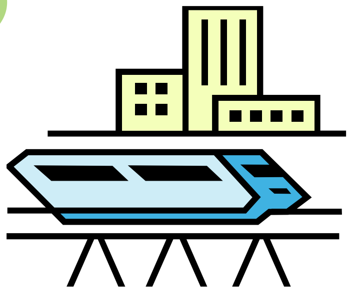
ベイズ統計入門 (11)

目標　一般モデルの選択

東京工業大学
渡辺澄夫

旅の地図

(1) ベイズ統計の定義
(2) 密度と条件つき密度
(3) 混合正規分布＋ギブスサンプラー
(4) 神経回路網＋ランジュバン方程式
(5) 真とモデルの関係
(6) 正則モデルの漸近理論
(7) AIC と BIC
(8) ハイパーパラメータ最適化
(9) 一般モデルの漸近理論
(10) 一般モデルの漸近理論
(11) 一般モデルの選択
(12) 条件つき独立 高次元
(13) 階層ベイズ
(14) 相転移
(15) まとめ



1　汎化損失を推測する規準

規準の定数倍について

この講座では、情報量規準を汎化損失の推定量と考えているため、
n に対するスケールが一般に使われているものと異なります。

統計ソフトなどで算出される情報量規準と値を比較したい場合には
2n 倍してください。

この講座で用いているスケールでは、真の分布がモデルで実現可能で
あれば n→∞において、情報量規準（交差損失も含む）は真の分布の
エントロピーに確率収束します。

その意味で異なるデータに対して情報量規準を比較することはできなく
はないのですが、異なるデータに対する情報量規準の揺らぎは汎化
損失の漸近挙動よりもオーダーとして大きいので、あまり意味のある
比較にはならないことにご注意ください。

クロスバリデーション

LOOCV

$$CV_n = -(1/n) \sum_{i=1}^{n} \log (p(X_i \mid X^n - X_i))$$

STONE, 1974. GEISSER, 1975.

データの独立性が必要である。他の情報量規準は条件つき
独立の場合に利用できるがCVは利用できない

重点サンプリングクロスバリデーション*(1992 Gelfand et.al.)*

$$ISCV_n = (1/n) \sum_{i=1}^{n} \log E_w[1 / p(X_i|w)]$$

Peruggia (1997) は X^n の中に影響力の大きなデータが
含まれているときには平均値 $E_w[1/p(x|w)]$ が存在しない、
あるいは分散が無限大になることを示した。

偏差情報量規準(DIC)

(2002　Spiegelhalter et.al.)

$$DIC_n = -(1/n) \sum_{i=1}^{n} \log(\, p(X_i|\mathbf{E}_w[w])\,)$$

$$+ (2/n) \sum_{i=1}^{n} \{ - E_w[\, \log(\, p(X_i|w))\,] + \log(\, p(X_i|\mathbf{E}_w[w])\,)\,\}$$

もしも真の分布がモデルで実現可能であり正則であれば

$$\mathbf{E}[G_n]=\mathbf{E}[DIC_n] + o(1/n).$$

$p(x|\mathbf{E}_w[w])$　を推測結果とする方法を事後平均プラグイン法と呼ぶことにする.
DICの目的は事後平均プラグイン法の評価を与えることかもしれない．ベイズと事後平均プラグインは、実現可能かつ正則な場合には漸近的に等価になるが一般には異なる．この規準は事後分布が正規分布でないときには利用できない。ハイパーパラメータの設計には使えない。

広く使える情報量規準

(2009 Watanabe)

