付けられる．

5.2　自家用乗合自動車の純保険料率の推定

ここでは 2 つ目の解析例として，自動車損害賠償責任保険の自家用乗合自動車の統計データを利用して，都道府県別の期待クレーム頻度・期待クレーム単価を推定し，それらの積により純保険料率すなわち期待クレーム総額の推定を行う．乗合自動車とは，バスなどの乗車定員 11 人以上の自動車で，貨物自動車等及び特種用途自動車等以外のものを指す．

本解析に自家用乗合自動車を選んだ理由は，クレーム頻度を算出するための既経過保険期間が推計しやすい点にある．自動車損害賠償責任保険では保険期間が長期に及び，かつ初度登録時など条件により保険期間が異

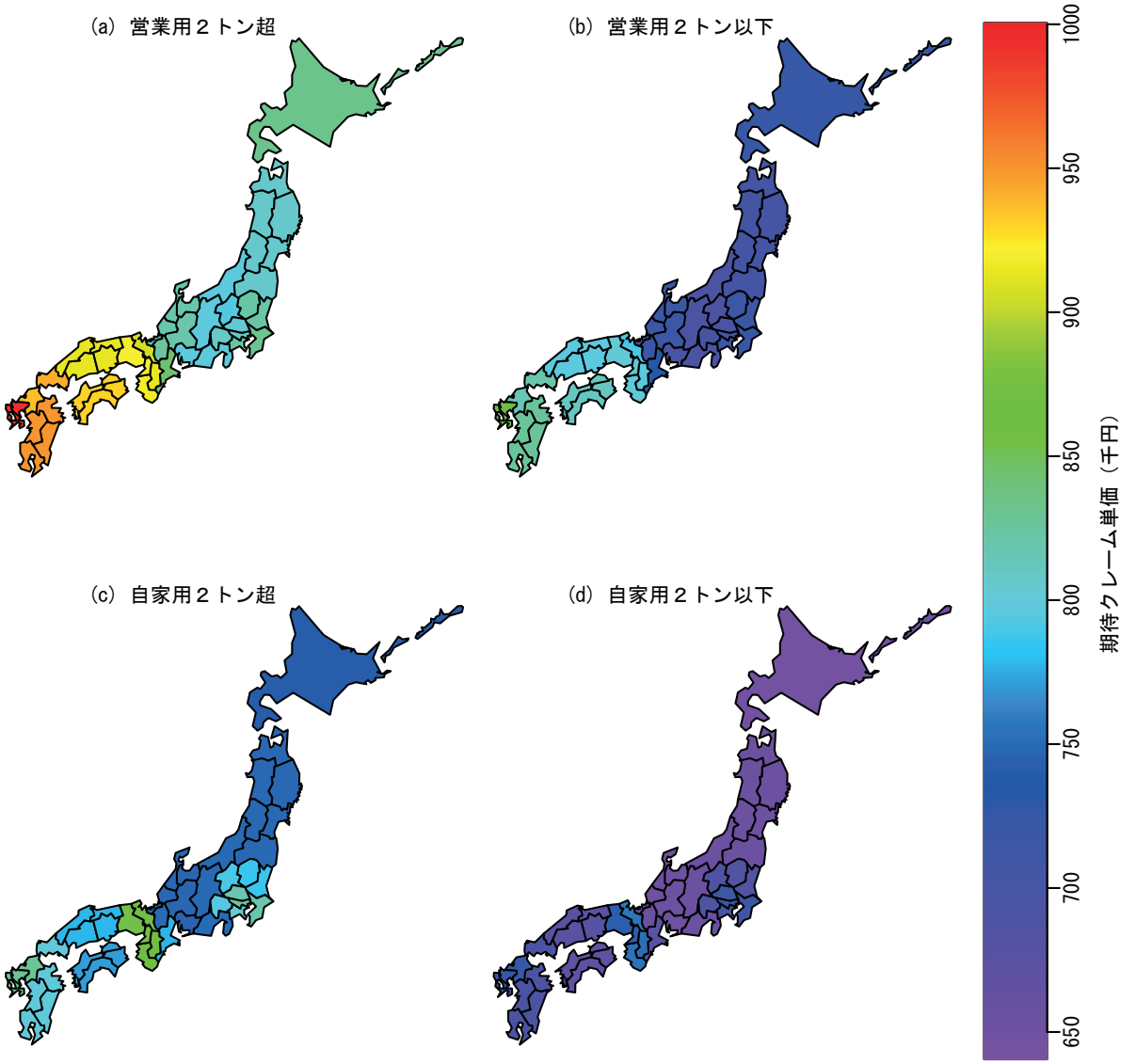

図 3　普通貨物自動車の期待クレーム単価の統計地図

なるような車種が多いため，損害保険料率算出機構統計集の数値のみから既経過保険期間を推計するのが難しい．しかし，自家用乗合自動車については保険期間が 1 ヵ月から 13 ヵ月までに限られ，さらに単年度の新契約保険料を新契約台数で割った保険料単価が保険期間 1 年（12 ヵ月）の基準料率と 1% 未満の差でほぼ一致しているため，単年度の既経過保険期間を推計するにあたりすべて保険期間 1 年で契約されているものと近似的に見なすことができる．よってここでは，ある単年度における既経過保険期間を，その年度と前年度の新契約台数の平均値によって推計することとする．

以上のようにして得られた平成 22 年度から平成 26 年度までの沖縄県および離島地域を除いた都道府県別の既経過保険期間と，前の解析と同じく死亡事故を除いた都道府県別のクレーム件数およびクレーム単価を用いて，式 (15) に示したクレーム件数・クレーム単価の同時モデルにより期待クレーム頻度および期待クレーム単価を推定し，そこから純保険料（＝期待クレーム頻度 × 期待クレーム単価）を算定することを目指す．クレーム件数には式 (2) のポアソン分布，クレーム単価には式 (3) のガンマ分布を当てはめ，対数リンク関数を用い

て都道府県間の対数期待クレーム頻度および対数期待クレーム単価の較差を推定する．推定の際には，図 1 の隣接する都道府県間の対数期待クレーム頻度較差と対数期待クレーム単価較差をグループ化した L1 正則化項を導入し，46 ある都道府県区分をより少ない区分数となるよう統合させる．

