情報・システム工学概論

統計モデルの数理

— 第 1 回：統計モデルの考え方 —

駒木　文保

工学部 計数工学科

2017 年 12 月 4 日

物理モデルと統計モデル

（広い意味の）物理モデル

- ニュートンの運動方程式，マクスウェルの方程式，シュレディンガー方程式, ...
- 回路，制御,...
- ロトカ・ヴォルテラ方程式，ホジキン・ハックスレー方程式,...

微分方程式を用いたモデルが多い

統計モデル

不確実な現象のモデリング
比較的新しいパラダイム

統計的モデリングの考え方の発展については，例えば Salsburg (2010)（日本語訳の文庫本）などが参考になる.

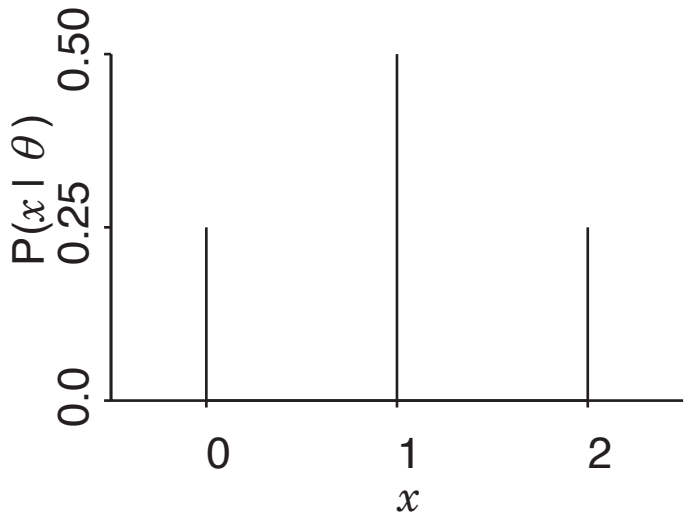
簡単な統計モデル１　２項分布モデル

θ: コイン投げで表の出る確率

x: コインを N 回 投げたとき表の出る回数 (確率変数)

x のしたがう確率分布:　２項分布 $\mathrm{Bin}(N, \theta)$

$$P(x;\theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}.$$



コインの表の出る回数の確率 $\mathrm{P}(x;\theta)$, $N = 2$, $\theta = 1/2$

ゆがんでいないコインであれば，$\theta = \frac{1}{2}$

コインの歪んでいるとき，$\theta \in [0, 1]$ の値は正確にはわからない.

θ: パラメータ
　未知であることを強調して，未知パラメータともいう

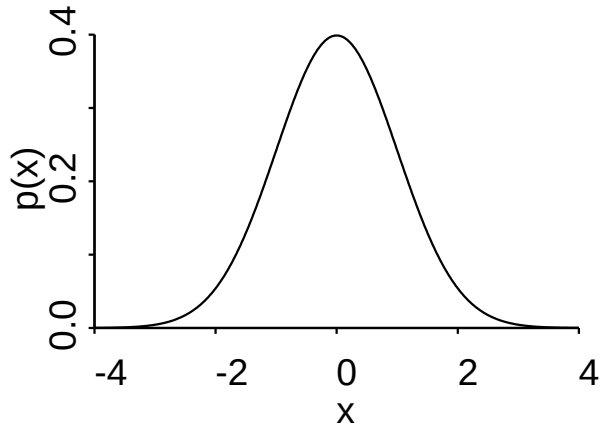
２項分布 $\mathrm{Bin}(N, \theta)$ 全体: パラメータ θ をもつ２項分布モデル

パラメータを変えて得られる確率分布全体を（パラメトリックな）統計モデルと呼ぶ.

x が観測されたとき，θ についてどのようなことがいえるのか.

簡単な統計モデル２ 正規分布モデル

正規分布 $\mathrm{N}(0,1)$　（0 は平均, 1 は分散）



正規分布 $\mathrm{N}(0,1)$ の確率密度関数.