$$\mathrm{WAIC}_n = T_n + (1/n) \sum_{i=1}^{n} \mathbf{V}_w[\log(p(X_i|w))]$$

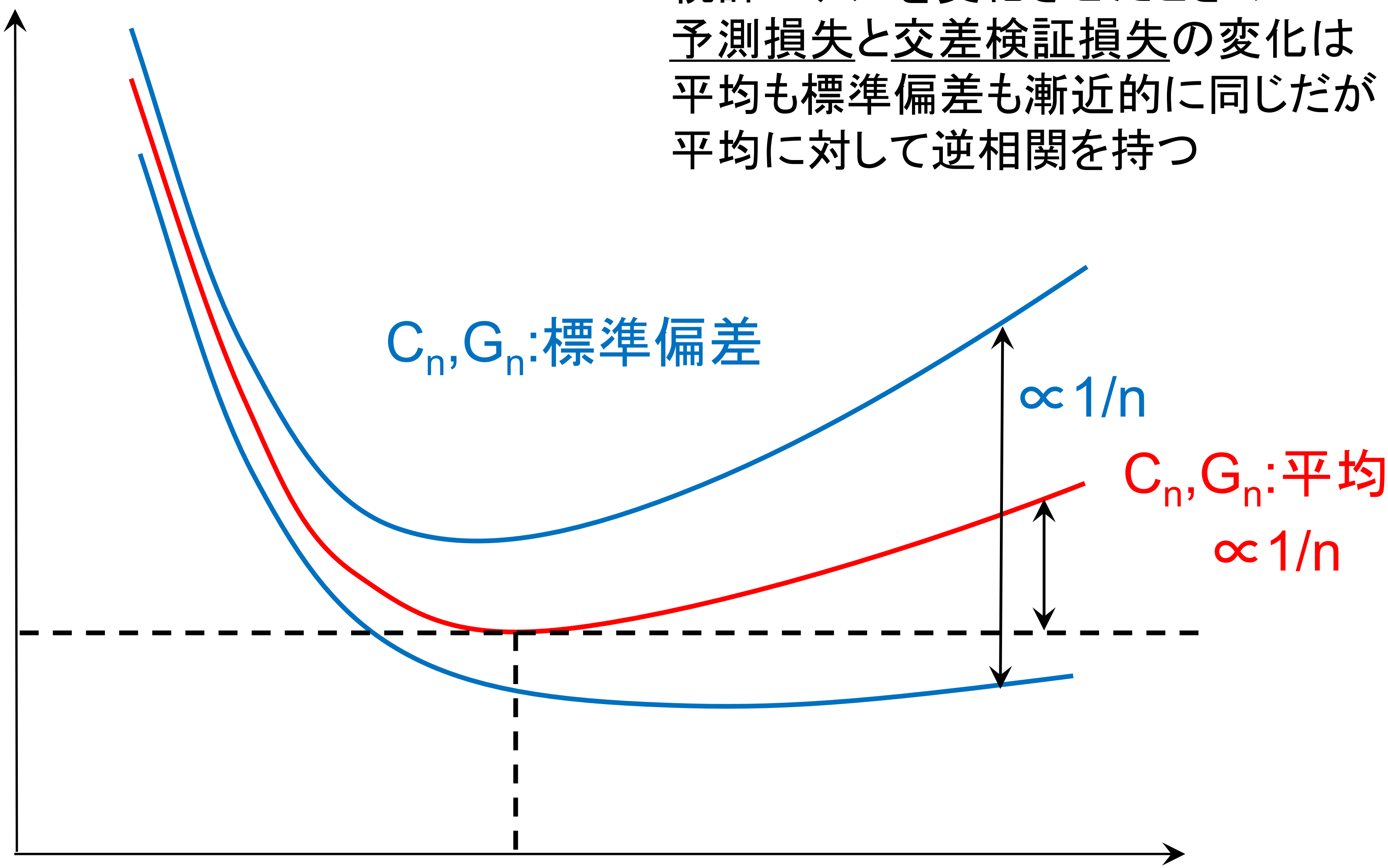
真がモデルで実現できなくても正則でなくても

$$\mathbf{E}[G_n]=\mathbf{E}[\mathrm{WAIC}_n] + o(1/n).$$

WAIC はベイズ法（フルベイズ）の汎化損失を推測する規準である。一個の正規分布などの簡単なケースでは ISCV とほとんど同じ値になるが、複雑な問題になると MCMC揺らぎによる分散が ISCVよりも小さいことが多い．影響力の大きなデータがあっても発散しない．
条件つき独立な場合にも適用できる。ISCVと両方を計算することを推奨する。両者の値が同じになっていないときは、影響力の大きなデータが存在して事後分布が不安定になっている可能性が高い。