グループ正則化項の正則化パラメータ κ は前の解析と同様に，都道府県の較差が完全にゼロとなる境界値 κ_0 を最大値とした $\kappa_i = \kappa_0 10^{-\frac{3i}{99}},\ (i = 99, 98, \ldots, 0)$ を候補値として，式 (16) の交差確認損失を用い，年度ごとにデータをグループ化した 5 分割交差確認法により選択することとする．ここで，較差がゼロとなる境界値 κ_0 は前の解析と同じ方法により求め，定数 ρ は正則化パラメータと同じ値 $\rho = \kappa$ へと設定した．なお，この解析では単一の車種・用途のみ扱うため，ファクターは都道府県のみであり交互作用は考慮する必要がない．

本解析で得られた，正則化パラメータの値ごとの交差確認損失と期待クレーム頻度および期待クレーム単価の較差を図 4 に示した．図 4(a) から (c) にかけて描かれた緑の縦点線は，交差確認損失を最小とする正則化パラメータの値 $\kappa = 4.42$ を表している．このときの交差確認損失は 1602.5 であり，正則化パラメータをゼロとおいて L1 正則化項を除外した場合の交差確認損失 1638.1 と比べて，新規データへの予測性能は大きく改善していると言える．なお，前の解析と同様に，46 都道府県を八地方区分に統合して通常の最尤法で一般化線形モデルを推定したときの交差確認損失は 1623.4 となり，提案手法の方が予測精度を大きく向上させていることが示された．

図 4(b) と (c) は，正則化パラメータに対する，北海道を基準とした対数期待クレーム頻度および対数期待クレーム単価の較差の変化を示している．図 2 のときと同様に，正則化パラメータの値が大きくなるにつれて 46 都道府県の較差は縮まり最終的に 1 つの値に統合されているが，図 4(b) のクレーム頻度よりも図 4(c) のクレーム単価の方が較差が縮まるのが早く，緑縦線の選択された正則化パラメータ値においては，対数期待クレーム頻度較差のばらつきが対数期待クレーム単価較差に比べて大きいことがわかる．ただし，グループ正則化項の働きにより，都道府県間の較差が完全にゼロとなり統合されるのはクレーム頻度とクレーム単価で同時であることに注意が必要である．緑縦線の交差確認損失を最小にする正則化パラメータ値においては，いずれも 11 の都道府県が他に統合されて，35 個の都道府県区分が残っている．

