Vol.2020-MPS-129 No.9
2020/7/27

相関のあるカウントデータの ベイジアンスパース共分散構造分析

一期崎 翔[1,2,a)]　川島 貴大[1]　庄野 逸[1,b)]

概要：本稿において、Bayesian Graphical Lasso を相関のあるカウントデータへ適用したモデルを提案する。提案モデルにおいて、Gaussian Graphical Model に従う潜在変数を仮定する。作成したサンプルデータを用いてシミュレーションを行い、最適なハイパーパラメータを決定する。さらに提案モデルによってカウントデータの潜在変数のスパースな共分散行列の逆行列を推定し、評価する。最後に提案モデルを犯罪データに適用し、犯罪においての推定された潜在変数の共分散行列の逆行列の偏相関係数について分かり易く表現するために可視化する。

キーワード：Bayesian Graphical LASSO，スパース推定，グラフィカルモデル，共分散構造分析, 犯罪データ分析

Bayesian Sparse Covariance Structure Analysis for Correlated Count Data

Sho Ichigozaki[1,2,a)]　Takahiro Kawashima[1]　Hayaru Shouno[1,b)]

Abstract: In this paper, we propose a Bayesian Graphical Lasso for correlated countable data and apply it to spatial crime data. In the proposed model, we assume a Gaussian Graphical Model for the latent variables which dominate the potential risks of crimes. To evaluate the proposed model, we determine optimal hyper-parameters which represent samples better. We apply the proposed model for estimation of the sparse inverse covariance of the latent variable and evaluate the partial correlation coefficients. Finally, we illustrate the results on crime spots data and consider the estimated latent variables and the partial correlation coefficients of the sparse inverse covariance.

Keywords: Bayesian Graphical Lasso, Sparse Estimation, Graphical Model, Covariance Structure Analysis, Crime Data Analysis

1. はじめに

データの特徴を分析する上で、変数間の相関関係を求めることは単純な方法であるがとても有効である。しかし、変数が多い場合は単純に相関を求めても、閾値の設定が難しくなり、重要な関係を見出すことはとても困難である。多変量の共分散構造分析において重要な変数間の関係だけを見出すために LASSO [1] を適用した Graphical Lasso [2] のようなスパースモデリングはとても有効であることが知られている。しかし、Graphical Lasso は Gaussian Graphical Model(GGM) を仮定しているため、カウントデータに適用できない。

本稿において、制限のある Graphical Lasso をスパースな共分散構造を持つポアソン分布型の観測データに適用す

[1] 電気通信大学大学院情報理工学研究科
The University of Electro-Communications, 1-5-1 Chofu-gaoka, Chofu, Tokyo 182-8585 Japan
[2] 警察庁情報通信局
National Police Agency Info-Communications Bureau 2-1-2 Kasumigaseki Chiyoda-ku, Tokyo 100-8974 Japan
[a)] ichigozaki.show@uec.ac.jp
[b)] shouno@uec.ac.jp

る階層ベイズモデルを提案する。提案モデルにおいて、斉次的ポアソン過程に従うイベントの発生に対する潜在的なリスクを、Bayesian Graphical Lasso [3] に従う潜在変数として定義する。最後に提案モデルの実データへの適用例として、犯罪データ分析に用いて数値実験の効果を調査する。

1.1 Graphical Lasso

Graphical Lasso [2] は GGM のスパース共分散構造分析にとても効果的であることが知られている。ここで平均 0、各要素 $\omega_{ij}(i,j=1,\ldots,A)$ で構成される精度行列 $\boldsymbol{\Omega}\in\mathbb{R}^{A\times A}$ で与えられるガウス分布に従うデータ行列を $\boldsymbol{Y}=(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T)\in\mathbb{R}^{T\times A}$ とする。

$$p(\boldsymbol{Y}|\boldsymbol{\Omega})=\prod_{t=1}^{T}\mathcal{N}(\boldsymbol{y}_t|\boldsymbol{0},\boldsymbol{\Omega}^{-1}) \tag{1}$$

Graphical Lasso は L_1 正則化によって $\boldsymbol{\Omega}$ の最尤推定を最適化する。

$$\boldsymbol{\Omega}=\underset{\boldsymbol{\Omega}'\in\mathbf{M}^+}{\arg\max}\log(\det(\boldsymbol{\Omega}'))-\mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{\Omega}')-\lambda\|\boldsymbol{\Omega}'\|_1 \tag{2}$$

