データ解析

第7回: 時系列分析

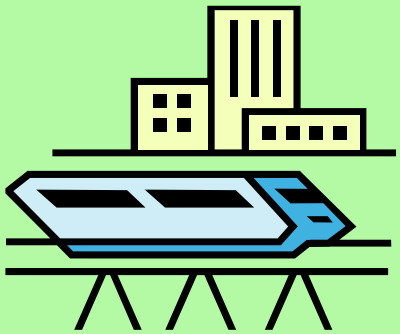
渡辺澄夫

過去から
未来を
予測
する

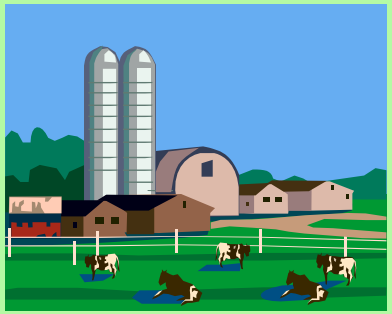
観測データ

回帰・判別分析

解析方法



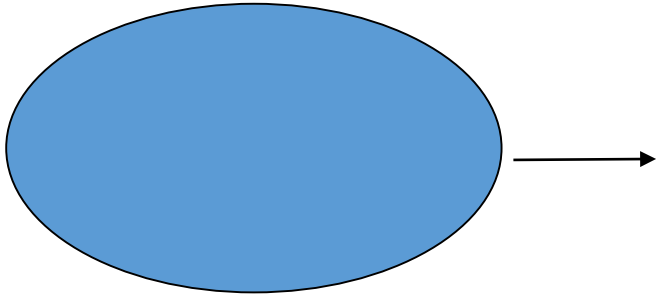
時系列予測



主成分・因子
・クラスタ分析

時系列を予測する

無限個の確率変数（確率変数が作る無限数列）{X(t) ; tは整数} を生成する情報源を考える。{X(t)} を確率過程という。確率過程について過去の値から未来を予測する問題を考える。



真の情報源

$$X(t-K), X(t-K+1), \ldots, X(t-1) \rightarrow X(t)$$

過去から未来を予測する

時系列の定常性･非定常性

定常性

確率過程 { X(t) ; tは整数} を考える（実数の数列に値をとる確率変数）。

定義. (1) X(t) が強定常過程であるとは、任意の整数 n について (X(t),X(t+1),…,X(t+n)) の同時確率密度関数が t によらないときをいう。

(2) X(t) が弱定常過程であるとは任意の t について $E[|X(t)|^2]$ が有限であり、任意の整数 n について E[X(t)X(t+n)] が t によらないときをいう。

（注意）任意の整数 t と n について (X(t),X(t+1),…,X(t+n)) が正規分布に従う場合には、強定常過程であることと弱定常過程であることは同値であるが、一般には同値ではない。

定常と非定常

時系列の予測を行なう場合には、何らかの意味で定常性が仮定されている（定常だと考えて予測している）ことが普通である。

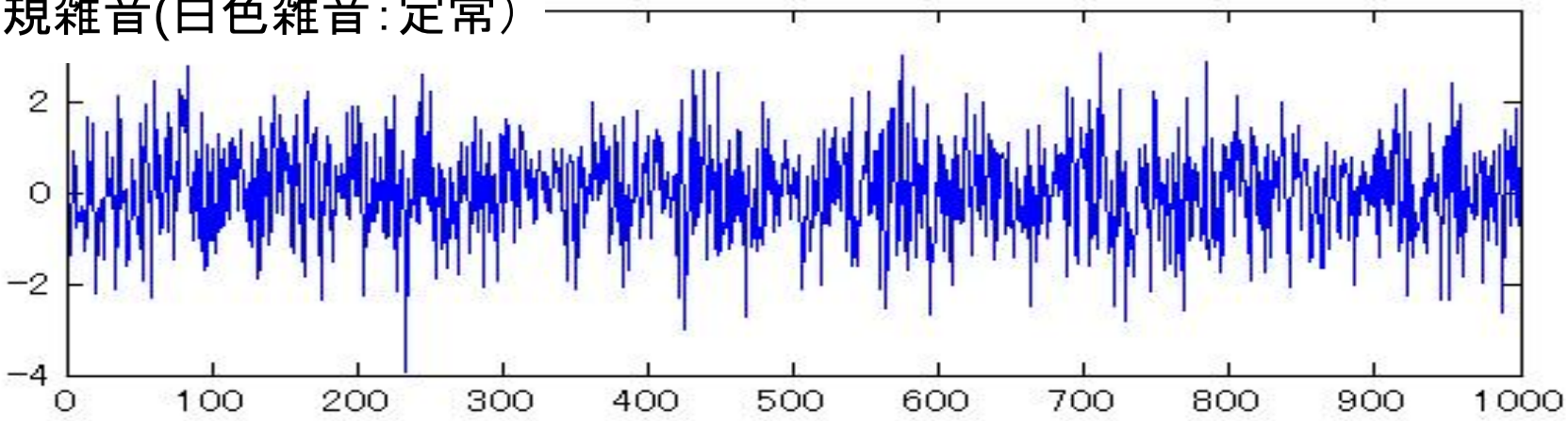
○　確率的でない関数 f(t) で X(t) – f(t) が定常であるようなものがあらかじめ分かっている場合には X(t) – f(t) を解析すればよい。

