論文解説：WAIC による過剰パラメータモデルの汎化誤差推定

奥野彰文 [*1,2], 矢野恵佑 [3]

[1] 統計数理研究所 統計思考院, [2] 理化学研究所 AIP センター, [3] 統計数理研究所 数理・推論研究系

要旨

本稿は Journal of Computational and Graphical Statistics 誌に採択された，過剰パラメータモデルの汎化誤差推定に関する我々の原著論文: Okuno and Yano (2023)[a]の解説です．解説の平易さを優先するため，厳密さに欠ける表現も一部含まれます．より厳密な記述については原著論文をご参照ください．

キーワード: WAIC，ニューラルネット，過剰パラメータモデル，特異モデル，自由度，汎化．

1 研究背景

1.1 回帰と汎化誤差

共変量 $X \in \mathbb{R}^d$ から目的変数 $Y \in \mathbb{R}$ を予測する問題を考えます．簡単のため，本稿では $Y \mid X = x$ が期待値 $f_*(x)$ と分散 1 の正規分布に従うよう生成した訓練データ $\mathcal{D}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ を用いて，

$$\text{訓練誤差} \quad \ell_n(\theta) := \frac{1}{2}\sum_{i=1}^{n}\{y_i - f_\theta(x_i)\}^2 \qquad (1)$$

の最小化によりパラメトリックモデル f_θ の最適パラメータ $\hat{\theta}_n$ を決定します．このような手続きを特に回帰と呼び，いくつかの仮定の下で漸近的には一致性 ($f_{\hat{\theta}_n} \to^p f_*, n \to \infty$) が成り立ちます．一方で，現実の n は有限であり，$f_{\hat{\theta}_n}$ は f_* と完全には一致しません．特に訓練誤差 (1) は特定の訓練データに対する当てはまりの良さを評価するので，訓練誤差が小さくても新規に取得しなおしたデータへの当てはまりは悪い場合があり，これを過学習などと呼びます．このような問題を回避するには，特定の訓練データに依存しないよう目的変数に関する期待値を取った

$$\text{汎化誤差} \quad \mathbb{E}_{Y_1, Y_2, \ldots, Y_n}\left(\ell_n(\theta)\right) \qquad (2)$$

が小さくできるとよいでしょう．

1.2 情報量規準 AIC

汎化誤差 (2) は Y に関する期待値で定義されますが，目的変数 Y の従う分布は推定の対象つまり未知なので，現実のデータでは期待値を取ることができません．代わりに訓練データ $\mathcal{D}_n$ から汎化誤差 (2) を推定する方法として，例えば Cross-Validation や情報量規準が知られています．統計学分野で特によく用いられる赤池情報量規準 (Akaike Information Criterion; AIC) は，パラメータ θ が $\Theta \subset \mathbb{R}^p$ を動くとき

$$\text{AIC} := \ell_n(\theta) + p \qquad (3)$$

で定義される量であり，ざっくり 2 つの仮定：

(i) 集合 $\{f_\theta : \theta \in \Theta\}$ が真のモデル f_* を含む
(ii) 対応する確率モデルが特異でない

を満たすとき，AIC は汎化誤差 (2) の漸近不偏推定量となることが知られています．したがって，実際には計算できない汎化誤差の代わりとして AIC を最小化することで，より予測の良い統計モデルを選択できると期待できます．一方で AIC は上記の仮定 (i), (ii) に依拠しているので，これらの仮定が満たされなくてもなお使える，適用範囲の広い情報量規準の研究が進められてきました．

1.3 特異モデルの情報量規準 WAIC

AIC が提案されて間もなく (i) を満たさない場合に使える情報量規準 Takeuchi Information Criterion (TIC) が提案されましたが，(ii) を満たさない，つまり特異モデルにも使える情報量規準の導出は長らく未解決の問題でした．特異なモデルの例として，単純な縮小ランク回帰モデル $f_\theta(x) = \langle a, Bx\rangle, \theta = (a, B) \in \mathbb{R}^k \times \mathbb{R}^{k\times d}$ ($k < d$) にさえ AIC や TIC を計算する理論的な妥当性は示されませんでしたが，2000 年代に入り (ii) を緩和する試みがいくつか報告されています．その中で

* 責任著者，okuno@ism.ac.jp

[a] https://doi.org/10.1080/10618600.2023.2197488

も有名なものに Widely-Applicable Information Criterion (WAIC; Watanabe, 2010) があります．汎関数分散 (Functional Variance, FV):

$$\mathrm{FV} := \sum_{i=1}^{n} \mathbb{V}_{\theta\sim\mathrm{Posterior}}\left(\{y_i - f_\theta(x_i)\}^2\right)$$

を用いて，WAIC は

$$\mathrm{WAIC} := \ell_n(\theta) + \mathrm{FV} \tag{4}$$

として定義されます．特異点解消定理を利用することで，WAIC は (ii) が成り立たない場合においても汎化誤差 (2) の事後期待値の漸近不遍推定量となる，つまり AIC の (ある種の) 拡張となることが示されています．つまり汎関数分散 (FV) は AIC でいうところのパラメータ数 p に対応しており，特異な確率モデルを扱う場合には，汎関数分散がモデルの実質的な自由度を表すことがわかります．

