Discussion Paper Series

RIEB

Kobe University

DP2017-J12

広く使えるベイズ情報量規準(WBIC)の応用について：線形回帰モデルを例に

渡辺　寛之

2017年9月4日

神戸大学 経済経営研究所

〒657-8501　神戸市灘区六甲台町 2-1

広く使えるベイズ情報量規準 (WBIC) の応用について：線形回帰モデルを例に

渡辺　寛之 *

2017 年 9 月

概 要

本稿は事前分布についての情報が無い場合におけるモデル選択について検討した研究ノートである。ベイズ統計学の文献で多く使われているモデル選択基準は周辺尤度である。ただし尤度関数が真のデータ生成過程と整合的であったとしても、事前分布を任意で設定し、それがかなり強い (informative な) 事前分布であるか、あるいはかなり弱い (noninformative な) 事前分布である場合、その尤度関数を含むモデルは周辺尤度によって支持されないことがある。一方、広く使えるベイズ情報量規準 (WBIC) では十分に弱い事前分布を用いれば “かろうじて無情報” な事前分布の下での周辺尤度の推定が自動的におこなわれるため、事前分布の情報が無い場合でも適切にモデル選択をおこなえる可能性がある。

キーワード:ベイズ情報量規準、周辺尤度、無情報事前分布、WBIC

*657-8501 兵庫県神戸市灘区六甲台町 2-1 神戸大学経済経営研究所　Email: watanabe1133@gmail.com

1　はじめに

計量経済学の応用において、精密な経済理論から導出した尤度関数を用いるが、それと同程度に事前分布については検討しない場合がある。その際、とりあえず先行研究と同様の事前分布を用いるか、弱い事前分布を用いて周辺尤度を計算し、モデル選択をおこなう。しかし、周辺尤度は事前分布までを含めてモデルを評価するため、尤度関数に重点を置いた評価をおこなうのは難しい。実際、弱い事前分布がモデル選択基準に与える影響は強い。このことは後で数値例を示す。先行研究では、事前分布とモデル選択基準の問題は Kass and Wasserman(1995), Berger etal.(2001), Wagenmakers (2007), Hoff (2010) などで研究されている。これらの研究では尤度関数に重点をおいて周辺尤度を計算できるような単位情報事前分布 (unit information prior) の利用について言及している。単位情報事前分布はデータ 1 単位が持っていると考えられる情報量（単位情報）を反映した事前分布であり、例えば線形回帰モデルの係数パラメーターであれば、その分散にデータの数 J を掛けたもの $N(\hat{\beta}_{ols}, \Sigma^{-1})$ が単位情報事前分布となる。ただし $\Sigma = (X'X)/(J\sigma^2)$ である。[1]しかし、このように簡単に単位情報を取り出せるのはパラメーターの（事後）平均、分散や分布型を容易に知ることができる場合に限られるだろう。他方、近年 Watanabe(2013b) においてパラメーターの正規性など基本的な条件に依存せず周辺尤度を近似できる “広く使えるベイズ情報量規準 (WBIC)” が提案されている。

本稿の目的は事前に事前分布の情報がない場合でも真のデータ生成過程と整合的なモデルを選ぶために、数値実験を通して WBIC の性質を確かめことである。得られた結果として、事前分布が強い場合では WBIC は周辺尤度を近似し、事前分布が弱い場合では WBIC は Schwarz Bayesian information criterion (BIC) を近似する。これは Friel etal.(2017) の報告を補完するものである。[2]また、弱い事前分布の代わりにフラットな（非正則な）無情報事前分を用いても WBIC は同様の結果をもたらす。本稿の構成は以下の通りである。第 2 章では真のデータ生成過程と整合的な尤度関数と任意の事前分布を組み合わせると、モデルが周辺尤度によって支持されないことを数値的に示す。第 3 章では事前分布の強さを変え

[1]詳しくは Hoff(2010) の 9 章を参照されたい。なお $\hat{\beta}_{ols}$ は最小二乗法による推定値である。

[2]Friel etal.(2017) は事前分布の分散を大きくすると、WBIC が周辺尤度を過剰に見積もることを報告している。

た場合における WBIC と周辺尤度と BIC の関係を確かめる。第 4 章で結論を述べる。

2　任意の事前分布と周辺尤度

まず観測データが以下のように真の確率分布から発生していると考える:

$$\boldsymbol{y} \sim Q(\boldsymbol{y}), \qquad Q(\boldsymbol{y}) = \prod_{n=1}^{N} q(y_n), \tag{1}$$