○　差分 X(t+1) - X(t) やさらに高階差分を取ることで定常になる場合には差分で得られる時系列を解析する。

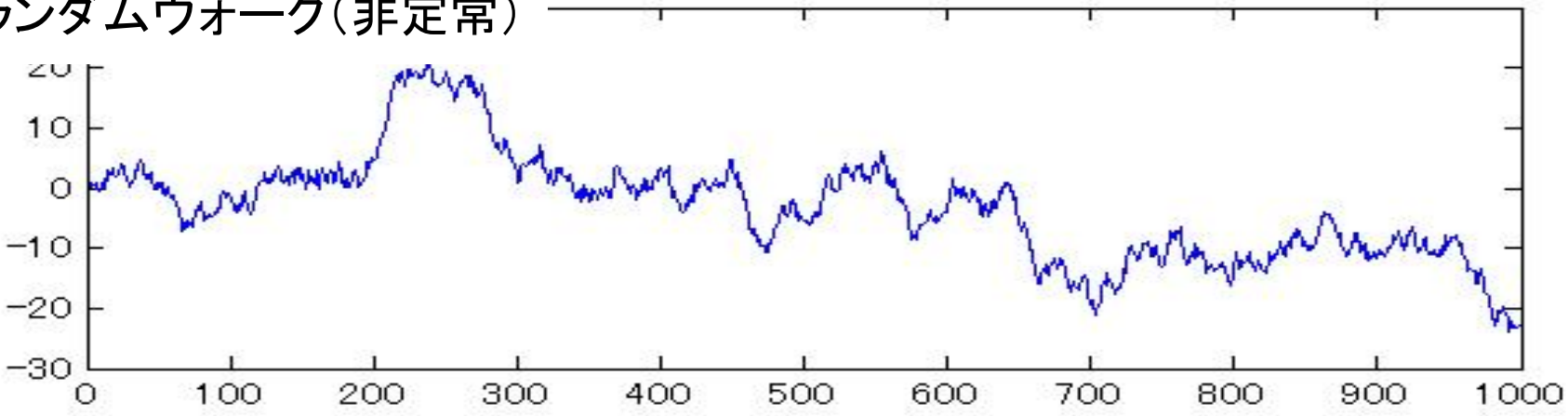
○　定常であるかどうかを調べたい場合には定常性をチェックする検定法もある。

例

正規雑音(白色雑音：定常）

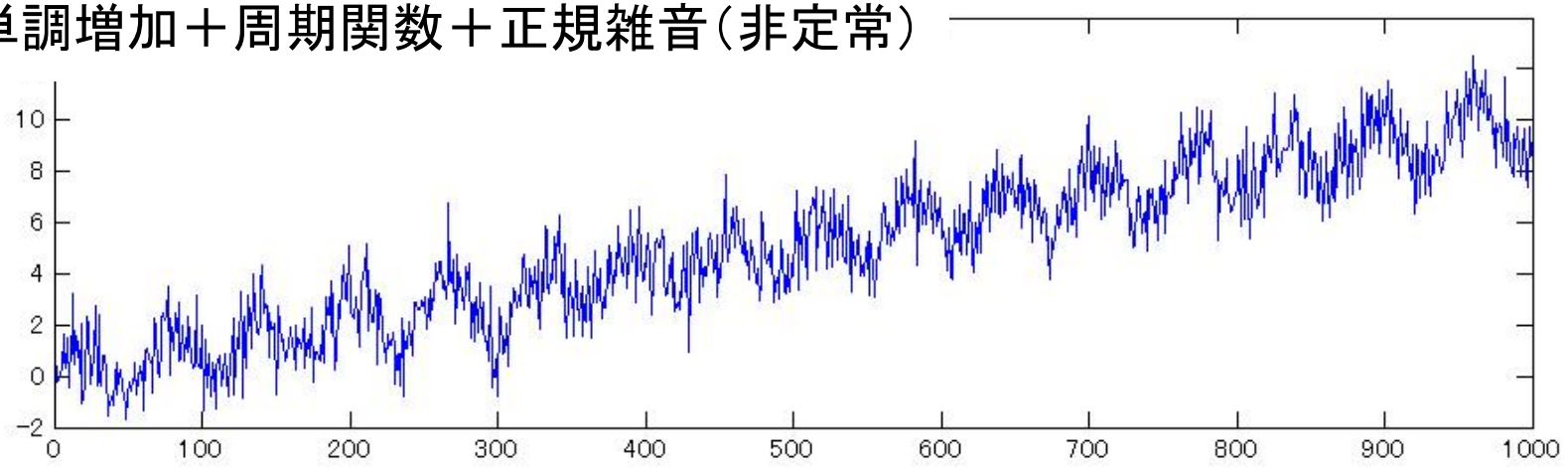


ランダムウォーク（非定常）

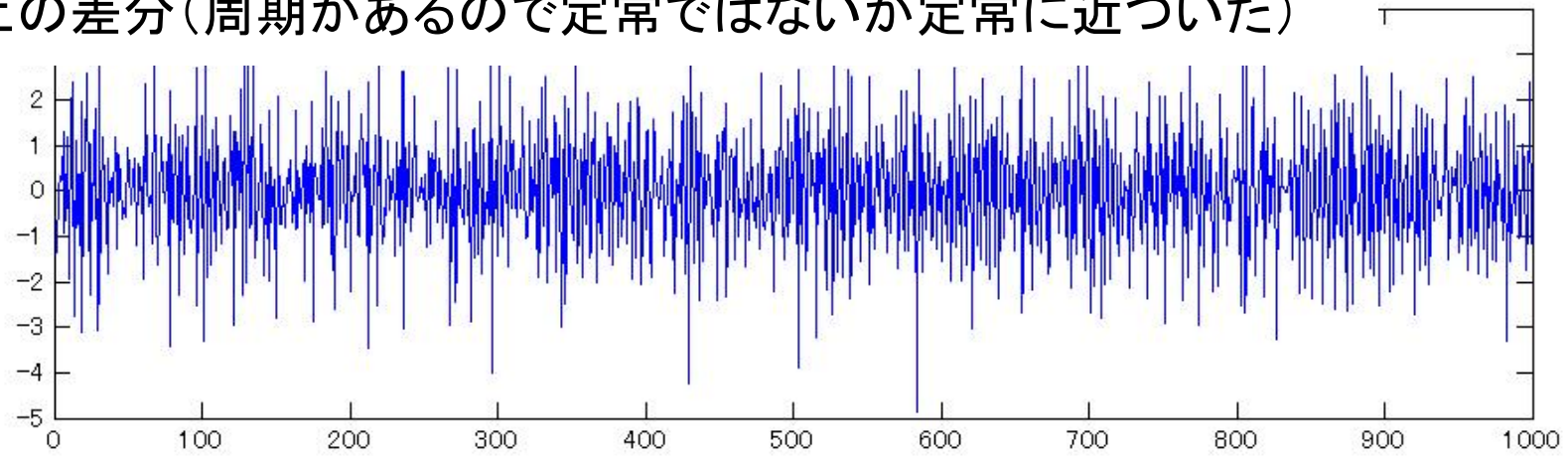


例

単調増加＋周期関数＋正規雑音（非定常）



上の差分（周期があるので定常ではないが定常に近づいた）



時系列の解析

Z変換と逆変換

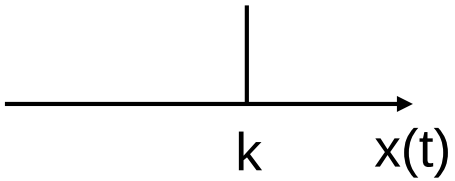
時系列 x(t) (t は整数) を考える。複素数 $z \neq 0$ の関数 (Fx)(z) を

$$(Fx)(z) = \sum_{t=-\infty}^{\infty} z^{-t}\, x(t)$$

定義して (Fx)(z) を x(t) のZ変換という。x(t) から (Fx)(z) への対応は線形写像である。特に $z=\exp(ik)$ や $z=\exp(k)$ の場合はフーリエ変換やラプラス変換になるので、Z変換はそれらの一般化である。逆変換

$$(F^{-1}f)(n) = \frac{1}{2\pi i} \oint z^{n-1} f(z)\, dz$$ （積分路はf(z)の極を含む）

が存在する。



(例1） $x(t) = \delta_{tk}$ のとき（$(Fx)(z) = z^{-k}$.

(例2） $y(t) = x(t-n)$ のとき $(Fy)(z) = z^{-n}\,(Fx)(z)$ 。

Z変換の応用例

〇Z変換により、「時間の移動」と「z^{-1} をかけること」が等価になる。

(例） 与えられた y(t) に対して $y(t) = x(t) + 2x(t-1) + 3x(t-2)$ を満たす x(t) を求めたい時はZ変換したものの関係