G_n:予測誤差
C_n:交差検証損失

(正則でも特異でも)
統計モデルを変化させたときの
予測損失と交差検証損失の変化は
平均も標準偏差も漸近的に同じだが
平均に対して逆相関を持つ

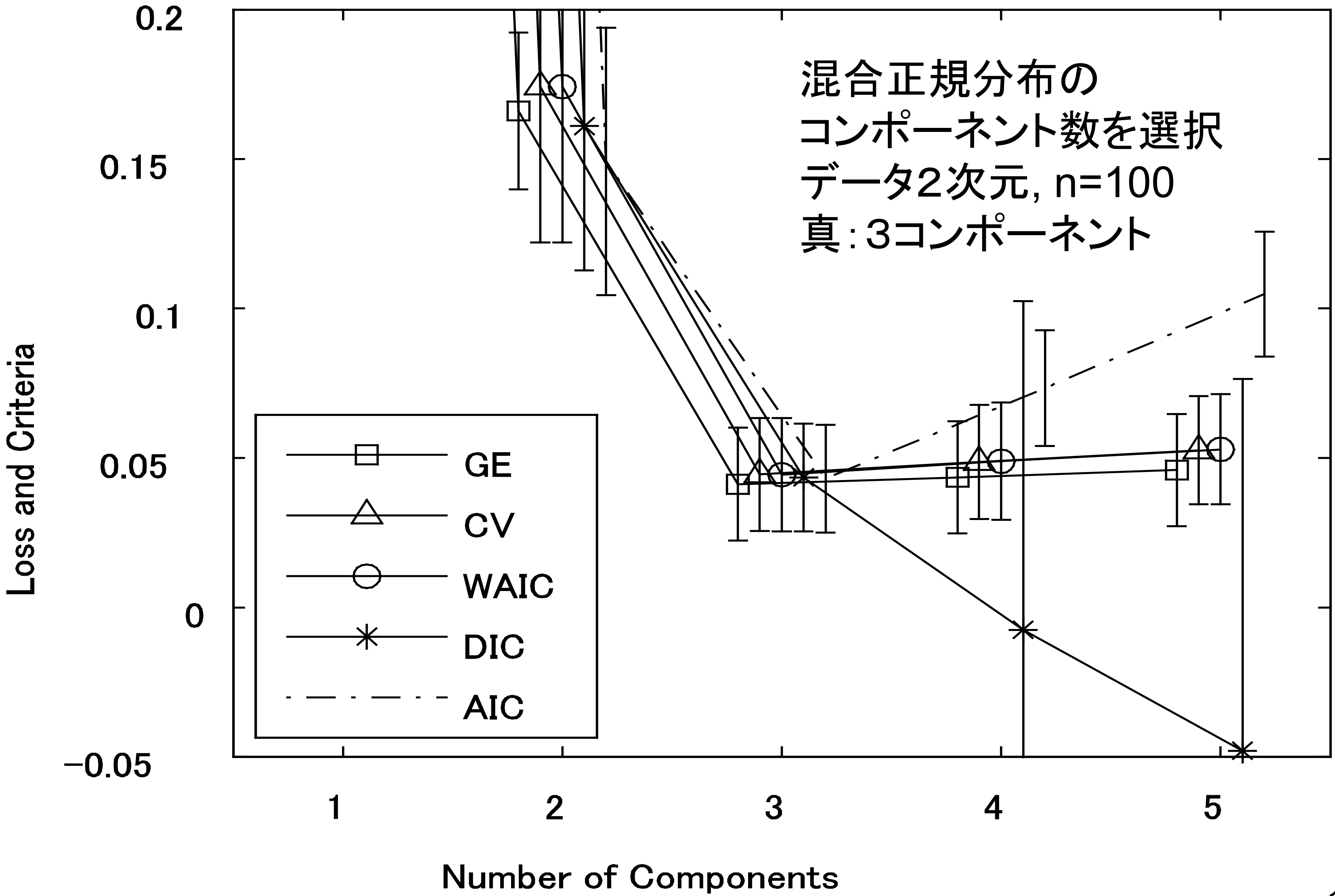


取り扱い説明書

	演算量	数値計算	漸近条件	実現可能かつ正則	データ独立性
CV	n				必要
ISCV	1	発散することがある			必要
AIC	1		必要	必要	
DIC	1		必要	強く必要	
WAIC	1		必要		

