データ解析

第14回: モデルの評価

渡辺澄夫

モデルの評価

評価法

実世界

観測データ

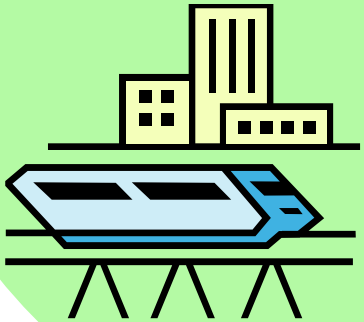
回帰・判別分析

推定と検定

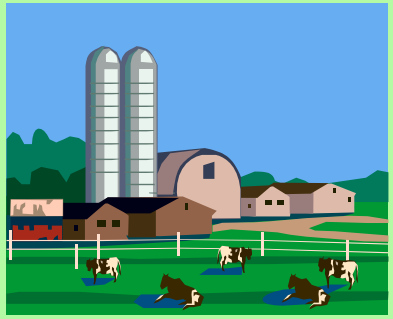
統計的検定

ベイズ法
・階層ベイズ法

解析方法



時系列予測



因子・主成分
・クラスタ分析

復習：統計学における評価とは

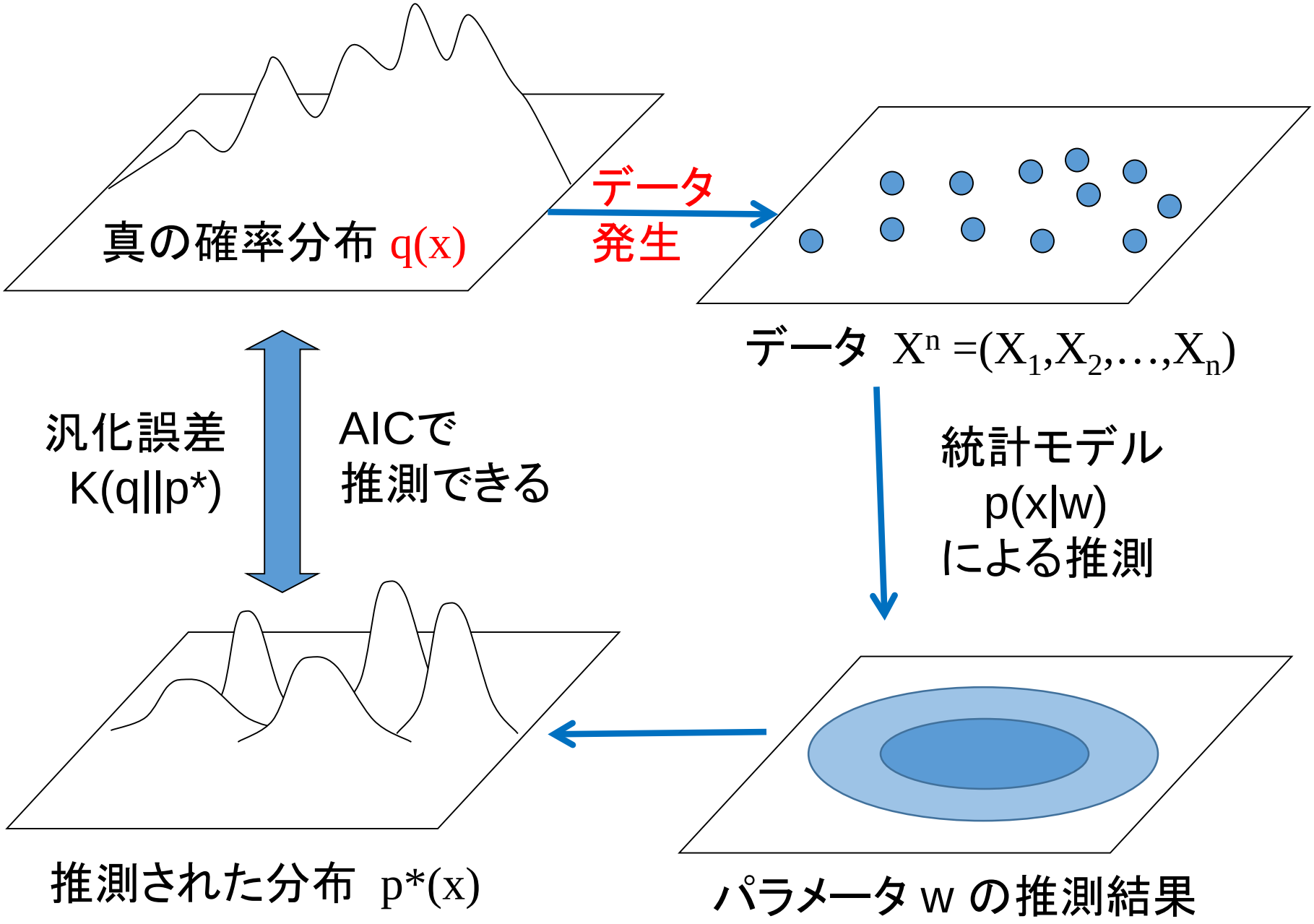
１．基本的な疑問：

「自分の用いたモデルは適切だったのだろうか」

２．モデルの適切さは、データを発生している真の分布を規準にして考察されるべきである。

３．しかしながら、真の分布はわからないので、較べることができない。

４．数学的な法則を利用すればよい。


真の確率分布 q(x)
データ
発生
データ $X^n = (X_1, X_2, \ldots, X_n)$
統計モデル
p(x|w)
による推測
パラメータ w の推測結果
推測された分布 p*(x)
汎化誤差
K(q||p*)
AICで
推測できる

復習：フィッシャー情報行列

定義．統計モデル $p(x|w)$ $(x \in R^N, w \in R^d)$ のフィッシャー情報行列を

$$I_{jk}(w) = \int [(\partial / \partial w_j) \log p(x|w)][(\partial / \partial w_k) \log p(x|w)]\, p(x|w)dx$$

と定義する。次のようにも書ける。

$$I_{jk}(w) = -\int [(\partial^2 / \partial w_j \partial w_k) \log p(x|w)]\, p(x|w)dx.$$

定理（最尤推定量の漸近正規性）．データ $X^n=(X_1,X_2,\ldots,X_n)$ が独立に $p(x|w_0)$ に従うものとする。統計モデル $p(x|w)$ のフィッシャー情報行列 $I(w)$ が正定値とする。データの数 n が∞に近づく極限で w^* は w_0 に確率収束し、$n^{1/2}(w^*-w_0)$ の確率分布は $N(0,I(w_0)^{-1})$ に収束する。従って (w^*-w_0) の確率分布は 漸近的に $N(0,I(w_0)^{-1}/n)$ に従う。

周辺尤度

周辺尤度

定義. $X^n=\{X_1,X_2,\ldots,X_n\}$ を独立な確率変数の集合とする。事前分布 $\varphi(w)$ と統計モデル $p(x|w)$ の周辺尤度を

