統計データ解析
第一回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義について

- 担当：加藤由幹（A423）
- 進め方
  - 基本的には資料を投影して進めます．板書もあります
  - 使用した資料はPDFでKUTLMSにアップロードします
  - ノートPCを使用するので，持参してください
  - 出欠はカードシステムで取ります
- 到達目標と成績評価



到達目標

1. 確率分布の性質を理解し，実験データを正しくグラフに表現し評価できる．
2. 回帰分析やフィルタ処理を用いてデータから知見を抽出できる．
3. 主成分分析や特異値分解を用いて多変量データを解析し，知見を抽出できる．

成績評価の方法・基準

試験は期末試験のみとする．
演習40%，期末試験60%とし，最終成績は得点率で評価する．

※演習（提出課題）は2回を予定しています

- その他

※補講が無い場合の休講は全員出席にします

体調不良や学会等による欠席は認める．最低でも10回以上は出席すること
大学院は研究が主なので，大量の課題等は出さないが，各自自習を怠らないようにすること
数学の基礎知識は各自持っているものとして進めます．（忘れている人は自習復習）

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

授業計画

第1回　統計データの性質とMatlabを用いたデータ解析の基礎
第2回　確率分布の性質
第3回　グラフと確率分布を用いた実験データの可視化
第4回　さまざまな分布
第5回　大数の法則と中心極限定理
第6回　時系列データ解析とフィルタ処理
第7回　時系列データ解析とフィルタ処理2
第8回　画像解析とフィルタ処理
第9回　最小二乗法，回帰分析
第10回　最尤推定，ベイズ推定
第11回　変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ
第12回　スパースモデリング
第13回　多変量データ解析の基礎と主成分分析
第14回　固有値分解，特異値分解
第15回　期末テスト
第16回　テスト返却と解説

統計解析の基礎
1次元，2変数
→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析
→ビッグデータ解析に必須のスキル

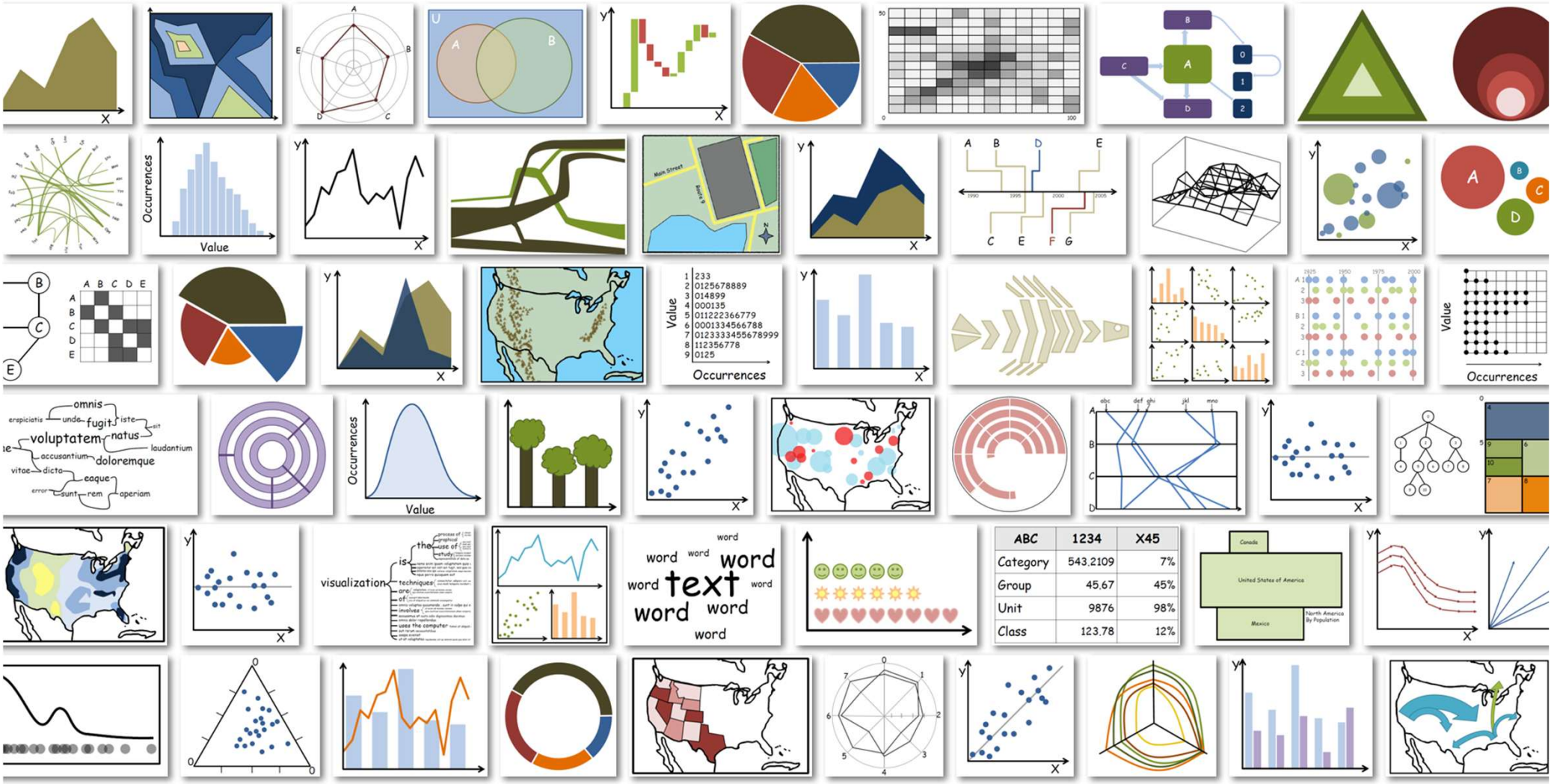
© 2020 Yuki Kato. All Rights Reserved

一週目

- 統計の復習（CLで見たことがあるはず・・）
  - グラフ作成について
  - 統計学について
  - 2変数データの分析と相関

© 2020 Yuki Kato. All Rights Reserved

What makes a data visualization memorable?