平均 μ, 分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ の確率密度関数

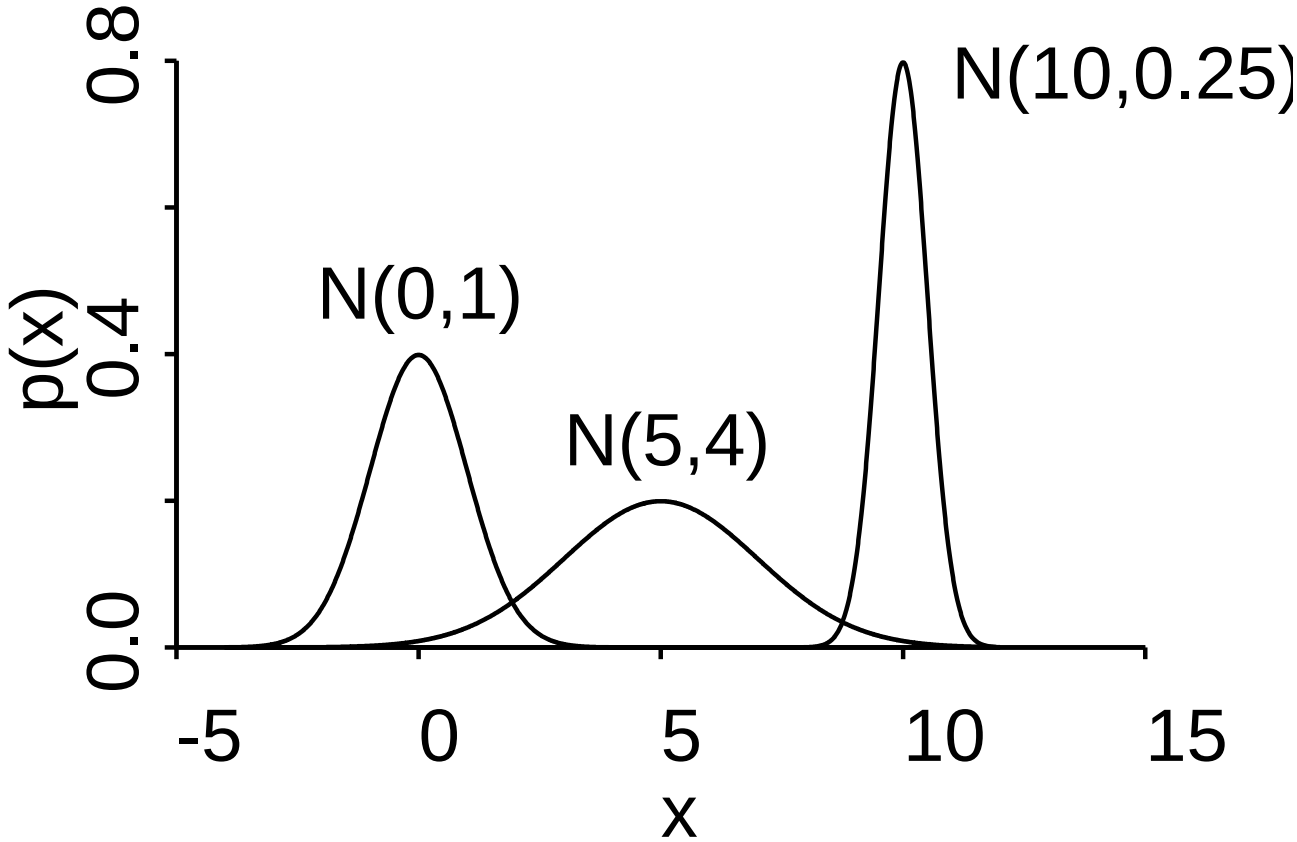
$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y-\mu)^2 \right\}.$$

最も基本的で重要な分布

中心極限定理

- 人間の身長の分布は正規分布でよく近似できる

いろいろな正規分布



$\mathrm{N}(0,1)$, $\mathrm{N}(5,4)$, $\mathrm{N}(10,0.25)$.

μ, σ: パラメータ

正規分布 $\mathrm{N}(\mu, \sigma^2)$ 全体: パラメータ μ, σ^2 をもつ統計モデル

物体の長さをある装置を使って測定

μ_0: 物体の真の長さ

ε: 装置の測定誤差

測定の結果得られる観測値

$$\mu_0 + \varepsilon$$

測定誤差 ε の分布 $\mathrm{N}(0, \sigma_0^2)$

測定の結果得られる観測値の分布：