$$Z(X^{\mathbf{n}}) = \int \varphi(w)\ \Pi_i\ p(X_i|w)\,dw$$

と定義する。「周辺尤度が大きいほどデータに対して統計モデルと事前分布が適切である」という考え方 (I.J.Good) が提案されている。

注意。この規準を用いる場合 $\int\varphi(w)dw=1$, $\int p(x|w)dx=1$ が必要です。

周辺尤度とは何か

もしもパラメータ w が $\varphi(w)$ に従って、X^n が $\Pi_i\, p(x_i|w)$ に従うならば、

$Z(X^{\mathbf{n}})$ は $\{\, \varphi(w), p(x|w) \,\}$ が与えられたとき X^nが得られる確率密度である。

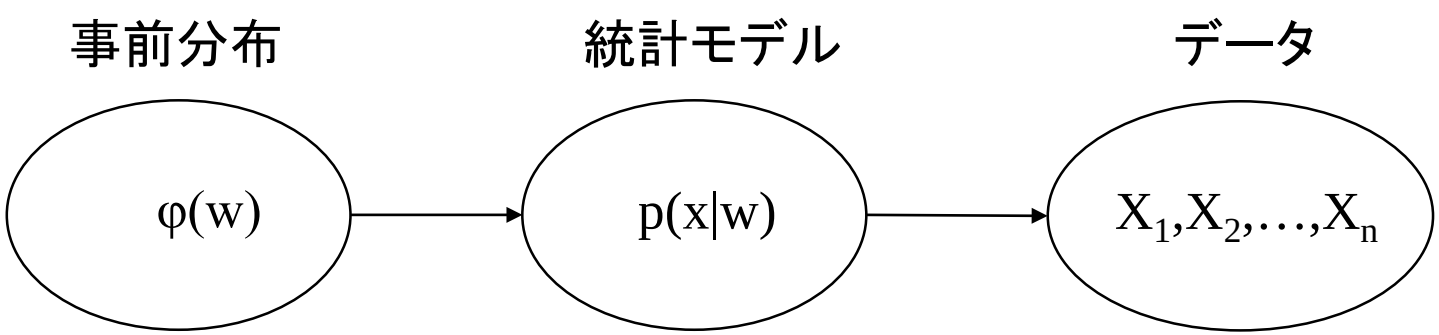
→　X^nが得られたとき $\{\, \varphi(w), p(x^n|w) \,\}$ から発生した確率は $Z(X^n)$ に

集合$\{\, \varphi(w), p(x|w) \,\}$ の上の事前確率をかけたものに比例する。

$$Z(X^n) = P(X^n|p,\varphi)$$

$$P(p,\varphi|X^n) \propto P(X^n|p,\varphi)P(p,\varphi) = Z(X^n)\, P(p,\varphi)$$

すなわち Z(Xn)は(p,φ)の尤度であり、これの最大化が提案されている。



周辺尤度とは何か:ベイズ法の検定から

$X^n=\{X_1,X_2,\ldots,X_n\}$ を確率変数の集合とする。

帰無仮説 :「w は $\varphi(w)$ から得られ、X^n は $q(x|w)$ から独立に得られた」

対立仮説 :「w は $\psi(w)$ から得られ、X^n は $p(x|w)$ から独立に得られた」

このときは

$$L(X^{\mathbf{n}}) = \int \psi(w)\, \Pi_i\, p(X_i|w)\, dw \;\Big/\; \int \varphi(w)\, \Pi_i\, q(X_i|w)\, dw$$

が最強力検定を与える。どちらを帰無仮説にしてどちらを対立仮説にしてもそれぞれの周辺尤度の大小を用いて比較を行うことになる。

対数周辺尤度

定義. データX^n、事前分布 $\varphi(w)$ 、統計モデル $p(x|w)$ が与えられたとき

$$F(X^n) = -\log Z(X^n) = -\log \int \varphi(w) \ \Pi_i \ p(X_i|w)\, dw$$

のことを対数周辺尤度の符号反転、確率的複雑さ、ベイズ符号長、統計的推測の自由エネルギーという。

（注意）数学的に同じものに多種類の名前がついているのは、この量が統計的推測、情報理論、統計力学においてまったく別の文脈の決定的なところで現れるからである。周辺尤度が大きいこととこの量が小さいことは等価である。情報科学や統計力学においては、この量が問題を解く鍵であることが知られている。すなわち、この量が計算できれば情報科学と統計力学の問題が解決することが知られている。しかしながら、この量の厳密値の計算は初等的には難しい場合が多く、現代数学の方法や理論物理学に基づく方法が発展している。

例：レプリカ法： $\mathbf{E}[\log Z] = \lim_{n\to+0} \mathbf{E}[Z^n-1]/n$. 例：ゼータ関数の極を使う方法.