DICが使える場合にはAICが使えます。AIC,DICは事前分布の評価には使えません。
CV 以外はほぼ同じ演算量で計算できるので、全部算出してみることを推奨します。

混合正規分布のモデル選択

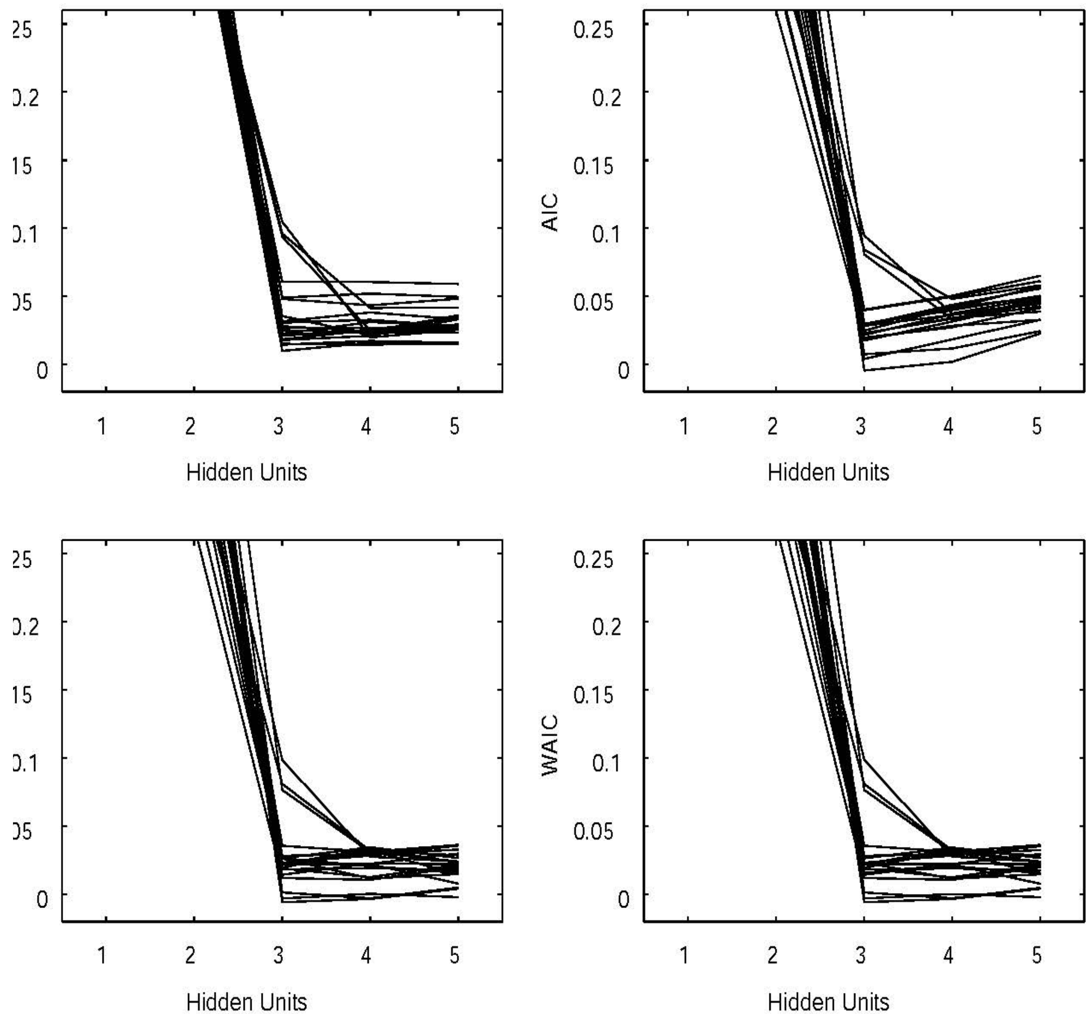


神経回路網のモデル選択

入力2出力1の神経回路網のモデル選択を行なった
n=200. 真のモデルはH=3.

神経回路網のようなモデルでは事後分布が正確にできないことがあり、真のモデルが選ばれていない例は、それが主な原因である。

ベイズ法を適用すると真の分布よりも複雑なモデルを用いても汎化損失はあまり大きくならないが、その分だけ真の分布があたりにくくなる。



モデル選択はいつでも汎化損失を小さくするわけではない

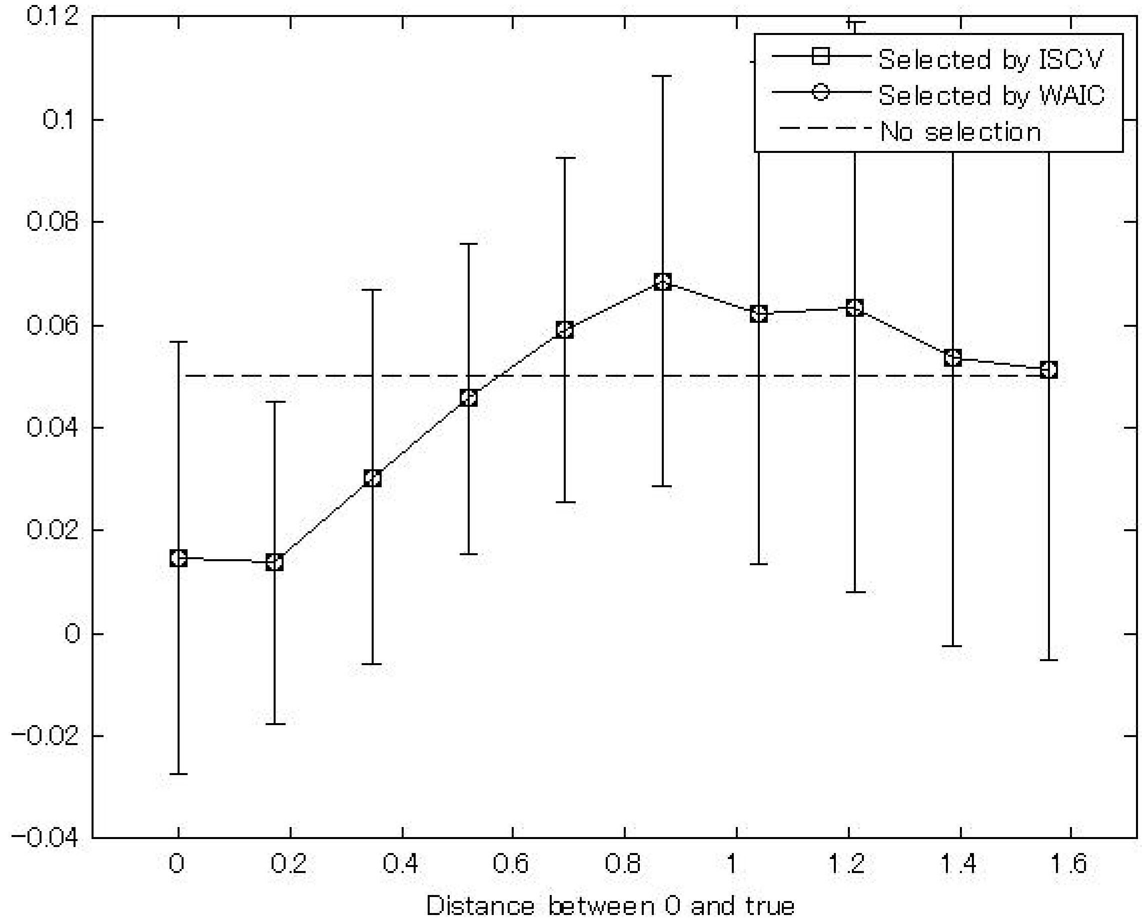
３次元のデータに対して二つの正規分布を比較する。d=3, n=30.

