2022 年 6 月 17 日 統計数理研究所 オープンハウス

# 事前分布強調型の情報量規準 Prior Intensified Information Criterion

二宮 嘉行　数理・推論研究系 教授

## 目的

- ベイズ予測におけるモデル選択の標準ツールとなりつつある WAIC (Watanabe '10) が，相当な大標本設定に依拠していることと事前分布の罰則を考慮していないことを鑑みつつ，Kullback-Leibler ダイバージェンスに基づく古典的なアプローチで PIIC を導く (Ninomiya '21)
- 基本的な回帰問題における数値実験で，特にパラメータ数の割にデータサイズが小さいとき，PIIC は WAIC を優越することを確認する
- 本過程の中で，ベイズ予測版 GIC (小西 '00) の妥当性をある程度示す

## 設定および記法

$$\tilde{z}_1, \ldots, \tilde{z}_n, z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} f(\cdot \mid \boldsymbol{\theta}); \qquad \boldsymbol{\theta} \sim \pi(\cdot \mid \boldsymbol{\xi})$$

- $\tilde{z}_1, \ldots, \tilde{z}_n$ は将来の値として用意，$\boldsymbol{\xi}$ は選択したいハイパーパラメータ
- WAIC は特異モデルを扱えることも強みだが，ここでは正則モデルのみ扱う（そのときの WAIC の良い性能については Watanabe '15 参照）
- $\boldsymbol{z} = (z_1, \ldots, z_n)$ として $\boldsymbol{\theta}$ の事後分布による期待値を $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\xi}}(\cdot)$ で表し，特にベイズ予測分布 $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\xi}}\{f(z \mid \boldsymbol{\theta})\}$ を $f(z \mid \boldsymbol{z}, \boldsymbol{\xi})$ と記すことにする

## WAIC の定義とそれに関する考察

$$-\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\xi}) + \sum_{i=1}^{n} (\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\xi}}[\{\log f(z_i \mid \boldsymbol{\theta})\}^2] - \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\xi}}\{\log f(z_i \mid \boldsymbol{\theta})\}^2)$$

- (ABIC のような) 事前分布に対する罰則項がなく，(ある意味で) 同じレベルの事前分布が複数存在したら，その中で最もデータにフィットしたものが選ばれるので，過適合になるはず
- たとえ同じレベルの事前分布がなくても，WAIC は高次の項をみて選択する話であり，つまり最尤推定に近い状況のみを扱っており，例えば LASSO のチューニングパラメータの選択にはうまくいかないはず

## 情報量規準の開発 (1)

- 罰則項（つまり事前分布）のオーダーを $\mathrm{O}(n)$ とする
- 具体的には，不格好でも $\pi(\boldsymbol{\theta} \mid \boldsymbol{\xi})^{n/n_0}$ という形の事前分布を考えるが ($n_0$：実際に現在得られているデータのサイズ)，データが増えたら事前分布が変わる，と実際に思っているわけではもちろんない
- 第一次近似との差

$$\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\xi}) - \sum_{i=1}^{n} \mathrm{E}_{\tilde{z}_i}\{\log f(\tilde{z}_i \mid \boldsymbol{z}, \boldsymbol{\xi})\}$$

に対し，AIC と同様，$\boldsymbol{z}$ についても期待値をとったものを漸近評価する

- 結果，以下の GIC (小西 '00) が得られ，そこでは与えられていなかった妥当性が示されたといえる（これを PIIC1($\boldsymbol{\xi}$) と記すことにする）

$$-\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\xi}) + \mathrm{tr}\left[ J_{\boldsymbol{z},\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\xi}})^{-1} \hat{\mathrm{E}} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} q_{z,\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\xi}}) \frac{\partial}{\partial \boldsymbol{\theta}'} q_{z,\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\xi}}) \right\} \right]$$

ここで $q_{z,\boldsymbol{\xi}}(\boldsymbol{\theta}) = \{\log f(z \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta} \mid \boldsymbol{\xi})^{n_0^{-1}}\}/n$，$q_{\boldsymbol{z},\boldsymbol{\xi}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} q_{z_i,\boldsymbol{\xi}}(\boldsymbol{\theta})$，$\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\xi}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, q_{\boldsymbol{z},\boldsymbol{\xi}}(\boldsymbol{\theta})$，$J_{\boldsymbol{z},\boldsymbol{\xi}}(\boldsymbol{\theta}) = -\partial^2 q_{\boldsymbol{z},\boldsymbol{\xi}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$