汎化誤差と自由エネルギー

汎化誤差:

n 個のデータで統計モデルと事前分布を用いて推測をしたとき

予測分布が真の分布とどれくらいずれているかを測る量

自由エネルギー:

n 個のデータで統計モデルと事前分布を用いて推測をしたとき

統計モデルと事前分布の確からしさを測る量

汎化誤差を最小にする(モデル、事前分布)と

自由エネルギーを最小にする(モデル、事前分布)は一般に同じではない。

問1

$x, w \in \mathbf{R}$ である。次の組に対して、自由エネルギーを求めよ。

統計モデル $p(x|w) = 1/(2\pi)^{1/2} \exp(-(x-w)^2/2)$

事前分布　$\varphi(w) = 1/(2\pi)^{1/2} \exp(-w^2/2)$

ただしデータを $\{X_1, X_2, \ldots, X_n\}$ とする。

ヒント:　確率変数 $x^* = 1/(n+1) \Sigma X_i$　とおく。

$$\varphi(w) \prod_{i=1}^{n} p(X_i|w) = 1/(2\pi)^{(n+1)/2} \exp(-(n+1) w^2/2 + (n+1) x^* \cdot w)$$

公式:　$a>0$ のとき $\int \exp(-au^2)du = (\pi/a)^{1/2}$

ラプラスの方法

ラプラスの方法

記号 f と φ を R^d から実数 R への関数とする。以下の条件を仮定する。

(1) $f(w) \geqq 0$ は3回連続微分可能で $w=w_0$ でのみ最小値 0 をとる。

(2) 任意の w について $(\nabla^2 f_{jk})(w) = (\partial^2 / \partial w_j \partial w_k) f(w)$ は正定値行列。

(3) $\varphi(w) \geqq 0$ は $\int \varphi(w) dw = 1$ を満たし、$\varphi(w_0) > 0$.

定理（ラプラス）. 正の整数 n について $Z(n) = \int \exp(-n f(w)) \varphi(w) dw$ とおくと $n \to \infty$ の極限において

$$- \log Z(n) = (d/2) \log n + (1/2) \log \det \nabla^2 f(w_0)$$

$$- \log \varphi(w_0) - (d/2) \log (2\pi) + o(1).$$

証明

関数 $\varepsilon = 1/n^{2/5}$ とおく。$Z(n) = Z_1(n) + Z_2(n)$ と分解する。ただし

$$Z_1(n) = \int_{\|w-w_0\|<\varepsilon} \exp(-n f(w))\, \varphi(w)\, dw,$$

$$Z_2(n) = \int_{\|w-w_0\|\geqq\varepsilon} \exp(-n f(w))\, \varphi(w)\, dw.$$

関数 $f(w)$ は $w=w0$ で最小値をとるので $f(w_0)=0, \nabla f(w_0)=0.$

$$f(w) = (1/2)(w-w_0)\cdot \nabla^2 f(w_0)(w-w_0) + O(\|w-w_0\|^3).$$

まず$Z_2(n)$を考える。$\nabla^2 f(w)$ は正定値行列なので最小固有値は正値。

集合 $\|w-w_0\|\geqq\varepsilon$ では $n f(w) \geqq c\, n^{1/5}$ となる定数 $c>0$ が存在する。

$$0 \leqq Z_2(n) \leqq \exp(-c\, n^{1/5}).$$

証明つづき

次に $Z_1(n)$を考える。$I(w)$ は正定値行列なので最小固有値は正値。

集合 $\|w-w_0\|<\varepsilon$ では

$$n\, f(w) = (n/2)(w-w_0)\cdot \nabla^2 f\,(w_0)(w-w_0) + O(n^{-1/5})$$

$$\varphi(w) = \varphi(w0) + O(n^{-1/2})$$

が成り立つので

$$Z_1(n) \;=\; \int_{\|w-w_0\|<\varepsilon} \exp(\,-\,n\, f(w)\,)\, \varphi(w)\, dw$$

$$= \int_{\|w-w_0\|<\varepsilon} \exp(\,-\,(n/2)(w-w_0)\cdot \nabla^2 f\,(w_0)(w-w_0)+O(n^{-1/5}))\,(\varphi(w_0)+O(n^{-1/2}))\, dw$$