次のいずれかのモデルを選ぶ。
モデル0 : $N_3(0,1)$
モデル1: N3(a,1)

真: $N(a_0,1)$

ISCV, WAIC を小さくするほうを選ぶという条件での汎化損失を調べた。

汎化損失ーエントロピー



原点と真の分布の平均の距離

2　自由エネルギー

自由エネルギーと調和平均の問題

自由エネルギー最小化（周辺尤度の最大化）でモデルや事前分布の最適化を行なう事は I.J. Good により提案された。周辺尤度は、データが与えられたときの（統計モデル、事前分布）の尤度になっている。

自由エネルギ: $p(X^n)$ と $q(X^n)$ を較べる
汎化損失 : $p(x|X^n)$ と $q(x)$ を較べる

（注）統計モデルと事前分布の二つの組が与えられたとき、一方を帰無仮説とし他方を対立仮説として検定を行なうことができますが、自由エネルギーの差を用いて行なう検定が最強検定になることが知られています。

自由エネルギーは計算上の問題がある。事後分布にしたがうパラメータのサンプル $\{w_k\}$ が得られてもそれだけでは自由エネルギーは計算できません。

$$F = \log E_w[\ \exp(nL_n(w))\ ]$$

は調和平均と呼ばれ、数式としては正しいのですが、数値計算に適していないことが知られています。

自由エネルギーの数値計算法

自由エネルギーを数値的に計算したいときの方法 (Ogata, 1990)

$$F(\beta) = -\log \int \exp(- \beta\, nL_n(w))\, \varphi(w)\, dw$$

とおいて F(1) を求めることが目標。F(0)=0 なので

$$F(1) = \int_0^1 (dF/d\beta)\, d\beta$$

$$(dF/d\beta) = \frac{\int nL_n(w) \exp(- \beta\, nL_n(w))\, \varphi(w)\, dw}{\int \exp(- \beta\, nL_n(w))\, \varphi(w)\, dw}$$

そこで βを細かく刻んで、各βごとに事後分布をつくり、(dF/dβ)を求めてから総和をとる。

中間値の定理

自由エネルギー F(1) は (dF/dβ) の積分で表される

$$F(1) = \int_0^1 (dF/d\beta)(\beta)\, d\beta$$

ので、積分の中間値の定理からある β* が存在して

$$F(1) = \frac{\int nL_n(w) \exp(-\beta^* nL_n(w))\, \varphi(w)\, dw}{\int \exp(-\beta^* nL_n(w))\, \varphi(w)\, dw}$$

正則であってもなくても β* = 1/log n + o(1/log n) であることが示せる。

WBICと自由エネルギー

真の分布がわからない場合に自由エネルギーの近似値を求める方法として次の方法もある。

$$E_w^{1/\log n}[\quad] = \frac{\int (\quad) \prod p(X_i|w)^{1/\log n} \varphi(w)\, dw}{\int \prod p(X_i|w)^{1/\log n} \varphi(w)\, dw}$$

$\mathrm{WBIC} = E_w^{1/\log n}[\, -\Sigma \log p(X_i|w) \,]$ とおくと

$$\mathrm{WBIC} = n L_n + \lambda \log n + O_p((\log n)^{1/2}).$$

sBIC （singular BIC)