Different types of visualizations may be intrinsically more or less memorable, according to a team of researchers from Harvard and MIT. Credit: Image courtesy of Michelle Borkin, Harvard SEAS.

Read more at: http://phys.org/news/2013-10-visualization.html#jCp

プロットグラフの書き方

- 基本は見やすく。
- 観測者の推測をできるだけ入れない。

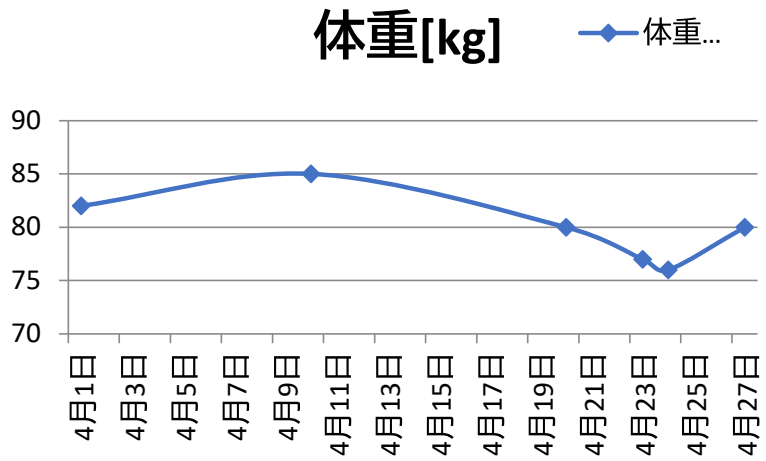


図1. 体重の時間変化
（最悪な例：醜い。見にくい）

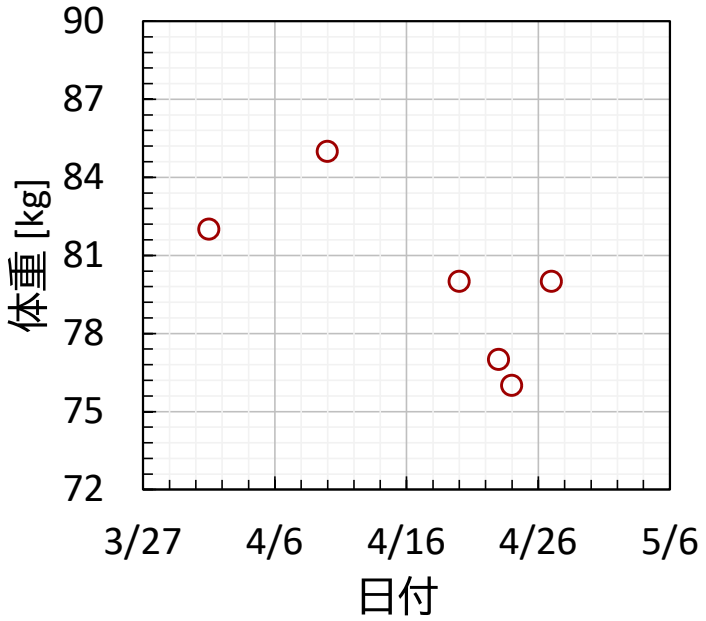


図2. 体重の時間変化
（工学的に正しいプロットの例）

グラフ作成の重要性：
レポート/論文でグラフを使う注意

- レポート/論文は、客観データに基づき、自分の主張をする文書

客観データ・・・・・反論できないデータ、数値
グラフはデータを図に表したもの

- つまり、「グラフには、主観を入れない。」

×印象操作するグラフ

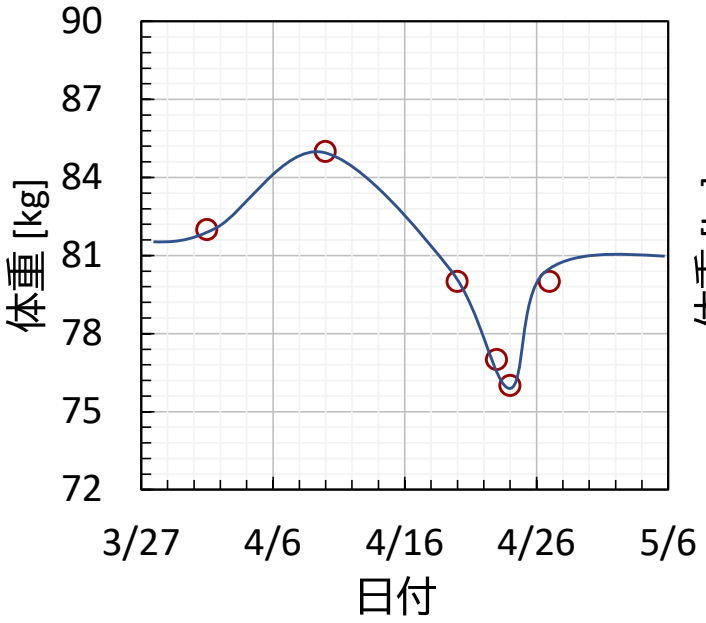




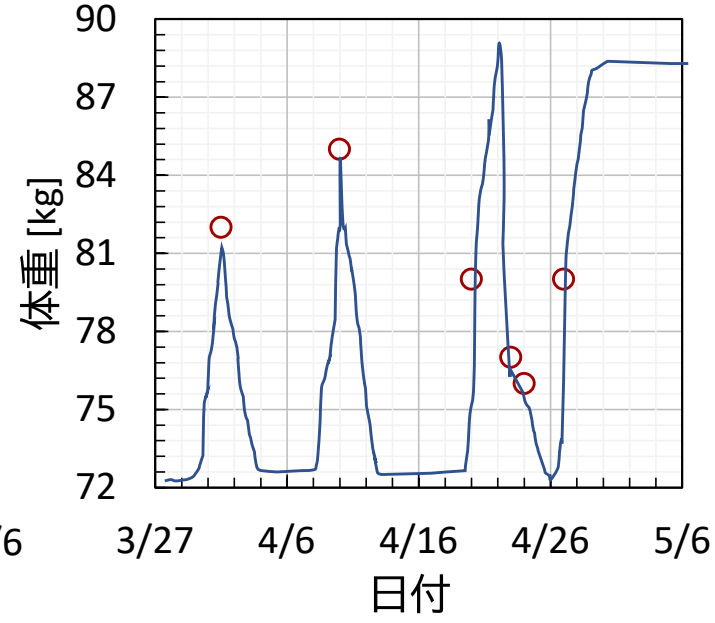