ここで $\mathbf{M}^+$ は $A\times A$ の半正定値行列、λ は正則化パラメータをそれぞれ表し、L_1 ノルムは $\|\boldsymbol{\Omega}\|_1=\sum_{ij}|\omega_{ij}|$ で定義される。また、$\boldsymbol{\Omega}$ の最尤推定は、$p(\omega_{ij})=\lambda/2\exp(-\lambda|\omega_{ij}|)$ で表されるラプラス分布 $\mathrm{DE}(\omega_{ij}\mid\lambda)$ に従う非対角要素と $p(\omega_{ii})=\lambda/2\exp(-\lambda\omega_{ii}/2)$ で表される指数分布 $\mathrm{Exp}(\omega_{ii}\mid\lambda/2)$ に従う対角要素で組み合わされたモデルから得られる。

$$\begin{aligned}&p(\boldsymbol{\Omega}\mid\lambda)\\&=C^{-1}\prod_{i<j}\{\mathrm{DE}(\omega_{ij}|\,\lambda)\}\prod_{i=1}^{A}\mathrm{Exp}\left(\omega_{ii}\left|\frac{\lambda}{2}\right.\right)1_{\boldsymbol{\Omega}\in\boldsymbol{M}^+}\end{aligned} \tag{3}$$

ここで C^{-1} は正規化項であり、$1_{\boldsymbol{\Omega}\in\boldsymbol{M}^+}$ は

$$1_{\boldsymbol{\Omega}\in\boldsymbol{M}^+}=\begin{cases}1 & (\boldsymbol{\Omega}\in\boldsymbol{M}^+)\\ 0 & (\text{otherwise})\end{cases} \tag{4}$$

を表す。

1.2 Bayesian Graphical Lasso

Bayesian Graphical Lasso は Graphical Lasso をフルベイズ化したものである。ラプラス分布が混合ガウス分布で表せることと指数分布によって、(3) は

$$\begin{aligned}&p(\boldsymbol{\Omega}|\boldsymbol{\tau},\lambda)\\&=C_{\boldsymbol{\tau}}^{-1}\prod_{i<j}\mathcal{N}(\omega_{ij}|0,\tau_{ij})\prod_{i=1}^{A}\mathrm{Exp}\left(\omega_{ii}\left|\frac{\lambda}{2}\right.\right)1_{\boldsymbol{\Omega}\in\boldsymbol{M}^+}\end{aligned} \tag{5}$$

$$p(\boldsymbol{\tau}|\lambda)\propto C_{\boldsymbol{\tau}}\prod_{i<j}\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_{ij}\right) \tag{6}$$

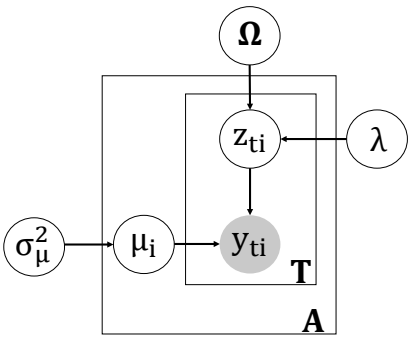

図 1 提案モデルにおけるパラメーターのグラフィカルモデル
Fig. 1 The graphical model of the proposed model

と表すことができる。ここで、$\boldsymbol{\omega}=\{\omega_{ij}\}_{i\le j}$ は $\boldsymbol{\Omega}$ の非対角要素と対角要素の上三角行列を表し、$\boldsymbol{\tau}=\{\tau_{ij}\}_{i<j}$ は補助変数を表す。補助変数 $\boldsymbol{\tau}$ の導入により、BGL モデル (5) の block Gibbs sampling アルゴリズムを構築できる。

2. 提案モデル

2.1 モデル

図 1 は提案モデルの Graphical model を示す。まず、非負の整数を $\mathbb{N}_0=\mathbb{N}\cup\{0\}$ で定義する。そして A 次元のデータが独立なポアソン分布に従って T 時点にわたり得られると仮定する.