$$(Fy)(z) = (1+2z^{-1}+3z^{-2})\,(Fx)(z)$$

から $1/(1+2z^{-1}+3z^{-2})\,(Fy)(z) = (Fx)(z)$ を導出し、逆Z変換を適用することで

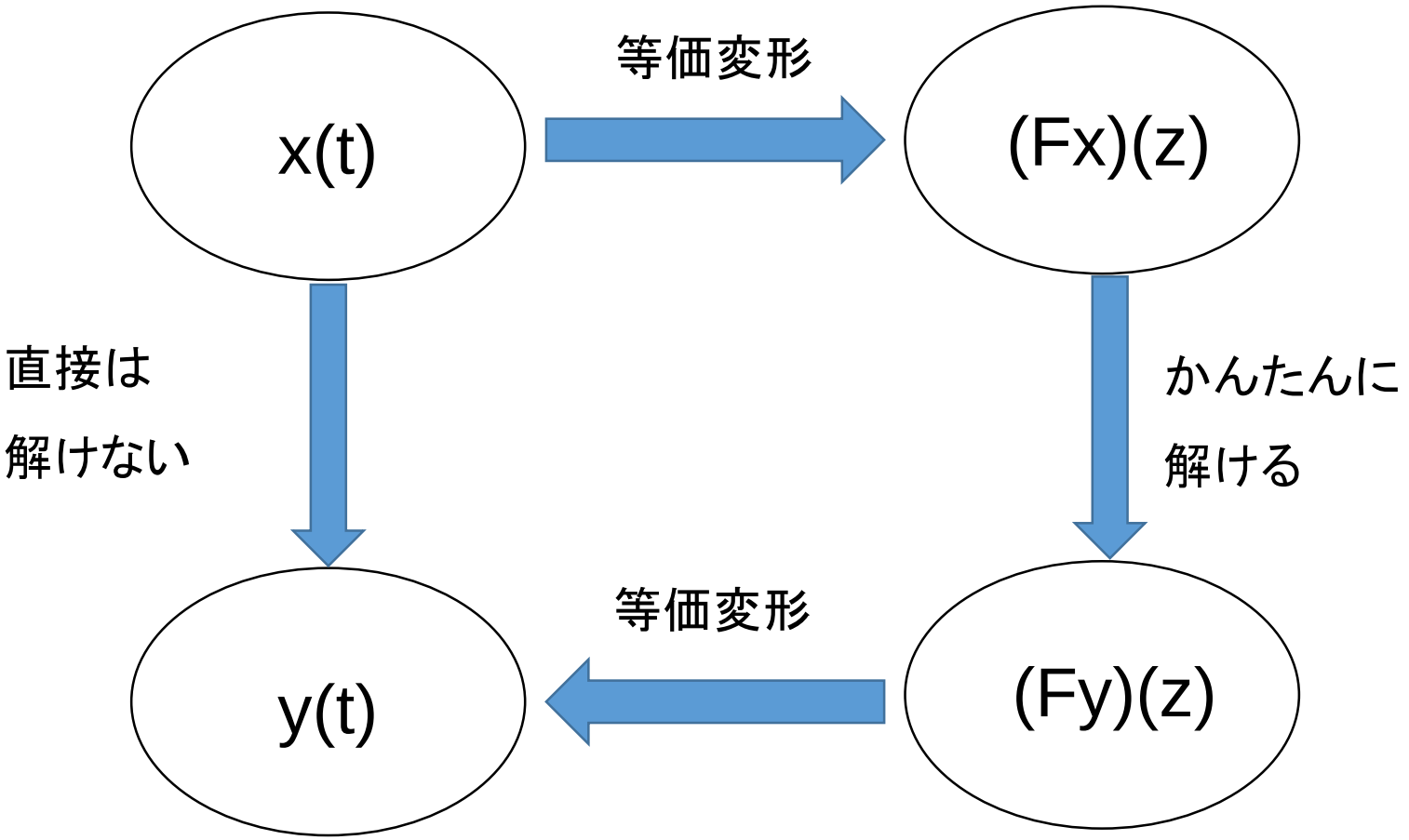
$$x(t) = F^{-1}\{\,(Fy)(z)\,/(1+2z^{-1}+3z^{-2})\,\}\,(t)$$

を得る。

「可解」な問題の構造

数学における「解ける問題」は下記の構造を持つことが多い。

例：行列のn乗、非線形可解方程式、統計的漸近理論など。



時系列モデル

実世界にある時系列は無限に複雑であることが多いが、
これを性質の良く分かっているモデルを用いて推測する
ことが行なわれている。まずモデルについて説明する。

自己回帰過程(AR, Auto Regressive Process)

時刻 t における X(t) はそれよりも K 時刻前までの重みつき和と雑音で生成される時系列のモデルである:

$$X(t) = -\sum_{k=1}^{K} a_k X(t-k) + e(t)$$

ここで $\{a_1,\ldots,a_K\}$ は定数である。次のように書くことができる。

$$X(t) + \sum_{k=1}^{K} a_k X(t-k) = e(t)$$

注意： 自己回帰過程(AR)

Z変換を計算すると $(1+\sum_{k=1}^{K} a_k Z^{-k})(Fx)(z) = (Fe)(z)$ が得られる。

K次代数方程式（zを変数とする） $1+\sum a_k Z^{-k}=0$ を特性多項式という。

代数方程式は複素数の範囲で多重度を含めてK個の解を持っている（代数学の基本定理）ので、$1+\sum a_k Z^{-k}$は因数数分解できる。その解の絶対値がすべて1よりも小さければ、分数関数$1/(1+\sum a_k Z^{-k})$ の無限部分分数展開は収束するが、そうでない場合は収束しない。このことは、対応「 $e(t) \rightarrow x(t)$ 」が発散し、定常性を持たないことを意味する。