正規分布 $\mathrm{N}(\mu_0, \sigma_0^2)$ （真の分布）

実際に測定を行う人は μ_0 の値を知らない.

装置の性能も分からない場合には σ_0^2 の値も未知.

正規分布モデル $\mathrm{N}(\mu, \sigma^2)$ を仮定して，μ_0 と σ_0^2 を推定

物体の長さと装置の測定誤差がわかる.

回帰モデル

データ：N 人についての身長と体重を組にした測定値.

データをもとにして，身長から体重を予測したい.

回帰モデルの応用は非常に広い.

一般的な傾向として身長の高い人ほど体重も重い傾向.

身長を x，体重を y として

$$y = bx + c + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

の直線状の関係を仮定してデータを解析.

ε は平均 0, 分散 σ^2 の正規分布 $\mathrm{N}(0, \sigma^2)$ にしたがう確率変数.

ε により，同じ身長のひとでも体重が違うことをモデル化できる.

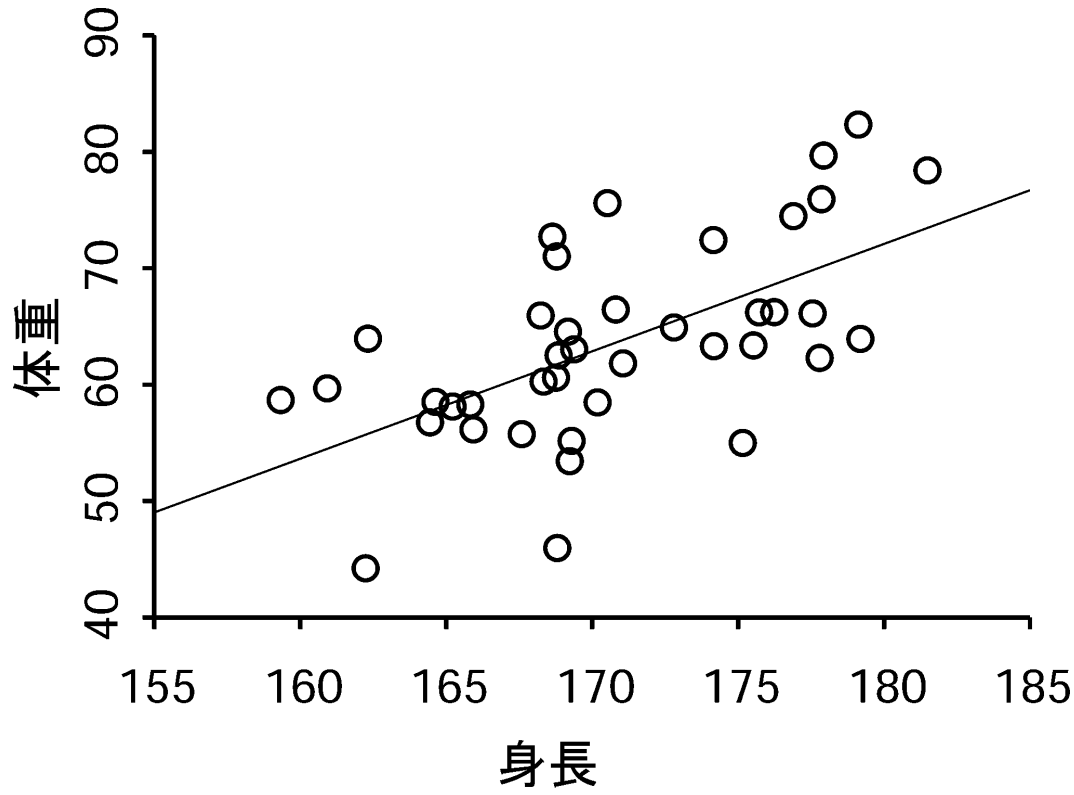
未知パラメータは b, c, σ.