## 情報量規準の開発 (2)

- ABIC と同様の考えで罰則項を求める
- 具体的には，PIIC1($\boldsymbol{\xi}$) を最小にする $\boldsymbol{\xi}$ である $\hat{\boldsymbol{\xi}}_{\boldsymbol{z}}$ のリスクも考え，つまり $-\sum_{i=1}^{n} \mathrm{E}_{\tilde{z}_i}\{\log f(\tilde{z}_i \mid \boldsymbol{z}, \hat{\boldsymbol{\xi}}_{\boldsymbol{z}})\}$ をターゲットとして次を評価する

$$\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \hat{\boldsymbol{\xi}}_{\boldsymbol{z}}) - \sum_{i=1}^{n} \mathrm{E}_{\tilde{z}_i}\{\log f(\tilde{z}_i \mid \boldsymbol{z}, \hat{\boldsymbol{\xi}}_{\boldsymbol{z}})\}$$

- 結果，以下の基準が得られ，これを PIIC2 と記すことにする

$$\mathrm{PIIC1}(\hat{\boldsymbol{\xi}}_{\boldsymbol{z}}) + \mathrm{tr}\left\{ \left( -\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} r_{z_i,\boldsymbol{z},\hat{\boldsymbol{\xi}}_z} \right)^{-1} \left( \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\xi}} r_{z_i,\boldsymbol{z},\hat{\boldsymbol{\xi}}_z} \frac{\partial}{\partial \boldsymbol{\xi}'} r_{z_i,\boldsymbol{z},\hat{\boldsymbol{\xi}}_z} \right) \right\}$$

ここで $r_{z_i,\boldsymbol{z},\boldsymbol{\xi}^*} \equiv \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\xi}^*)$

## 数値実験

$$y_i = \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i = \boldsymbol{\theta}' \boldsymbol{x}_i + \varepsilon_i \qquad (1 \le i \le n)$$

- WAIC1 と PIIC1：$\mathrm{N}(\mathbf{0}_p, \mathrm{diag}(\xi \mathbf{1}_p))$ を $\boldsymbol{\theta}$ の事前分布として用いたもの ⇒ 事前分布強調の効果だけをみる
- WAIC2 と PIIC2：$\mathrm{N}(\mathbf{0}_p, \mathrm{diag}(\xi_1 \mathbf{1}_{p/3}, \xi_2 \mathbf{1}_{p/3}, \xi_3 \mathbf{1}_{p/3}))$ を $\boldsymbol{\theta}$ の事前分布として用いたもの ⇒ 事前分布の複雑度考慮の効果もみる
- どのモデルでも（真でも）$\boldsymbol{x}_i$ は $\mathrm{N}(\mathbf{1}_p, \boldsymbol{I}_p)$ に，$\varepsilon_i$ は $\mathrm{N}(0,1)$ に独立にしたがっているとし，$\boldsymbol{\theta}$ の真値は $\boldsymbol{\theta}^* = (\theta_1^* \mathbf{1}_{p/3}, \theta_2^* \mathbf{1}_{p/3}, \theta_3^* \mathbf{1}_{p/3})$ とする

IC1 や IC2：真の分布と推定分布の Kullback-Leibler ダイバージェンス
Rate1 や Rate2：「WAIC による K-L $<, =, >$ PIIC による K-L」の割合