ただし、y_n は n 番目の観測データであり、$n = 1, 2, ..., N$ である。$\boldsymbol{y}$ は $y_{1:N}$ を意味する。q は確率分布である。続いて、以下の任意の同時確率分布を考える：

$$P(\boldsymbol{y}, \theta) = f(\boldsymbol{y}|\theta) f(\theta), \tag{2}$$

ただし、θ はパラメーターベクトル、$f(\boldsymbol{y}|\theta)$ は尤度関数とする。また $f(\theta)$ は事前分布である。この時、$\int P(\boldsymbol{y}, \theta) d\theta$ の $Q(\boldsymbol{y})$ に対する近似の良さを知るために $\int P(\boldsymbol{y}, \theta) d\theta$ それ自体が情報量として参照されるので以下のように定義する：

$$m(\boldsymbol{y}) = log\left(\int P(\boldsymbol{y}, \theta) d\theta \right), \tag{3}$$

$m(\boldsymbol{y})$ を周辺尤度と呼ぶ。[3]周辺尤度は $Q(\boldsymbol{y})$ に対するカルバックライブラ情報量に対応する。[4]したがって $\boldsymbol{y}$ が $Q(\boldsymbol{y})$ からあたえられた下で、競合するいくつかの $P(\boldsymbol{y}, \theta)$ を比較することが可能である。

ここで、事前分布が強い、または弱い場合に尤度関数が真のデータ生成過程と整合的であっても周辺尤度によって支持されない例を確認しておく。数値実験の設定は以下の通りである。引き続き真のデータ生成過程を $Q(\boldsymbol{y})$、確率モデルを $P_*(\boldsymbol{y}, \theta)$ とする。なお $P_*(\boldsymbol{y}, \theta)$ の尤度関数は $Q(\boldsymbol{y})$ と整合的である。もう一つの確率モデルを $P_\dagger(\boldsymbol{y}, \theta)$ とする。なお $P_\dagger(\boldsymbol{y}, \theta)$ の尤度関数は $Q(\boldsymbol{y})$ と非整合である：

[3]正確には対数周辺尤度であるが、本稿では省略する。

[4]厳密には周辺尤度の期待値がカルバックライブラ情報量と、モデルの設定に依存しない定数との和と一致する。詳しくは渡辺 (2012) を参照のこと。

データを生成している分布 $Q(\boldsymbol{y})$

$$Q(\boldsymbol{y}) = (2\pi\sigma^2)^{-N/2} exp\left\{ -\frac{\sum_{n=1}^{N}(y_n - \mu_n)^2}{2\sigma^2} \right\}, \tag{4}$$

$$\mu_n = 1 - 3x_n - 0.3x_n^2 + x_n^3 - 0.1x_n^4 + 0.1x_n^5 - 0.1x_n^6 + 0.1x_n^7, \tag{5}$$

ただし $\sigma = \sqrt{50}$、 $N = 300$ であり、 $x_{1:N}$ は -2 から 2 までを当分割した N 個の外生変数である.

$Q(\boldsymbol{y})$ と整合的な尤度関数を持つ分布 $P_*(\boldsymbol{y}, \theta)$

$$P_*(\boldsymbol{y}, \theta) = f_*(\boldsymbol{y}|\theta) f_*(\theta), \tag{6}$$

$$f_*(\boldsymbol{y}|\theta) = (2\pi\sigma^2)^{-N/2} exp\left\{ -\frac{\sum_{n=1}^{N}(y_n - \mu_n)^2}{2\sigma^2} \right\}, \tag{7}$$

$$\mu_n = \alpha + \beta_1 x_n + \beta_2 x_n^2 + \beta_3 x_n^3 + \beta_4 x_n^4 + \beta_5 x_n^5 + \beta_6 x_n^6 + \beta_7 x_n^7, \tag{8}$$

$$f_*(\theta) = N(\mathbf{0}, \tau^2 I), \tag{9}$$

ただし σ は $\sqrt{50}$ に固定している。 $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)$ は推定されるパラメーターであり、 I は 8×8 の単位行列である. また、対応する事後分布を $P_*(\theta|\boldsymbol{y})$ 周辺尤度を $m_*(\boldsymbol{y})$ と定義しておく。