$b > 0$ ならば，身長が増えると体重も増える傾向をもつ.

データから b c σ の推定値 $\hat{b}$, $\hat{c}$, $\hat{\sigma}$ を得ることにより，身長と体重の関係式

$$y = \hat{b}x + \hat{c} + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \hat{\sigma}^2)$$

を利用して，身長から体重が予測できる.



回帰モデル．身長と体重の仮想的なデータのプロットと，データに当てはめた直線 $y = \hat{b}x + \hat{c}$.

マルコフ連鎖モデル

簡単のために，天気に晴と雨しか無いと仮定.

第 n 日目が晴であれば $X_n = 0$, 雨であれば $X_n = 1$ と表す.

確率変数の列 $X_0, X_1, X_2, \ldots$ を考える.

p: 晴れた日の翌日に晴れる確率,
（晴れた日の翌日に雨が降る確率は $1 - p$）,

q: 雨が降った日の翌日に晴れる確率,
（雨が降った日の翌日に雨が降る確率は $1 - q$）

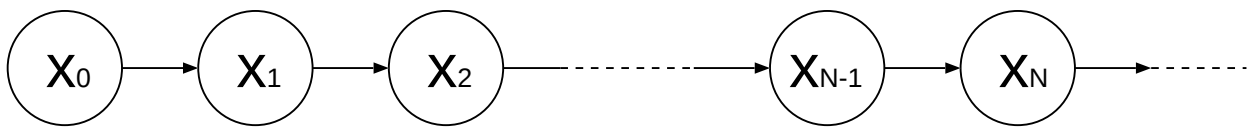
マルコフ連鎖と呼ばれるモデルのクラスの簡単な例.

p, q: モデルのパラメータ.

過去のデータから p と q の推定値 $\hat{p}$, $\hat{q}$ を構成して，今日の天気から明日の天気が予測できる.

マルコフ連鎖モデルを一般化した隠れマルコフモデルは音声認識やアミノ酸配列・塩基配列の解析（遺伝子解析）等で広く利用される．

マルコフ連鎖は 1 次元の構造をもっている．これを多次元に拡張したマルコフ場モデルは，画像解析や空間統計学などで利用される．



マルコフ連鎖

ベイジアンネットワーク

マルコフ連鎖は 1 次元の構造をもつ.

ベイジアンネットワーク，グラフィカルモデル，
確率ニューラルネットワーク

多くの確率変数が影響を及ぼし合うことを考慮したモデル

簡単な例 (Cowell 他, 1999)

計算機が動作しないときに考えられる２つの原因

停電 or 計算機故障

二つとも原因として考えらえるが，室内の照明も点灯しなければ，
原因が停電である可能性が高くなる．

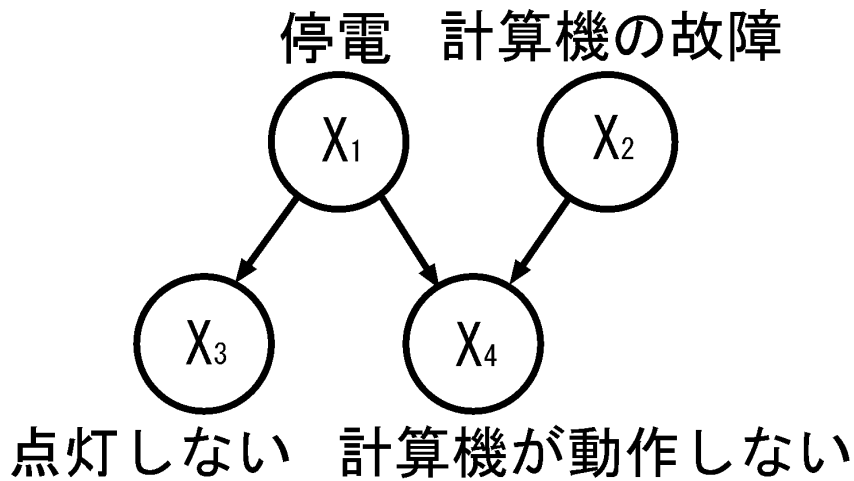
X_1: 停電であるかないか
X_2: 計算機が故障しているかいないか
X_3: 照明が点灯するかしないか
X_4: 計算機が動作するかしないか

それぞれ 1 か 0 かで表す．

確率変数 X_1, X_2, X_3, X_4 が互いに影響を及ぼしあっている程度を
数値化して，パラメトリックな統計モデルを構築．

このモデルを利用することにより，計算機の故障の原因に関する推論が自動的にできる．

このような統計モデルは，人工知能，パターン認識，データ圧縮，符号理論などの分野で利用される．



ベイジアンネットワーク

統計モデルのパラメータ推定

統計モデルのうちで最も簡単な正規分布モデル $\mathrm{N}(\mu, \tau)$ を考える.