- グラフを傾けているので、手前のデータが大きく「見える」
  ・・・グラフを操作して、主張をしている。

観測者の推定　（＝主観）をできるだけ入れない



こういう近似線を入れたくなるけど



ほんとはこうかもしれない。

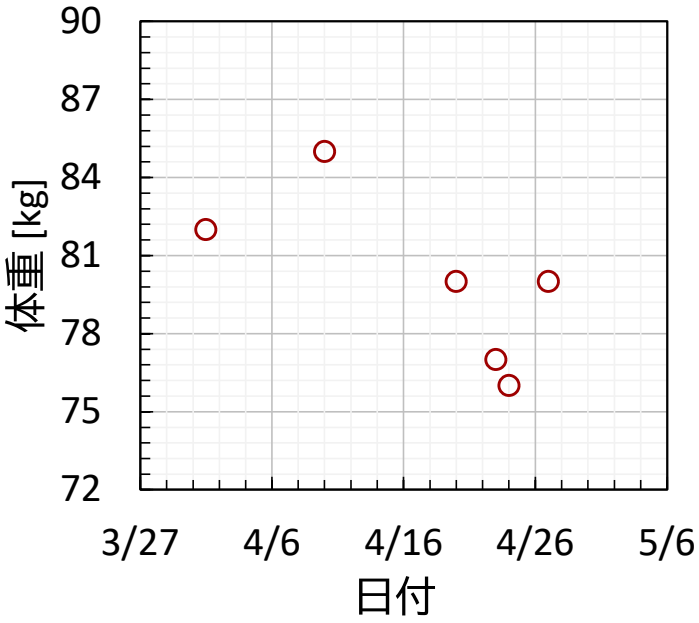


図2. 体重の時間変化
（工学的に正しいプロット例）

観測者の推定をプロットにできるだけ入れない：半導体ＰＮ接合における量子トンネル効果の発見

江崎玲於奈

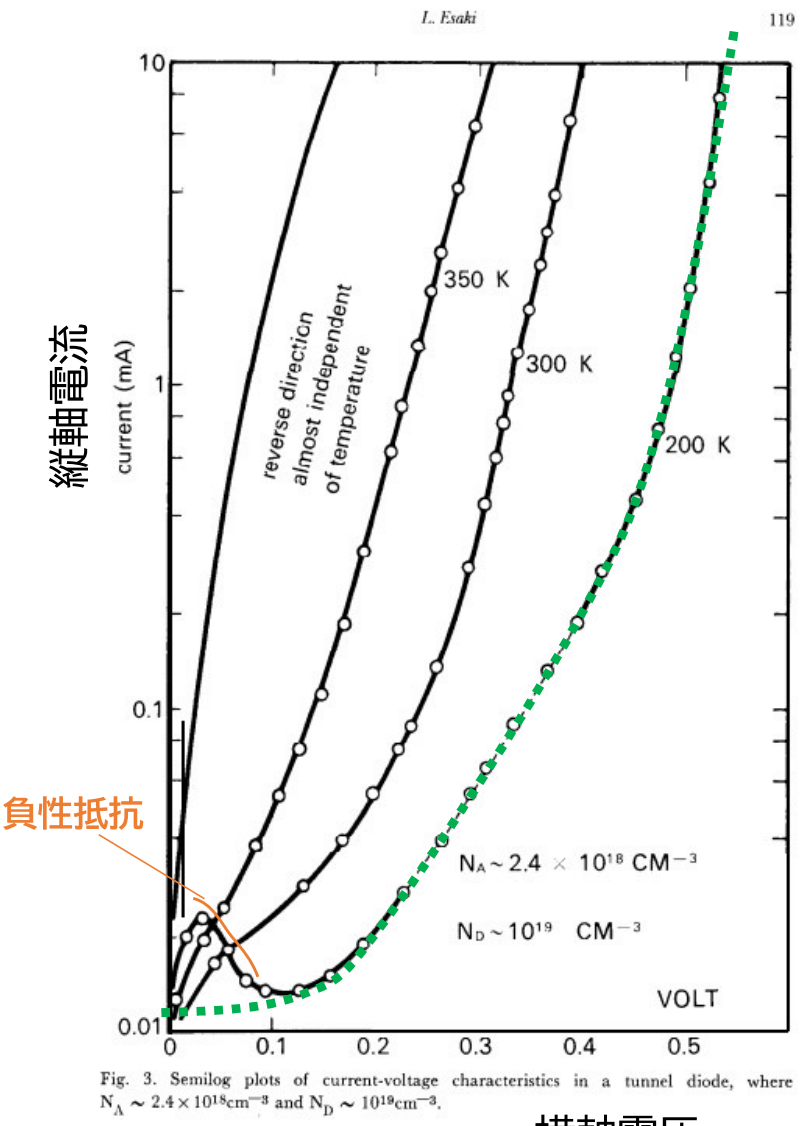


Fig. 3. Semilog plots of current-voltage characteristics in a tunnel diode, where $N_A \sim 2.4 \times 10^{18}$cm^{-3} and $N_D \sim 10^{19}$cm^{-3}.

縦軸電流

負性抵抗

横軸電圧



PN接合ダイオードにおいて，電圧を増したときに流れる電流が減少する「負性抵抗」を発見。

実験しながらグラフを描く。

0.1Vおきに0.1V、0.2V、0.3Vと測定し終えた後、実験道具を片付けた後にグラフを書くと、あとからデータ点が増やせない。

大発見を見逃さないためにも、安易にデータ点を曲線で結ばない。

0~0.1Vのデータ点を、測定ミスと考えて、図中緑点線のように曲線を引いてしまったら、負性抵抗を見逃してしまう。

曲線を引くときは、曲線で結んでも問題が起こらないか十分に注意する。曲線に見せたいのならデータ点数を増やす。

