

## 科学研究費助成事業 研究成果報告書

令和 2 年 6 月 11 日現在

機関番号：12608

研究種目：基盤研究(C) (一般)

研究期間：2015～2019

課題番号：15K00331

研究課題名(和文)代数幾何学と構造学習理論に基づく周辺尤度と汎化誤差のモデル評価法の相違の解明

研究課題名(英文) Difference between marginal likelihood and generalization error as statistical model evaluation based on algebraic geometry and structure learning theory

研究代表者

渡邊 澄夫 (WATANABE, SUMIO)

東京工業大学・情報理工学院・教授

研究者番号：80273118

交付決定額(研究期間全体)：(直接経費) 3,200,000円

研究成果の概要(和文)：ベイズ推測において統計モデルと事前分布の最適化のための規準として周辺尤度と汎化誤差が知られているが、その相違について十分には明らかにされていなかった。本研究では主に4つの研究成果が得られた。(1)事後分布が正規分布で近似できるとき、WAICと交差検証を最小化するハイパーパラメータは平均汎化誤差を最小化する値に近づくが周辺尤度の最大化では近づかない。(2)事後分布が正規分布で近似できない例として非負値行列分解の汎化誤差と周辺尤度を解明した。(3)同じ問題設定において変分自由エネルギーの漸近挙動を解明した。(4)混合正規分布の最強検定を与える自由エネルギーの定数オーダー項の分布を導出した。

研究成果の学術的意義や社会的意義

データの分析において統計モデルと事前分布をどのように設計したらよいかという問題は実社会においても常に必要になる課題である。この課題に対して情報量規準、交差検証、周辺尤度は、既に広く実用に用いられているが、それらの相違については必ずしも明確にはされていなかった。本研究により、次のことが明らかになった。(1)WAICと交差検証の最小化は平均汎化誤差を最小にするが、周辺尤度最大化ではそうならない。(2)非負値行列分解における汎化誤差と周辺尤度の漸近挙動の上界の値を求めることができる。(3)そのモデルの変分自由エネルギーの値を知ることができる。(4)混合正規分布のベイズ最強検定を作ることができる。

研究成果の概要(英文)：In Bayesian statistics, two criteria, the generalization error and the marginal likelihood are well known for optimizing a model and a prior, however, their difference was not clarified. In this research, the following four results were obtained, (1) If the posterior can be approximated by a normal distribution, the optimal hyperparameters for WAIC and cross validation converge to the parameter that minimizes the average generalization error, whereas the parameter which maximizes the marginal likelihood does not. (2) Nonnegative matrix factorization is an example of singular statistical models, whose asymptotic generalization error and marginal likelihood were clarified. (3) its variational free energy was also clarified, (4) The asymptotic distribution of the constant order term of the free energy, which is necessary to construct the most powerful test in Bayesian statistics, was clarified.

研究分野：情報数理

キーワード：情報量規準 交差検証 周辺尤度 WAIC WBIC

## 様式 C - 19、F - 19 - 1、Z - 19 (共通)

### 1. 研究開始当初の背景

ある確率分布が存在し、その確率分布に独立に従う確率変数の有限個の集合をサンプルという。サンプルから確率分布(以下では真の分布と呼ぶ)を推測することを統計的推測という。統計的推測を行う方法には様々なものがあるが、統計モデルと事前分布を用意することで定まる事後分布により統計モデルを平均として定義される予測分布を推測の結果とする方法をベイズ推測という。真の分布からみた予測分布との誤差は、設定された統計モデルと事前分布に依存する。この誤差を定量的に調べる際に用いられる指標として汎化誤差と周辺尤度が知られているが、二つの指標の数学的な性質を明らかにすることは、統計学の理論的な基盤として大切であるだけでなく、統計学が応用される領域におけるモデルの評価や選択においても重要な役割を果たすことが期待されていた。

### 2. 研究の目的

本研究では、ベイズ推測において(真の分布、統計モデル、事前分布)の三組を固定したときに、汎化損失と周辺尤度の確率的な挙動を明らかにする。また、真の分布を固定したとき、統計モデルあるいは事前分布の候補が複数であるときに汎化損失と周辺尤度のそれぞれを最小化する方法により統計モデルあるいは事前分布を選ぶ方法の性質を数学的に解明する。