以下，σ^2 を τ と書き換える.

真の分布 $p_0(y)$ にしたがうデータ $x_1, x_2, \ldots, x_n$ が得られたとき，データを基にして，なるべく真のパラメータ μ_0, τ_0 に近い推定値 $\hat{\mu}, \hat{\tau}$ を得たい.

$p_0(y)$ をよく近似する $p(y; \hat{\mu}, \hat{\tau})$ をデータに基づいて選ぶことは，パラメータ推定と呼ばれる重要な問題.

最尤推定: さまざまな統計モデルに応用できるパラメータ推定法

Kullback–Leibler ダイバージェンス

推定の良さを評価するためには真の分布の確率密度関数 $p_0(y)$ と推定した確率密度関数 $p(y; \hat{\mu}, \hat{\tau})$ との近さを評価する必要がある.

定義. 確率密度関数 $p(y)$ から $q(y)$ への **Kullback–Leibler** ダイバージェンス（相対エントロピーとも呼ばれる）

$$D(p, q) = \int p(y) \log \frac{p(y)}{q(y)} \mathrm{d}y$$

$p(y)$, $q(y)$ がどのくらい離れているかを表す.

統計学や情報理論で本質的な役割を果たす. 重要!

例. 正規分布 $\mathrm{N}(\mu_1, \tau_1)$ から $\mathrm{N}(\mu_2, \tau_2)$ への Kullback-Leibler ダイバージェンス

$$D(p(y;\mu_1,\tau_1), p(y;\mu_2,\tau_2)) = \frac{1}{2}\left\{\left(\frac{\tau_1}{\tau_2} - \log\frac{\tau_1}{\tau_2} - 1\right) + \frac{1}{\tau_2}(\mu_1-\mu_2)^2\right\}.$$

Kullback-Leibler ダイバージェンスは非負の量で，$p = q$ のときのみ 0 になるという距離に似た性質を持つ.

距離の公理は満たさない.

$D(p, q) = D(q, p)$ は成立しない.

真の分布 $p_0(y)$ から推定した分布 $p(y; \hat{\mu}, \hat{\tau})$ への Kullback-Leibler ダイバージェンス

$$D(p_0(y), p(y; \hat{\mu}, \hat{\tau}))$$

を最小にする $\hat{\mu}$, $\hat{\tau}$ を選ぶことができれば良い.

$p_0(y)$ は未知なので工夫が必要.

真の分布 $p_0(y)$ からモデルに属する分布 $p(y;\mu,\tau)$ への Kullback-Leibler ダイバージェンスを

$$
\begin{aligned}
D(p_0(y), p(y;\mu,\tau)) &= \int p_0(y) \log \frac{p_0(y)}{p(y;\mu,\tau)} \mathrm{d}y \\
&= \int p_0(y) \log p_0(y) \mathrm{d}y - \int p_0(y) \log p(y;\mu,\tau) \mathrm{d}y
\end{aligned}
$$

のように変形.

第 1 項はパラメータの値によらない項なので，$D(p_0, p(y;\mu,\tau))$ を最小化することは

$$
\int p_0(y) \log p(y;\mu,\tau) \mathrm{d}y
$$

を最大化することに帰着.

$\int p_0(y) \log p(y;\mu,\tau) \mathrm{d}y$ は $\log p(y;\mu,\tau)$ の p_0 に関する期待値.

真の分布 $p_0(y)$ はわからないため，p_0 に関する期待値をデータ，$x_1, x_2, \ldots, x_n$ に対する平均

$$\frac{1}{n}\sum_{i=1}^{n} \log p(x_i; \mu, \tau)$$

におきかえる.

この量は対数尤度関数（パラメータ μ, τ の関数とみなす）と呼ばれるものになっている.

これを最大化する μ, τ の値 $\hat{\mu}$, $\hat{\tau}$ が 最尤推定量.

パラメータ μ, τ の最尤推定量 $\hat{\mu}, \hat{\tau}$ の具体的な形は

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \hat{\tau} = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n}.$$

最尤推定はさまざまなモデルに対して汎用的に用いることのできる推定法.

複雑な統計モデルに対して，最尤推定量を求めるためには計算機を利用した最適化手法の利用が必要.

モデル選択

統計的モデルを利用したデータ解析を行う場合，最初からひとつのモデルが特定できていることは少ない.

いくつかのモデルの候補のうちから一番よいと思われるモデルを選択するのが普通.

どのようにしてモデルを選択するのかは統計的手法を利用する際の重要な問題.

- データの特性を忠実に表現するにはある程度複雑なモデルを利用することが必要.
- あまり複雑なモデルを採用するとパラメータの推定の精度がおちる.