$Q(\boldsymbol{y})$ と非整合な尤度関数を持つ分布 $P_*(\boldsymbol{y}, \theta)$

$$P_\dagger(\boldsymbol{y}, \theta) = f_\dagger(\boldsymbol{y}|\theta) f_\dagger(\theta), \tag{10}$$

$$f_\dagger(\boldsymbol{y}|\theta) = (2\pi\sigma^2)^{-N/2} exp\left\{ -\frac{\sum_{n=1}^{N}(y_n - \mu_n)^2}{2\sigma^2} \right\}, \tag{11}$$

$$\mu_n = \alpha + \beta x_n, \tag{12}$$

$$f_\dagger(\theta) = N(\mathbf{0}, \tau^2 I), \tag{13}$$

ただし σ はOLSにより得られた最適な値があたえられており、$\sqrt{75.2}$ である。 (α, β) は推定されるパラメーターであり、 τ は式 (9) と同様である。 I は 2×2 の単位行列である。 また、対応する事後分布を $P_\dagger(\theta|\boldsymbol{y})$、周辺尤度を $m_\dagger(\boldsymbol{y})$ と定義しておく。

これらの設定の下で τ の値を $\sqrt{10^{z_1}}$ から $\sqrt{10^{z_{21}}}$ まで動かす。ただし $z_1 = -10, z_2 = -9, ..., z_{20} = 9$, そして $z_{21} = 10$ である。 以下のベイズの

定理を使って、それぞれの z_i に対して $m_*(\boldsymbol{y})$ と $m_\dagger(\boldsymbol{y})$ を解析的に計算した:

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)f(\theta)}{f(\theta|\boldsymbol{y})}, \tag{14}$$

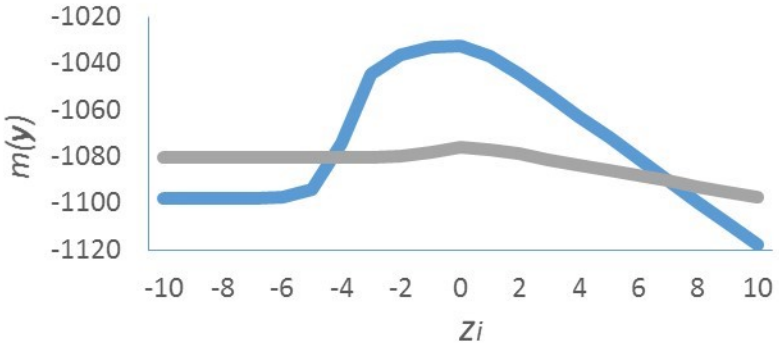
ただし $f(\theta|\boldsymbol{y})$ は事後分布であり、 $\boldsymbol{y}$ は $Q(\boldsymbol{y})$ から得られたものである。図 1 は計算した周辺尤度の値をプロットしている。

図 1 によれば $z_i < -5$ あたりで $m_*(\boldsymbol{y})$ は $m_\dagger(\boldsymbol{y})$ より小さく、非整合な尤度関数をもつ分布が支持される。 当該の事前分布は強い事前分布であり事後分布に対する影響が大きく、尤度関数の影響を凌駕している。また式 (2) と式 (3) を思い出すと、$m(\boldsymbol{y})$ の値は事前分布のモード周辺で評価した $f(\boldsymbol{y}|\theta)$ の値に近づくことがわかる。よく知られている通り、任意の点で尤度関数を評価したとしてもモデル選択の観点で意味があるとは言えない。このような意味で尤度関数がデータ生成過程と整合的であっても周辺尤度は必ずしも、そのモデルを支持しないことがわかる。

本稿の焦点は $z_i = 10$ のあたりで図 1 が示していることである。そこでは $m_*(\boldsymbol{y})$ が $m_\dagger(\boldsymbol{y})$ より小さい。事前分布は弱く、事後分布に対してほとんど無情報である。この事前分布の下では、事後分布はほとんど尤度関数に比例した形になり、事前分布の影響をあまり受けない。図 2 は z_i に対する $f_*(\theta)$ と $f_*(\theta|\boldsymbol{y})$ をプロットしている。[5]図 3 は z_i に対する $f_\dagger(\theta)$ と $f_\dagger(\theta|\boldsymbol{y})$ をプロットしている。両方の場合において事後密度は事前分布の増加する分散に対して一定である。対照的に事前分布の密度は単調に低下している。重要なのは事前分布の密度低下のスピードが事前分布の次元に依存することである。$f_*(\theta)$ の次元は $f_\dagger(\theta)$ よりも高いので事前分布の分散が大きくなるにつれ $m_*(\boldsymbol{y})$ は必ず $m_\dagger(\boldsymbol{y})$ よりも低くなる。[6]データ生成過程と整合的な尤度関数を使っていたとしても弱い事前分布を用いた場合、周辺尤度という基準では逆の結果を得るという一例である。