### 3. 研究の方法

まず問題の設定を説明する。

実ユークリッド空間上の確率密度関数  $q(x)$  に独立に従う  $n$  個の確率変数  $X_1, X_2, \dots, X_n$  の集合を  $D_n$  と書いてサンプルと呼ぶ。サンプルが与えられたとき  $q(x)$  の推測を行うという問題を考える。パラメータ  $w$  に対して定義される  $x$  の確率密度関数  $p(x|w)$  を統計モデルあるいは学習モデルという。またパラメータの集合上に定義された確率密度関数  $p(w)$  を事前分布という。与えられたサンプルと統計モデルおよび事前分布に対して事後分布  $p(w|D_n)$  を定義し、事後分布による平均と分散を  $E_w$  と分散  $V_w$  と表記する。統計モデル  $p(x|w)$  を事後分布を用いて平均することで得られる  $E_w[p(x|w)]$  を  $p(x|D_n)$  と書いて予測分布という。汎化損失  $G$ 、学習損失  $T$ 、汎関数分散  $V$ 、情報量規準  $W$ 、交差検証  $C$  を次の式で定義する。

$$\begin{aligned} G &= E_x [\log p(X|D_n)], \\ T &= (1/n) \sum \log p(X_i|D_n), \\ V &= (1/n) \{ V_w [(\log p(X_i|w))^2] - V_w [\log p(X_i|w)]^2 \}, \\ W &= T+V, \\ C &= (1/n) \sum \log E_w [1/p(X_i|w)]. \end{aligned}$$

このとき  $W, C$  のサンプルに関する平均は  $G$  のサンプルに関する平均と漸近的に等しくなることが知られている。すなわち、サンプルの現れ方に関する平均を  $E[\ ]$  と書くことにすると

$$\begin{aligned} E[W] &= E[G] + O(1/n^2), \\ E[C] &= E[G] + O(1/n^2), \end{aligned}$$

が成り立つ。また周辺尤度  $Z$  と自由エネルギー  $F$  を次式で定義する。

$$\begin{aligned} Z &= \int p(w) p(X_i|w) dw, \\ F &= -\log Z. \end{aligned}$$

以上のように定義された  $W, C, F$  は(統計モデル、事前分布)の組の汎関数であると考えることができ、 $W, C, F$  のいずれかを最小とするように(統計モデル、事前分布)を最適化する方法が統計学および機械学習においてしばしば用いられている。しかしながら  $C, W, F$  の最小化は一般には等価ではなく、それらがどのように異なるかについては十分には調べられていなかった。

本研究では、上記の問題を解決するために、次の点を解明する方法を用いた。

- (1)事後分布が正規分布で近似できる場合において  $C, W, F$  による事前分布の最適化の相違を比較する。
- (2)事後分布が正規分布で近似できないモデルの例として非負値行列分解を考察し、その実対数閾値を解明し、 $C, W, F$  の漸近挙動を比較する。
- (3)非負値行列分解の統計的推測の方法として変分ベイズ法を用いた場合の漸近挙動を求めてベイズ法の場合と比較する。
- (4)ベイズ法を用いた統計的検定においては  $F$  の分布を調べ有意水準に対応する棄却域を定める

必要があるが、正則でないモデルの例として混合正規分布における F の確率分布を解明する。

#### 4. 研究成果

研究目的(1)～(4)に合わせてそれぞれの研究成果を説明する。

- (1) 事後分布が正規分布で近似できる場合に事前分布を最適化するという問題を考察し、次の結果が得られた。事前分布がハイパーパラメータをもち、ハイパーパラメータを最適化するという場合を考える。
- (1.1) ハイパーパラメータを W が最小になるように定めた場合、そのハイパーパラメータは、n が大きくなるにつれて E[G] を最小化する値に収束する。
  - (1.2) ハイパーパラメータを C が最小になるように定めた場合、そのハイパーパラメータは、n が大きくなるにつれて E[G] を最小化する値に収束する。
  - (1.3) ハイパーパラメータを F が最小になるように定めた場合、そのハイパーパラメータは、n が大きくなっても E[G] を最小化する値に一般には収束しない。
  - (1.4) G を最小化するハイパーパラメータは n が大きくなってもある固定の値には収束しない。すなわち G を最小化するパラメータは E[G] を最小化するパラメータに近づかない。従って(1.1), (1.2), (1.3)のいずれの場合も、G は最小化されない。
  - (1.5) W を最小化するハイパーパラメータと C を最小化するハイパーパラメータは漸近的な挙動は等しいが、具体的な例による実験では W を最小化するほうが分散が小さいことが多い。