続く表 3 には，35 個の都道府県区分の構成と，区分ごとの期待クレーム単価の推定値を示した．東北地方，中部地方，中国地方などで幾つかの都道府県が統合されているが，前の普通貨物自動車の解析例と比べるとあまり統合されずに単一の都道府県からなる区分が多い．都道府県区分ごとの推定値を比較すると，期待クレーム単価の較差は最大でも 1.5 倍程度であるのに対して，期待クレーム頻度の較差は大きく，大阪府と青森県とで 2.4 倍もの開きがある．さらに，それらを掛け合わせた期待クレーム総額では，大阪府と島根県・山口県との間に約 3.2 倍もの較差が生じており，本州内でも地域によってリスクにかなり大きな差があることがわかる．

図 5 には，都道府県別の期待クレーム頻度，期待クレーム単価，そしてその積である期待クレーム総額の推定値について，地図上にカラースケールで示している．図 5(a) に示された期待クレーム頻度は，東京，大阪とその近郊が高い他，北海道や愛知，福岡などの比較的大都市のある地域が比較的高い傾向にある．図 5(b) の期待クレーム単価については，東京と大阪とその周辺で高い傾向は同じであるが，北海道では逆にとても低くなっているなど，期待クレーム頻度との違いも見受けられる．ただし，前述のように期待クレーム単価の較差は期待クレーム頻度に比べて小さいため，これらを乗じた図 5(c) の期待クレーム総額すなわち純保険料率の地域的傾向は，期待クレーム頻度と概ね近いものとなっている．

6　終わりに

本稿では，代表的なクラス料率算定法である一般化線形モデルの最尤推定に際し，対数尤度に L1 正則化項を加えて最適化することにより，料率区分の適切なグルーピングと料率算定を同時に実行できる手法を提案した．特に，ファクター間の交互作用を考える場合や，クレーム頻度とクレーム単価に対して同じグルーピングを課したい場合に対応するための group fused lasso を導入した．モデルパラメータの推定アルゴリズムには，