そこで、自己回帰過程では（定常性を満たす必要がある場合には）上記の特性方式の解の絶対値がすべて1よりも小さいことを要請する。

問1

次の方程式を満たす時系列が定常であるような A の範囲を求めよ。

$$X(t) + A\,x(t-1) - 0.25\,x(t-2) - 0.25A\,x(t-3) + e(t) = 0$$

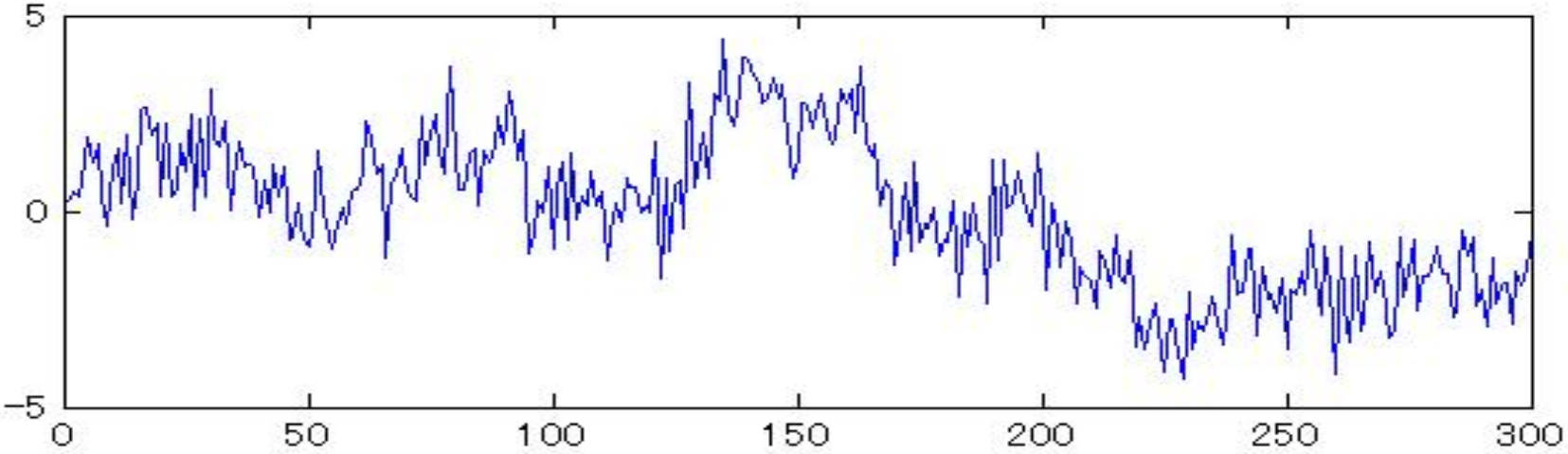
次の方程式を満たす時系列が定常であるような A の範囲を求めよ。

ただし次の恒等式を用いてもよい。

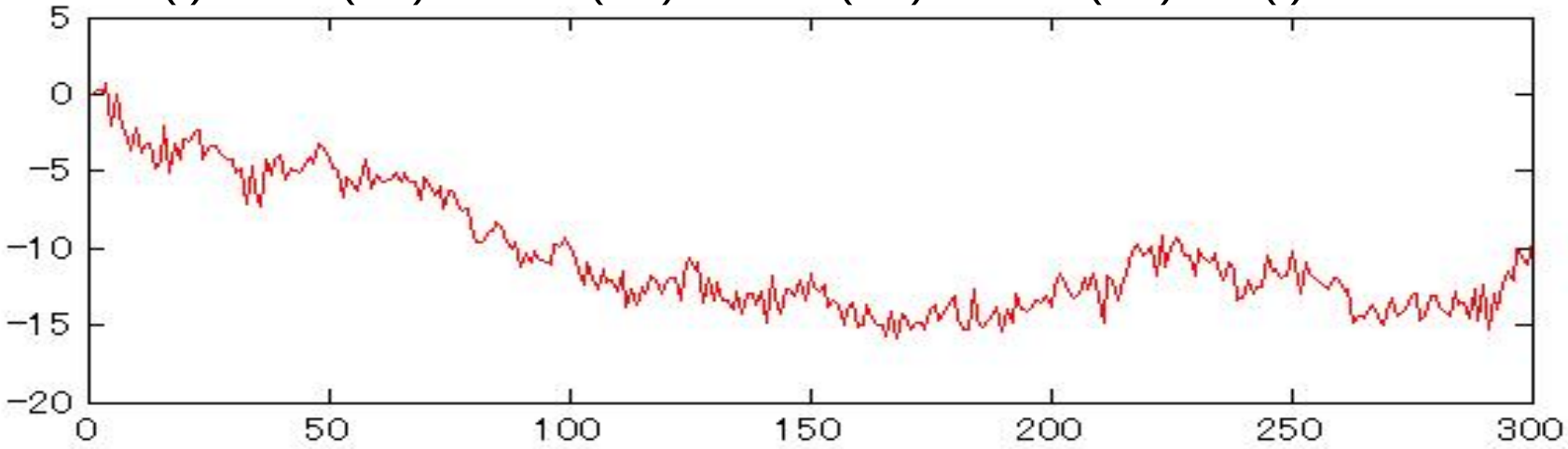
$$z^3 + A\,z^2 - 0.25\,z - 0.25A = (z - 0.5)(z+0.5)(z+A)$$

自己回帰過程の例

X(t)=0.3X(t-1) + 0.3X(t-2) + 0.2X(t-3) + 0.1X(t-4) + e(t)

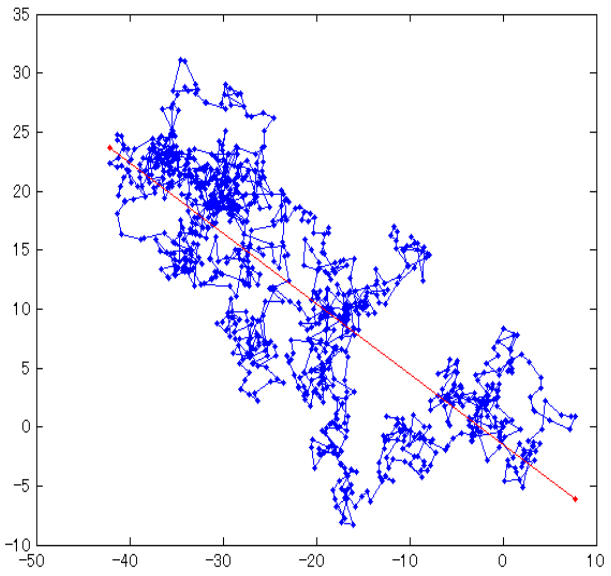


X(t)=0.3X(t-1) + 0.3X(t-2) + 0.3X(t-3) + 0.1X(t-4) + e(t)



注意： 単位根の問題

AR過程で特性方程式の解の絶対値が1（単位根）を含む場合には、定常過程ではなく、ランダムウォークになる。単位根を持つ二つの過程 X(t) と Y(t) があるとき、X(t) から Y(t) への回帰を行なうと、本来は存在しない回帰関係が存在するように見えてしまうことからしばしば実問題における解析ミスの原因になることが知られている。



移動平均過程 (MA, Moving Average Process)

時刻 t における X(t) はそれよりも K 時刻前までの雑音 e(t)の

重みつき和で生成される時系列のモデルである:

$$X(t) = \sum_{k=1}^{K} b_k \, e(t-k) + b_0 e(t)$$

ここで $\{b_0, b_1, \ldots, b_K\}$ は定数である。

注意： 移動平均過程(MA)

Z変換を計算すると $(Fx)(z) = (b_0 + \sum_{k=1}^{K} b_k Z^{-k})(Fe)(z)$ が得られる。

今度は、もしも特性方程式 $(b_0 + \sum_{k=1}^{K} b_k Z^{-k}) = 0$ の解の絶対値が1より

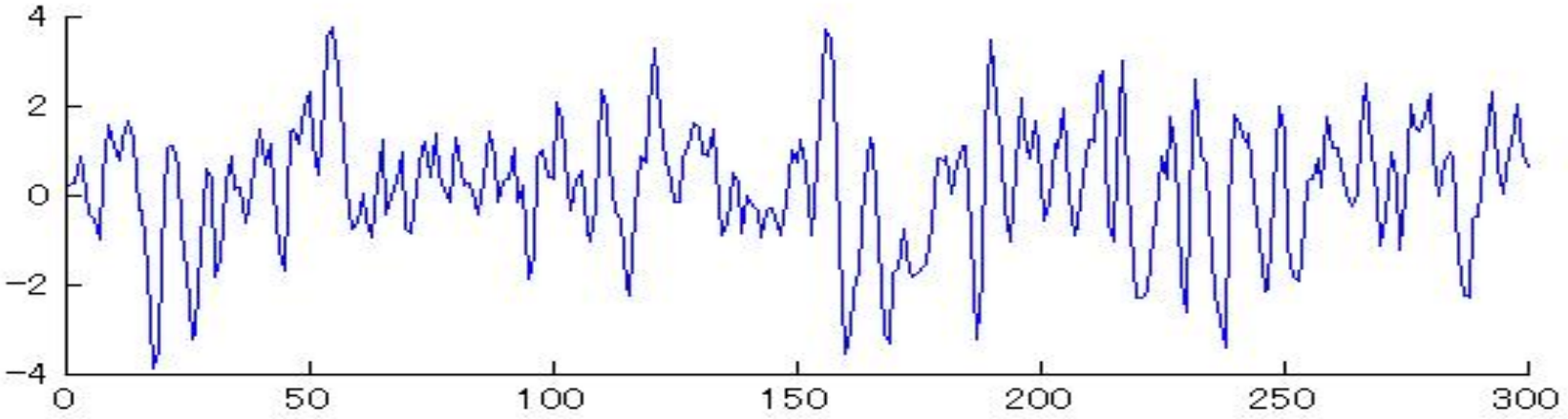
小さいものがあると対応「 $x(t) \rightarrow e(t)$ 」が発散し、$x(t)$ から $e(t)$ が