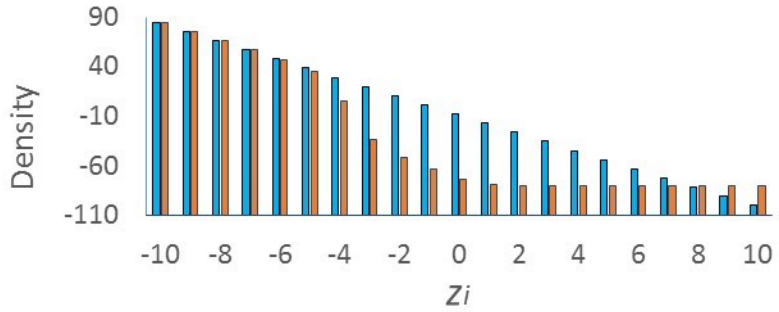
[5] $(\alpha, \beta_1, ..., \beta_7) = \mathbf{0}$ で評価している。
[6] 式 (14) を念頭に議論している。ただし、そこでは尤度関数は一定である。

図 1: $m_*(\boldsymbol{y})$ と $m_\dagger(\boldsymbol{y})$



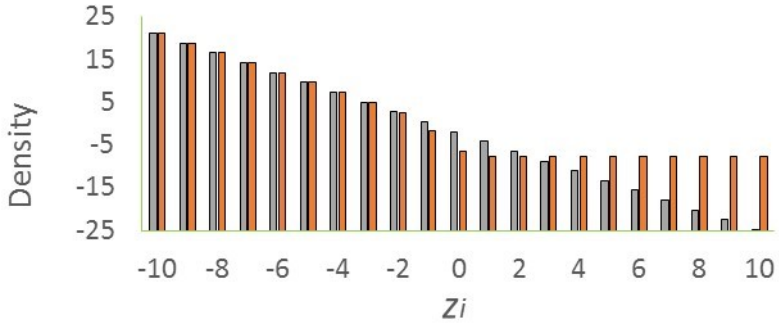
注: 青い線が $m_*(\boldsymbol{y})$、 グレーの線が $m_\dagger(\boldsymbol{y})$ を示している。

図 2: $f_*(\theta)$ と $P_*(\theta|\boldsymbol{y})$



注: 青いバーが事前分布の密度、赤いバーが事後分布の密度を示している。

図 3: $f_\dagger(\theta)$ と $P_\dagger(\theta|\boldsymbol{y})$



注: グレーのバーが事前分布の密度、赤いバーが事後分布の密度を示している。

3　WBICと周辺尤度およびBICの関係

これまでの例が示すように、事前分布についての事前の情報が無いままモデル選択をおこなう場合、事前分布は強すぎても弱すぎても効果的でない。尤度関数と整合的でかつ適切な強さが望ましいと考えられる。このような事前分布の一例として単位情報事前分布 (unit information prior) があり、Kass and Wasserman (1995)、 Berger etal. (2001)、 Wagenmakers (2007)、Hoff (2010, Section 9) で言及されている。特に Hoff (2010, Section 9) がわかりやすい。単位情報事前分布はデータ 1 単位が持っていると考えられる情報量（単位情報）を反映した事前分布であり、例えば線形回帰モデルの係数パラメーターであれば、その分散にデータの数 J を掛けたもの $N(\hat{\beta}_{ols}, \Sigma^{-1})$ が単位情報事前分布となる。ただし $\Sigma = (X'X)/(J\sigma^2)$ である。しかし、このように簡単に単位情報を取り出せるのはパラメーターの（事後）平均、分散や分布型を容易に知ることができる場合に限られるだろう。他方、近年 Watanabe(2013b) においてパラメーターの正規性に依存せず周辺尤度を近似できる “広く使えるベイズ情報量規準 (WBIC)” が提案されている。Watanabe(2013b) によれば WBIC は以下のように計算できる。