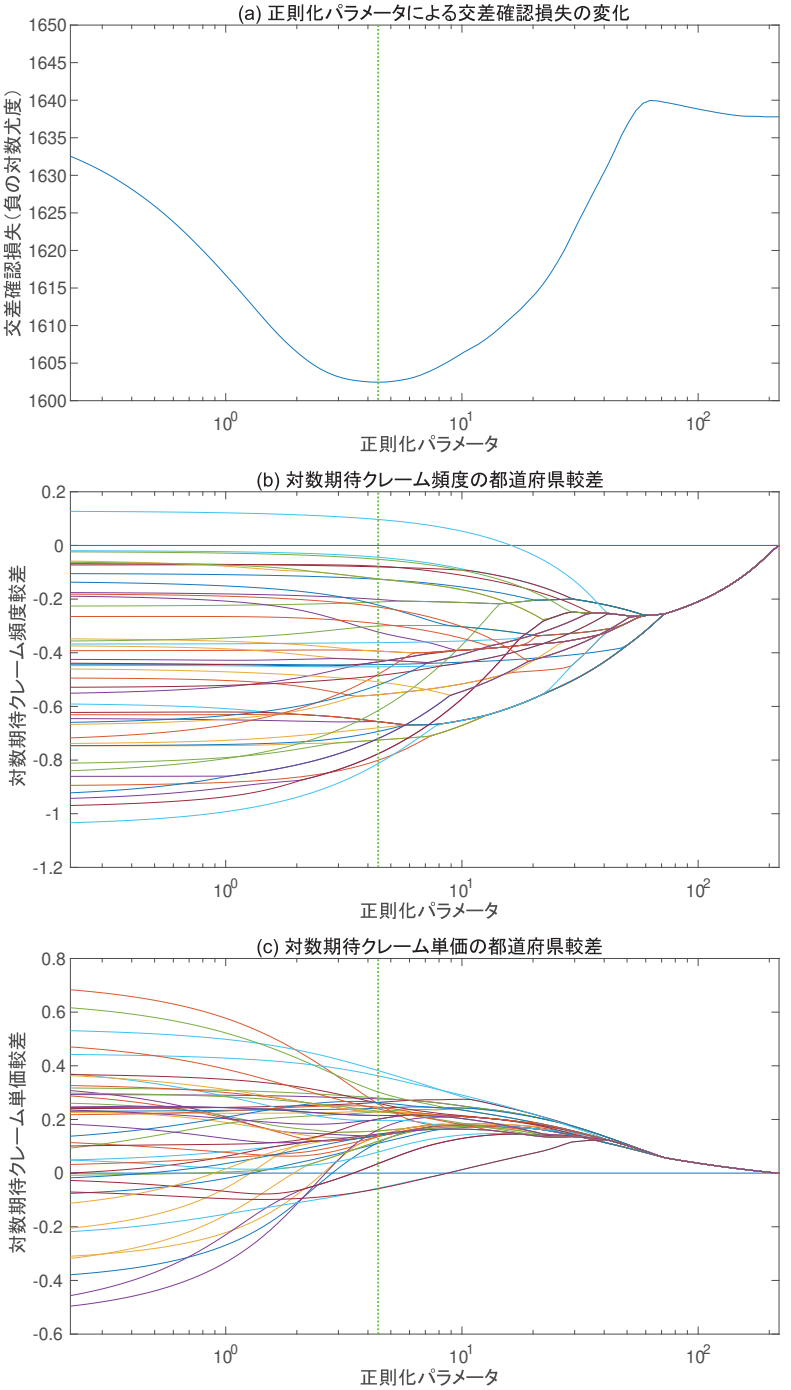

図 4　正則化パラメータによる交差確認損失と期待クレーム頻度・期待クレーム単価較差の変化

正則化項付き対数尤度の最適化手法として交互方向乗数法を用いており，本稿のガンマ分布やポアソン分布のような強凸性を満たす対数尤度に対して線形収束が保証される．

また，解析例では自動車損害賠償責任保険に提案モデルを適用し，隣接する都道府県の較差に対するグループ正則化項を設けて，都道府県の料率適用上のグルーピングを行った．普通貨物自動車の解析では，都道府県区分と用途，重量との交互作用の有無を検討した結果，営業用か自家用かの別によってクレーム単価の地域的傾向が異なることが明らかになった．さらに，自家用乗合自動車について純保険料（期待クレーム頻度×期待クレーム単価）を推定した結果，都道府県間の純保険料の較差が最大で 3 倍超となり，地域による較差が非常

に大きいことが浮き彫りとなった．現行の沖縄・離島などによる料率区分とは異なり，都道府県別料率を適用するには登録地域の実態と異なる申告を防ぐなどの運営面での課題が予想されるものの，都道府県の較差を踏まえた料率形態には検討の価値があるものと考えられる．

本解析例の都道府県以外にも，例えば被保険者年齢や車両の経過年数などのファクターは多数の区分を有しており，本稿のモデルを適用することで適切な年齢条件のセグメンテーションを行うこともできる．複数のファクターについて区分統合を行いたい場合，ファクターごとに適当なグラフ構造を定めて fused lasso を適用すればよく，さらに，それらのグラフの直積グラフ上の fused lasso を与えることで，複数ファクターの交互作用について区分統合することも可能である．また実用に際しては，例えば運転者等級をファクターとする場合，等級の上昇に従い料率が減少するよう単調性が保証されるべきであり，そのような実務的制約を満たすためには正則化項の改変が必要となることもある．本稿で用いた lasso と呼ばれる手法は，一般にはファクターや契約データ量が非常に膨大になるようなビッグデータに対して高速に変数選択しながら推定する手法であり，今後は任意保険のリスク細分型商品に対してその真価を発揮することを期待したい．

謝辞

本稿の執筆にあたり，査読者の先生方より本稿の改訂のための有益なご助言を数多くいただいた．ここに，深く感謝申し上げたい．

参考文献

[1] Bailey, R. A. (1963), "Insurance rates with munimum bias," In *Proceedings of the Casualty Actuarial Society*.