安定的に推測できないことを意味する。そこで、上記の特性方式の解の

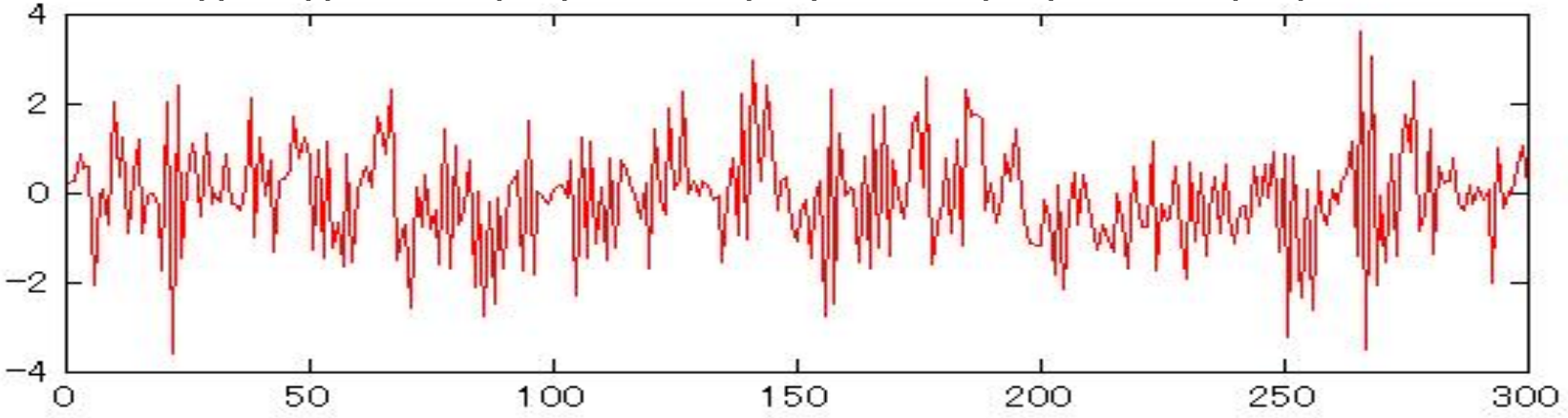
絶対値がすべて1よりも大きいことを要請する。

移動平均過程の例

X(t)=0.5e(t)+0.5e(t-1) + 0.5e(t-2) + 0.5e(t-3) + 0.5e(t-4)



X(t)=e(t) - 0.8e(t-1) + 0.6e(t-2) - 0.4e(t-3) + 0.2e(t-4)



自己回帰移動平均過程（ARMA)

AR過程とMV過程の両方を持つ時系列のモデルである:

$$X(t) + \sum_{k=1}^{K} a_k X(t-k) = \sum_{k=1}^{K} b_k e(t-k) + b_0 e(t)$$

ここで $\{a_1,\ldots,a_K,b_0,b_1,\ldots,b_K\}$ は定数である。

さらに差分するとARMA過程になるような時系列をARIMA過程という。

(Auto Regressive Integrated Moving Average Process).

時系列の推定

実世界の時系列はARIMAなどのモデルで記述できる
ものではない。性質がよくわかっている時系列モデルを
使って実世界にある時系列の予測を行なうことで
どんなことがどのくらいにできるのだろうか。

自己回帰過程を用いた推定

AR過程を用いて推測する場合、二乗誤差を最小にすることでパラメータの推定を行なうことができる。$X(t) = \Sigma\ a_k X(t-k)$ (符号注意）の場合には

$$\sum_{t=K+1}^{T} \{ X(t) - \sum_{k=1}^{K} a_k X(t-k) \}^2 \text{ を } \{a_1,\ldots,a_K\} \text{ について最小化}$$

○　最小二乗法で得られるパラメータは、T が無限大に近づく時、真の時系列に対して最適なパラメータに収束するか？

一般的には難しい問題であるが、{X(t)}が正規分布であるときには、「$n\to\infty$のとき$E[X(t)X(t+n)]\to 0$」が十分条件である。

計算手順

(1) { X(t) ; t=1,2,…,T} を得る。

(2) $\sum_{t=K+1}^{T} \{ X(t) - \sum_{k=1}^{K} a_k X(t-k) \}^2$ を $\{a_1,\ldots,a_K\}$ について最小化

行列 $S_{km} = \sum_{t=K+1}^{T} X(t-k)\, X(t-m)$ $(1 \leqq k,m \leqq K)$

ベクトル $v_k = \sum_{t=K+1}^{T} X(t)\, X(t-k)$ $(1 \leqq k \leqq K)$

とおくと、推定されたパラメータは $a^* = S^{-1}v$

(3) 予測式 $X(t) = \sum_{k=1}^{K} (a^*)_k X(t-k)$ を得る。

(4) 方程式 $1 - \sum (a^*)_k z^{-k} = 0$ (符号注意)の解の絶対値で1よりも大きいものがあるときは推定された時系列は定常ではない。(定常でなくても予測はできるが注意が必要である)。

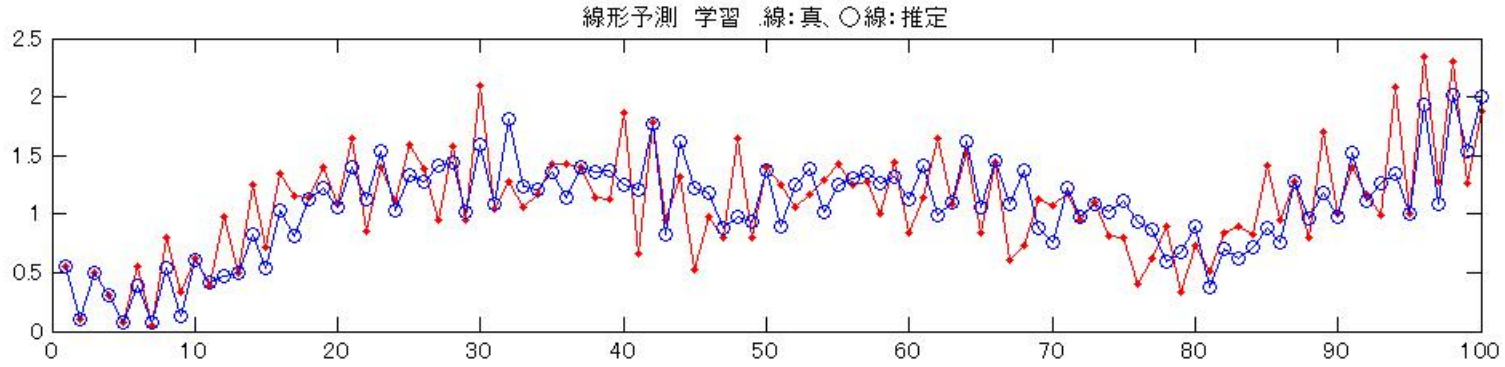
AR過程以外を用いた推定

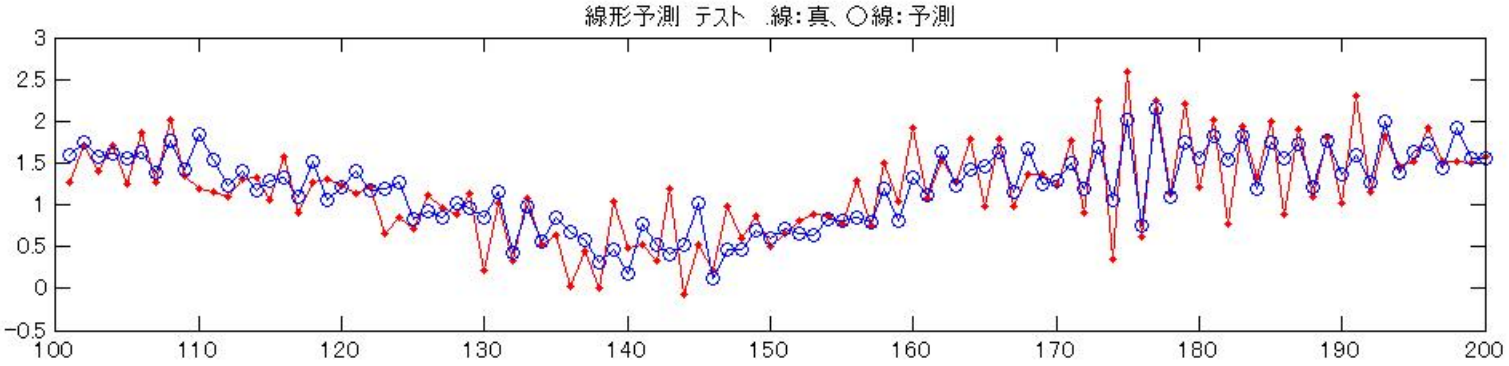
MV, ARMA,ARIMA においてはパラメータの推定をデータから解析的に行なうことはできない。また最小二乗法でない方法についても研究がなされている。