証明つづき

変数を変換して $u = n^{1/2}(w-w_0)$ とおくと $du = n^{d/2}\,dw$ であり、積分領域は

集合 $\|w-w_0\| < 1/n^{2/5}$ から集合 $\|u\| < n^{1/5}$ に変換されるので $Z_1(n)$ は

$$\int_{\|w-w_0\|<\varepsilon} \exp(-(n/2)(w-w_0)\cdot\nabla^2 f(w_0)(w-w_0) + O(n^{-1/5}))\,(\varphi(w_0)+O(n^{-1/2}))\,dw$$

$$= 1/n^{d/2} \int_{\|u\|< n1/5} \exp(-(1/2)u\cdot\nabla^2 f(w_0)\,u + O(n^{-1/5}))\,(\varphi(w_0)+O(n^{-1/2}))\,du$$

$$= 1/n^{d/2}\ \varphi(w_0)\,\{\int \exp(-(1/2)u\cdot\nabla^2 f(w_0)\,u)\ du + o(1)\}$$

$$= 1/n^{d/2}\ \varphi(w_0)\,\{(2\pi)^{d/2} / (\det \nabla^2 f(w_0))^{1/2} + o(1)\}$$

公式

$$\int \exp(-u\cdot Su/2)du$$

$$= (2\pi)^{d/2} / (\det S)^{1/2}$$

以上より $-\log Z(n) = -\log(Z_1(n) + Z_2(n))$ であるから定理が得られた（証明終）。

周辺尤度の漸近挙動

漸近挙動

データX^n、事前分布 $\varphi(w)$ 、統計モデル $p(x|w)$ が与えられたとき

周辺尤度を $Z(X^{\mathbf{n}}) = \int \varphi(w) \ \Pi_i \ p(X_i|w)\, dw$ とおく。

自由エネルギーを $F(X^n) = -\mathrm{Log}\, Z(X^{\mathbf{n}})$ とおく。

経験誤差関数（= -対数尤度）を $H(w) = -(1/n)\, \Sigma_i \ \log p(X_i|w)$ とおく。

最尤推定量 w^* は $H(w)$ を最小にするパラメータである。

定理. 次式が成立する（$n\to\infty$において漸近的に）。

$$F(X^{\mathbf{n}}) = nH(w^*) + (d/2)\log n + (1/2)\log\det(\nabla^2 H(w^*))$$

$$- \log\varphi(w^*) - (d/2)\log(2\pi) + o_p(1).$$

証明

周辺尤度を書き換えると　$Z(X^{\mathbf{n}}) = \int \varphi(w) \exp(-nH(w))\, dw$ となる。

そこで $f(w) = H(w) - H(w^*)$ とおくと

$$Z(X^{\mathbf{n}}) = \exp(-nH(w^*)) \int \varphi(w) \exp(-n f(w))\, dw$$

となる。これにラプラスの方法を適用すると定理が得られる。（証明終）。

(注意）厳密に残余項の $o_p(1)$ が$n\to\infty$で0に確率収束することを示すには

確率論的な議論が必要である。

多項分布の例

{A,B,C} の目が確率 (w_1,w_2,w_3)で出るサイコロの確率分布(多項分布）は

変数を $x \in \{(1,0,0),(0,1,0),(0,0,1)\}$ とすると $p(x|w) = (w_1)^{x_1}(w_2)^{x_2}(w_3)^{x_3}$.