[2] Bleakley, K. and J. P. Vert (2011), "The group fused lasso for multiple change-point detection," Working paper, Available at arXiv: http://arxiv.org/abs/1106.4199v1.

[3] Jung, J. (1968), "On automobile claim data," *ASTIN Bulletin*, **5**, 41–48.

[4] Nelder, J. A. and R. W. N. Wedderburn (1972), "Generalized linear models," *Journal of the Royal Statistical Society: Series A*, **135**, 370–384.

[5] Ohlsson, E. and B. Johansson (2010), *Non-Life Insurance Pricing with Generalized Linear Models*, EAA Series, Springer.(岩沢宏和監訳, 日本アクチュアリー会 ASTIN 関連研究会訳 (2014). 『一般化線形モデルを使用した損害保険料率の算定』, 日本アクチュアリー会.)

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

[7] Tibshirani, R., M. Saunders, S. Rosset, J. Zhu and K. Knight (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B*, **67**, 91–108.

[8] Tweedie, M. (1984), "An index which distinguishes between some important exponential families," In J.Ghosh and J.Roy (eds.), *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, 579–604. Indian Statistical Institute, Calcutta.

[9] Wahlberg, B., S. Boyd, M. Annergren and Y. Wang (2011), "An ADMM algorithm for a class of total variation regularized estimation problems," *IFAC Proceedings Volumes*, **45**, (16), 83–88.

[10] Wytock, M., S. Sra and J. Z. Kolter (2014), "Fast Newton methods for the group fused lasso," In*Uncertainty in Artificial Intelligence*.

表 3　自家用乗合自動車の期待クレーム頻度・期待クレーム単価・期待クレーム総額の推定値

都道府県の区分	期待クレーム頻度（件／年）	期待クレーム単価（千円）	期待クレーム総額（千円／年）
北海道	0.0214	463	9.9
青森	0.0096	599	5.7
岩手，秋田	0.0104	541	5.6
宮城，山形，福島，栃木	0.0111	532	5.9
茨城	0.0137	517	7.1
群馬	0.0108	533	5.8
埼玉	0.0175	610	10.7
千葉	0.0173	612	10.6
東京	0.0205	678	13.9
神奈川	0.0197	602	11.9
新潟	0.0107	518	5.5
富山，石川	0.0123	531	6.5
福井	0.0155	544	8.4
山梨	0.0116	582	6.8
長野	0.0095	563	5.3
岐阜，滋賀	0.0139	566	7.9
静岡	0.0188	589	11.1
愛知，三重	0.0144	581	8.4
京都	0.0158	626	9.9
大阪	0.0236	665	15.7
兵庫	0.0198	600	11.9
奈良	0.0171	602	10.3
和歌山	0.0132	592	7.8
鳥取，岡山	0.0189	521	9.8
島根，山口	0.0098	479	4.7
広島	0.0149	500	7.4
徳島，高知	0.0104	574	6.0
香川	0.0170	530	9.0
愛媛	0.0129	581	7.5
福岡	0.0203	583	11.8
佐賀	0.0136	437	5.9
長崎	0.0132	436	5.8
熊本	0.0127	533	6.8
大分	0.0160	588	9.4
宮崎，鹿児島	0.0140	535	7.5

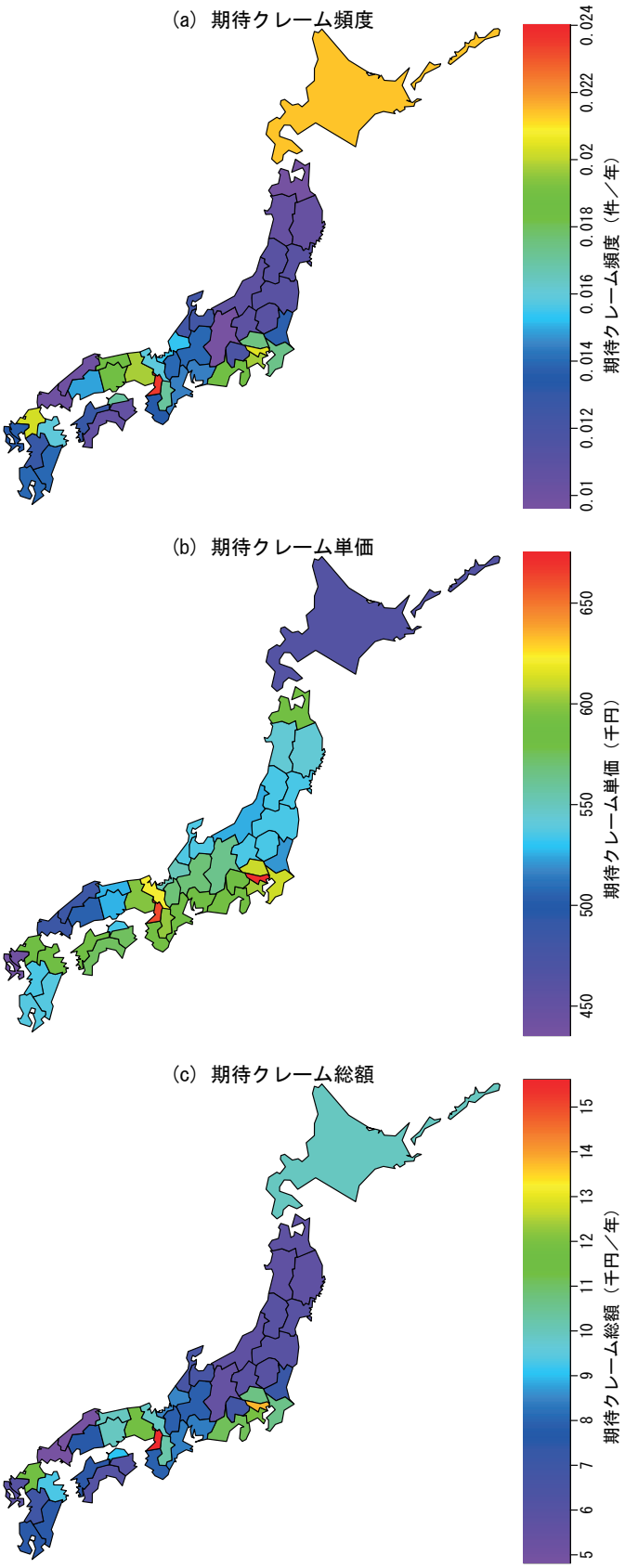

図５　自家用乗合自動車の期待クレーム頻度・期待クレーム単価・期待クレーム総額の統計地図

Automatic segmentation of rating classes via the group fused lasso

Shunichi Nomura*

Received 24 January 2017
Accepted 24 October 2017

Abstract :

In this paper, we propose a new insurance ratemaking method automatically clustering rating classes of risk factors into groups with the same risk levels using group fused lasso, a sparse regression method. In tariff analysis, rating factors with large numbers of classes are often regrouped into smaller number of categories to obtain reliable best estimates. However, the number of combination of the rating-class segmentation is often so huge that it is computationally difficult to try and compare all the possible segmentation. L1 regularization methods, called the fused lasso, are very appropriate in such situations because they integrate neighboring classes with non-significant difference in their risk levels automatically in their estimation processes. Here, we introduce a grouped regularization terms into generalized linear models for ratemaking especially in the cases of considering interaction of some rating factors and modeling claim frequency and severity simultaneously. We use the alternating direction multiplier method to estimate model parameters from log-likelihood with regularization term and select regularization parameter by the cross-validation method. As an illustration, we apply the proposed model to claim datasets of Voluntary Automobile insurance in Japan and group Japanese prefectures into clusters with the same levels of claim frequency or severity.

Keyword : Tariff analysis, Generalized linear model, Group fused lasso, Alternating direction multiplier method, Voluntary Automobile insurance

* The Institute of Statistical Mathematics E-mail: nomura@ism.ac.jp