Minimax Predictive Densities for Sparse Statistical Models with Sample Size Heterogeneity

（不均一な観測数を持つスパース統計モデルにおけるミニマックス予測分布）

数理情報学専攻 48176204　金子 亮也
指導教員　駒木 文保 教授

1　はじめに

予測分布とは，既に得られた観測値に基づいて構成される将来の観測値の分布のことを指す．予測分布の構成は統計学における重要な問題の一つであり，有限次元パラメータを持つ代表的な統計モデルに対して性質の良い予測分布が提案されている．一方，近年の観測データ次元の増加に伴って，高次元統計モデルに対する研究が急速に発展している．高次元データの多くは何らかの変換を通じて，意味のある情報が一部の次元にしか存在しないデータ (スパースデータ) に帰着するため，高次元スパースデータに対する統計的推測は特に重要である．しかし，高次元スパースデータに対する予測分布の構築に関する研究は未だ数少ない ([1], [2])．

本研究では，連続・離散各々の代表的な統計モデルである Gauss モデルおよび Poisson モデルに対し，パラメータ空間にスパース制約を課した下で適切な予測分布の構成を考える．特に本研究では観測数の不均一性に着目する．観測数の不均一性とは，データの各次元ごとに観測数が異なる状態を意味する．具体的には観測地点ごとに異なった数の観測データが得られる等の状況に対応しており，観測数の均一性 (データの各次元ごとに観測数が等しい状態) を包含するより一般の設定となっている．[1] および [2] が観測数が均一なモデルに対するスパース予測の結果を与えたのに対して，本研究では観測数の不均一性を表現する Missing-Completely-At-Random という構造を仮定し，上記 2 つのモデルに対してそれぞれ漸近的にミニマックス最適な予測分布を構成する．また実データに対する応用を通して，提案手法および既存手法の性能を比較する．

2　問題設定

既に得られた観測値を表す変数を X, 将来の観測値を表す変数を Y とし，各々は独立に統計モデル $\{p(x|\theta) : \theta \in \Theta\}$，$\{q(y|\theta) : \theta \in \Theta\}$ に従うとする．$\theta = (\theta_i)_{i=1}^n$ は共通の未知パラメータである．観測変数 X に基づいて Y の予測分布 $\widehat{q}(y;x)$ を構成することを考える．このとき，予測分布 $\widehat{q}$ の真の分布 $q(\cdot|\theta)$ からの乖離度を以下の Kullback–Leibler リスク $R(\theta, \widehat{q})$ で測る：

$$R(\theta, \widehat{q}) \coloneqq \int\int p(x|\theta)q(y|\theta)\log\frac{q(y|\theta)}{\widehat{q}(y;x)}\mathrm{d}x\mathrm{d}y.$$

この時ミニマックス予測分布 $q^*(y;x)$ は

$$\inf_{\widehat{q}}\sup_{\theta\in\Theta} R(\theta, \widehat{q}) = \sup_{\theta\in\Theta}(\theta, q^*)$$

を満たす予測分布として定義される．ただし inf は全ての予測分布の中で取るものとする．

上記の統計モデルの例として，本研究で扱う 2 つのスパース統計モデルを導入する．スパース Gauss モデルとは，$D \coloneqq \mathrm{diag}\{r_1, \ldots, r_n\}$ $(r_i \ldots, r_n > 0)$ として

$$\begin{aligned} &p(x \mid \theta) = \mathrm{N}_n(\theta, I_n),\ q(y \mid \theta) = \mathrm{N}_n(\theta, D), \\ &\theta \in \Theta[s] \coloneqq \{\theta \in \mathbb{R}^n : ||\theta||_0 \le s\} \end{aligned} \tag{1}$$

で表されるモデルを指す．ただし $\mathrm{N}_n(\cdot,\cdot)$ は n 次元 Gauss 分布を表し，$||\cdot||_0$ は非ゼロ要素の個数とする．

次に，スパース Poisson モデルとは

$$\begin{aligned} &p(x \mid \theta) = \otimes_{i=1}^n \mathrm{Po}(r_i\theta_i),\ q(y \mid \theta) = \otimes_{i=1}^n \mathrm{Po}(\theta_i), \\ &\theta \in \Theta[s] \coloneqq \{\theta \in \mathbb{R}_+^n : ||\theta||_0 \le s\} \end{aligned} \tag{2}$$

で表されるモデルを指す．ただし $\mathrm{Po}(\cdot)$ は Poisson 分布を表す．

Gauss，Poisson 両モデルの十分性により，本研究における主眼である観測数の不均一性は，$r_1, \ldots, r_n$ の値が互いに等しいとは限らないという仮定と対応付けられることに注意する．さらに，$r_1, \ldots, r_n$ に対して観測数の不均一性の表現の 1 つである Missing-Completely-At-Random という構造を仮定する：$r_1, \ldots, r_n$ は独立同一に確率分布 G に従うとする．

3　本研究の成果

本研究における成果のうち，観測数の不均一性を伴うスパース Gauss, Poisson 両モデルに対する予測分布の漸近的ミニマックス最適性に関する結果を述べる．

定理 1. スパース Gauss モデル（1）を考える．G に関して $\mathbb{E}_G[r_1] < \infty$，$\mathbb{E}_G[1/r_1] < \infty$ を仮定する．このとき，$\lim_{n\to\infty} s_n/n = 0$ ならば以下が成り立つ：

$$\operatorname*{plim}_{n\to\infty} \frac{\inf_{\widehat{q}} \sup_{\Theta[s_n]} R(\theta, \widehat{q})}{\mathbb{E}_G[\mathcal{C}_1(r_1)] s_n \log(n/s_n)} = 1.$$