$$P(\boldsymbol{Y}|\boldsymbol{\mu},\mathbf{Z})=\prod_{i=1}^{A}\prod_{t=1}^{T}\mathrm{Poisson}(y_{ti}|\exp(\eta(\mu_i,z_{ti}))) \tag{7}$$

$$\eta(\mu_i,z_{ti})=\mu_i+z_{ti} \tag{8}$$

さらに、事前分布は

$$p(\boldsymbol{\mu})=\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{0},\sigma_\mu^2\boldsymbol{I}) \tag{9}$$

$$p(\boldsymbol{z}_t|\boldsymbol{\Omega})=\mathcal{N}(\boldsymbol{z}_t|\boldsymbol{0},\boldsymbol{\Omega}^{-1})\ \text{ for } t=1,\ldots,T \tag{10}$$

$$p(\boldsymbol{\omega}|\lambda)\propto\prod_{i<j}\mathrm{DE}(\omega_{ij}|\lambda)\prod_{i=1}^{A}\mathrm{Exp}\left(\omega_{ii}\left|\frac{\lambda}{2}\right.\right)1_{\boldsymbol{\Omega}\in\boldsymbol{M}^+} \tag{11}$$

$$p(\lambda)=\mathrm{Gamma}(\lambda|a_\lambda,b_\lambda) \tag{12}$$

とする。ここでイベント発生の潜在的なリスクとして線形予測子 $\eta(\mu_i,z_{ti})=\mu_i+z_{ti}$ を仮定する。μ_i は次元 i の潜在リスクの平均を表し、z_{ti} は μ_i からのばらつきを表す。$\boldsymbol{z}_t$ は BGL の事前分布に従うため、スパースかつ重要なカウントデータの共起構造を抽出することができる。よって、事後分布は

$$\begin{aligned}&p(\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Omega},\lambda|\boldsymbol{Y})\\&\propto P(\boldsymbol{Y}|\boldsymbol{\mu},\boldsymbol{Z})p(\boldsymbol{\mu})\left[\prod_{t=1}^{T}p(\boldsymbol{z}_t|\boldsymbol{\Omega})\right]p(\boldsymbol{\omega}|\lambda)p(\lambda)\end{aligned} \tag{13}$$

と表すことができる。

2.2 Sampling

提案モデルのパラメーターの推定は 2 つのサンプリング手法を使い分けている。まず初めに、BGL での $\boldsymbol{\Omega}$ や λ の推定において block Gibbs sampling を行う。次に潜在的なリスクの $\boldsymbol{\mu}$ と $\boldsymbol{Z}$ は Metropolis-Hastings 法 [4] によって推定する。例えば $\boldsymbol{\mu}$ の完全条件付き分布のサンプリングは、$\boldsymbol{\mu}$ から $\boldsymbol{\mu}'$ に遷移するときの採択率 r を

$$r = \min\left\{1, \frac{p(\boldsymbol{\mu}'|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Omega},\lambda)p(\boldsymbol{\mu}|\boldsymbol{\mu}')}{p(\boldsymbol{\mu}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Omega},\lambda)p(\boldsymbol{\mu}'|\boldsymbol{\mu})}\right\} \tag{14}$$

として表すことができる。サンプリング手順は以下の通りである。

Step 1

完全条件付き分布 $\boldsymbol{\mu}$ について Newton-Rhapson 法を使って

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \log p(\boldsymbol{\mu}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Omega},\lambda) \tag{15}$$

で最適化する。

Step 2

多変量 t 分布として定義される提案分布

$$p(\boldsymbol{\mu}'|\hat{\boldsymbol{\mu}}) = \text{Multi-t}(\boldsymbol{\mu}'|\hat{\boldsymbol{\mu}}, \boldsymbol{H}_{\hat{\mu}}^{-1}, \nu) \tag{16}$$

から候補 $\boldsymbol{\mu}'$ をサンプルする。ここで $\boldsymbol{H}_{\mu} \in \boldsymbol{R}^{A\times A}$ は (15) のヘシアン行列、$\nu(>0)$ は自由度を表す任意のハイパーパラメータである。$\boldsymbol{\mu}'$ は確率 r において採択の場合は $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}'$ で更新するか、もしくは棄却される。$p(\boldsymbol{z}_t|\boldsymbol{Y},\boldsymbol{Z}_{\setminus t},\boldsymbol{\mu},\boldsymbol{\Omega},\lambda)$ についても $\boldsymbol{\mu}$ と同様の方法でサンプリングする。

3. シミュレーション

3.1 人工データ

提案モデルの性能を評価するため、サイズがそれぞれ $(A,T) = (10,30),(50,60),(100,60),(200,60)$ の人工データを作成する。また、$\mu_i = 0.2 \quad (i = 1,\ldots,A)$、要素が $\omega_{ii} = C_1, \omega_{i,i-A/2} = \omega_{i-A/2,i} = C_2$ かつその他が 0 である $\boldsymbol{M}^+$ の $\boldsymbol{\Omega}$ を作成する。ここで、C_1 と C_2 は定数を表す。それらを真値として与え、サンプル $\boldsymbol{Z}$ とサンプル観測データ行列 $\boldsymbol{Y}$ を式 (7) と (10) に従ってそれぞれ作成する。

3.2 最適なハイパーパラメータ

各サイズの人工データにおいて、各ハイパーパラメータを変更して実験を行う。推定結果を真値との MAE で評価して、最適なハイパーパラメータを決定する。

Metropolis-Hastings 法でのサンプリングにおいて、t 分布の自由度 ν が重要な影響を与える。t 分布の ν について、$\nu = 1$ の時はコーシー分布になり、$\nu \to \infty$ では正規分布に近づくため、効果的な値は $\nu = 5$ である。次に、正規化項のハイパーパラメータ λ に寄与する a_λ については、次元が $A = 10, 50, 100$ である時は $a_\lambda = A$、比較的に多次元である $A = 200$ の時は $a_\lambda = 0.01$ を適用する。さらに提案モデルにおいて、$\boldsymbol{\mu}$ と $\boldsymbol{Z}$ は指数関数内にあるポアソン分布のパラメーターであるため、μ の絶対値の大きさは小さな値であるべきである。故に事前分布 $p(\boldsymbol{\mu})$ については、$\sigma_\mu^2 = 0.05$ とする。

3.3 結果

人工データ $(A,T) = (50,60)$ において、ハイパーパラメーターを $(a_\lambda, \sigma_\mu^2, \nu) = (A, 0.05, 3)$ として推定した $\boldsymbol{\Omega}$ の結果を示す。偏相関係数行列 $\boldsymbol{P}$ は精度行列 $\boldsymbol{\Omega}$ の推定結果から

$$p_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}}\sqrt{\omega_{jj}}} \tag{17}$$

で得られる。図 2 は推定結果の $\boldsymbol{\Omega}$ と真値の $\boldsymbol{\Omega}$ から得られた偏相関係数行列 $\boldsymbol{P}$ をそれぞれ示す。真値が 0 の要素であるのに対して、推定結果の対応する各要素はほとんどが 0 に近いとても小さい値である。また真値が 0 以外の値をもつ要素では、推定結果の対応する各要素はほとんどが 0 要素よりも大きな値である。

4. 犯罪データ分析

4.1 犯罪データ

提案モデルのカウントデータへの実用例として、National Institute of Justice（NIJ）[5] が公開しているアメリカオレゴン州のポートランドでの犯罪発生データに適用する。犯罪データにおいて、潜在変数 μ_i は i 番目のエリアの犯罪発生の平均の潜在的なリスクを表す。また潜在変数 $\boldsymbol{Z}$ については、単位時間ごとの犯罪発生の潜在的なリスクのばらつきを表す。故に、$\boldsymbol{Z}$ は相互作用の構造を表す。

使用するデータは 2016 年に発生した犯罪データであり、犯罪発生スポットを 60 エリアに分けてまとめる。さらに、犯罪発生件数を 1 週間の時間単位で分割したのもので、データ行列 $\boldsymbol{Y}$ は $(A,T) = (60,52)$ である。

4.2 犯罪における $\boldsymbol{\Omega}$ の偏相関係数

推定された $\boldsymbol{\Omega}$ は時間的な変動による効果 $\boldsymbol{Z}$ から構築されたスパースな共分散行列の逆行列である。そのため、$\boldsymbol{\Omega}$ から偏相関係数を導き出すことは、エリア間の犯罪発生リスクのスパースな相関関係を表すことになる。

推定された $\boldsymbol{\Omega}$ から偏相関係数を (17) によって計算する。エリア間の犯罪発生リスクの相関関係を視覚的に見やすくするために、ポートランドのマップ上に表したものを図 3 に示す。表した係数は絶対値で全体の上位約 2% である。