| $n$ | $p$ | $\boldsymbol{\theta}^*$ | WAIC1 | PIIC1 | Rate1 | WAIC2 | PIIC2 | Rate2 |
|---|---|---|---|---|---|---|---|---|
| 12 | 6 | (2,2,2) | 0.456 | 0.456 | (48,4,48) | 0.484 | 0.460 | (31,1,68) |
| 12 | 6 | (3,2,1) | 0.459 | 0.457 | (41,14,45) | 0.477 | 0.463 | (43,1,56) |
| 12 | 6 | (3,1,-1) | 0.462 | 0.460 | (41,13,46) | 0.475 | 0.464 | (43,1,56) |
| 12 | 9 | (2,2,2) | 0.515 | 0.513 | (45,9,46) | 0.622 | 0.525 | (25,2,73) |
| 18 | 9 | (2,2,2) | 0.518 | 0.524 | (59,0,41) | 0.546 | 0.524 | (37,0,63) |
| 18 | 12 | (2,2,2) | 0.495 | 0.495 | (50,4,46) | 0.535 | 0.497 | (33,0,67) |
| 18 | 12 | (3,2,1) | 0.509 | 0.505 | (48,7,45) | 0.522 | 0.507 | (41,0,59) |
| 18 | 12 | (3,1,-1) | 0.509 | 0.500 | (24,41,35) | 0.524 | 0.501 | (36,4,60) |
| 18 | 15 | (2,2,2) | 0.447 | 0.433 | (43,9,48) | 0.526 | 0.441 | (20,1,79) |
| 24 | 15 | (2,2,2) | 0.454 | 0.464 | (53,1,46) | 0.478 | 0.464 | (39,1,60) |
| 24 | 18 | (2,2,2) | 0.519 | 0.531 | (56,2,42) | 0.572 | 0.532 | (36,0,64) |
| 24 | 18 | (3,2,1) | 0.543 | 0.528 | (40,14,46) | 0.563 | 0.534 | (35,0,65) |
| 24 | 18 | (3,1,-1) | 0.559 | 0.543 | (18,41,41) | 0.585 | 0.543 | (34,6,60) |

## 実データ解析

Diabetes データに対する線形回帰分析において Laplace 分布を事前分布にしたときの回帰係数の推定量の比較（13 分割したうちの最初の 7 組）

| | age | sex | bmi | map | tc | ldl | hdl | tch | ltg | glu |
|---|---|---|---|---|---|---|---|---|---|---|
| WAIC | 0.00 | 0.00 | 0.00 | 0.43 | -44.17 | 17.66 | 20.85 | 33.89 | 47.88 | -7.65 |
| PIIC | 0.00 | 0.00 | 8.44 | 0.00 | 0.00 | 0.00 | -3.18 | 0.00 | 24.31 | 0.00 |
| WAIC | 6.06 | -9.91 | 13.85 | 6.87 | 0.00 | -1.11 | 2.11 | 0.00 | 19.48 | -4.58 |
| PIIC | 0.00 | 0.00 | 7.00 | 3.16 | 0.00 | 0.00 | 0.00 | 0.00 | 13.14 | 0.00 |
| WAIC | 0.00 | -16.30 | 5.35 | 0.00 | 73.40 | -48.12 | -59.28 | -19.91 | -1.43 | 5.93 |
| PIIC | 0.00 | -8.92 | 6.53 | 0.00 | 0.00 | -2.29 | -15.41 | 0.00 | 11.62 | 0.56 |
| WAIC | -7.33 | -5.04 | 2.67 | 9.15 | 0.00 | -1.54 | -5.81 | -7.05 | 23.78 | 10.63 |
| PIIC | -8.90 | -9.70 | 8.38 | 31.12 | 0.00 | 0.00 | 0.00 | 0.00 | 1.54 | 0.00 |
| WAIC | 7.84 | -6.67 | 30.60 | 8.76 | 0.00 | 0.00 | 0.00 | 9.96 | 5.10 | 1.07 |
| PIIC | 1.75 | 0.00 | 25.07 | 4.57 | 0.00 | 0.00 | 0.00 | 6.22 | 9.20 | 4.40 |
| WAIC | 3.64 | -0.59 | 16.17 | 0.91 | -0.40 | -5.99 | -2.32 | 0.00 | 10.86 | 5.51 |
| PIIC | 0.00 | 0.00 | 14.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 9.50 | 4.82 |
| WAIC | 0.00 | 1.53 | 3.34 | 0.00 | 0.00 | 0.33 | -8.10 | 1.39 | 15.93 | 9.43 |
| PIIC | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | -7.69 | 3.39 | 18.31 | 13.16 |

## 引用文献


- 小西貞則. (2000). 統計的モデリングと情報量規準構成の理論 – 汎関数に基づくアプローチ. 数学, 52, 128–141.
- Ninomiya, Y. (2021). Prior intensified information criterion. arXiv preprint 2110.12145.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11 (12).
- Watanabe, S. (2015). Bayesian cross validation and WAIC for predictive prior design in regular asymptotic theory. arXiv preprint 1503.07970.


大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics