クレジット:
UTokyo Online Education 数理手法Ⅶ 2019 北川源四郎

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

時系列解析（３）

東京大学 数理・情報教育研究センター
北川　源四郎

概 要

- 統計的モデリングとモデル評価
- 予測の視点とK-L情報量
- K-L情報量の推定と最尤法
- バイアス補正と情報量規準
- AICとTICの関係
- AICによるモデル選択例
- その他の情報量規準

統計的モデリング

モデルを通して情報抽出が実現できる

統計的モデル

統計的モデルは確率分布で表現する

基本的な形

- 連続型確率分布
  - 正規分布
  - コーシー分布
  - ピアソン分布族
  - 指数分布
  - 二重指数分布
  - χ^2分布
  - 一様分布
- 離散型確率分布
  - 二項分布
  - 多項分布
  - ポアソン分布

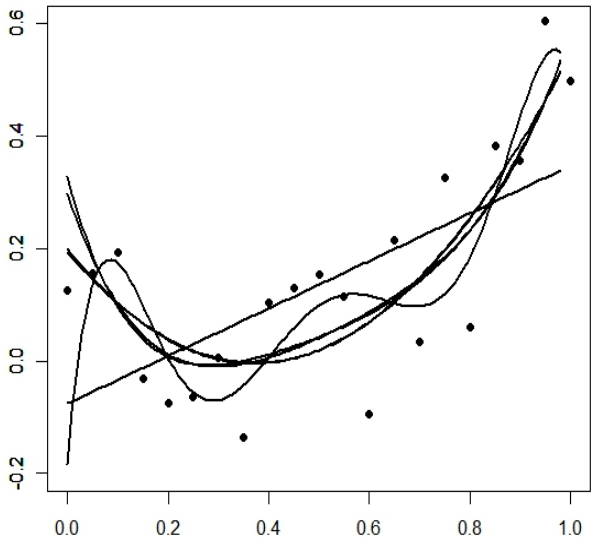
実際の統計モデルは様々な情報を取り入れて構築される

- 他の変数の情報
- 過去の変動の情報
- 時間経過による変化

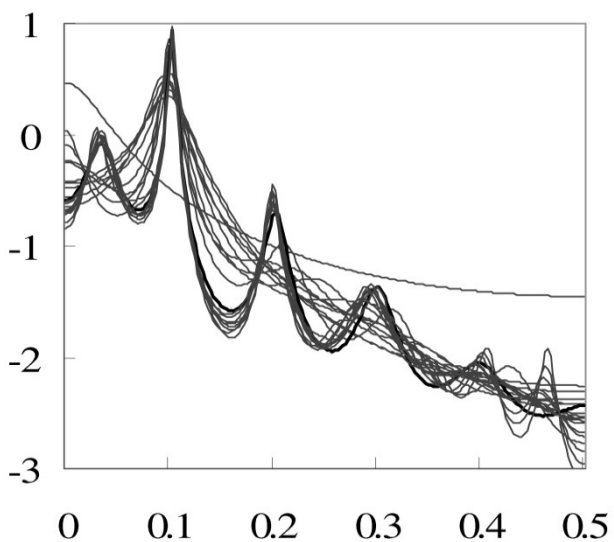
これらの情報の有効活用がカギ

利用するモデルの影響

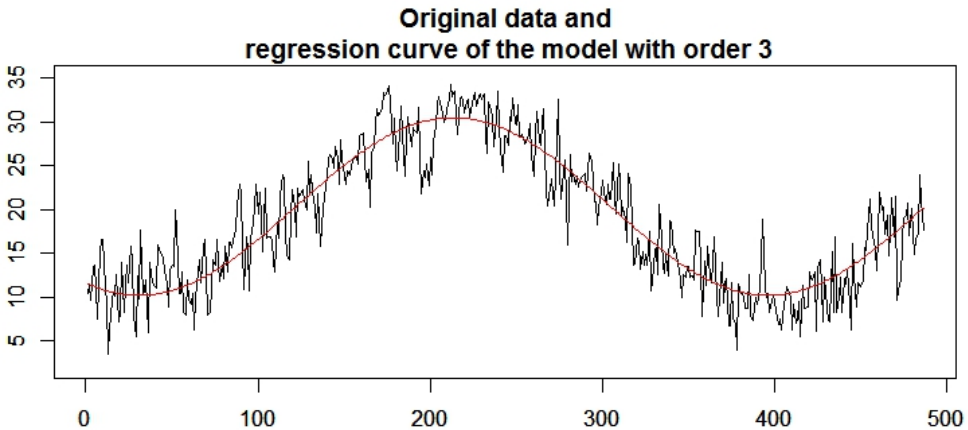
- 多項式回帰モデル

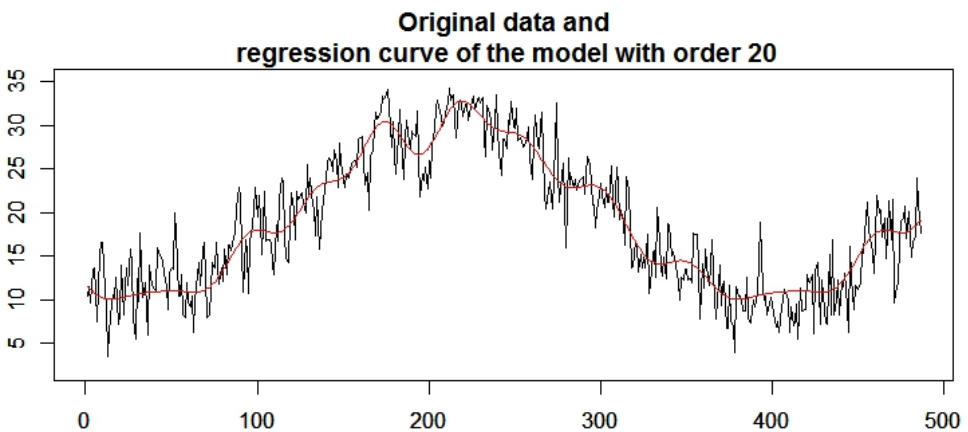


- 時系列モデルによるスペクトル推定



- 時系列モデルによるトレンド推定





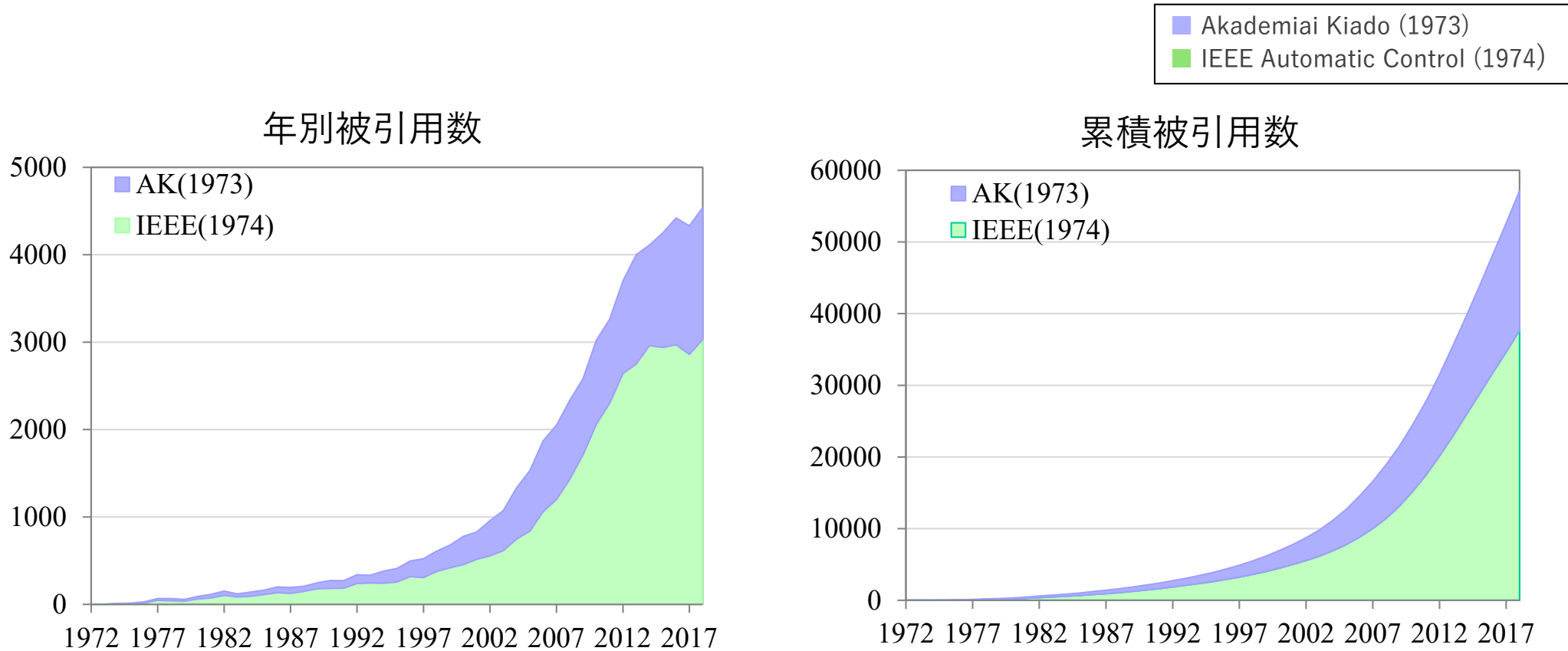
利用するモデルによって，予測や情報抽出の結果は著しく異なる．

モデル評価の重要性

- モデルの利用によって，予測や情報抽出ができる
- 統計的推論の結果は利用するモデルに依存する
- モデル評価・選択が重要

➡ モデル評価のための情報量規準

情報量規準：論文被引用回数



- 赤池弘次氏　Googleのトップロゴになる 11/6/2017

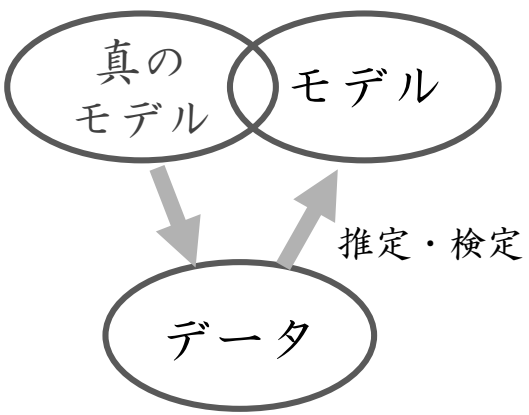
  Google Doodle

  https://www.google.com/doodles/hirotugu-akaikes-90th-birthday

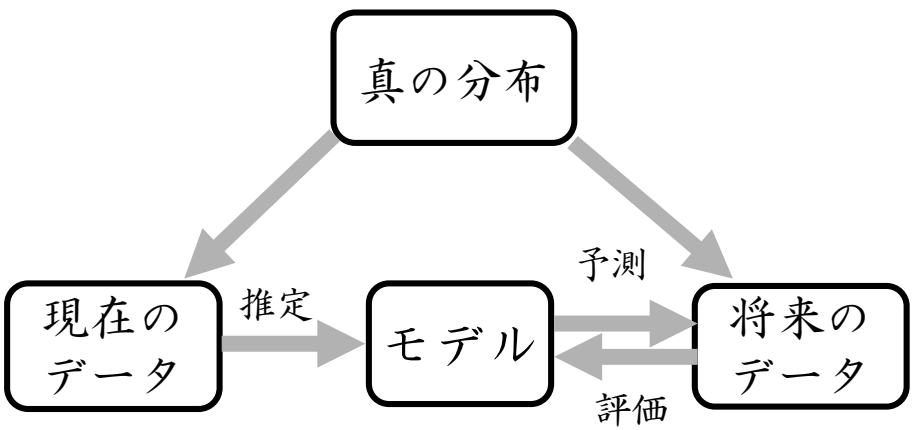
情報量規準への道

- モデルのよさを予測能力で評価する
- 予測は点推定ではなく予測分布で行う
- 分布の近さをK-L情報量で評価する

従来の視点



予測の視点



K-L情報量によるモデルの評価

モデル　$g(y)$：　真の分布

$f(y)$：　モデルの分布

Kullback-Leibler情報量（K-L ダイバージェンスともいう）

真の分布とモデルの分布の乖離を測る尺度

$$I(g;f) = E_Y \log\left\{\frac{g(Y)}{f(Y)}\right\} = \int \log\left\{\frac{g(x)}{f(x)}\right\} dG(x)$$

$$= \begin{cases} \displaystyle\int \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx & \text{連続分布モデル} \\ \displaystyle\sum_{i\in J} g_i \log\left\{\frac{g_i}{f_i}\right\} & \text{離散分布モデル} \end{cases}$$

Akaike(1973,1974), 小西・北川 3.1

K-L情報量の性質

$$(i)\ \ I(g;f) \geq 0$$

$$(ii)\ I(g;f) = 0 \ \Leftrightarrow\ g(x) = f(x)$$

（注意）K-L情報量は距離ではない．距離の公理の(2)対称性，(3)三角不等式を満たさない．

（距離の公理）

$$(1)\quad d(g,f) \geq 0, \qquad d(g,f) = 0 \ \Leftrightarrow\ g(x) = f(x)$$

$$(2)\ \ d(g,f) = d(f,g)$$

$$(3)\ \ d(f,g) + d(g,h) \geq d(f,h)$$

その他の尺度

- ヘリンジャー距離 $\int\left\{\sqrt{f(x)}-\sqrt{g(x)}\right\}^2 dx$
- 一般化情報量 $\frac{1}{\lambda}\int\left\{\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}g(x)dx$
- ダイバージェンス $\int u\left(\frac{g(x)}{f(x)}\right)g(x)dx$
- L_1ノルム $\int\left|g(x)-f(x)\right|dx$
- L_2ノルム $\int\left(g(x)-f(x)\right)^2 dx$

K-L情報量とエントロピー

Boltzmannのエントロピー

$$B(g;f) = -I(g;f) = \sum g_i \log \frac{g_i}{f_i}$$

モデル	$f=(f_1,\ldots,f_k)$
n個の独立な観測値	$(n_1,\ldots,n_k)$ $n_1+\cdots+n_k=n$
相対度数	$(g_1,\ldots,g_k)$
$g_i=n_i/n$	$(n_1,\ldots,n_k)$が得られる確率

$$B(g;f) \sim \frac{1}{n}\log W$$

W: 想定したモデルから得られたサンプルの相対度数が真の分布と一致する確率

$$W = \frac{n!}{n_1!\cdots n_k!} f_1^{n_1}\cdots f_k^{n_k}$$

スターリングの近似

$$\log n! \sim n\log n - n$$

$$\log W! = \log n! - \sum_{i=1}^{k}\log n_i! + \sum_{i=1}^{k} n_i \log f_i$$

$$\sim n\log n - n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i \log f_i$$

$$= -\sum_{i=1}^{k} n_i \log \frac{n_i}{n} + \sum_{i=1}^{k} n_i \log f_i$$

$$= \sum_{i=1}^{k} n_i \log \frac{f_i}{g_i}$$

$$= n\sum_{i=1}^{k} g_i \log \frac{f_i}{g_i} = nB(g;f)$$

小西・北川(2004) 30p

K-L情報量：計算例

正規分布

$$g(y) \sim N(\mu, \sigma^2), \qquad f(y) \sim N(\xi, \tau^2)$$

$$I(g;f) = \int_{-\infty}^{\infty} g(x) \log g(x) dx - \int_{-\infty}^{\infty} g(x) \log f(x) dx$$

$$\int_{-\infty}^{\infty} g(x) \log f(x) dx = E_X\left[-\frac{1}{2}\log 2\pi\tau^2 - \frac{(X-\xi)^2}{2\tau^2} \right]$$

$$= -\frac{1}{2}\log 2\pi\tau^2 - \frac{\sigma^2 + (\mu-\xi)^2}{2\tau^2}$$

$$\int_{-\infty}^{\infty} g(x) \log g(x) dx = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}$$

$$I(g;f) = \frac{1}{2}\left\{ \log\frac{\tau^2}{\sigma^2} + \frac{\sigma^2 + (\mu-\xi)^2}{\tau^2} - 1 \right\}$$

多項分布

$$f_1 = \{0.20, 0.12, 0.18, 0.12, 0.20, 0.18\}$$
$$f_2 = \{0.18, 0.12, 0.14, 0.19, 0.22, 0.15\}$$
$$g = \{\ 1/6,\ 1/6,\ 1/6,\ 1/6,\ 1/6,\ 1/6\}$$

$$I(g;f) = E_Y \log\frac{g}{f} = \sum_{i=1}^{6} g_i \log\left(\frac{g_i}{f_i}\right)$$

$I(g;f_1)$=0.023, $I(g;f_2)$= 0.020

応用：比例代表制における議席配分

g：得票分布，f：議席配分

小西・北川(2004) 30p

K-L情報量の推定

統計的モデリングでは通常，K-L情報量は直接計算できない

- 理由：　真のモデル $g(y)$ は未知
- 対策：　$I(g(y); f(y))$ をデータから推定する

$$I(g; f) = E_Y \log \frac{g}{f} = E_Y \log g - E_Y \log f$$

$E_Y \log g$ と $E_Y \log f$ を分離できることがK-L情報量のメリット

平均対数尤度

$$I(g;f) = E_Y \log g - E_Y \log f$$

$E_Y \log f(Y)$　　平均対数尤度

$E_Y \log g$ は未知だが f に関係なく一定

➡ $E_Y \log f(Y)$：　大 $\Leftrightarrow I(g:f)$：小

モデリングにはK-L情報量の代わりに平均対数尤度を使える

注意：　$I(g:f)$　　絶対評価

$E_Y \log f(Y)$　相対評価

平均対数尤度 $E \log f$ の推定

平均対数尤度も未知の分布を含む ➡ 推定が必要

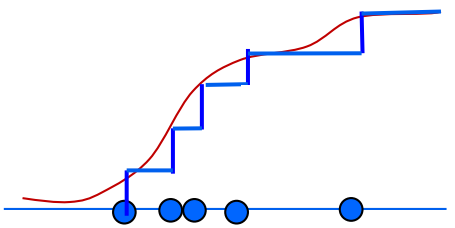
平均対数尤度

$$E_Y \log f(Y) = \int \log f(y) dG(y) \qquad dG(y) = g(y)dy$$

Data

$$G(y) \Rightarrow \hat{G}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(y, X_i)$$

(経験分布関数)



対数尤度

$$\ell = n\int \log f(y) d\hat{G}_n(y) = \sum_{i=1}^{n} \log f(X_i)$$

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) \Rightarrow E_Y \log f(Y) \qquad \text{大数の法則}$$

最尤法

パラメトリックモデル

$$f(y\,|\,\theta),\quad \theta \equiv (\theta_1,\ldots,\theta_k)'$$

対数尤度　$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i\,|\,\theta) \equiv \log f(X\,|\,\theta)$

$$I(g;f)\text{: Min} \quad \Leftrightarrow \quad E_Y \log f(Y)\text{:Max} \quad \cong \quad \ell\text{:Max}$$

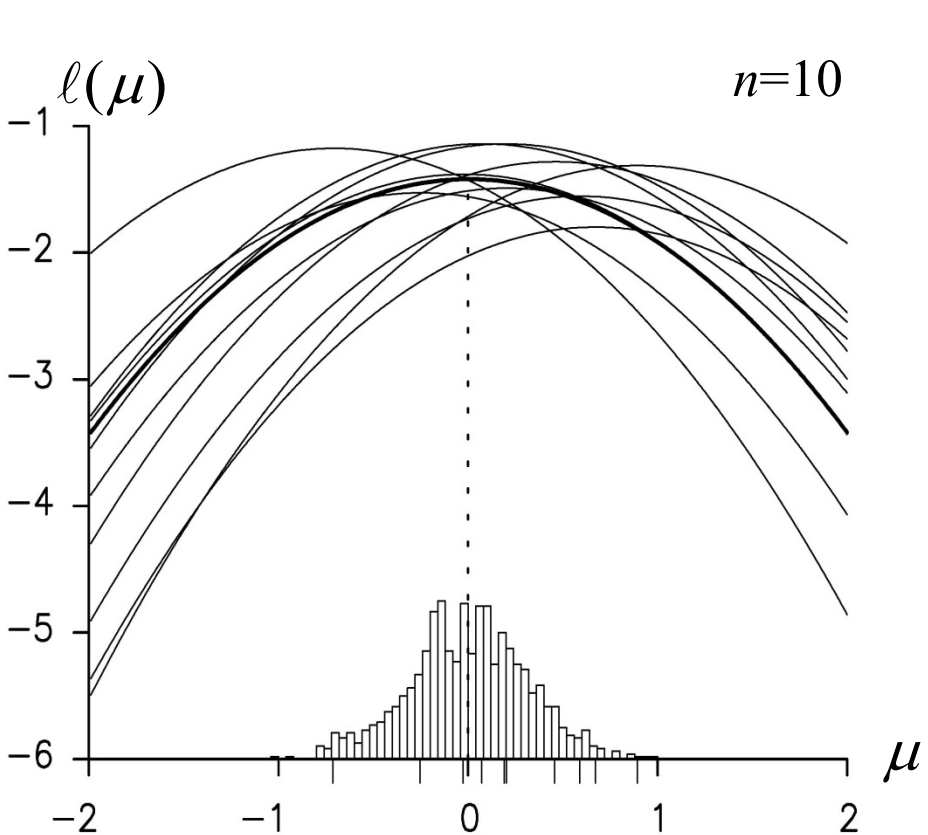
最尤法　$\max_{\theta} \ell(\theta)$ ⇒ $\hat{\theta} = \hat{\theta}(X)$

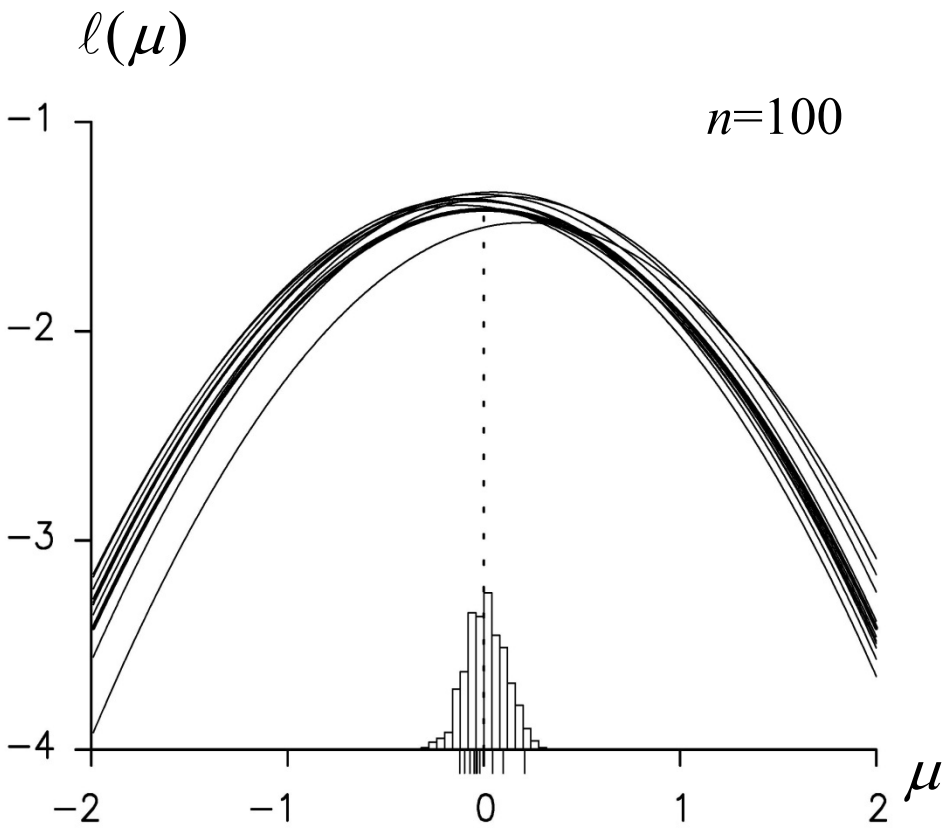
$\hat{\theta}$	最尤推定量
$\ell(\hat{\theta})$	最大対数尤度
$f(y\,|\,\hat{\theta})$	最尤モデル

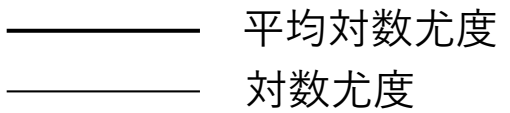
最尤法は近似的にK-L情報量を最小化

最尤推定値の例(平均)

$$y \sim N(0,1), \qquad \text{model}: N(\mu,1)$$

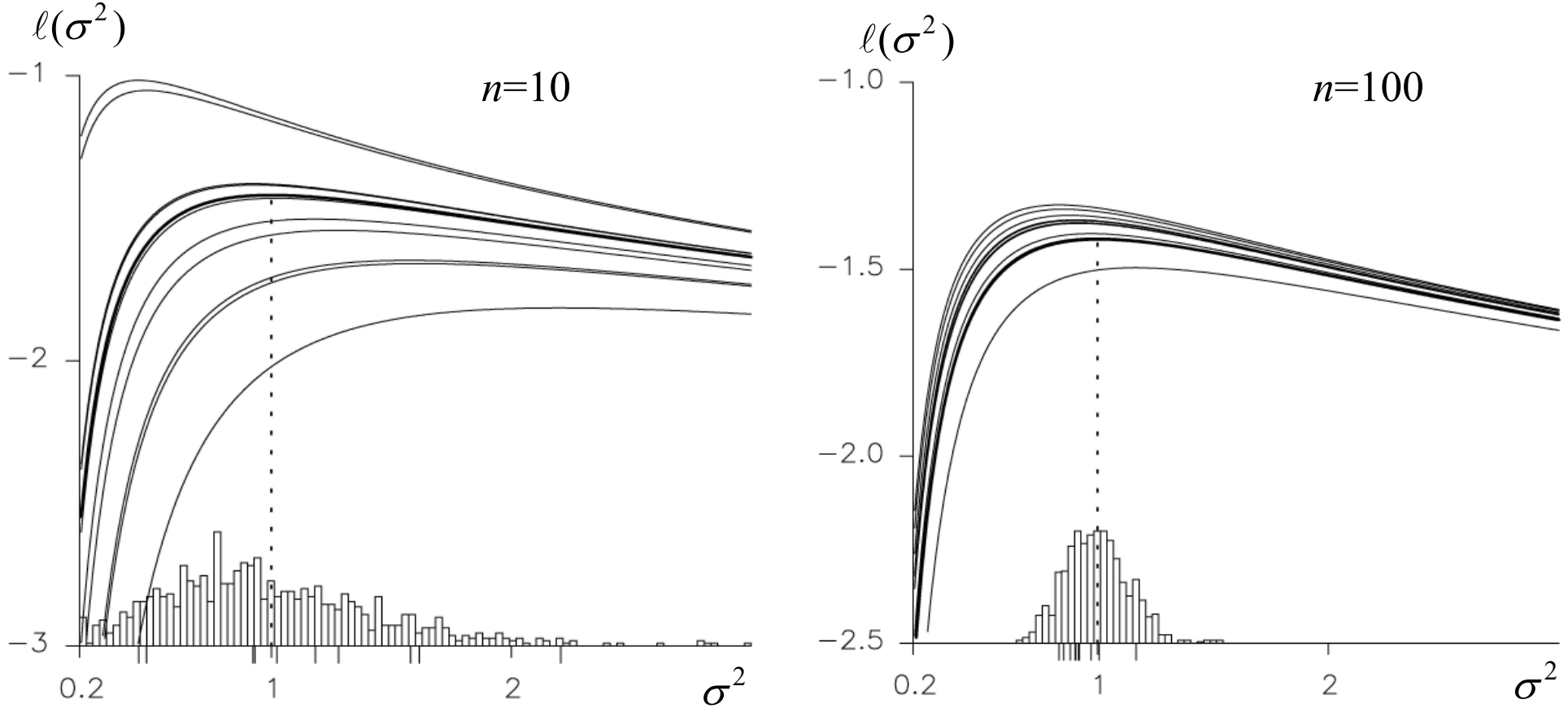






最尤推定値の例(分散)

$$y \sim N(0,1), \quad \text{model}: N(0,\sigma^2)$$



最尤推定値

最尤推定値の求め方

（１）尤度方程式を解く

（２）数値的最適化による

$$\max_{\theta} \ell(\theta) = \ell(\hat{\theta})$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

最尤推定量の性質

（１）尤度方程式 $\frac{\partial \ell(\theta)}{\partial \theta} = 0$ はθ_0に収束する解を持つ

（２）$\hat{\theta}_n$ は $n \to +\infty$ のときθ_0に確率収束

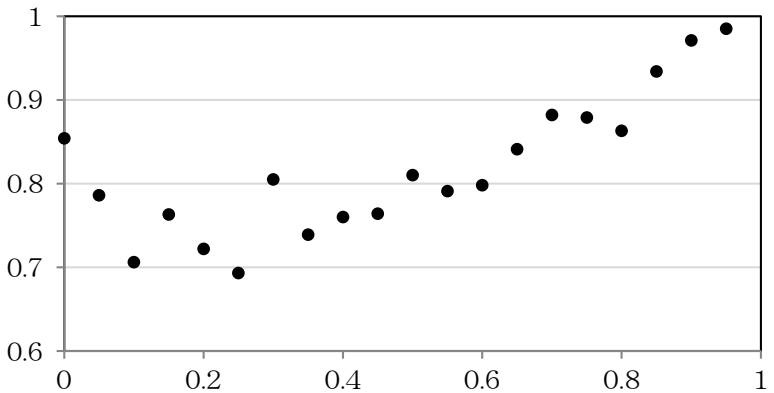
（３）$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, I(\theta_0)^{-1})$

複数モデルの比較

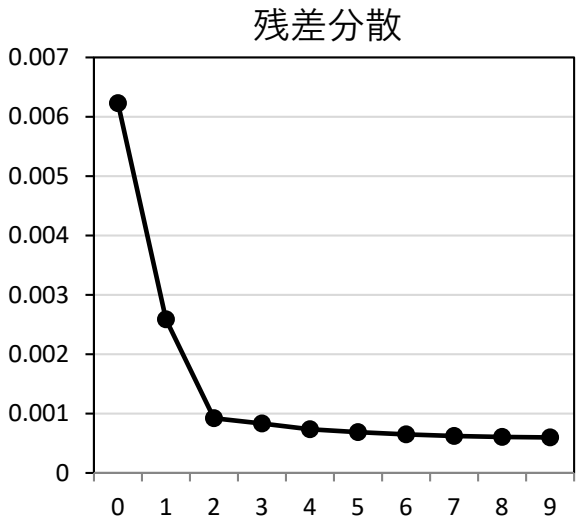
M_1	…	M_k	モデル
θ_1	…	θ_k	パラメータ
$\ell_1(\hat{\theta}_1)$	…	$\ell_k(\hat{\theta}_k)$	最大対数尤度

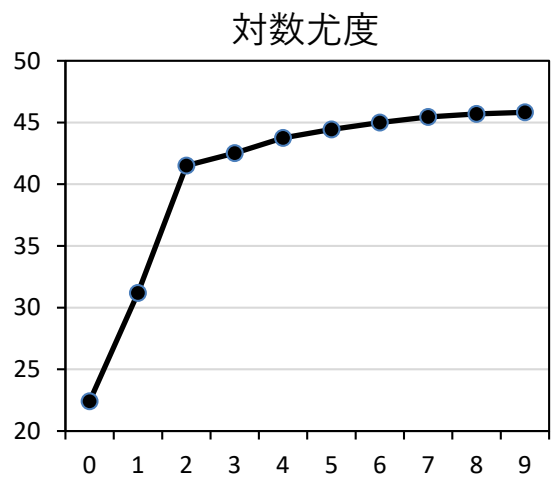
最大対数尤度 $\ell_j(\hat{\theta}_j)$ を比較して，最大となる j を探せばよい？

多項式回帰の次数と残差分散，対数尤度



次数	残差分散	対数尤度
–	0.678301	-24.50
0	0.006229	22.41
1	0.002587	31.19
2	0.000922	41.51
3	0.000833	42.52
4	0.000737	43.75
5	0.000688	44.44
6	0.000650	45.00
7	0.000622	45.45
8	0.000607	45.69
9	0.000599	45.83





$\ell_j(\hat{\theta}_j)$ はそのままではモデル選択に使えない

理由・原因と対策

理由： $\ell(\hat{\theta})$ が $E\log f(x|\hat{\theta})$ の推定値としてバイアスを持ち
しかも，バイアス量がモデルによって異なる

原因： パラメータ推定と平均対数尤度の推定に同じデータ
を2回用いたため

対策： バイアスを評価し補正する

記号と準備

$g(x)$ 真のモデル　　$f(x|\theta)$ パラメトリックモデル

θ_0 「真」の値　　$E_Y \log f(Y|\theta_0) = \max E_Y \log f(Y|\theta)$

$$\Rightarrow \frac{\partial}{\partial\theta} E_Y \log f(Y|\theta_0) = 0$$

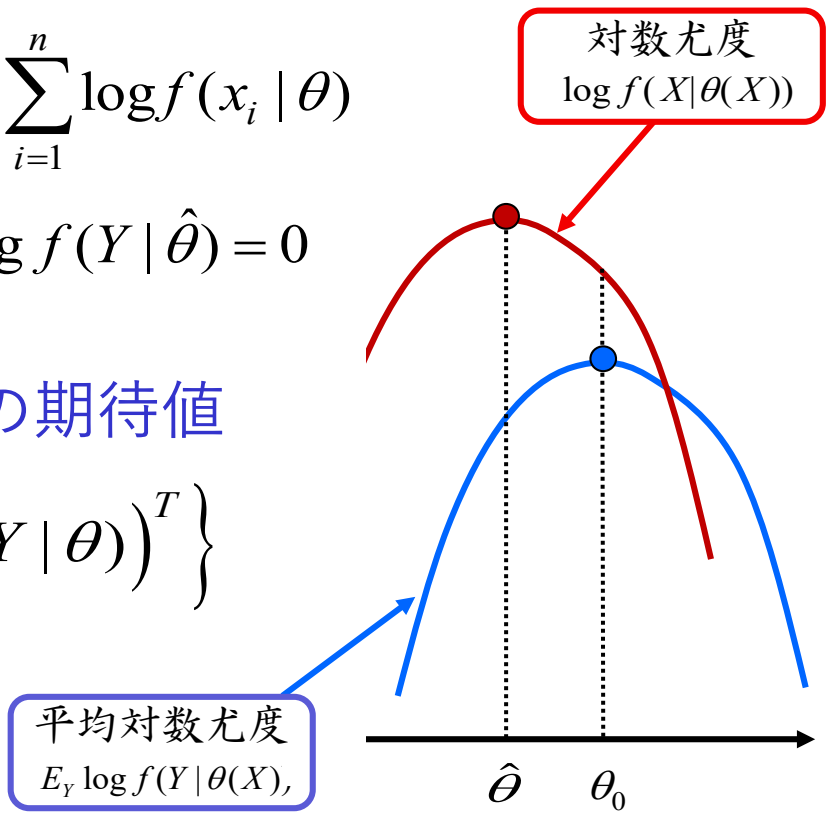
$\hat{\theta}$ 最尤推定値　　$\sum_{i=1}^{n} \log f(x_i|\hat{\theta}) = \max \sum_{i=1}^{n} \log f(x_i|\theta)$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y|\hat{\theta}) = 0$$

$I(\theta)$：Fisher情報行列，$J(\theta)$：Hessianの期待値

$$I(\theta) \equiv E\left\{\left(\frac{\partial}{\partial\theta}\log f(Y|\theta)\right)\left(\frac{\partial}{\partial\theta}\log f(Y|\theta)\right)^T\right\}$$

$$J(\theta) = -E\left\{\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(Y|\theta)\right\}$$



記号と準備

標準的中心極限定理

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, I(\theta_0)^{-1})$$

$g(x) = f(x \mid \theta_0)$ となる $\theta_0 \in \Theta$ が存在しない場合でも

$$\hat{\theta}_n \to \theta_0$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, J(\theta_0)^{-1} I(\theta_0) J(\theta_0)^{-1})$$

$$\begin{aligned} E\left[(\hat{\theta}-\theta_0)^T J(\theta_0)(\hat{\theta}-\theta_0)\right] &= \mathrm{tr}\left\{J(\theta_0) E\left[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T\right]\right\} \\ &= \mathrm{tr}\left\{J(\theta_0)\frac{1}{n} J(\theta_0)^{-1} I(\theta_0) J(\theta_0)^{-1}\right\} \\ &= \frac{1}{n}\mathrm{tr}\left\{I(\theta_0) J(\theta_0)^{-1}\right\} \end{aligned}$$

最大対数尤度のバイアス補正

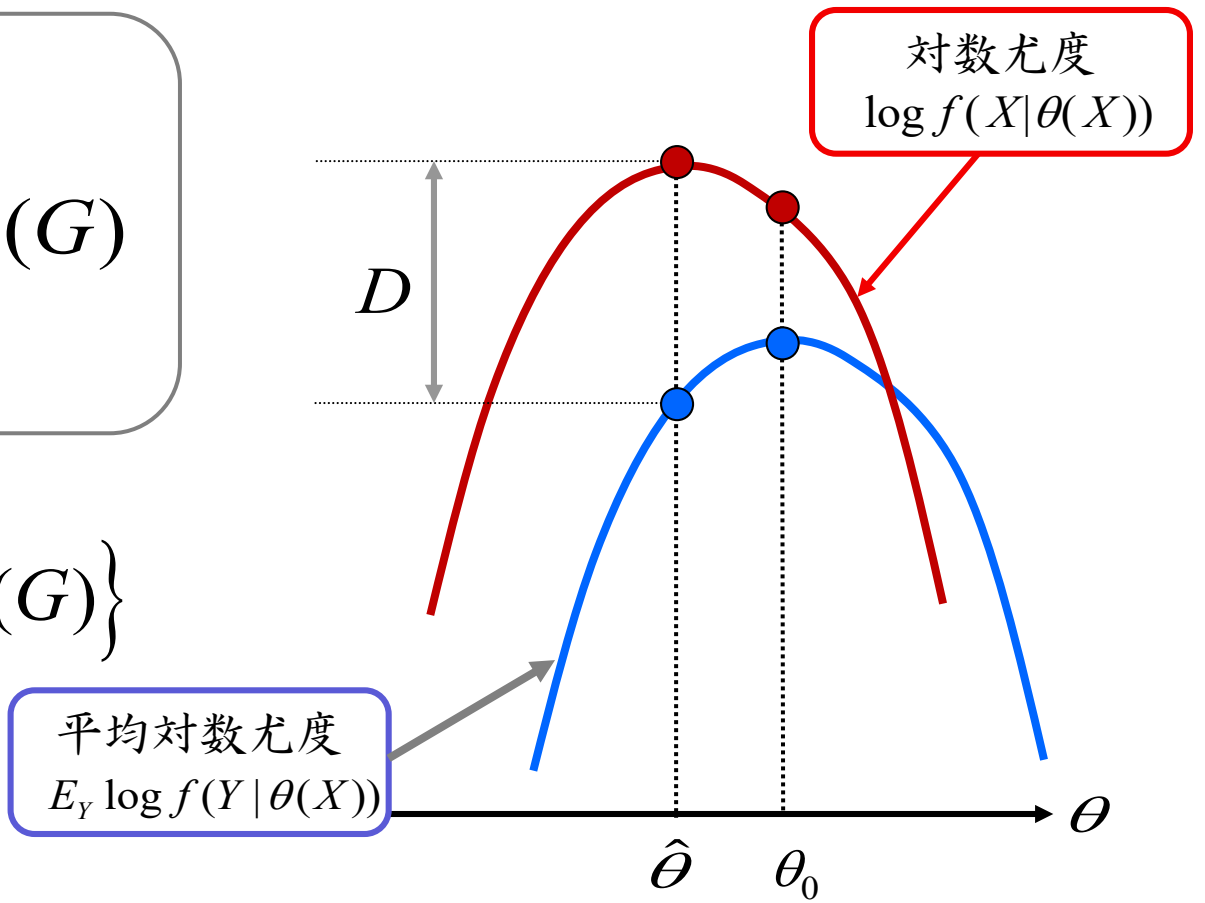
$$b(G) = E_X\{D\} = E_X\left\{ \frac{1}{n}\log f(X \mid \hat{\theta}(X)) - E_Y \log f(Y \mid \hat{\theta}(X)) \right\}$$

バイアス補正

$$\log f(X \mid \hat{\theta}(X)) - nb(G)$$

（Dの期待値を補正）

$$E_X\left\{\log f(X \mid \hat{\theta}(X)) - nb(G)\right\}$$
$$= E_Y \log f(Y \mid \hat{\theta}(X))$$

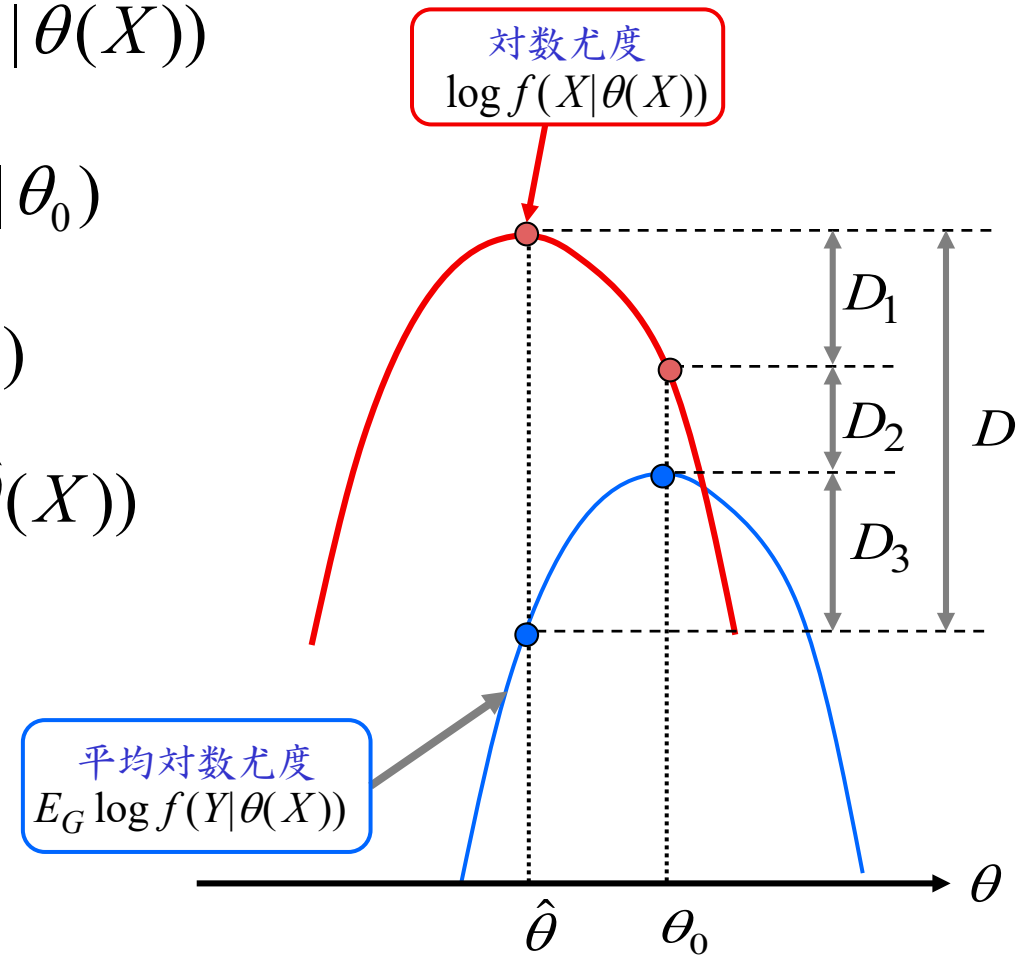


バイアスの評価

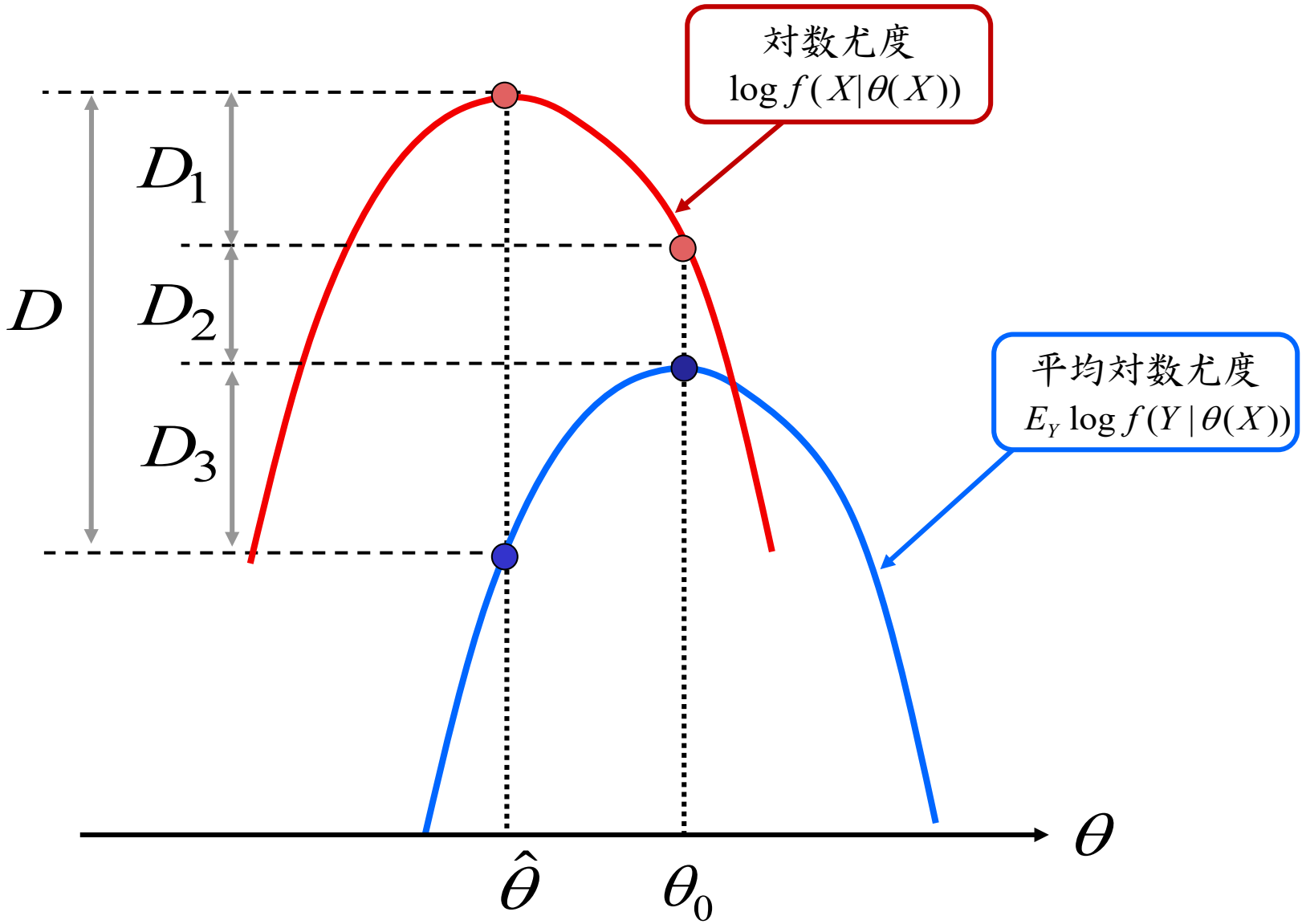
$$X = (x_1, \ldots, x_n)^T$$
$$\log f(X \mid \theta) \equiv \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

対数尤度　　平均対数尤度

$$D = \frac{1}{n} \log f(X \mid \hat{\theta}(X)) - E_Y \log f(Y \mid \hat{\theta}(X))$$
$$= \frac{1}{n} \log f(X \mid \hat{\theta}(X)) - \frac{1}{n} \log f(X \mid \theta_0)$$
$$+ \frac{1}{n} \log f(X \mid \theta_0) - E_Y \log f(Y \mid \theta_0)$$
$$+ E_Y \log f(Y \mid \theta_0) - E_Y \log f(Y \mid \hat{\theta}(X))$$
$$= D_1 + D_2 + D_3$$



バイアスの構造


対数尤度
$\log f(X|\theta(X))$
平均対数尤度
$E_Y \log f(Y|\theta(X))$
D
D_1
D_2
D_3
θ
$\hat{\theta}$
θ_0

バイアスの評価

D の期待値を計算する

$$E[D] = E[D_1] + E[D_2] + E[D_3]$$

D_3 の評価

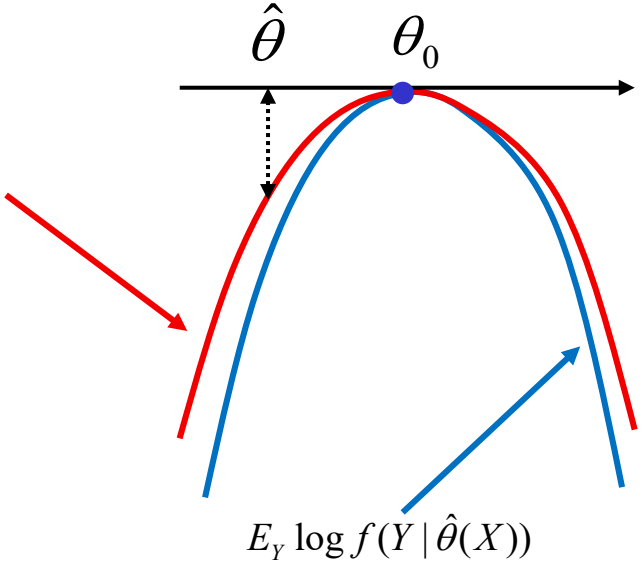
$$E_Y \log f(Y \mid \hat{\theta}(X))$$

$$\approx E_Y \log f(Y \mid \theta_0) + \underbrace{\frac{\partial}{\partial \theta} E_Y \log f(Y \mid \theta_0)}_{=0} (\hat{\theta} - \theta_0)$$

$$+ \frac{1}{2} (\hat{\theta} - \theta_0)^T \boxed{E_Y \frac{\partial^2}{\partial \theta \partial \theta} \log f(Y \mid \theta_0)} (\hat{\theta} - \theta_0)$$

$$= E_Y \log f(Y \mid \theta_0) - \frac{1}{2} (\hat{\theta} - \theta_0)^T \boxed{J(\theta_0)} (\hat{\theta} - \theta_0)$$

$$E_X \left\{ (\hat{\theta} - \theta_0)^T J(\theta_0) (\hat{\theta} - \theta_0) \right\} = \frac{1}{n} \mathrm{tr} \left\{ I(\theta_0) J^{-1}(\theta_0) \right\}$$



$$E_X \left\{ D_3 \right\} = E_X \left\{ E_Y \log f(Y \mid \theta_0) - E_Y \log f(Y \mid \hat{\theta}(X)) \right\} \approx \frac{1}{2n} \mathrm{tr} \left\{ I J^{-1} \right\}$$

D_1 の評価

$$\log f(X \mid \theta_0)$$

$$\approx \log f(X \mid \hat{\theta}(X)) + \frac{\partial}{\partial \theta} \log f(X \mid \hat{\theta}(X))(\hat{\theta} - \theta_0)$$

$$+ \frac{1}{2}(\hat{\theta} - \theta_0)^T \frac{\partial^2}{\partial \theta \partial \theta} \log f(X \mid \hat{\theta}(X))(\hat{\theta} - \theta_0)$$

$$= \log f(X \mid \hat{\theta}(X)) - \frac{1}{2}(\hat{\theta} - \theta_0)^T J(\theta_0)(\hat{\theta} - \theta_0)$$

$$E_X\left\{(\hat{\theta} - \theta_0)^T J(\theta_0)(\hat{\theta} - \theta_0)\right\} = \mathrm{tr}\left\{I(\theta_0)J^{-1}(\theta_0)\right\}$$

$$E_X\{D_1\} = E_X\left\{\log f(X \mid \hat{\theta}(X)) - \log f(X \mid \theta_0)\right\} \approx \frac{1}{2n}\mathrm{tr}\left\{IJ^{-1}\right\}$$

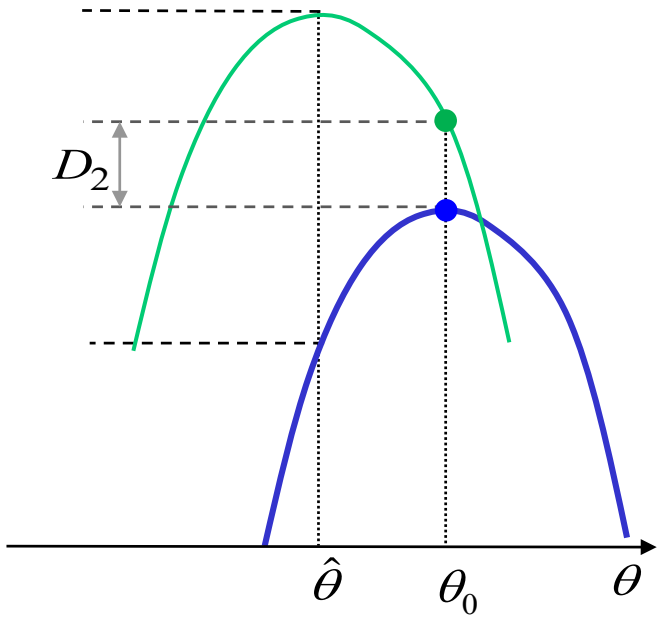
D_2 の評価

$$X = (X_1, \ldots, X_n)$$

$$\log f(X \mid \theta_0) = \log f(X_1, \ldots, X_n \mid \theta_0) = \sum_{j=1}^{n} \log f(X_j \mid \theta_0)$$

$$E_X \frac{1}{n} \log f(X \mid \theta_0) = \frac{1}{n} \sum_{j=1}^{n} E_X \log f(X_j \mid \theta_0)$$

$$= E_Y \log f(Y \mid \theta_0)$$



$$E_X\{D_2\} = E_X\left\{\frac{1}{n}\log f(X \mid \theta_0) - E_Y \log f(Y \mid \theta_0)\right\} = 0$$

バイアス補正量

$$b(G) = E_X\left[D\right] = E_X\left[D_1 + D_2 + D_3\right]$$

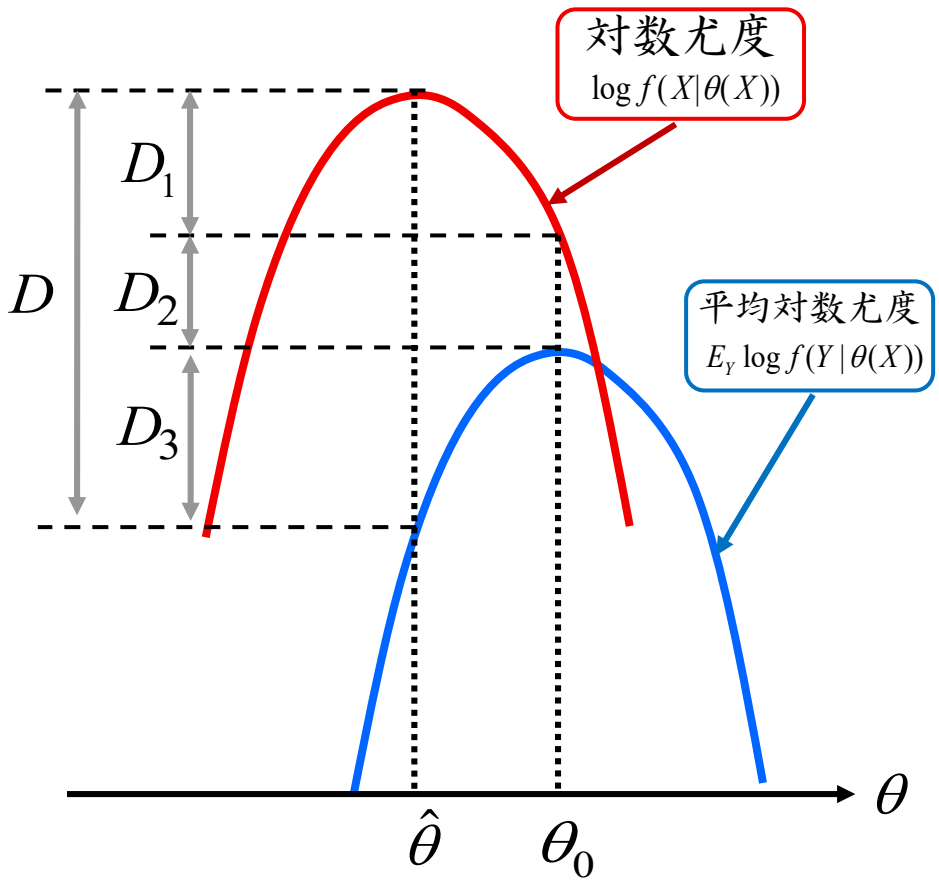
$$b_{\mathrm{TIC}}(G) = \mathrm{tr}\left\{I(G)J(G)^{-1}\right\}$$

$$I(G) = E_X\left[\frac{\partial \log f(X\,|\,\theta)}{\partial\theta}\frac{\partial \log f(X\,|\,\theta)}{\partial\theta'}\right]$$