赤池情報量規準 (Akaike's Information Criterion, AIC)

データに基づいて適切なモデルを選択するための規準

定義

$\mathrm{AIC} = -2 \times$ モデルの最大対数尤度 $+ 2 \times$ モデルのパラメータ数.

最大対数尤度が大きければモデルがデータに良く当てはまっていることになる.

- モデルを複雑にすると第 1 項は小さくなる（最大対数尤度は大きくなる）.
- モデルを複雑にするとモデルのパラメータ数が大きくなる.

ＡＩＣを小さくするモデルを選ぶことにより，データに対するあてはまりの良さとモデルの複雑さとのバランスをとる.

日本語の解説書: 坂元・石黒・北川 (1983)，小西・北川 (2004)

google ロゴ (2017 年 11 月 5 日)

Hirotugu Akaike's 90th Birthday

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

例. 多項式回帰モデル

$y_i,\ i = 1, 2, \ldots, N$: 正規分布 $\mathrm{N}(f(x_i), \sigma^2)$ にしたがう観測値.

$f(x)$: なめらかな関数で σ^2 とともに未知.

k 次多項式回帰モデル

$$y_i = a_0 + a_1 x_i + a_2 {x_i}^2 + \cdots + a_k {x_i}^k + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$$

を仮定して解析する.

$f(x)$ は高次の多項式を使えば原理的にはいくらでも精密に近似できる.

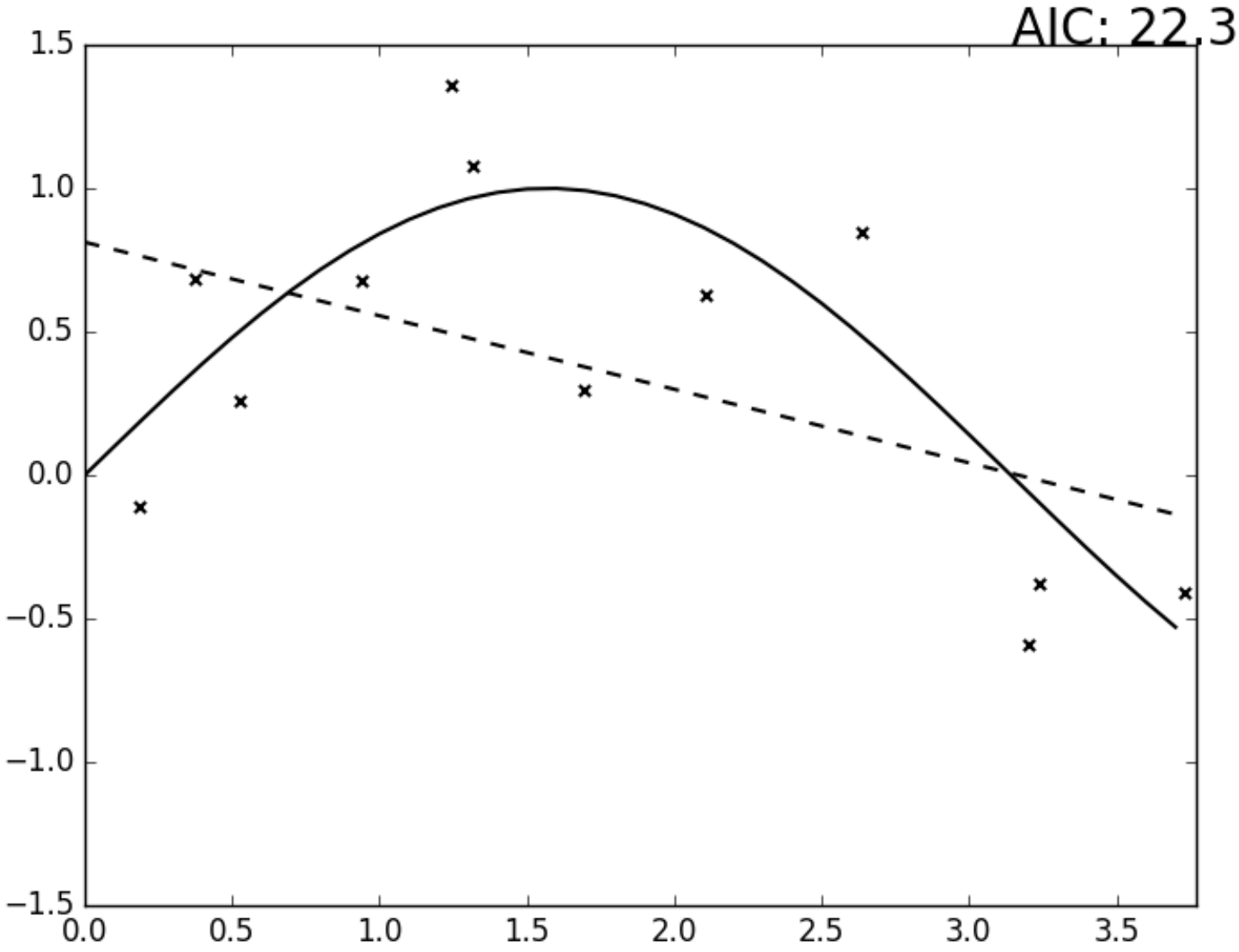
高次の多項式を使うと推定するパラメータ $a_0, a_1, \ldots, a_k, \sigma^2$ の数が多くなり，観測値の数が限られているので，パラメータ推定の精度が悪くなる

$\Rightarrow$ $f(x)$ の近似は必要以上に高次のモデルを利用するとかえって悪くなる.