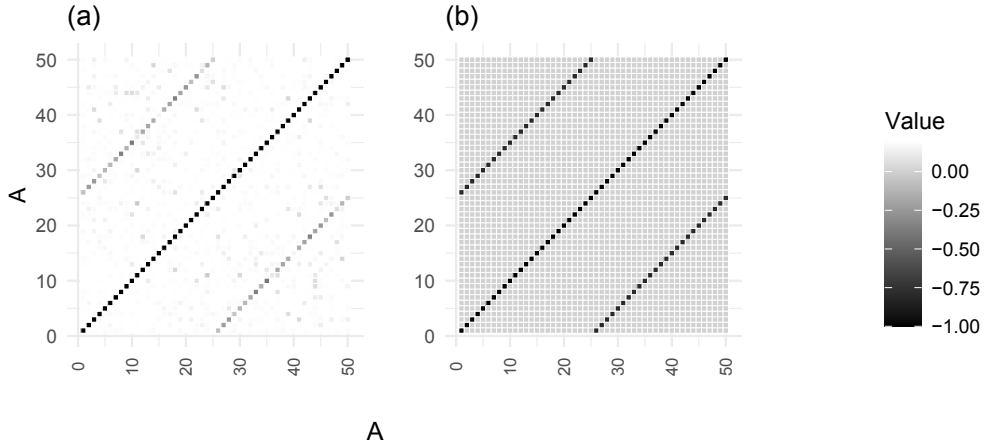

図 2 (a) は推定結果から得られた $\boldsymbol{\Omega}$ の偏相関係数、(b) は作成した真値である $\boldsymbol{\Omega}$ の偏相関係数を表す。

Fig. 2 (a) shows the estimated partial correlation coefficients of $\boldsymbol{\Omega}$ and (b) shows the true partial correlation coefficients of $\boldsymbol{\Omega}$.

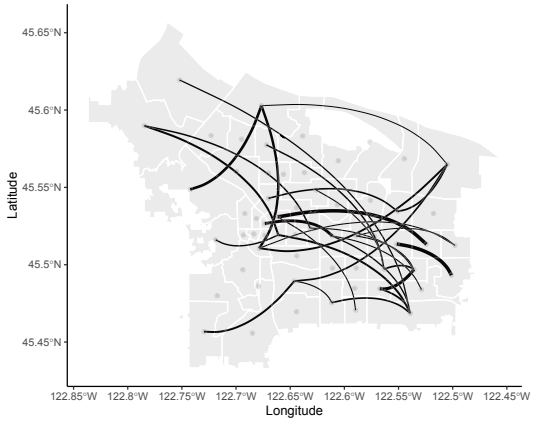

図 3 エリア間のスパースな偏相関係数をポートランドのマップ上に可視化
エリア間を結ぶ線の太さは偏相関係数の大きさを表す。

Fig. 3 The visualization of sparse partial correlation coefficients between areas on Portland city's map. The thickness of the curved black lines shows the magnitudes of the coefficients.

5. 結論

提案モデルは有効な $\boldsymbol{\Omega}$ と潜在変数 $\boldsymbol{\mu}$ と $\boldsymbol{Z}$ を推定することができる。一方で提案モデルでは、ポアソン過程においてパラメーターが 1 以下の場合、イベントの発生数は 0 になりやすいため、潜在変数の負の真値を推定することが困難である。この問題を避けるためには、潜在変数の範囲 $\mathbb{R}$ を $[0, \infty)$ と制限することなどが考えられる。

また、提案モデルによって推定された潜在変数の結果を分析することで、犯罪データなどのカウントデータにおいてスパースな相関関係な有意な特徴を見出すことができる。今後の展望としては、一般的にポアソン過程は時間に関係した例が多いため、時系列を考慮したモデルを構築したい。

参考文献

[1] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288 (online), available from ⟨http://www.jstor.org/stable/2346178⟩ (1996).

[2] Friedman, J., Hastie, T. and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, Vol. 9, pp. 432–441 (2008).

[3] Wang, H. et al.: Bayesian graphical lasso models and efficient posterior computation, *Bayesian Analysis*, Vol. 7, No. 4, pp. 867–886 (2012).

[4] Chib, S. and Jeliazkov, I.: Marginal likelihood from the Metropolis–Hastings output, *Journal of the American Statistical Association*, Vol. 96, No. 453, pp. 270–281 (2001).

[5] Homepage, N.: Real-Time Crime Forecasting Challenge, `https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge` Last accessed 18 May 2020.