Drton-Plummer (1997) は複数のモデル族に対して実対数閾値を当てはめることにより、自由エネルギーを推測する方法を提案した。

$$\mathrm{sBIC} = n\,L_n(w^*) + \lambda \log n$$

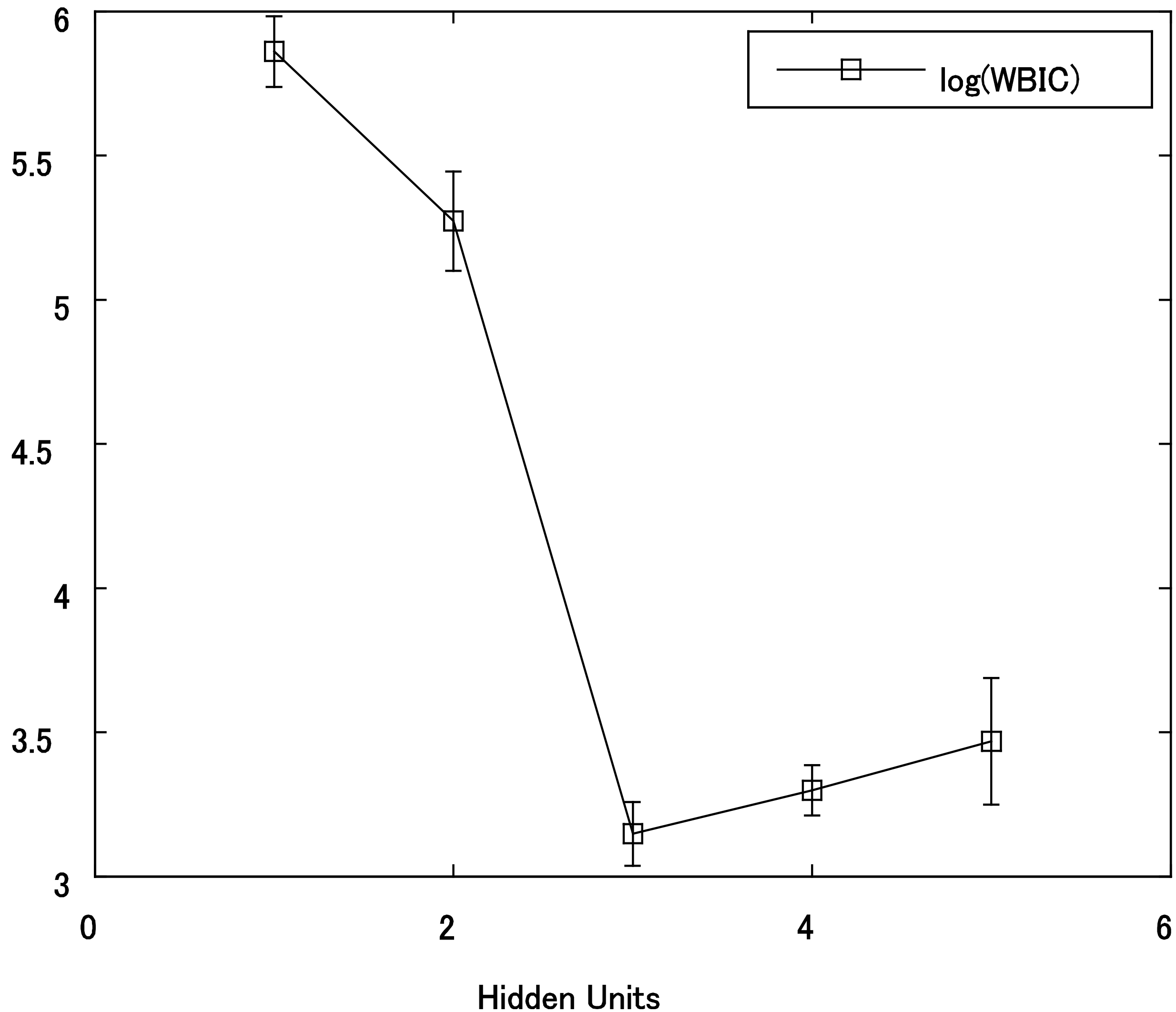
実対数閾値の理論研究の成果を活用できるため、揺らぎが小さいという利点がある。

取り扱い説明書

	演算量	漸近 条件	実現可能 かつ正則	実対数 閾値
F	逆温度数×MCMC			
BIC	1	必要	必要	
sBIC	1	必要		必要
WBIC	MCMC1回	必要		

実験例

log(自由エネルギーの推定値) = (log(WBIC))



3層神経回路網
入力ユニット2、
出力1
真の分布
中間ユニット数3

WBIC を計算して
その対数を表示
した。

F_n:自由エネルギー

(正則でも特異でも)
統計モデルを変化させたときの
自由エネルギーの変化は
平均が log n であるのに対して
標準偏差は1である。