https://www3.nhk.or.jp/news/special/nobelprize/2023/physics/article_14.html

https://www.shmj.or.jp/museum2010/exhibi302.htm

一週目

- 統計の復習（CLで見たことがあるはず・・）
  - グラフ作成について
  - 統計学について
  - 2変数データの分析と相関

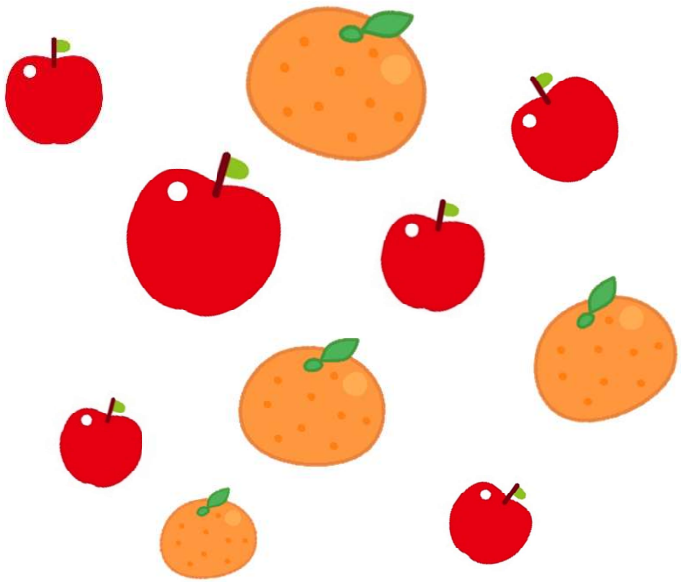
© 2020 Yuki Kato. All Rights Reserved

統計学とは？

- データを統計量（平均など）や図・表にまとめて、その特徴を捉える学問・・・栗原伸一・丸山敦史　統計学図鑑　オーム社
- 現象の集団的把握を目的とする学問・・・ブリタニカ百科事典

データ

わかりにくい

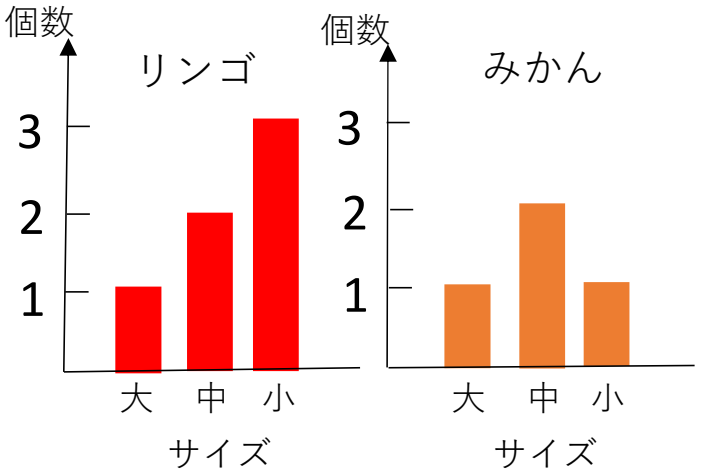


表

果物	個数
リンゴ	6
みかん	4

図

ヒストグラム



図や表にまとめるとわかりやすい

・統べる・・・多くのものを一つにまとめる

・計る・・・・ある基準をもとにして物の度合いを調べる。測定する。

ページ　ノート　　閲覧　編集　履歴表示　Wikipedia内を検索

ウィキペディア
フリー百科事典

メインページ
コミュニティ・ポータル
最近の出来事
新しいページ
最近の更新
おまかせ表示
練習用ページ
アップロード (ウィキメディア・コモンズ)