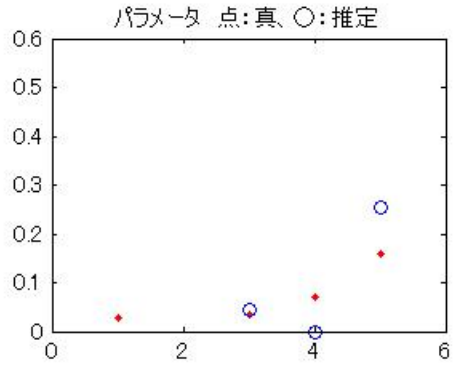
最近の動向

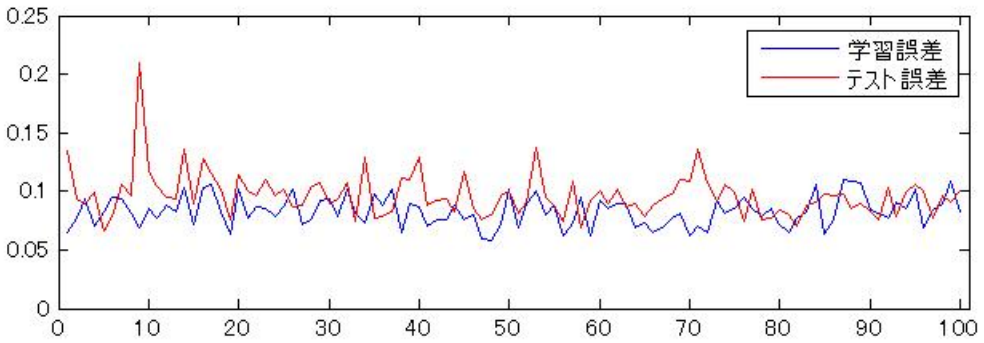
時系列の予測に深層学習などの新しい方法が適用されるようになってきた。

線形予測（5次式）


線形予測　学習　.線：真、○線：推定
2.5
2
1.5
1
0.5
0
0
10
20
30
40
50
60
70
80
90
100


線形予測　テスト　.線：真、○線：予測
3
2.5
2
1.5
1
0.5
0
−0.5
100
110
120
130
140
150
160
170
180
190
200


パラメータ　点：真、○：推定
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6


0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100
学習誤差
テスト誤差

線形予測（5次式）

問2

$$\sum_{t=3}^{T} \{ X(t) - a_1 X(t-1) - a_2 X(t-2) \}^2$$ を最小にする $\{a_1, a_2\}$ を求めよ。

例

ある月の白菜の値段を、それよりも30ヶ月前までの白菜の値段を
用いて予測するための関数

$$x(t) = f(x(t-1),x(t-2),\ldots,x(t-30))$$

を推定するという問題を考える（30よりも少ない場合も含めて考える）。

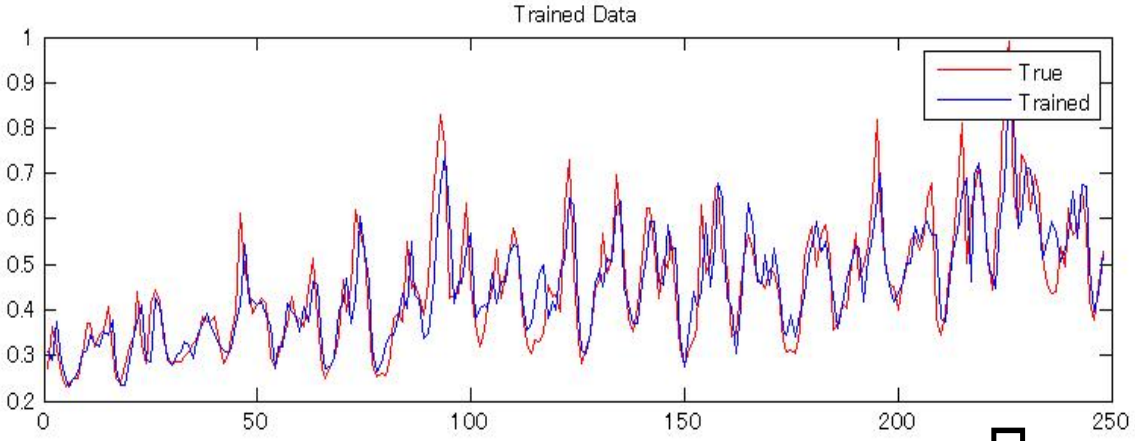
時系列予測の例：1970年1月から2013年12月までの白菜の値段
「政府統計の総合窓口」のデータを使用しています。
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

（注意） 時系列では、季節、曜日、昼夜の影響を受けると分かっている
場合にはそれらの情報も入力に使うとより精度の良い予測ができますが、
ここでは単純な予測問題を扱います。

例 白菜の値段を　AR過程(30)で予測

定常ではないようです・・・。たぶんバブルの前と後で法則が変化していると思います。
しかし、これは演習なので、AR過程でどのくらい予測できるかやってみましょう。
特性方程式の解の最大絶対値は 1.0023 でした（定常ではない）。