$$m(\boldsymbol{y}) = w + O_p(\sqrt{log(N)}), \tag{15}$$

$$w = \int log(f(\boldsymbol{y}|\theta))g(\theta)d\theta, \tag{16}$$

$$g(\theta) = f(\boldsymbol{y}|\theta)^{\lambda} f(\theta)/c, \tag{17}$$

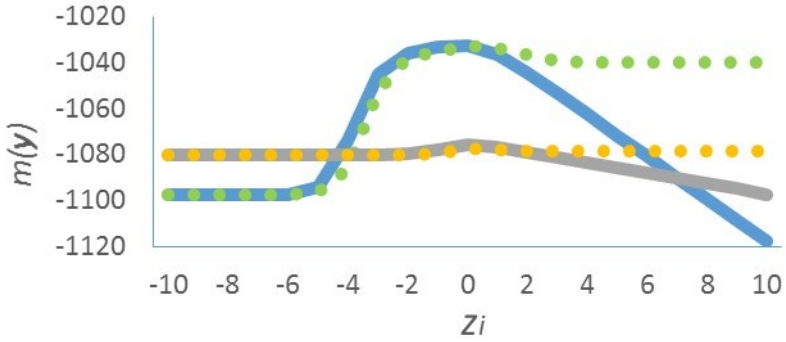
ただし w は WBIC、 $\lambda = 1/log(N)$、 $g(\theta)$ は事後分布であり c は基準化定数である。[7] $g(\theta)$ からサンプリングした MCMC ドローで対数尤度関数を平均することにより周辺尤度が推定される。

ここで注意すべきは WBIC は事前分布の密度低下に起因する周辺尤度の推定値の低下を妨げる働きがあるということである。[8] この特徴は事前分布が弱い場合に $g(\theta)$ からの MCMC ドローが事前分布の影響を受けないことに由来する。つまり事前分布の分散を大きくすることにより事前分布の密度が低下していったとしても、MCMC ドローへの影響が無くなった（無情報になった）時点で WBIC は周辺尤度に対する事前分布の密度

[7] 本稿では WBIC と BIC の符号を逆転させている。これは自由エネルギー、$-m(\boldsymbol{y})$ ではなく周辺尤度に対応させるためである。

[8] Friel et al.(2017) は 事前分布の分散が大きい時に WBIC は周辺尤度と一致しないことを報告しているが、このことは本稿での議論と整合的である。

図 4: WBIC と周辺尤度



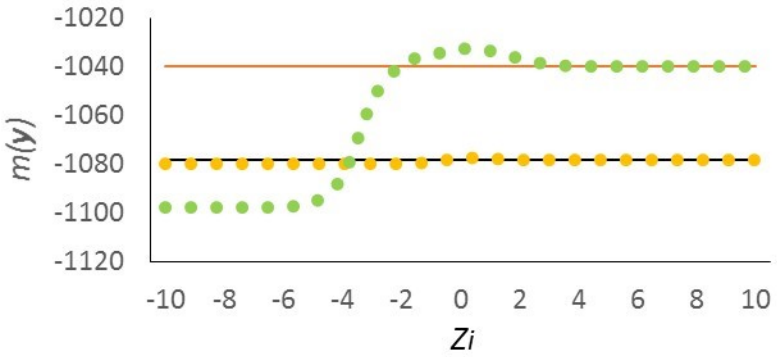
注: 緑の点線は $P_*(\boldsymbol{y},\theta)$ に対する WBIC を、 黄色の点線は $P_\dagger(\boldsymbol{y},\theta)$ に対する WBIC を示している。青い実線は $m_*(\boldsymbol{y})$ を、グレーの実線は $m_\dagger(\boldsymbol{y})$ を示している。

低下の効果を無視するようになり、周辺尤度の推定値の低下をある時点で止めることになる。したがって、WBIC は我々が $g(\theta)$ に対して “かろうじて無情報” な事前分布を用いているかのような結果をもたらす。図 4 は $P_*(\boldsymbol{y},\theta)$ と $P_\dagger(\boldsymbol{y},\theta)$ について計算した WBIC の値および $m_*(\boldsymbol{y})$ と $m_\dagger(\boldsymbol{y})$ を重ねてプロットしている。事前分布が $g(\theta)$ に対して影響していると思われる $z_i < 1$ において WBIC は周辺尤度を正確に推定している。一方、事前分布が $g(\theta)$ に対して無情報であると思われる $z_i > 1$ において WBIC は低下を止めている。そして $z_i = 10$ に至るまで WBIC はほとんど一定であり “かろうじて無情報” な事前分布を用いた下で計算される周辺尤度の値を示し続けることになる。後述するが、この低下の止まった WBIC は Schwarz Bayesian Information Criterion (BIC) と一致する。[9]