- (2) 統計モデルが正則であってもなくても、G、W、C、F は次の漸近挙動を持つことが知られている。

$$\begin{aligned} E[G] &= S + \frac{1}{n} + o(1/n), \\ E[W] &= S + \frac{1}{n} + o(1/n), \\ E[C] &= S + \frac{1}{n} + o(1/n), \\ F &= n S n + \log n - (m-1) \log \log n + o_p(1). \end{aligned}$$

ここで S, S<sub>n</sub> はそれぞれ真の分布のエントロピー、経験エントロピーであり、 $\frac{1}{n}$  は実対数閾値であり、m はその多重度である。非負値行列分解ではこの値が不明であったが、本研究において次のように上界が与えられた。M×N の非負値行列のサンプルが与えられて M×H の非負値行列と H×N の非負値行列の積に統計的に分解するとき

$$(1/2) \{ (YH - H_0) \min(M, N) + H_0(M+N-1) \}$$

であることが解明された。ここで H<sub>0</sub> は真の行列のランクである。

- (3) ベイズ推測は事後分布を実現するために多大な演算量を必要とすることが多く、少ない演算量で事後分布を近似する方法として変分ベイズ法(平均場近似法)が知られている。変分ベイズ法によって得られる変分ベイズ自由エネルギーは非負値行列分解では不明であったが、本研究により、パラメータの事前分布が原点の近傍で零に近づかない場合には

$$F = (1/2)(M+N)H \log n + o_p(\log n)$$

であることが解明された。また原点の近傍で零に近づく場合に生じる相転移についても解明することができた。

- (4) ベイズ推測における検定では、自由エネルギーは最強検定を与えることが知られているが、自由エネルギーを用いた検定では自由エネルギーの定数オーダーの項が従う確率分布が必要となるため、正則でないモデルにおける検定統計量を作ることは困難であった。本研究では混合正規分布における検定統計量が、平均と分散の推定量から計算が可能であることを明らかにした。

5. 主な発表論文等

〔雑誌論文〕 計9件（うち査読付論文 5件／うち国際共著 0件／うちオープンアクセス 0件）

1. 著者名 Naoki Hayashi, Sumio Watanabe.	4. 巻 1
2. 論文標題 Asymptotic Bayesian Generalization Error in Latent Dirichlet Allocation and Stochastic Matrix Factorization.	5. 発行年 2020年
3. 雑誌名 SN Computer Science	6. 最初と最後の頁 1-22
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42979-020-0071-3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 渡邊匠、渡辺澄夫	4. 巻 IBISML2019-18
2. 論文標題 混合多項分布のベイズ汎化誤差の漸近挙動	5. 発行年 2020年
3. 雑誌名 信学技報	6. 最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 田中来輝、渡辺澄夫	4. 巻 IBISML2019-18
2. 論文標題 Swish 関数を用いた階層型神経回路網の実対数閾値	5. 発行年 2020年
3. 雑誌名 信学技報	6. 最初と最後の頁 9-5
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 永安修也、渡辺澄夫	4. 巻 NC2019-14
2. 論文標題 最適な確率分布が一意でないときのベイズ学習曲線	5. 発行年 2020年
3. 雑誌名 信学技報	6. 最初と最後の頁 107-112
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Sumio Watanabe	4. 巻 25210.1007/978-3-319-97798-0
2. 論文標題 Higher Order Equivalence of Bayes Cross Validation and WAIC	5. 発行年 2018年
3. 雑誌名 Springer Proceedings in Mathematics and Statistics	6. 最初と最後の頁 pp.47-73
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-3-319-97798-02	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Hayashi Naoki, Watanabe Sumio	4. 巻 266
2. 論文標題 Upper bound of Bayesian generalization error in non-negative matrix factorization	5. 発行年 2017年
3. 雑誌名 Neurocomputing	6. 最初と最後の頁 21 ~ 28
掲載論文のDOI (デジタルオブジェクト識別子) doi.org/10.1016/j.neucom.2017.04.068	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Kohjima Masahiro, Watanabe Sumio	4. 巻 Vol.2
2. 論文標題 Phase Transition Structure of Variational Bayesian Nonnegative Matrix Factorization	5. 発行年 2017年
3. 雑誌名 Proc. of ICANN	6. 最初と最後の頁 146 ~ 154
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-3-319-68612-7_17	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Hayashi Naoki, Watanabe Sumio	4. 巻 Vol.1
2. 論文標題 Tighter upper bound of real log canonical threshold of non-negative matrix factorization and its application to Bayesian inference	5. 発行年 2017年
3. 雑誌名 IEEE Symposium Series on Computational Intelligence	6. 最初と最後の頁 1-8
掲載論文のDOI (デジタルオブジェクト識別子) 10.1109/SSCI.2017.8280811	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Sumio Watanabe	4. 巻 Vol.1
2. 論文標題 Difference between Bayes Cross Validation and WAIC for Conditional Independent Samples	5. 発行年 2017年
3. 雑誌名 Proc. of 10th WITMSE2017	6. 最初と最後の頁 38-41
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