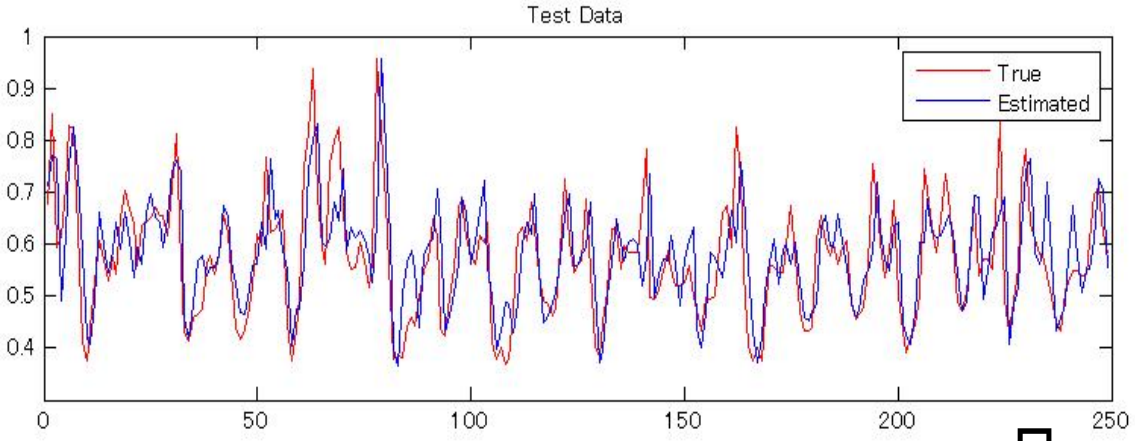
値段



月

学習データ
250個
赤：真
青：予測

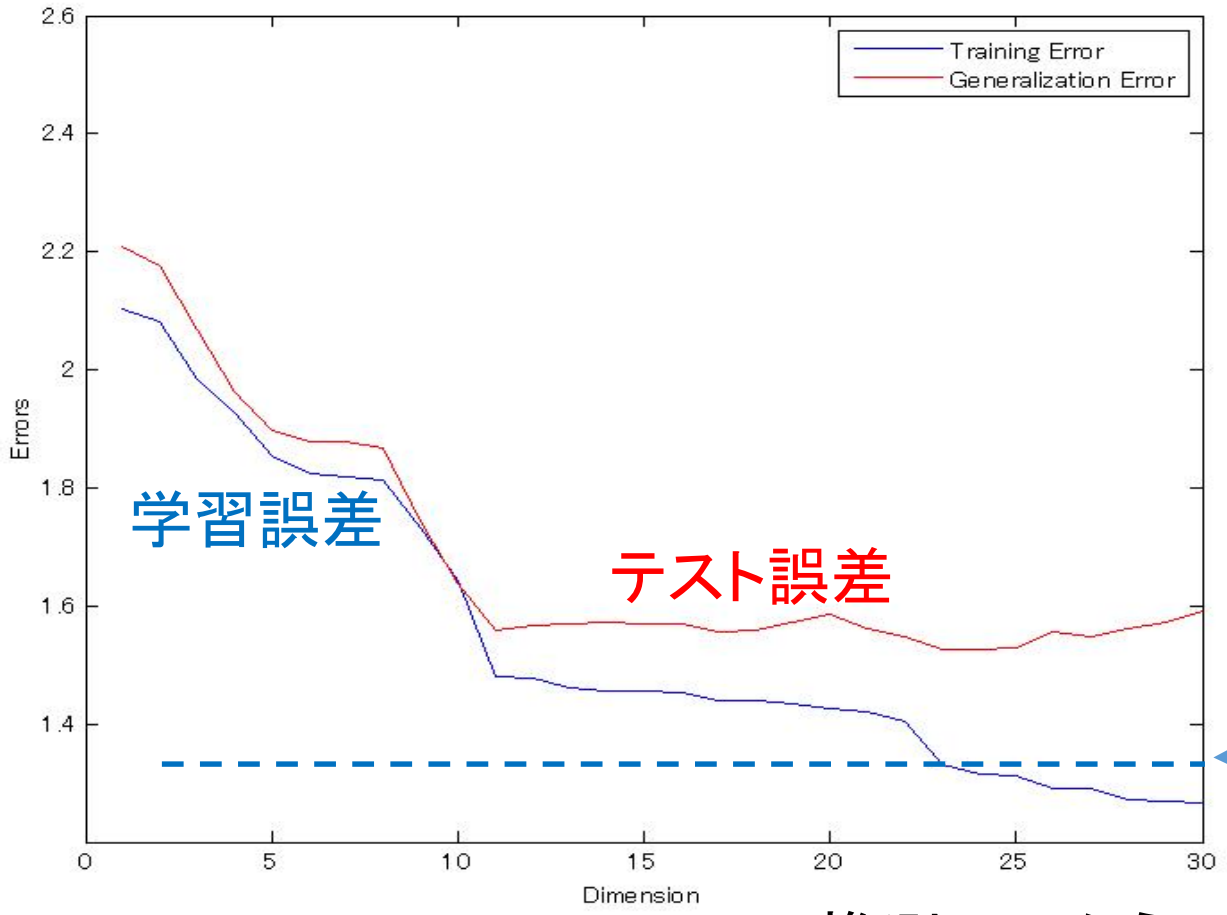
値段



月

テストデータ
250個
赤：真
青：予測

どのくらい過去の月まで使うのがよいか



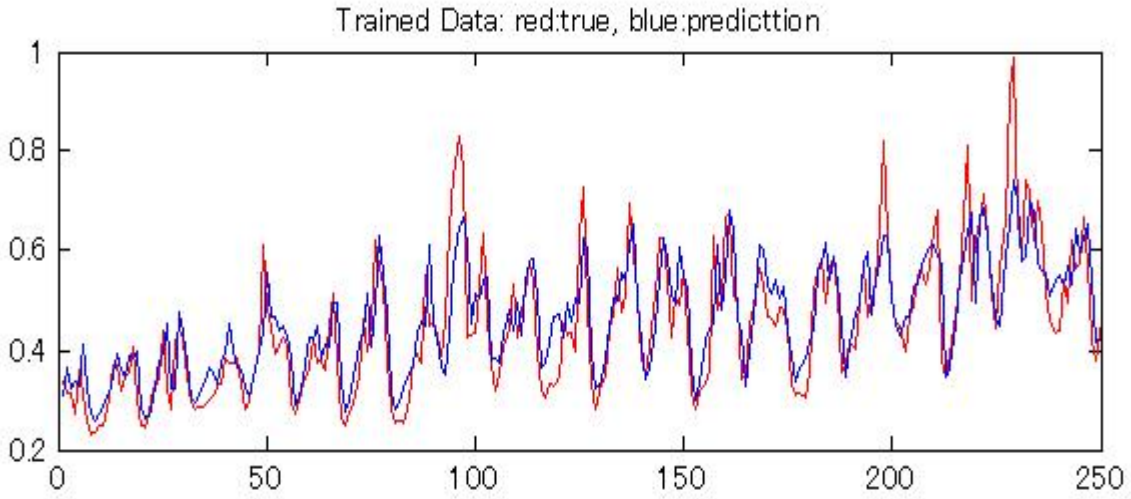
AR(次元) 解析
学習誤差
とテスト誤差

特性方程式の
解の絶対値の
最大値が
1よりも小さく
なるのは9次元
までだった。

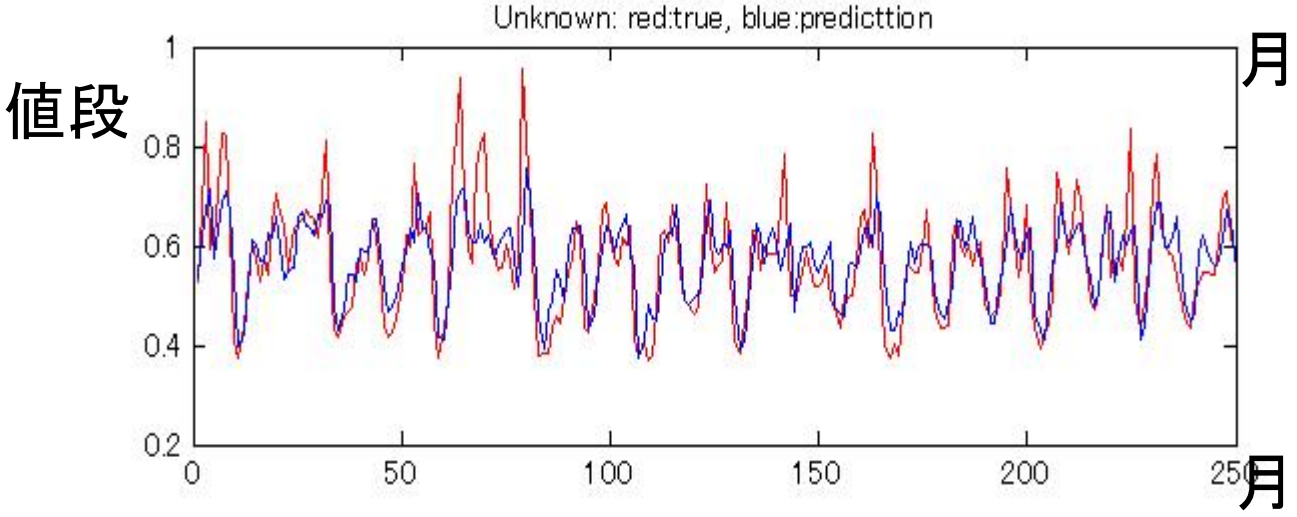
深層学習