以上のような WBIC の特徴を利用することにより、十分弱い事前分布の下で計量経済モデルを比較することが可能である。ただし現実的な応用においては我々は事前分布が十分弱いかどうか事前にわからないこともある。したがって実用的な方法としてフラットな無情報事前分布の利用が考えられる。例えば $f(\theta) = 1$ や $f(log(\theta)) = 1$ などである。フラットな無情報事前分布を用いたとしても WBIC は “かろうじて無情報” な事前分布を用いた下で計算される周辺尤度の値を自動的に推定することになる。図 5 はこれまでと同じモデル $P_*(\boldsymbol{y},\theta)$ と $P_\dagger(\boldsymbol{y},\theta)$ に対して WBIC と BIC を計算し、重ねてプロットしたものである。この数値例が示すと

[9]ただし、BIC と一致するのはパラメーターが正規分布するようなケースであり、特異モデルと呼ばれる混合正規モデルなどは該当しない。

図 5: WBIC と BIC



注: 緑の点線は $P_*(\boldsymbol{y},\theta)$ に対する WBIC を、黄色の点線は $P_\dagger(\boldsymbol{y},\theta)$ に対する WBIC を示している。ピンクの実線は $P_*(\boldsymbol{y},\theta)$ に対する BIC を、黒の実線は $P_\dagger(\boldsymbol{y},\theta)$ に対する BIC を示している。

おり事前分布が十分に弱ければ WBIC は BIC に対応する。この関係性は Watanabe(2013b) で理論的にサポートされている：

$$w = b + o_p(1), \tag{18}$$

ここで b は BIC である。ただし BIC は常に周辺尤度に対応しているわけではないので、(18) と (15) は必ず同時に成立するわけではないということに注意が必要である。図 4 と図 5 が示唆していることをまとめると：強い事前分布を使うと WBIC は周辺尤度を近似し、弱い (noninformative な) 事前分布を使うと BIC を近似するということである。したがって例えば (i) 経済理論から要請されるような情報を事前分布に取り込むような形で事前に事前分布に関する情報が得られている時や、(ii) 尤度関数だけが手元にあり事前分布についての情報が得られていない時、その両方のケースで WBIC は妥当な結果をもたらすことが期待される。

4 結語

計量モデルをベイジアンの文脈で評価する場合、事前分布についての情報がわからない状態で周辺尤度を用いると誤った結論を導く恐れがある。本稿では線形回帰モデルを例に数値実験を行うことで、そのような状況を回避するための一手段としての WBIC の応用を議論した。WBIC はその汎用性からモデル選択の有望な手段となりうると考えられる。

参考文献

[1] Berger, J. O., Pericchi, L. R., Ghosh, J. K., Samanta, T., De Santis, F., Berger, J. O., and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. Lecture Notes-Monograph Series, 135-207.

[2] Friel, N., McKeone, J. P., Oates, C. J., and Pettitt, A. N. (2017). Investigation of the widely applicable Bayesian information criterion. Stat. Comput., 27(3), 833-844.

[3] Hoff, P. D. (2010). A first course in Bayesian statistical methods. Springer Sci. Bus. Media.

[4] Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J. Am. Stat. Assoc., 90(431), 928-934.

[5] Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. Statistics and Computing, 26(4), 881-897.

[6] Mononen, T. (2015). A case study of the widely applicable Bayesian information criterion and its optimality. Statistics and Computing, 25(5), 929-940.

[7] Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. Psychon. bull. rev., 14(5), 779-804.

[8] Watanabe, S. (2013a). WAIC and WBIC are information criteria for singular statistical model evaluation. Discussion paper.

[9] Watanabe, S. (2013b). A widely applicable Bayesian information criterion. J. Mach. Learn. Res., 14(Mar), 867-897.

[10] Watanabe, S. (2009). Algebraic geometry and statistical learning theory (Vol. 25). Cambridge University Press.

[11] 渡辺澄夫 (2012) ベイズ統計の理論と方法　コロナ社