事前分布を集合 $\{w=(w_1,w_2,w_3)\,;\, w_i \geqq 0,\ w_1+w_2+w_3=1\}$ 上の一様分布

$\varphi(w)=\Gamma(3)$ とする。試行 n 回で{A,B,C} の出た回数を(n_1,n_2,n_3) とすると、

自由エネルギーの厳密値は（*この例では計算できる*）

$$F(X^{\mathbf{n}}) = \log \Gamma(n+3) - \Sigma_{j=1}^{3} \log \Gamma(n_j+1) - \log \Gamma(3).$$

一方、ラプラス近似を用いると $w^*=(n_1/n,n_2/n,n_3/n)$ でパラメータは2個であり

$$F(X^{\mathbf{n}}) \fallingdotseq nH(w^*) + \log n + (1/2) \log \det(\nabla^2 H(w^*)) - \log \Gamma(3) - \log (2\pi).$$

ここで $H(w) = -\Sigma_{j=1}^{3} (n_j/n) \log w_j$ である。真のエントロピーは $S_n=H(w_0)$で

$$\nabla^2 H(w^*) = \begin{pmatrix} n_1/(n\, w_1^2) + n_3/(nw_3^2) & n_3/(nw_3^2) \\ n_3/(nw_3^2) & n_2/(n\, w_2^2) + n_3/(nw_3^2) \end{pmatrix}$$

多項分布の例：自由エネルギー真値とラプラス近似

問2

統計モデル $p(x|a,s) = (s/2\pi)^{1/2} \exp(-(s/2)(x-a)^2)$ とする。

事前分布を $\varphi(a,s)$ とする。このとき

$$\log p(x|a,s) = -(s/2)(x-a)^2 + (1/2)\log s + \text{定数}$$

(1) $H(w)$ を求めよ。

(2) 最尤推定量 w^* を求めよ。

(3) 自由エネルギーのラプラス近似を求めよ。

ベイズ情報量規準

BICの定義

赤池情報量規準 (AIC, 1974) を

$$AIC = -2\,\Sigma_i \log p(X_i|w^*) + 2d$$

と定義すれば　$E[AIC] = 2n\,E[\text{汎化誤差}] + o(n)$ が成り立つ.

シュワルツのベイズ情報量規準 (BIC, 1978) を

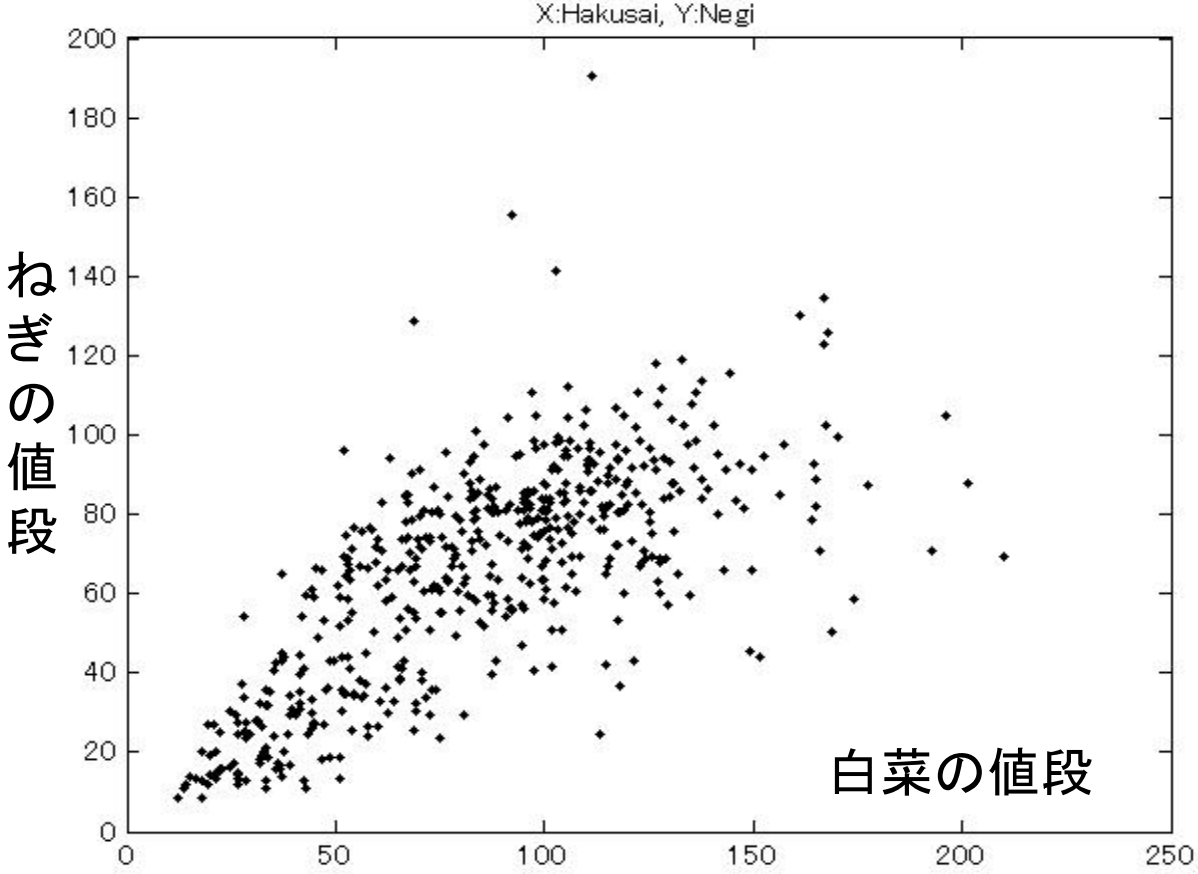
$$BIC = -2\,\Sigma_i \log p(X_i|w^*) + d \log n$$