による
テスト
誤差

推測につかう
モデルの次元

参考： 深層学習で学習し予測してみた

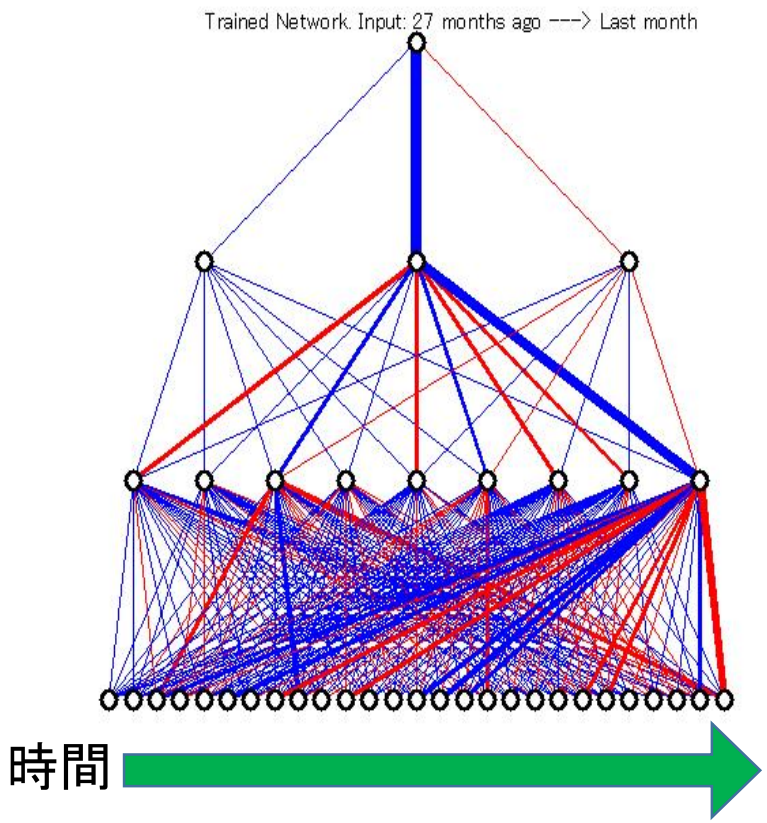


学習データ
赤：真
青：予測

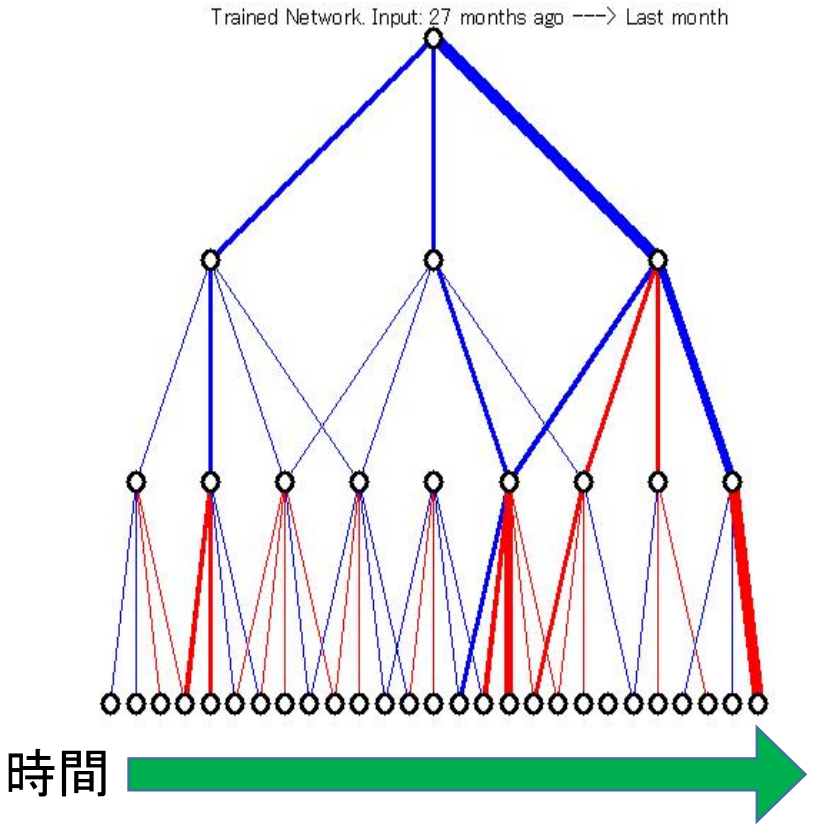


テストデータ
赤：真
青：予測

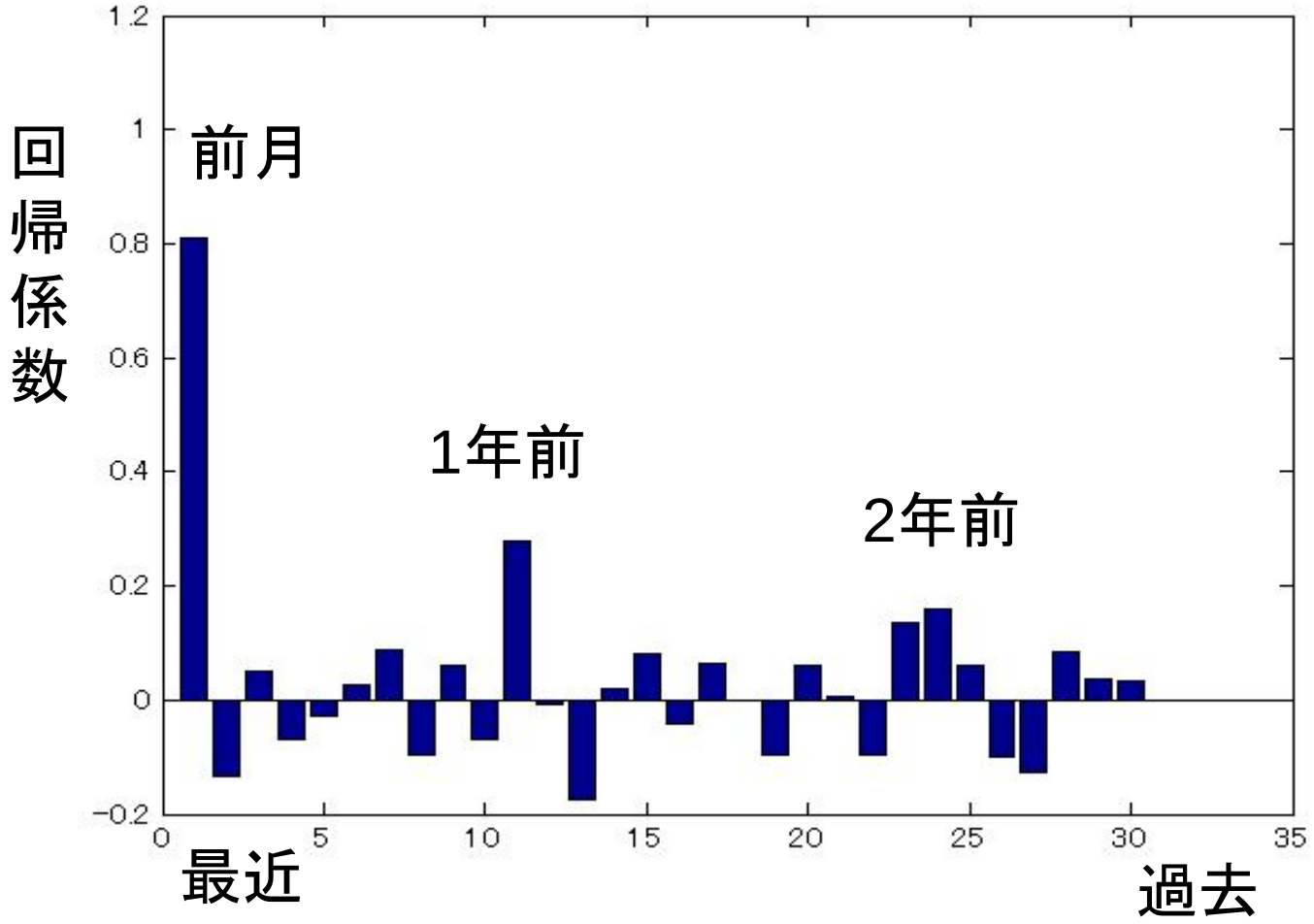
全結合



畳み込みネットワーク



（注）第１層から第２層までの結合の初期値に線形予測で得られた
結合加重を用いています。

問3

白菜の値段の線形予測における回帰係数は下記のようになった。
ここからわかることを述べよ。