ヘルプ
ヘルプ
井戸端
お知らせ
バグの報告
寄付
ウィキペディアに関するお問い合わせ

ツール
リンク元
関連ページの更新状況
ファイルをアップロード
特別ページ
この版への固定リンク

中立的な観点の方針の大幅改訂に関して、調査投票を実施中です。　[非表示]

フローレンス・ナイチンゲール

出典: フリー百科事典『ウィキペディア（Wikipedia）』

出典は列挙するだけでなく、**脚注**などを用いて**どの記述の情報源であるかを明記**してください。記事の信頼性向上にご協力をお願いいたします。（2015年12月）

フローレンス・ナイチンゲール（Florence Nightingale、1820年5月12日 - 1910年8月13日）は、イギリスの看護婦、社会起業家、統計学者、看護教育学者。近代看護教育の母。病院建築でも非凡な才能を発揮した。クリミア戦争での負傷兵たちへの献身や統計に基づく医療衛生改革で著名[1]。国際看護師の日（5月12日）は彼女の誕生日である。

ロンドンの聖トーマス病院に付属してナイチンゲール看護学校を設立、これは世界初の宗教系でない看護学校であり、現在はキングス・カレッジ・ロンドンの一部となっている。

ギリシア哲学についても造詣が深く、オックスフォード大学のプラトン学者、ベンジャミン・ジョウェット（英語版）とも親しく交流した。

目次 [非表示]

1 人物
　1.1 青年期
　1.2 クリミア戦争へ

フローレンス・ナイチンゲール



生誕　1820年5月12日
トスカーナ大公国フィレンツェ

フローレンス・ナイチンゲール　(1820-1910)

ページ　ノート　　閲覧　編集　履歴表示　Wikipedia内を検索

ウィキペディア
フリー百科事典

メインページ
コミュニティ・ポータル
最近の出来事
新しいページ
最近の更新
おまかせ表示
練習用ページ
アップロード (ウィキメディア・コモンズ)

ヘルプ

ヘルプ
井戸端
お知らせ
バグの報告
寄付
ウィキペディアに関するお問い合わせ

ツール

リンク元
関連ページの更新状況
ファイルをアップロード
特別ページ
この版への固定リンク

中立的な観点の方針の大幅改訂に関して、調査投票を実施中です。　[非表示]

フローレンス・ナイチンゲール

出典: フリー百科事典『ウィキペディア（Wikipedia）』

出典は列挙するだけでなく、脚注などを用いてどの記述の情報源であるかを明記してください。記事の信頼性向上にご協力をお願いいたします。（2015年12月）

フローレンス・ナイチンゲール（Florence Nightingale、1820年5月12日 - 1910年8月13日）は、イギリスの看護婦、社会起業家、統計学者、看護教育学者。近代看護教育の母。病院建築でも非凡な才能を発揮した。クリミア戦争での負傷兵たちへの献身や統計に基づく医療衛生改革で著名[1]。国際看護師の日（5月12日）は彼女の誕生日である。

ロンドンの聖トーマス病院に付属してナイチンゲール看護学校を設立、これは世界初の宗教系でない看護学校であり、現在はキングス・カレッジ・ロンドンの一部となっている。

ギリシア哲学についても造詣が深く、オックスフォード大学のプラトン学者、ベンジャミン・ジョウェット（英語版）とも親しく交流した。

目次 [非表示]