以上をまとめますと，特異なモデルに対して汎関数分散を計算することで，汎化およびそのモデルの自由度 (複雑さ) が評価できます[b]．

2　本研究の問いと貢献

WAIC はニューラルネットにも使えるか？

旧来の AIC は特にモデルのパラメータ数 p が小さく固定されている状況を主眼に置いていましたが，昨今の機械学習分野で用いられるニューラルネットなどは内包するパラメータ数がとても大きく，我々はそのモデルの汎化の良さおよび実質的な自由度 (複雑さ) の評価に興味がありました．例えば昨今話題の大規模言語モデル (Large Language Model; LLM) は数百億個のパラメータを内包している一方で，内部構造は特殊な形に限定されているので，実際の自由度はパラメータ数よりも遥かに小さいと考えられます．

ニューラルネットは素子の入れ替えなどに対称性を持ち，特異なモデルですので AIC ではなく WAIC を使うとよさそうです．では WAIC を使ってニューラルネットの自由度が評価できるのでしょうか？実は既存の WAIC 理論は，パラメータ数 p が固定された状況で $n \to \infty$ となる状況を考えているので，非常に小規模なニューラルネット (例えば中間素子数 3 のパーセプトロン) を評価できることは示されていましたが，大規模モデル (過剰パラメータモデル) の設定：

$$p > n \quad (\to \infty) \tag{5}$$

で WAIC が利用できるのかわかっていませんでした．上述の状況から，本研究が提起した問いは

(A) WAIC は大規模モデルに利用できるのか，また
(B) 大規模モデルで効率的に WAIC を計算できるか

の 2 つであり，以下の回答を与えています．

(A) 理論的貢献

本研究では，大規模ニューラルネットなどの過剰パラメータモデルの線形近似を f_θ とみなして，高次線形モデル f_θ に対して WAIC が汎化誤差 (2) の事後期待値の漸近不偏推定量となることを示しました．

通常のニューラルネットは理論に登場するいくつかの仮定を満たすことが既存研究によって示されていて，上述の結果は WAIC が過剰パラメータモデルの設定 (5) でも利用できることを示唆しています．

(B) 手法的貢献

すでに説明した通り，WAIC は比較的小規模なモデルを念頭に開発され，大規模モデルで利用するには効率的な計算法が必要です．本研究ではニューラルネットの訓練に用いられる勾配法の亜種として知られている Langevin 過程を利用することで，PyTorch[c]など既存の Wrapper を用いて効率的に WAIC を計算できる Langevin FV (LFV) を提案しました[d]．

実際のニューラルネットで LFV と汎化ギャップを計算した結果が表 1です．LFV が実際の汎化ギャップをよく推定できていることがわかります．本稿では割愛しますが，UCI のレポジトリ[e]から取得した実データセットで，LFV と Cross-Validation 統計量の強い相関も確認しています．

[b] WAIC の詳細を日本語で学べる書籍として，「ベイズ統計の理論と方法」(渡辺澄夫・著) や，より平易な説明が付された「渡辺澄夫ベイズ理論 100 問 with Python/Stan」(鈴木譲・著) などがあります．

[c] https://pytorch.org/

[d] 本研究に関連して，ニューラルネットで巨大なフィッシャー情報行列を計算し，その結果を利用して更に TIC を計算する試みが Bengio グループの Thomas ら (https://proceedings.mlr.press/v108/thomas20a.html) や横田グループの長沼ら (https://openreview.net/forum?id=FH_mZOKFX-b) によって進められています．より厳密にはフィッシャー情報行列が退化してしまい通常の TIC は計算不可能で，固有値をクリップした一般化逆行列を代用しています．詳細はOkuno and Yano (2023) の 3.4 節にあります．

[e] https://archive.ics.uci.edu/

表 1: 中間層 1 層のパーセプトロン (活性化関数は tanh) について，LFV と汎化ギャップ $\tilde{\Delta}$ を人工データで計算したもの．$n = 1000$ で，各値は 25 回の実験の平均と標準偏差．過剰パラメータに該当する設定を灰色で強調した．

	中間素子数 $= 50$		中間素子数 $= 100$		中間素子数 $= 150$	
	LFV	$\tilde{\Delta}$	LFV	$\tilde{\Delta}$	LFV	$\tilde{\Delta}$
共変量次元 $= 5$	8.86 ± 1.20	9.13	9.99 ± 1.32	9.80	10.78 ± 1.38	10.15
共変量次元 $= 10$	17.16 ± 1.56	23.46	19.09 ± 1.55	23.87	21.50 ± 2.41	23.99
共変量次元 $= 15$	25.42 ± 2.04	31.70	30.25 ± 2.21	31.82	32.81 ± 2.30	31.87

3　今後の課題

以上の結果により，WAIC を大規模ニューラルネットなどに適用する準備が整いましたが，課題はまだたくさん残されています．

例えば我々の実験に利用したニューラルネットはパラメータ数が高々 1000–2000 程度であって，昨今用いられている大規模モデルより遥かに小さいです．より大規模なニューラルネットで WAIC(および FV, LFV) がどのように振舞うかについては，新たな共同研究者を加えて現在数値実験を行っています．

また，本研究が提案した理論はあくまで過剰パラメータモデルの線形近似に対して証明を与えているだけなので，非線形なモデルではどうなるのか厳密には証明できていません．より広範な設定をカバーできる理論の登場を期待しています．

参考文献

Okuno, A. and Yano, K. (2023). A generalization gap estimation for overparameterized models via the langevin functional variance. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2023.2197488.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594.