但し $\mathcal{C}_1(r) \coloneqq 1/(r+1)$ である．また，以下を満たす予測分布 q^* を具体的に構成できる：

$$\operatorname*{plim}_{n\to\infty} \frac{\sup_{\Theta[s_n]} R(\theta, q^*)}{\inf_{\widehat{q}} \sup_{\Theta[s_n]} R(\theta, \widehat{q})} = 1.$$

定理 2. スパース Poisson モデル（2）を考える．G に関して $\mathbb{E}_G[r_1] < \infty$，$\mathbb{E}_G[1/r_1^2] < \infty$ を仮定する．このとき，$\lim_{n\to\infty} s_n/n = 0$ ならば以下が成り立つ：

$$\operatorname*{plim}_{n\to\infty} \frac{\inf_{\widehat{q}} \sup_{\Theta[s_n]} R(\theta, \widehat{q})}{\mathbb{E}_G[\mathcal{C}_2(r_1)] s_n \log(n/s_n)} = 1.$$

但し $\mathcal{C}_2(r) \coloneqq \{r/(r+1)\}^r \{1/(r+1)\}$ である．また，以下を満たす予測分布 q^* を具体的に構成できる：

$$\operatorname*{plim}_{n\to\infty} \frac{\sup_{\Theta[s_n]} R(\theta, q^*)}{\inf_{\widehat{q}} \sup_{\Theta[s_n]} R(\theta, \widehat{q})} = 1.$$

4 数値実験

都内の犯罪データに対する応用例を示す．警視庁による都内のスリ件数データを用いる．図 1 は都内 8 区における 2012 年〜2017 年の町丁別スリ件数を表している．色の濃淡は件数の多寡を表し，白色の地域は当該期間にスリが起きなかった町丁を示す．多くの地域ではスリ件数が 0 もしくは 0 に近い一方，いくつかの地域では多くのスリが起きており，本データはスパースを有していると言える．更に，本データには無報告を原因とする観測数の不均一性が見られることにも注意する．

本実験では，本データに対して観測数の不均一性を伴うスパース Poisson モデルを仮定し，予測分布を用いて 2018 年前半のスリ件数の予測を行う．評価指標には予測分布からのサンプル平均と 2018 年前半の実測値との重み付き ℓ_1 距離 (W-ℓ_1 distance)，予測対数尤度 (PLL)，そして予測分布からのサンプルによる次元ごとの 90% 被覆区間が実測値を被覆する割合 (90%CPP) を用いる．

比較結果を表 1 に示す．表 1 第 2 段より，提案予測分布を用いると重み付き ℓ_1 距離の意味で実測値に近いサンプルを平均的に得られることが分かる．すなわち，サンプリングによる点予測において提案手法が有用であることが確認できる．また第 3 段によれば，予測対数尤度の意味で提案手法は既存手法 (GH，K04，ℓ_1-pen)

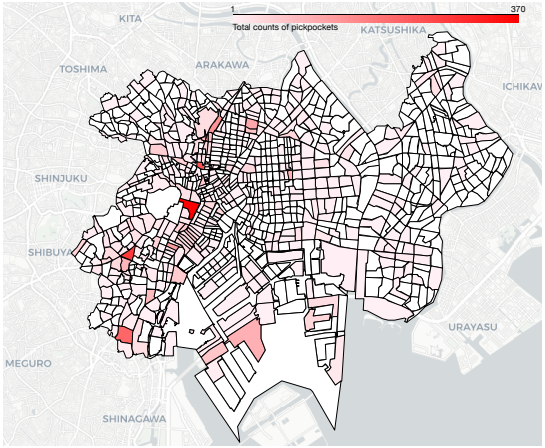

図 1. 都内 8 区における 2012 年〜2017 年の町丁別スリ件数．白色は件数 0 の地域を表す．

表 1. 各評価指標に基づく各予測分布手法の比較．下線は全手法の中で最良の評価値を示す．

	提案手法	GH	K04	ℓ_1-pen
W-ℓ_1 distance	<u>273</u>	293	<u>273</u>	297
PLL	<u>-394</u>	-399	-429	-Inf
90%CCP (%)	<u>93.0</u>	27	84.2	<u>93.0</u>

を改善していることが見て取れる．さらに第 4, 5 段によると，提案手法の 90% 被覆区間が水準に近い被覆割合を示している．この結果から，提案手法が将来データの不確実性の定量化にも有効であることが示唆される．

5 結論

本研究により，Missing-Completely-At-Random を伴うスパース Gauss モデル，スパース Poisson モデルにおいて漸近的にミニマックス最適な予測分布を構成できることが明らかになった．さらに，実データへの適用により提案予測分布の応用上の有効性を確認した．本稿で報告した内容に加え，本研究では θ の推定量に基づく plug-in 予測分布のクラスに限定したミニマックスリスクを導出し，ミニマックス最適性の意味で予測と推定を比較した．また，スパース Poisson モデルに関しては，スパース性の規模を表す s_n に対して適応的な漸近的ミニマックス予測分布の構成についても議論した．

参考文献

[1] Gourab Mukherjee and Iain Johnstone. Exact minimax estimation of the predictive density in sparse gaussian models. *Ann. Statist.*, 43:937–961, 2015.

[2] Keisuke Yano, Ryoya Kaneko, and Fumiyasu Komaki. Asymptotically minimax predictive density for sparse count data. arXiv:1812.06037, 2018.