Fisher情報量

$$J(G) = -E_X\left[\frac{\partial^2 \log f(X\,|\,\theta)}{\partial\theta\,\partial\theta'}\right]$$

ヘッセ行列の期待値



情報量規準

情報量規準の一般形

$$\mathrm{IC} = -2\log f(x\,|\,\hat{\theta}) + 2nb(G)$$

$$b(G) = E[D_1] + E[D_2] + E[D_3] = \frac{1}{n}\,\mathrm{tr}\left\{I(G)J(G)^{-1}\right\}$$

$$\mathrm{TIC} = -2\log f(x\,|\,\hat{\theta}) + 2\mathrm{tr}\{I(G)J(G)^{-1}\}$$

$I(G)$　Fisher 情報行列

$J(G)$　－（Hessianの期待値）

竹内(1976)

赤池情報量規準 AIC

$$\mathrm{AIC} = -2\log f(x \mid \hat{\theta}) + 2k$$

k: 自由パラメータ数（θの次元）

$\hat{\theta}$ 最尤推定量

$f(x \mid \hat{\theta})$ 最大対数尤度

$$f(x \mid \hat{\theta}) = \max_{\theta} f(x \mid \theta)$$

Akaike (1973)

行列 $I(\theta)$ と $J(\theta)$ の関係

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x|\theta)=\frac{\partial}{\partial\theta_i}\left\{\frac{\partial}{\partial\theta_j}\log f(x|\theta)\right\}$$

$$=\frac{\partial}{\partial\theta_i}\left\{\frac{1}{f(x|\theta)}\frac{\partial}{\partial\theta_j}f(x|\theta)\right\}$$

$$=\frac{1}{f(x|\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x|\theta)-\frac{1}{f(x|\theta)^2}\frac{\partial}{\partial\theta_i}f(x|\theta)\frac{\partial}{\partial\theta_j}f(x|\theta)$$

$$=\frac{1}{f(x|\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x|\theta)-\frac{\partial}{\partial\theta_i}\log f(x|\theta)\frac{\partial}{\partial\theta_j}\log f(x|\theta)$$

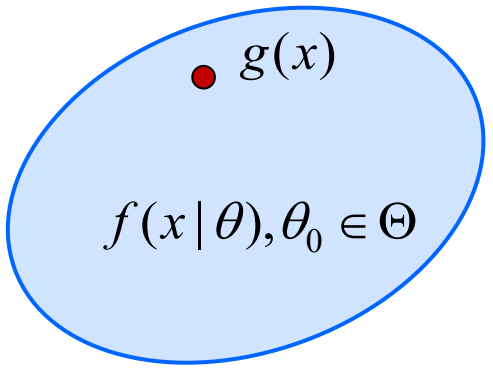
$$E_G\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x|\theta)\right]=E_G\left[\frac{1}{f(x|\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x|\theta)\right]-E_G\left[\frac{\partial}{\partial\theta_i}\log f(x|\theta)\frac{\partial}{\partial\theta_j}\log f(x|\theta)\right]$$

一般に　$I(\theta)\neq J(\theta)$

小西・北川(2004)　45p

AICとTICの関係

モデル族が真の分布を含む場合



$$\exists \theta_0 \in \Theta \text{ s.t. } g(x) = f(x|\theta_0)$$

$$E_G\left[\frac{1}{f(x|\theta_0)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x|\theta_0)\right] = \int \frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x|\theta_0)dx$$

$$= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f(x|\theta_0)dx = 0$$

$$I(\theta) = J(\theta)$$

$$nb(G) = \mathrm{tr}\left\{ I(G)J(G)^{-1} \right\}$$
$$= \mathrm{tr}\left\{ I_k \right\} = k$$

AICとTICの関係

モデルが真の分布を含む場合： TIC=AIC

TICがAICより優れていることを意味しない

AICの補正項は真の分布Gを含まない.

1．TICの補正項の計算はやや面倒

2．TICの補正項は実際には未知 ➡ データから推定

3．高次モーメントを含む. ➡ 分散が大きい.

TICの補正項：正規分布の場合

$$f(x\,|\,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$\log f(x\,|\,\theta)=-\frac{1}{2}\log 2\pi\sigma^2-\frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu}\log f(x\,|\,\theta)=\frac{x-\mu}{\sigma^2}$$

$$\frac{\partial}{\partial\sigma^2}\log f(x\,|\,\theta)=-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{\sigma^4}$$

$$\frac{\partial^2}{\partial\mu^2}\log f(x\,|\,\theta)=-\frac{1}{\sigma^2}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(x\,|\,\theta)=-\frac{x-\mu}{\sigma^4}$$

$$\frac{\partial^2}{(\partial\sigma^2)^2}\log f(x\,|\,\theta)=\frac{1}{2\sigma^4}-\frac{(x-\mu)^2}{\sigma^6}$$

$$J(\theta_0)=-\begin{bmatrix} E\left[\frac{\partial^2}{\partial\mu^2}\log f(X\,|\,\theta)\right] & E\left[\frac{\partial^2}{\partial\sigma^2\partial\mu}\log f(X\,|\,\theta)\right] \\ E\left[\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(X\,|\,\theta)\right] & E\left[\frac{\partial^2}{(\partial\sigma^2)^2}\log f(X\,|\,\theta)\right]\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{bmatrix}$$

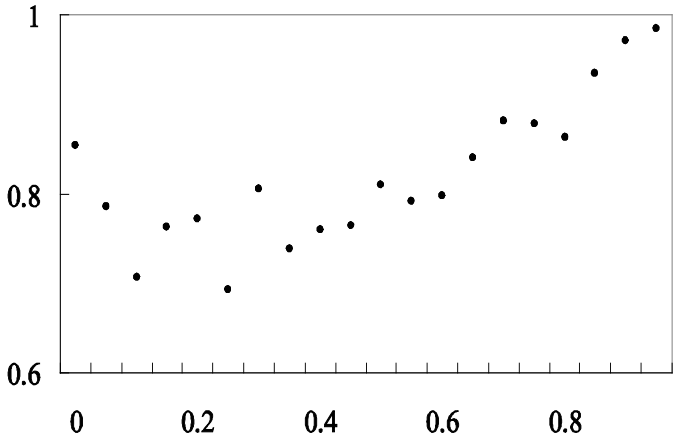
$$I(\theta_0)=E\left[\begin{bmatrix}\frac{X-\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}+\frac{(X-\mu)^2}{2\sigma^4}\end{bmatrix}\begin{bmatrix}\frac{X-\mu}{\sigma^2} & -\frac{1}{2\sigma^2}+\frac{(X-\mu)^2}{2\sigma^4}\end{bmatrix}\right]=\begin{bmatrix}\frac{1}{\sigma^2} & \frac{\mu_3}{2\sigma^6}\\ \frac{\mu_3}{2\sigma^6} & \frac{\mu_4}{4\sigma^8}-\frac{1}{4\sigma^4}\end{bmatrix}$$

$$I(G)J(G)^{-1}=\begin{bmatrix}\frac{1}{\sigma^2} & \frac{\mu_3}{2\sigma^6}\\ \frac{\mu_3}{2\sigma^6} & \frac{\mu_4}{4\sigma^8}-\frac{1}{4\sigma^4}\end{bmatrix}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}=\begin{bmatrix}1 & \frac{\mu_3}{\sigma^2}\\ \frac{\mu_3}{2\sigma^4} & \frac{\mu_4}{2\sigma^4}-\frac{1}{2}\end{bmatrix}$$

$$\mathrm{tr}\{I(G)J(G)^{-1}\}=1+\frac{\mu_4}{2\sigma^4}-\frac{1}{2}=\frac{1}{2}\left(1+\frac{\mu_4}{\sigma^4}\right)$$

モデル選択例：多項式回帰の次数

データ



$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

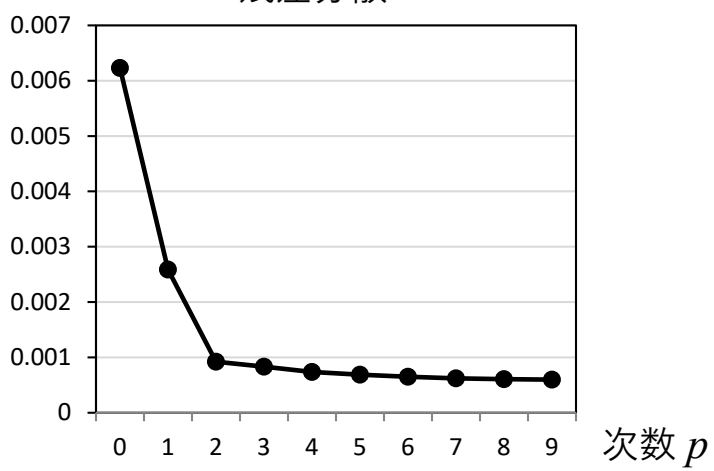
$$\theta = (\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$$

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\pi\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j y_{i-j}\right)^2$$

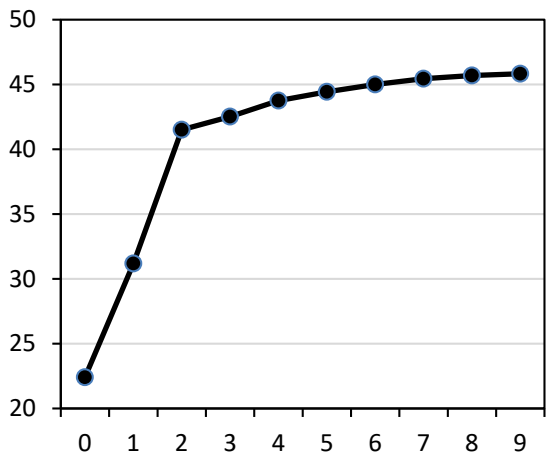
$$\ell(\hat{\theta}) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}$$

$$\mathrm{AIC}_p = n(\log 2\pi + 1) + n\log\hat{\sigma}^2 + 2(p+2)$$

残差分散

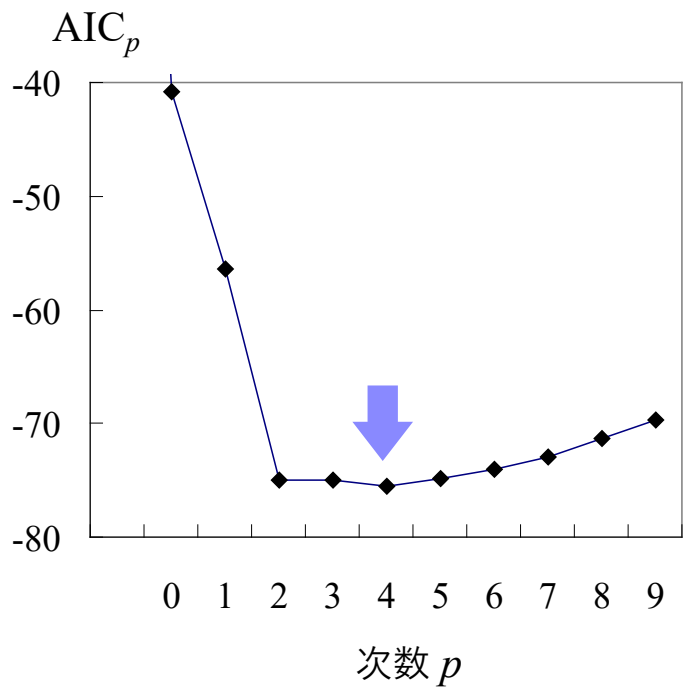


対数尤度

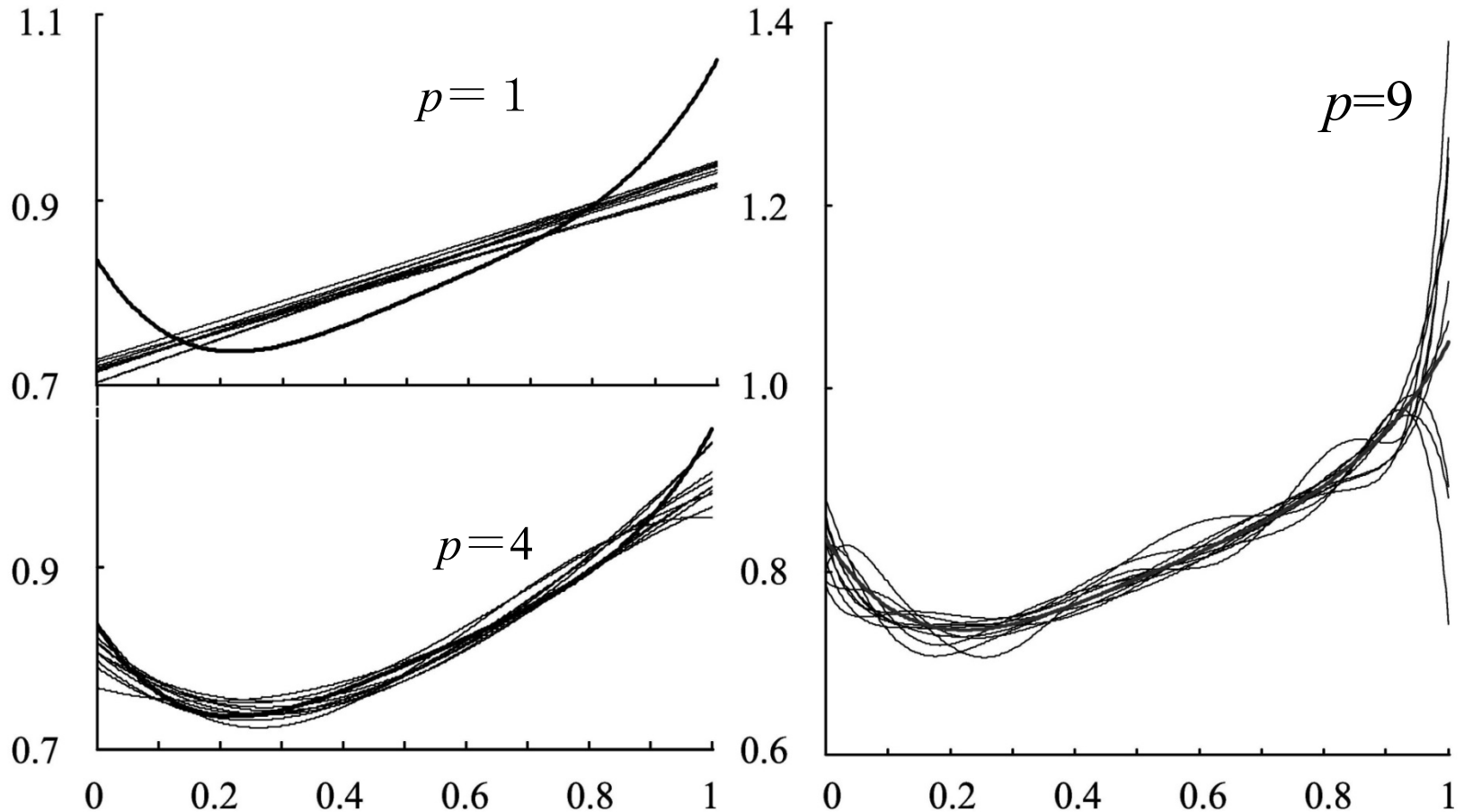


モデル選択例：多項式回帰の次数

次数	残差分散	対数尤度	AIC	AICの差
–	0.678301	-24.50	50.99	126.49
0	0.006229	22.41	-40.81	34.68
1	0.002587	31.19	-56.38	19.11
2	0.000922	41.51	-75.03	0.47
3	0.000833	42.52	-75.04	0.46
4	0.000737	43.75	-75.50	0.00
5	0.000688	44.44	-74.89	0.61
6	0.000650	45.00	-74.00	1.49
7	0.000622	45.45	-72.89	2.61
8	0.000607	45.69	-71.38	4.12
9	0.000599	45.83	-69.66	5.84



多項式回帰の次数と安定性



次数小：バイアスが大きい
次数大：変動が大きい

モデルの予測誤差分散

予測誤差 ＝ バイアス ＋ 分散

バイアス　～　モデルの悪さ
分散　～　モデルの不安定さ

AIC 最小モデル（$p = 4$）

- バイアスと分散を適度に小さくしたモデル
- 期待予測誤差最小のモデル

その他の情報量規準

AIC$_c$　有限修正

$$nb(G) = \frac{n(p+1)}{n-p-2}$$

GIC　統計的汎函数で定義される任意の推定量

$$nb(G) = \mathrm{tr}\left\{ \int T^{(1)}(x;G) \frac{\partial \log f(x|\theta)}{\partial \theta} dG(x) \right\}$$

EIC　Bootstrap法によるバイアス推定

$$b^*(G) = \frac{1}{n} E_{X^*} \left\{ \log f(X^* \mid \hat{\theta}(X^*)) - \log f(X \mid \hat{\theta}(X^*)) \right\}$$

ABIC ベイズ型情報量規準

$$\mathrm{ABIC} = -2 \max_{\lambda} \log \int f(x \mid \theta) \pi(\theta \mid \lambda) d\theta + 2q$$

NIC, BIC, WAIC, PIC, RIC

情報量規準の有限修正

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n)$$

$$nb_{\mathrm{c}}(G) = \frac{n(p+1)}{n-p-2}$$

$$\mathrm{AIC}_c = -2\log f(X \mid \hat{\theta}(X)) + 2\frac{n(p+1)}{n-p-2}$$

$p=1$ の場合の $nb_c(G)$

n	4	6	8	12	18	25	50	100	200
$b_{\mathrm{AICc}}(G)$	8.0	4.0	3.2	2.7	2.4	2.27	2.13	2.06	2.03

Sugiura (1978), 小西・北川(2004) p63, Konishi-Kitagawa(2008) p181

一般情報量規準：　GIC

$$\mathrm{GIC} = -2\log f(X \mid \hat{\theta}(X)) + 2nb(\hat{G}_{\mathrm{GIC}})$$

$\theta = T(G)$　　統計的汎関数

$\hat{\theta} = T(\hat{G})$

特長

- 統計的汎関数として定義できる任意の推定量に適用可能
  （最尤推定量，ロバスト推定量，正則化法，ベイズ推定の一部）
- EIC等の理論解析にも有用
- 高次補正も可能

弱点

- 汎関数微分の計算が面倒

一般情報量規準： GIC

$$E_X(D_1) = \frac{1}{n}\left(\int T^{(1)}_{p\alpha} f_p^{\alpha} dG + \frac{1}{2} S^{(2)}_{p} F_{p} + \frac{1}{2} S^{(11)}_{pq} F_{pq} \right)$$

$$E_X(D_3) = -\frac{1}{2n}\left(S^{(2)}_{p} F_{p} + S^{(11)}_{pq} F_{pq} \right)$$

$$nb(G) = \int T^{(1)}_{p\alpha} f_p^{\alpha} dG(x)$$

$$= \mathrm{tr}\left\{ \int T^{(1)}(x;G) \frac{\partial \log f(x|\theta)}{\partial \theta} dG(x) \right\}$$

Bootstrap情報量規準：　EIC

$$\mathrm{EIC} = -2\log f(X \mid \hat{\theta}(X)) + 2nb(G^*)$$

バイアス補正量を解析的にではなく，ブートストラップによって数値的に求める．

特長

- 解析的近似が不要
- 計算実装も比較的容易
- 最尤推定量以外の広範な推定量やモデルに適用可能

弱点

- データ生成・推定を繰り返すため計算量が多い

Bootstrap 情報量規準： EIC

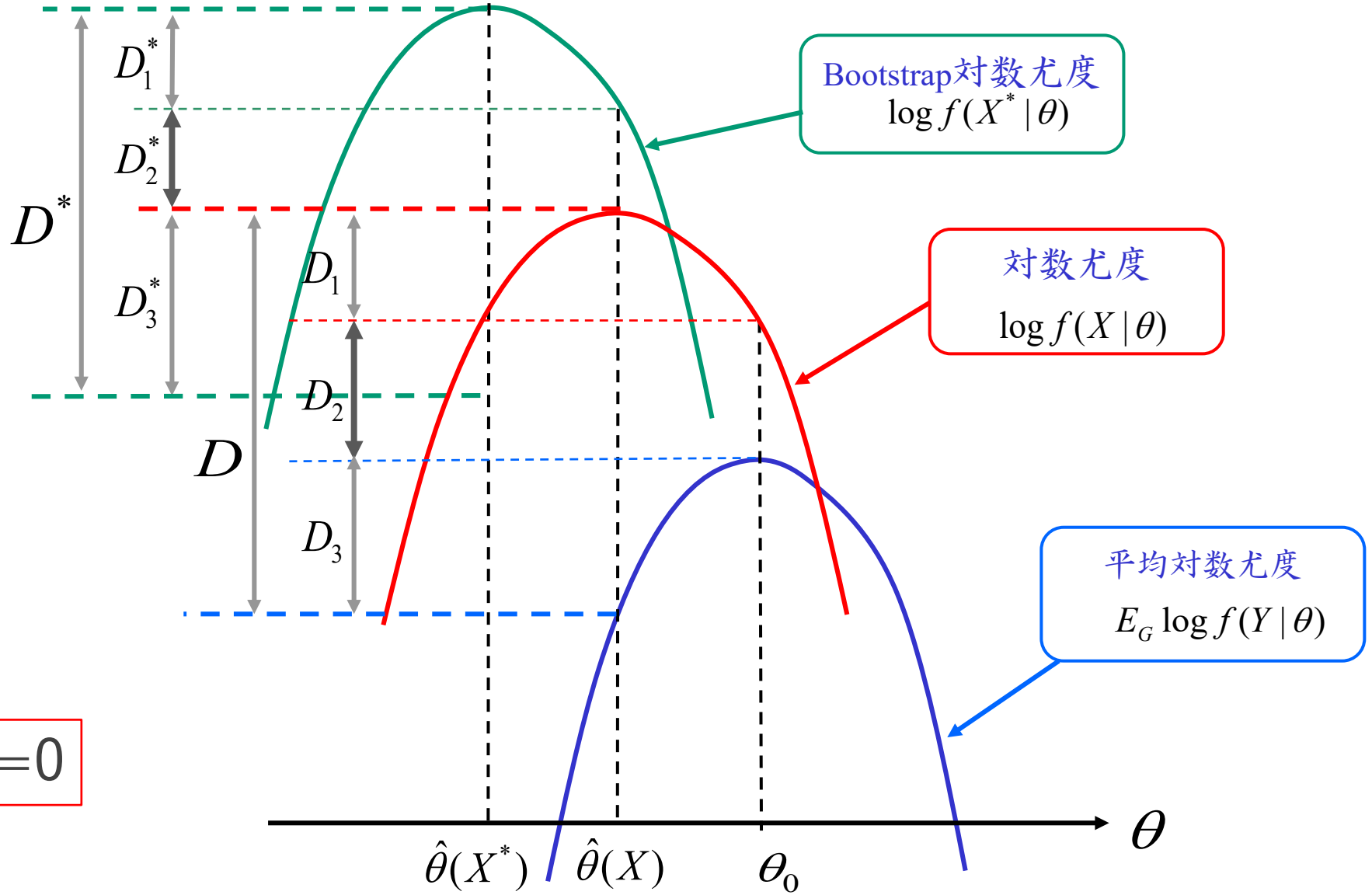
$$b(G) = E_X\left\{\frac{1}{n}\log f(X\,|\,\hat{\theta}(X)) - E_Y \log f(Y\,|\,\hat{\theta}(X))\right\}$$

- データ $X = (X_1, \ldots, X_n) \sim G(x)$
- 経験分布関数 $G_n(x) = \frac{1}{n}\sum_{j=1}^{n} I(x, X_j)$
- Bootstrap標本 $X^* = (X_1^*, \ldots, X_n^*) \sim \hat{G}_n(x)$

$$b^*(G) = E_{X^*}\left\{\frac{1}{n}\log f(X^*\,|\,\hat{\theta}(X^*)) - E_{Y^*} \log f(Y^*\,|\,\hat{\theta}(X^*))\right\}$$

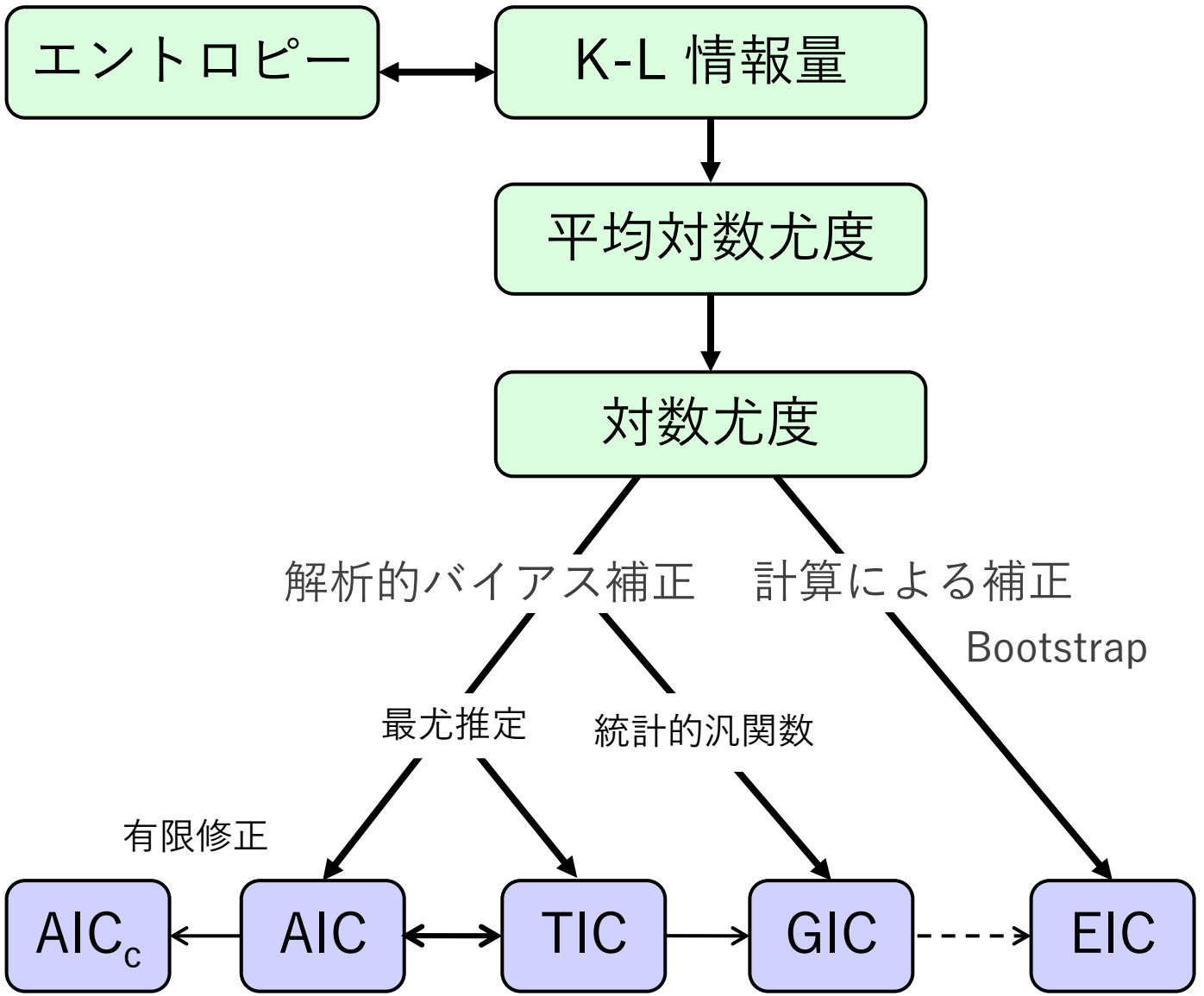
$$= \frac{1}{n} E_{X^*}\left\{\log f(X^*\,|\,\hat{\theta}(X^*)) - \log f(X\,|\,\hat{\theta}(X^*))\right\}$$

Bootstrap法によるバイアス補正



$\mathrm{E}[D_2]=0$

情報量規準の系譜


エントロピー
K-L 情報量
平均対数尤度
対数尤度
解析的バイアス補正
計算による補正
Bootstrap
最尤推定
統計的汎関数
有限修正
AIC_c
AIC
TIC
GIC
EIC

よいモデルを求める方法

複雑な現実　・・・　有限のデータ

- パラメータ数を少なくする・・・MAICE
- パラメータに制約を課す・・・Bayesモデル

モデル選択だけではない.

ABIC（ベイズ型情報量規準）

ベイズモデル

$f(x|\theta);\ \theta \in \Theta$　パラメトリックモデル

$\pi(\theta|\lambda)$　事前分布

λ　超（ハイパー）パラメータ（q 次元）

$$p(x|\lambda) = \int f(x|\theta)\pi(\theta|\lambda)d\theta$$　周辺分布

これをパラメータを λ とするモデルとみなす

$$\begin{aligned}\mathrm{ABIC} &= -2\log\left\{\max_{\lambda} p(x|\lambda)\right\} + 2q \\ &= -2\max_{\lambda}\log\int f(x|\theta)\pi(\theta|\lambda)d\theta + 2q\end{aligned}$$

北川(2004), Konishi-Kitagawa(2007)

BIC

ベイズモデル

$f(x|\theta_m);\ \ \theta_m \in \Theta \subset R^m$　　パラメトリックモデル

$\pi(\theta_m)$　　事前分布

$$p(x) = \int f(x|\theta_m)\pi(\theta_m)d\theta_m$$　　周辺尤度

$$\mathrm{BIC}_m = -2\log p(x) = -2\log \int f(x|\theta)\pi(\theta)d\theta$$
$$\approx -2\log f(x|\hat{\theta}) + m\log N$$

小西・北川(2004)

交差検証法（Cross Validation）

$$\{y_1,\ldots,y_N\} = \{x_1,\ldots,x_m\} \oplus \{z_1,\ldots,z_\ell\} \qquad m+\ell = N$$

推定用データ　　評価用データ

1. 全データを推定用データと評価用データに分ける
2. 推定用データでモデルを推定
3. 評価用データでモデルを評価（予測2乗誤差など）
4. 1の分割の仕方を変えて，すべての場合について　2, 3を繰り返し，評価量の平均を求める

分割の仕方：

Leave-one-out：　1個のデータだけ評価に用いる

k 分割法：　全体のデータをk 分割し，そのうちの1つを評価に用いる

AICに関する批判について

次数の一致性が最も重要な問題ではない.

1. モデリングの目的は，「よい」モデルを求めることで，「真の」モデルを求めることではない.
2. 次数の一致性は良いモデルを求めるための必要条件でも十分条件でもない.
3. 「真」の次数は一般に存在しない．存在する場合でも真の次数の推定されたモデルが予測によいとは限らない.
4. 「真」の次数より高くてもパラメータが一致性を持てばモデルは一致する.

シミュレーションの設定自体が不適切なことが多い

参考書

- 坂元慶行, 石黒真木夫, 北川源四郎(1983). 「情報量統計学」, 共立出版, 情報科学講座　A.5.4
- Y.Sakamoto, M.Ishiguro and G.Kitagawa (1986) *Akaike Information Criterion Statistics,* D.Reidel, Dordrecht.
- Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.
- 小西貞則, 北川源四郎(2004)「情報量規準」, 朝倉書店, 予測と発見の科学 2
- 竹内・下平・伊藤・久保川(2004)：モデル選択, 統計科学のフロンティア, 岩波書店
- 赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿, 編者 室田一雄・土谷隆(2007)「赤池情報量規準AIC－モデリング・予測・知識発見」共立出版
- S. Konishi and G. Kitagawa (2008). *Information Criteria and Statistical Modeling*, Springer Verlag

関連論文リスト

- Akaike, H. (1973), “Information theory and an extension of the maximum likelihood principle.” *Proc. 2nd International Symposium on Information Theory* , B. N. Petrov and F. Csaki eds., Akademiai Kiado, Budapest, 267-281.
- Akaike, H. (1974), “A new look at the statistical model identification.” *IEEE Trans. Automat. Contrl.*, AC-19, No. 6, 716-723.
- 竹内啓, (1976). 情報統計量の分布とモデルの適切さの規準, ＜特集>情報量規準. 数理科学, 14(3), 12-18.
- Konishi and Kitagawa (1996), “Generalized Information Criteria in Model Selection”, *Biometrika*, Vol. 83, No.4, 875-890.
- Ishiguro, Sakamoto and Kitagawa (1997), “Bootstrapping Log Likelihood and EIC, an Extension of AIC”, *Annals of the Institute of Statistical Mathematics*, Vol. 49, No. 3, 411-434.