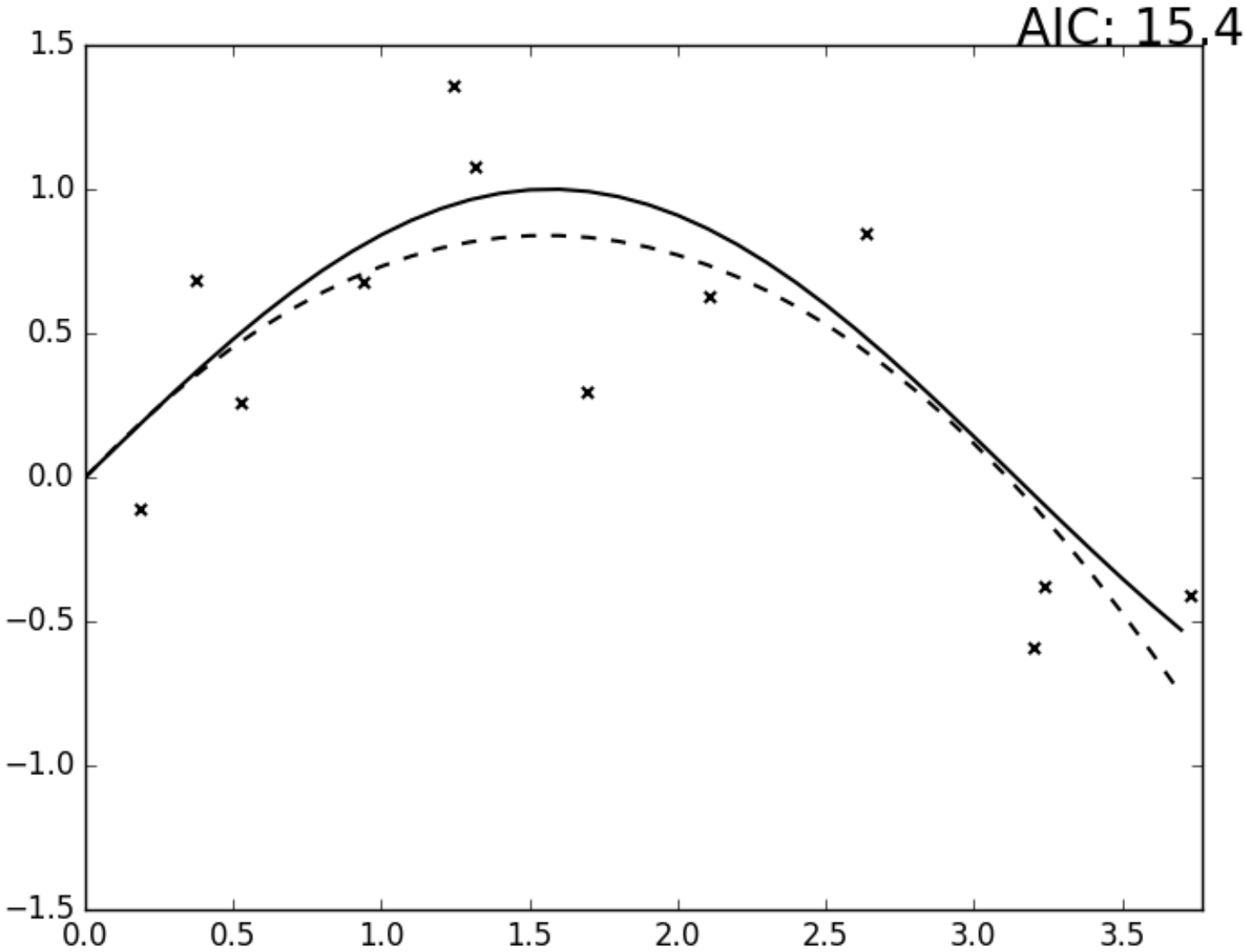
数値例：

$$f(x) = \sin x, \quad \sigma = 0.3$$

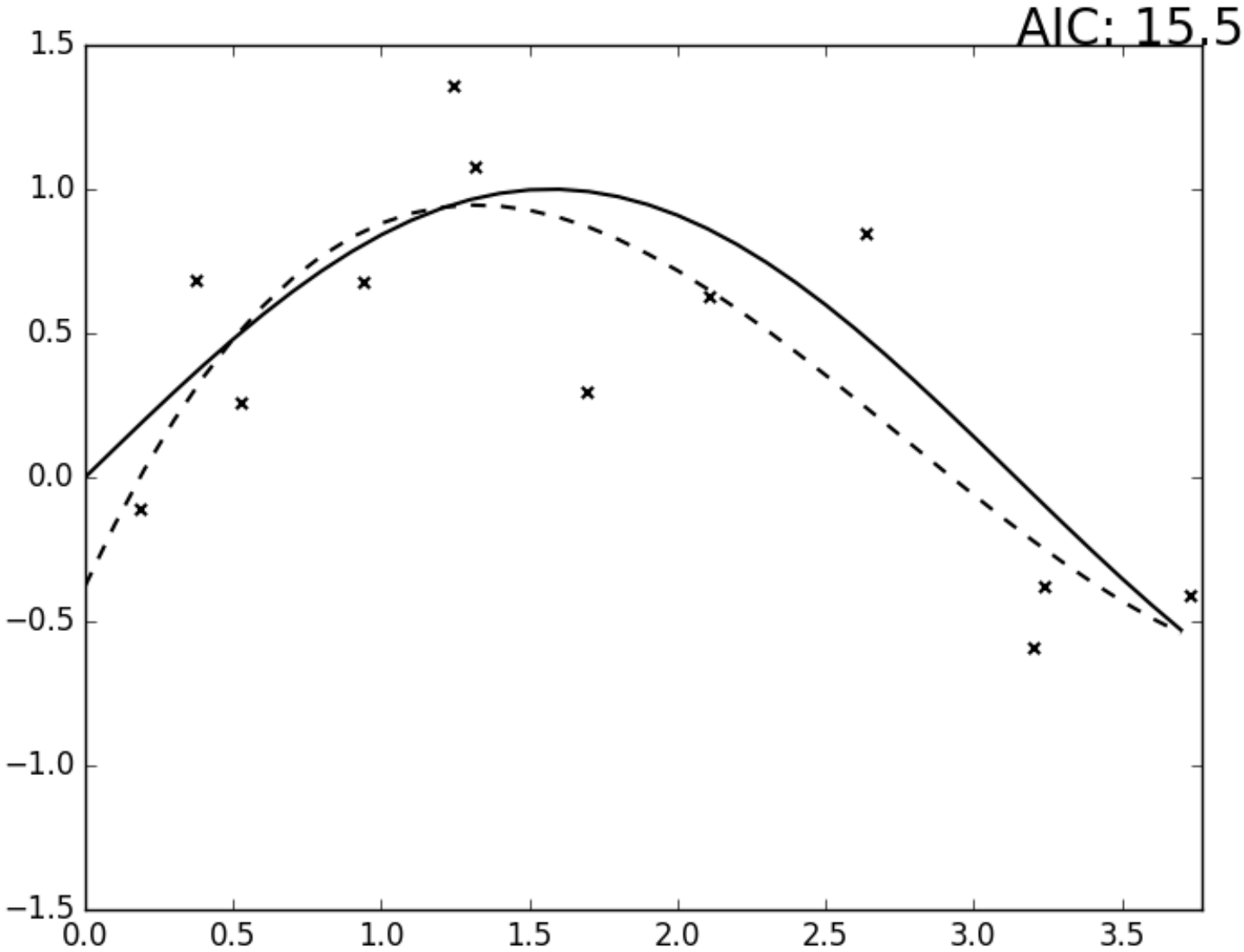
$f(x)$ を $1 \sim 5$ 次の多項式モデルを用いて推定.