1 人物
　1.1 青年期
　1.2 クリミア戦争へ

フローレンス・ナイチンゲール

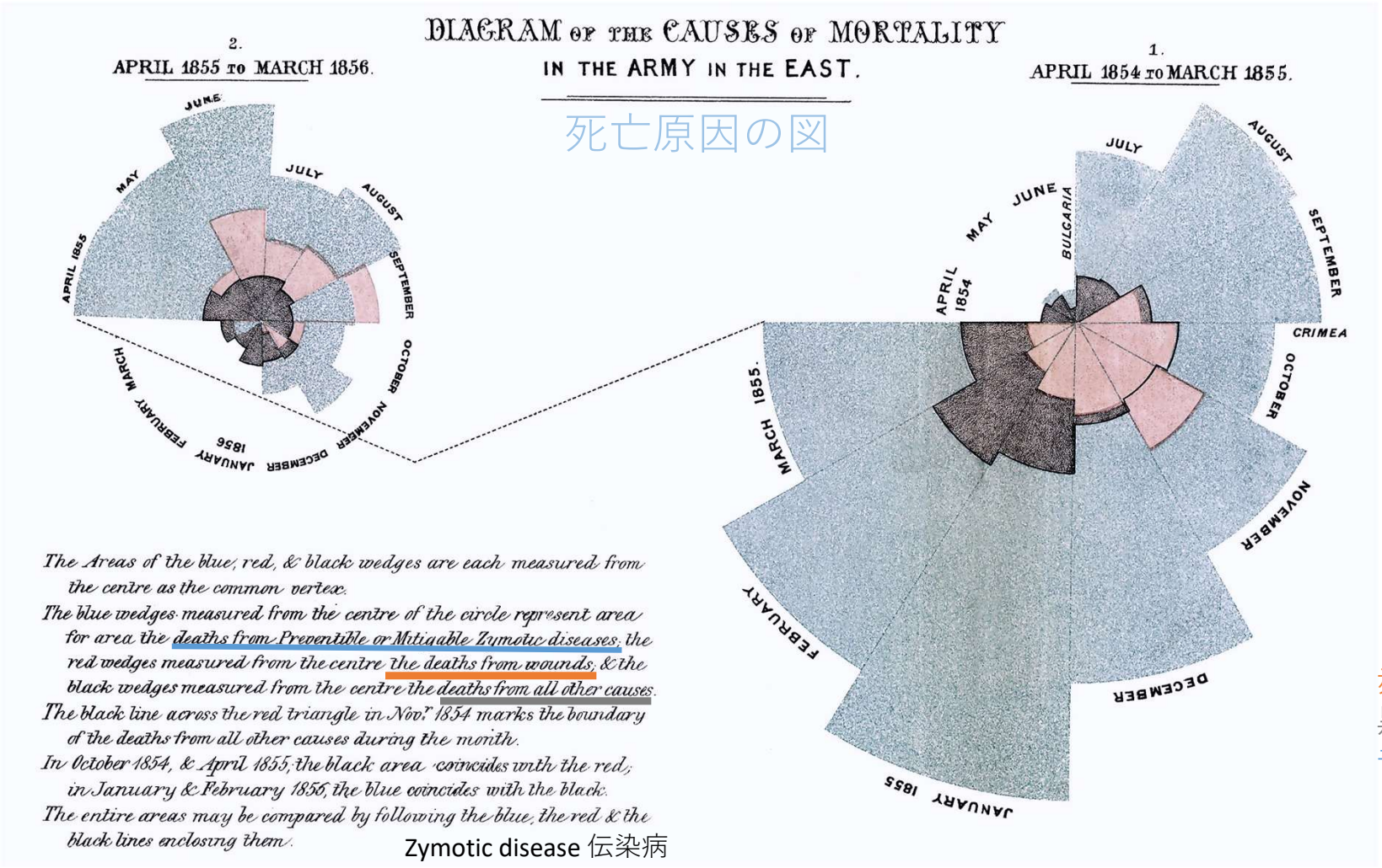


生誕　1820年5月12日
トスカーナ大公国フィレンツェ

フローレンス・ナイチンゲール　(1820-1910)

ナイチンゲールと統計





Zymotic disease 伝染病

赤：傷病による死者
黒：その他
青：予防できた死者

ナイチンゲールは、クリミア戦争(1853-1856)での”イギリス軍の戦死者・傷病者に関する膨大なデータを解析し、ほとんどが戦闘による死亡ではなく、傷を負った後の治療、病院の不衛生が原因で死亡したことを明らかにした。“[1]

→統計の力を使って医療改革を先導

[1]統計学習の指導のために　総務省統計局
http://www.stat.go.jp/teacher/c2epi3.html

統計学の種類

- 記述統計と推測統計に大きく分類できる

記述統計学

・データを整理し、データの特徴(記述統計量)をわかりやすく表す
・データの特徴や傾向を捉える

調べたデータを、全部(母集団)とする。

推測統計学

・標本を使って、その抽出元である母集団の特徴を推測する[2]

全部調べるのは大変だから一部（標本）を調べる。

頻度論的統計

従来の古典的統計学の考え方。手元のデータはたくさん実験したうちの一つであり、事前に立てた仮説が正しいとした場合、そのデータが観測される確率が何％なのかを考える。[2]

TVの視聴率など，600世帯から推定

ベイズ統計[4-5]

知識や経験、新しいデータを柔軟に取り込んで、より正確な分析を目指す。[2]

説明に比重

ベイズ推定

ベイズ推定（Bayesian inference）とは、ベイズ確率の考え方に基づき、観測事象（観測された事実）から、推定したい事柄（それの起因である原因事象）を、確率的な意味で推論することを指す[3]

予測に比重

機械学習へ

[2] 統計学図鑑　栗原伸一・丸山敦史　オーム社
[3] Wikipedia/ベイズ推定
[4] https://to-kei.net/basic-study/bayes/bayes-test-1/
[5] https://www.slideshare.net/matsukenbook/15-59154892

記述統計量

- **平均（average, mean）**
- **中央値（メジアン, median）**
- **標準偏差（SD, standard deviation）**
- 最頻値（モード, mode）
- 標準誤差（SE, standard error）
- 分散（variance）
- 尖度（kurtosis）
- 歪度（skewness）
- 範囲（range）
- 最小（min）
- 最大（max）
- 合計（sum）
- 標本数（size）

平均

- **算術平均** (arithmetic mean, 相加平均)

$$\bar{X} = \frac{(X_1 + X_2 + \cdots + X_n)}{n} = \sum_{i=1}^{n} X_i \,/n$$

Excel関数

AVERAGE(　)

  - 例：月別の電気代の年間の平均、平均体温(平熱)

中央値

Excel関数

MEDIAN(　)

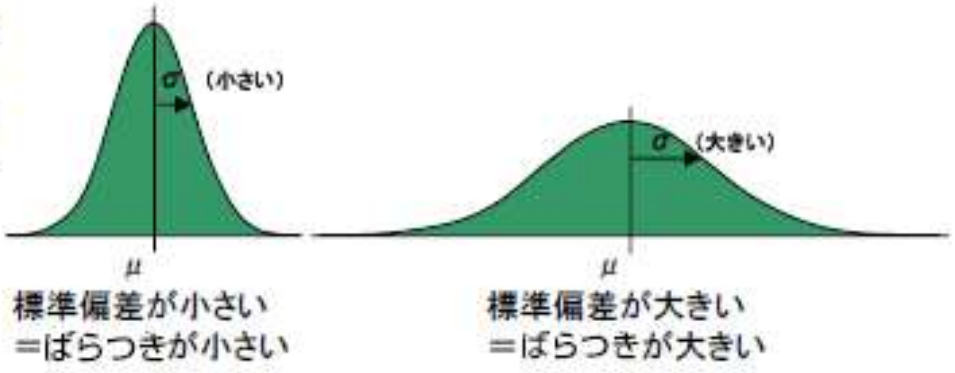
Ｎ個のデータを大きさの順に並べたとき、
中央に位置する値。データが偶数個なら
中央の順位に隣り合う２数の平均値

標準偏差

分散　$: \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

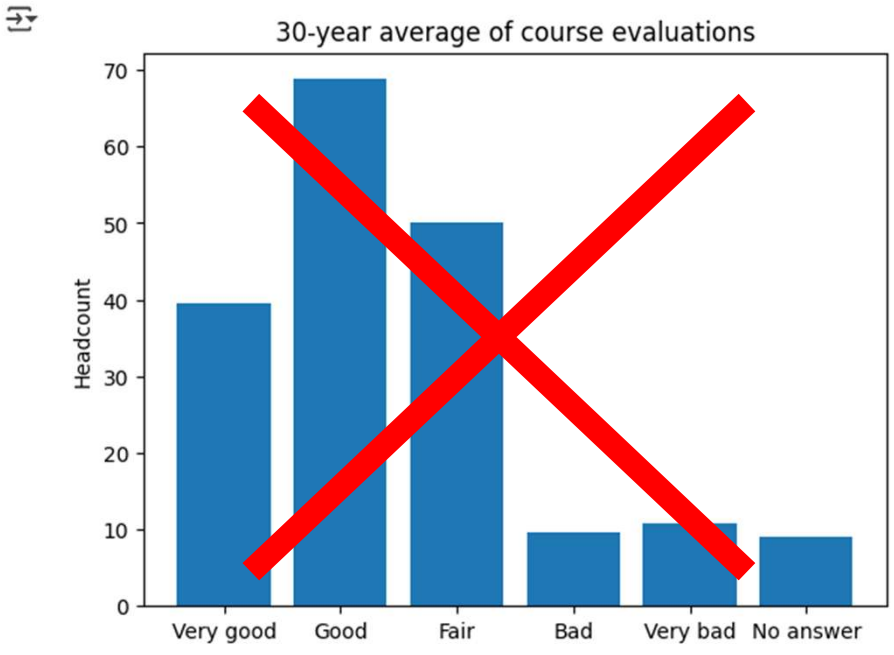
標準偏差 $: \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$