と定義すれば $BIC = 2 \times \text{自由エネルギー} + O_p(1)$ が成り立つ.

(注意）AIC も BIC も機械学習で現れるモデル（カーネル法、深層学習、ボルツマンマシンなど）では前提となる条件（フィッシャー情報行列の正定値性）が成り立たないので利用できない。→最新の研究を検索してみよう。

実データに使ってみる

時系列予測の例：1970年1月から2013年１２月までの白菜とねぎの値段
「政府統計の総合窓口」のデータを使用しています。
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do



白菜の値段からねぎの値段を知りたいとき、
どんなモデルを使うとよいのだろうか。

いろいろなモデル

真の関係はわからない。多項式モデルを複数考えてみる。

$$p(y|x,w,\sigma) = 1/(2\pi\sigma^2)^{1/2} \exp(- (1/2\sigma^2)(y-f(x,w))^2)$$

$Y=a+$雑音$(0,\sigma^2)$

$Y=a+bX+$雑音$(0,\sigma^2)$

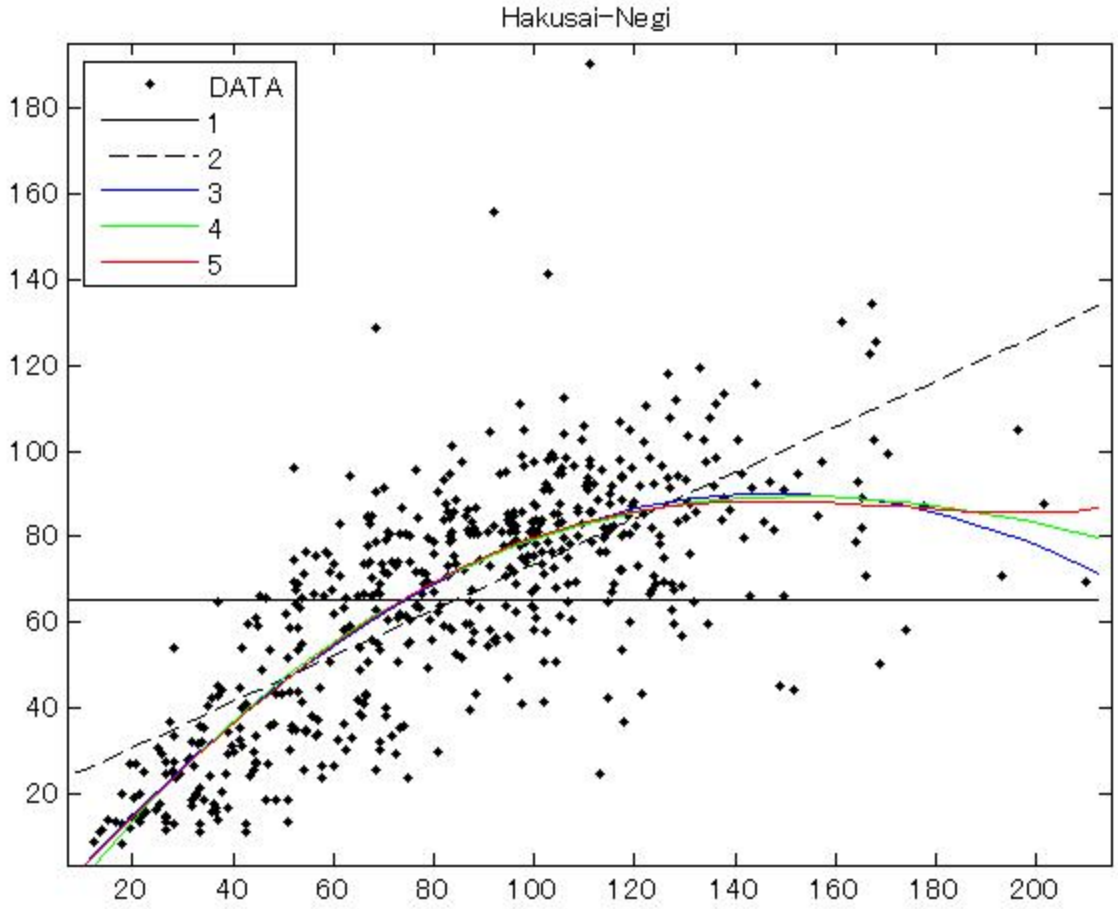
$Y=a+bX+cX^2+$雑音$(0,\sigma^2)$

$Y=a+bX+cX^2+dX^3+$雑音$(0,\sigma^2)$

$Y=a+bX+cX^2+dX^3+eX^4+$雑音$(0,\sigma^2)$

二乗誤差を最小にするパラメータは簡単に求められるが・・・。

学習してみた



グラフを見てもどれが良いのかわからない・・・。
高次元だとグラフを見ることもできない・・・。

AICとBICの計算法

統計モデル