〔学会発表〕 計15件 (うち招待講演 6件 / うち国際学会 4件)

1. 発表者名 渡辺澄夫
2. 発表標題 神経回路網研究における代数幾何学的方法とその成果について
3. 学会等名 神経回路学会 (招待講演)
4. 発表年 2019年

1. 発表者名 飯屋夏樹 渡辺澄夫
2. 発表標題 混合正規分布の均一性検定におけるベイズ検定統計量の漸近挙動
3. 学会等名 2018年度統計関連連合大会
4. 発表年 2018年

1. 発表者名 Sumio Watanabe
2. 発表標題 Cross Validation and WAIC in layered neural networks
3. 学会等名 Deep Learning: Theory, Algorithms, and Applications (招待講演) (国際学会)
4. 発表年 2018年

1. 発表者名 佐藤件一郎、渡辺澄夫、
2. 発表標題 混合ポアソン分布の実対数閾値とベイズ汎化誤差について、
3. 学会等名 情報論的学習理論と機械学習研究会
4. 発表年 2018年

1. 発表者名 林 直輝、渡辺澄夫、
2. 発表標題 確率行列分解の実対数閾値とBayes学習への応用、
3. 学会等名 情報論的学習理論と機械学習研究会
4. 発表年 2018年

1. 発表者名 渡邊澄夫
2. 発表標題 Higher Order Analysis of Bayesian Cross Validation in Regular Asymptotic Theory
3. 学会等名 Information Geometry and its Applications IV (招待講演) (国際学会)
4. 発表年 2016年

1. 発表者名 渡邊澄夫
2. 発表標題 Bayesian sparse regression and information criteria for singular models. International Workshop
3. 学会等名 Algebraic and Geometric Methods in Statistics (国際学会)
4. 発表年 2017年

1. 発表者名 渡邊澄夫
2. 発表標題 ベイズ推論
3. 学会等名 電子情報通信学会ソサイエティ大会（招待講演）
4. 発表年 2016年

1. 発表者名 林直輝 渡辺澄夫
2. 発表標題 非負値行列分解の実対数閾値とBayes学習への応用
3. 学会等名 電子情報通信学会情報論的学習理論研究会
4. 発表年 2016年

1. 発表者名 林直輝 渡辺澄夫
2. 発表標題 非負値行列分解における実対数閾値の実験的考察
3. 学会等名 電子情報通信学会NC研究会
4. 発表年 2017年

1. 発表者名 香田夏輝 渡邊澄夫
2. 発表標題 スパース表現を用いた非線形多層主成分分析における 学習結果の分類法について
3. 学会等名 情報論的学習理論と機械学習研究会
4. 発表年 2015年



1. 発表者名 渡邊澄夫
2. 発表標題 Cross Validation and WAIC in Statistical Model Evaluation
3. 学会等名 Symposium What is a good model? (招待講演) (国際学会)
4. 発表年 2016年

1. 発表者名 渡邊澄夫
2. 発表標題 潜在変数を持つモデルの評価について
3. 学会等名 データ科学シンポジウム (招待講演)
4. 発表年 2016年

1. 発表者名 中村文士 渡邊澄夫
2. 発表標題 ハミルトニアンモンテカルロ法を用いた神経回路網の学習と汎化誤差の推定について
3. 学会等名 信学技報
4. 発表年 2016年

1. 発表者名 須貝将士 渡邊澄夫
2. 発表標題 非ガウス可解モデルを用いたレプリカモンテカルロ法 による自由エネルギーの計算精度の評価
3. 学会等名 信学技報
4. 発表年 2016年

〔図書〕 計1件

1. 著者名 Sumio Watanabe	4. 発行年 2018年
2. 出版社 CRC Press	5. 総ページ数 300
3. 書名 Mathematical theory of Bayesian statistics	

〔産業財産権〕

〔その他〕

渡辺澄夫 <a href="http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/index-j.html">http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/index-j.html</a>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------

6. 研究組織

氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
---------------------------	-----------------------	----