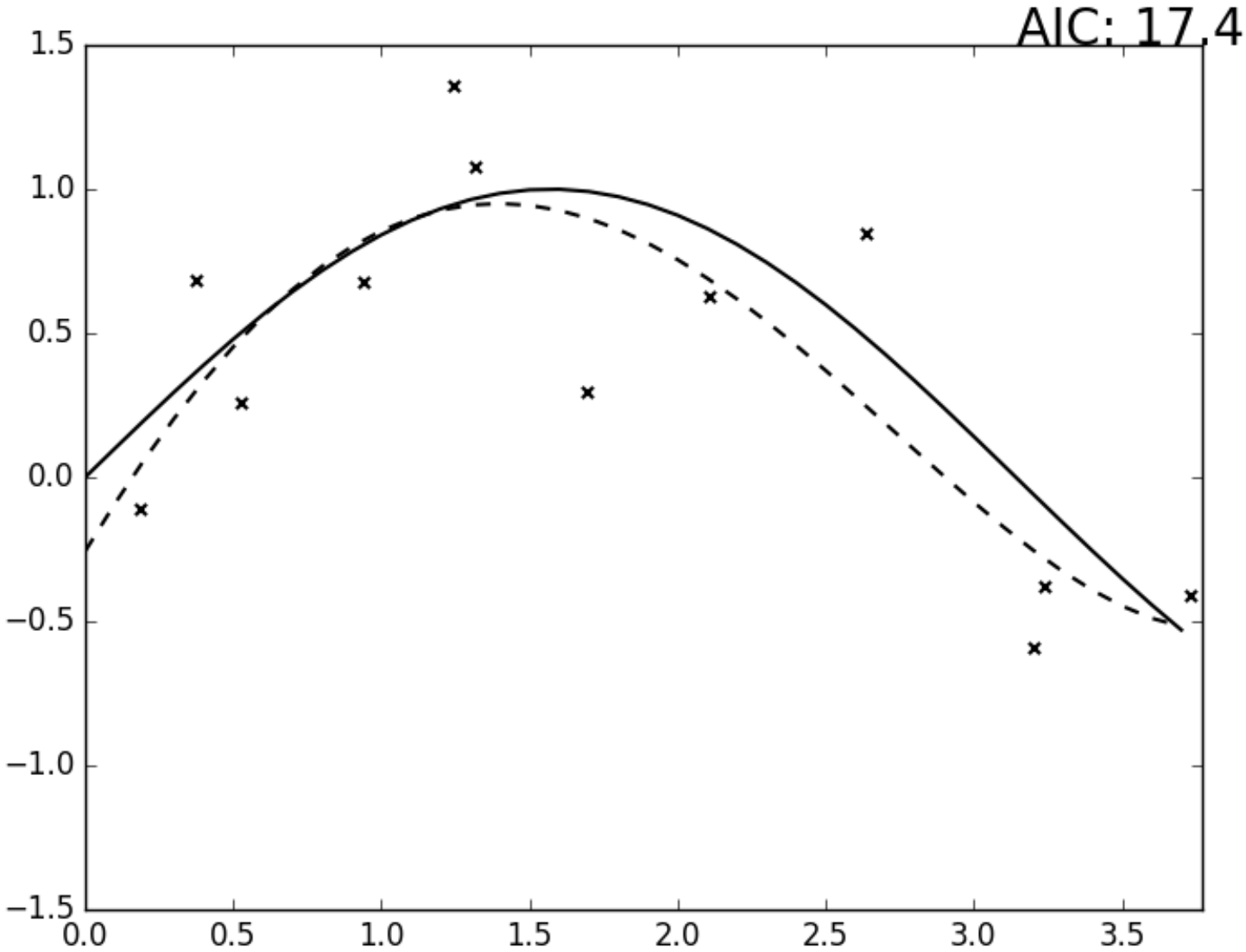
実線：真の $f(x)$, 点線：1 次式を用いた推定結果



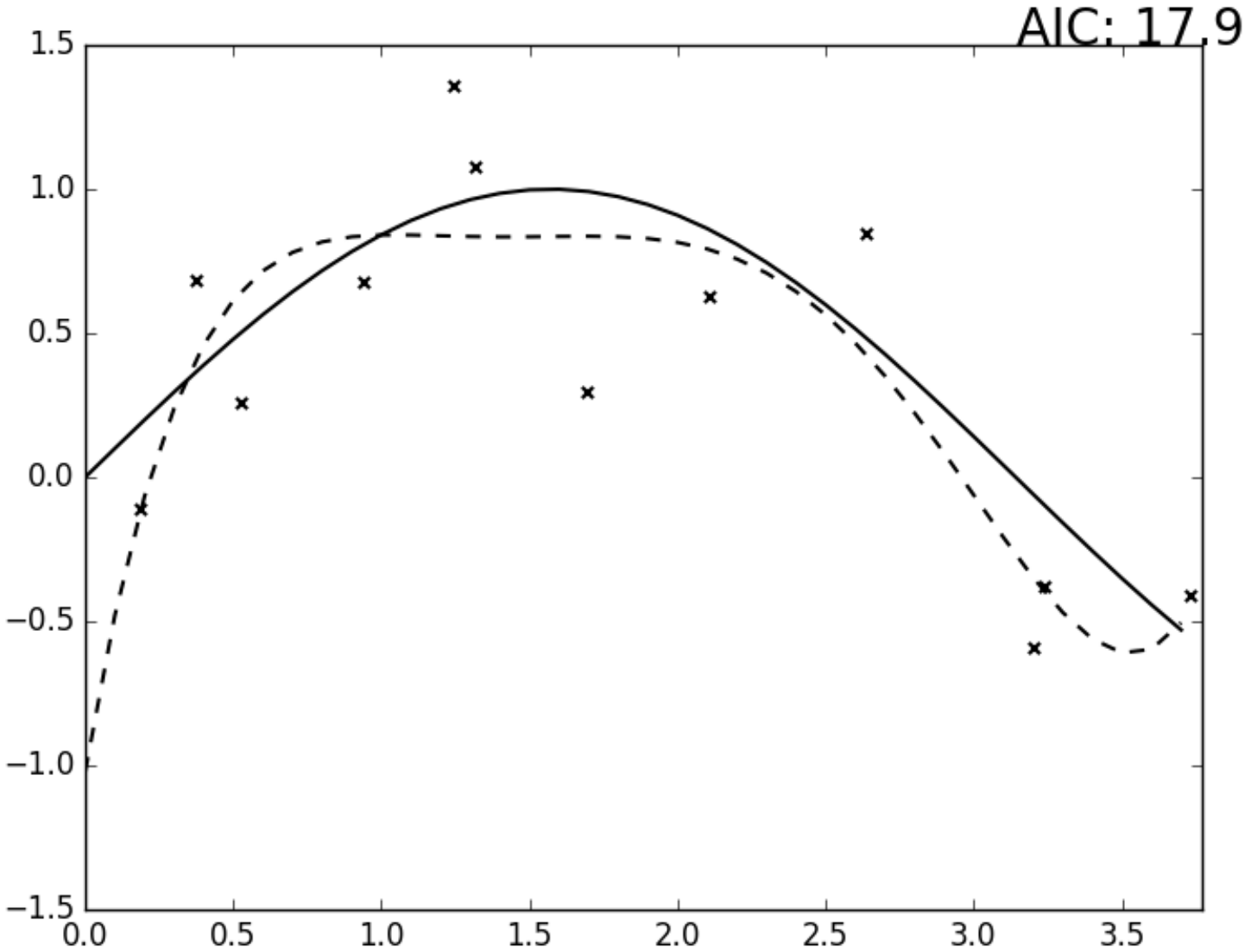
実線：真の $f(x)$, 点線：2 次式を用いた推定結果



実線：真の $f(x)$,　点線：3 次式を用いた推定結果



実線：真の $f(x)$, 点線：4 次式を用いた推定結果



実線：真の $f(x)$, 点線：5 次式を用いた推定結果

参考文献

Salsburg, D. S. (2010) 統計学を拓いた異才たち，竹内・熊谷訳，日本経済新聞出版社

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*, New York: Springer-Verlag.

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.

小西貞則，北川源四郎 (2004). 情報量規準，朝倉書店.