$$p(y|x,w,\sigma) = 1/(2\pi\sigma^2)^{1/2} \exp(- (1/2\sigma^2)(y-f(x,w))^2)$$

に対して最尤推定量 (w^*,σ^*) を計算して

$$AIC = -2 \ \Sigma \log p(Y_i|X_i,w^*,\sigma^*) + 2d$$

$$BIC = -2 \ \Sigma \log p(Y_i|X_i,w^*,\sigma^*) + d \log n$$

を求めればよい。

モデル番号	AIC	BIC
1	AIC=1751.5,	BIC=1760.1
2	AIC=1560.2,	BIC=1573.0
3	AIC=1508.2,	BIC=1525.3
4	AIC=1509.6,	BIC=1531.1
5	AIC=1511.3,	BIC=1537.0

AICとBICの性質

AICは汎化誤差を推定する指標であり、BICはモデルと事前分布の確率を推定する指標である。目的が違うので、数式が異なることに矛盾はない。

☆　想定しているモデルの集合の中に真のデータを発生している分布があるとき、データの数 $n \to \infty$のとき、真のデータを発生している（パラメータ数が一番少ない）モデルを選ぶことができる時、そのモデル選択法を「一致性を持つ」という。

☆　考察している有限のモデルの集合の中にデータを発生している分布がないときデータの数 $n \to \infty$のとき、最も汎化誤差を小さくするモデルを選ぶことができるときそのモデル選択法を「有効性を持つ」という。

AICは一致性は持たないが、有効性をもつケースがあることが知られている。

BICは一致性を持つが、有効性は持たないことが知られている。

現実の問題では両方計算して、それぞれの観点から考察しよう。

問3

白菜の値段からねぎの値段を推定する問題で AICとBIC を計算することで何がわかったかを述べよ。