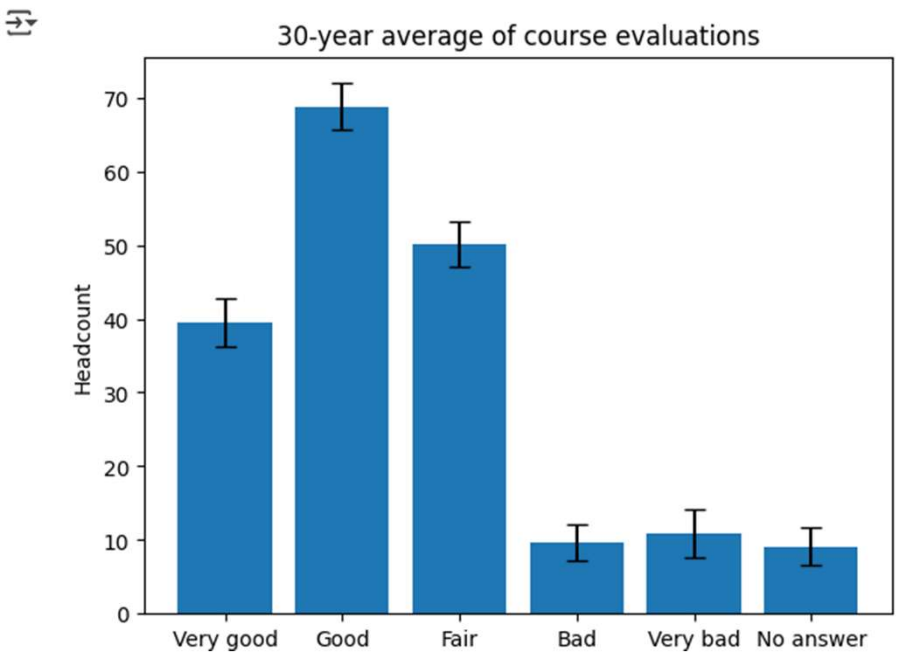
n ：実測回数
x_i ：i回目の実測データ
μ ：実測データの平均値



標準偏差が小さい
＝ばらつきが小さい

標準偏差が大きい
＝ばらつきが大きい





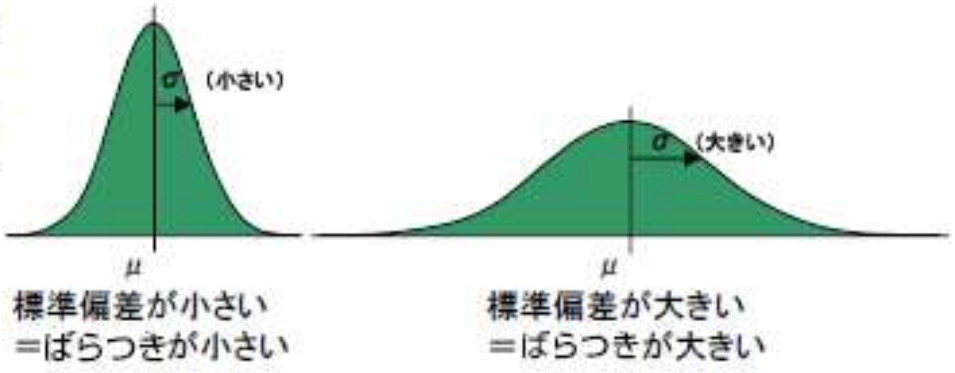
※箱ひげ図だとなお良いが，工学的には標準偏差で十分なことも多い

標準偏差

分散　$: \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

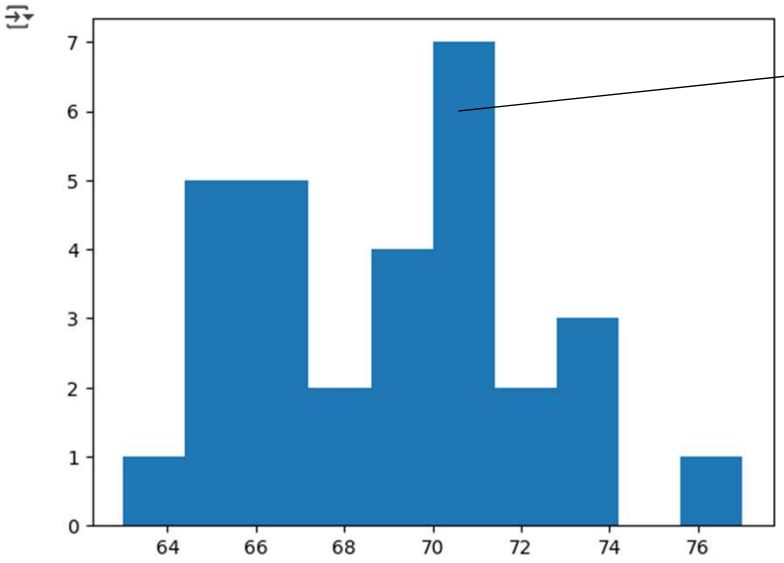
標準偏差 $: \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$

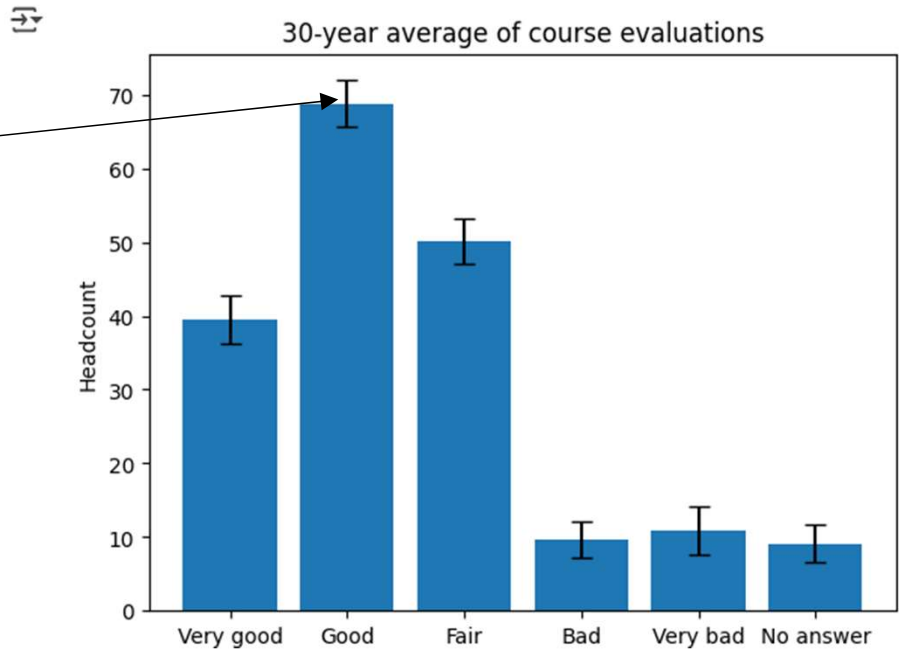
n ：実測回数
x_i ：i回目の実測データ
μ ：実測データの平均値



標準偏差が小さい
＝ばらつきが小さい

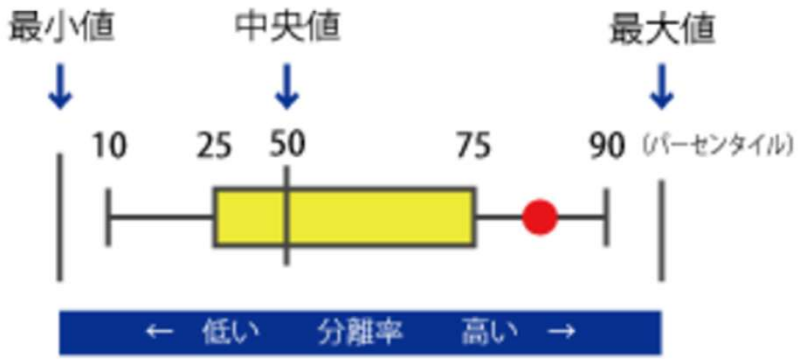
標準偏差が大きい
＝ばらつきが大きい





※箱ひげ図だとなお良いが，工学的には標準偏差で十分なことも多い

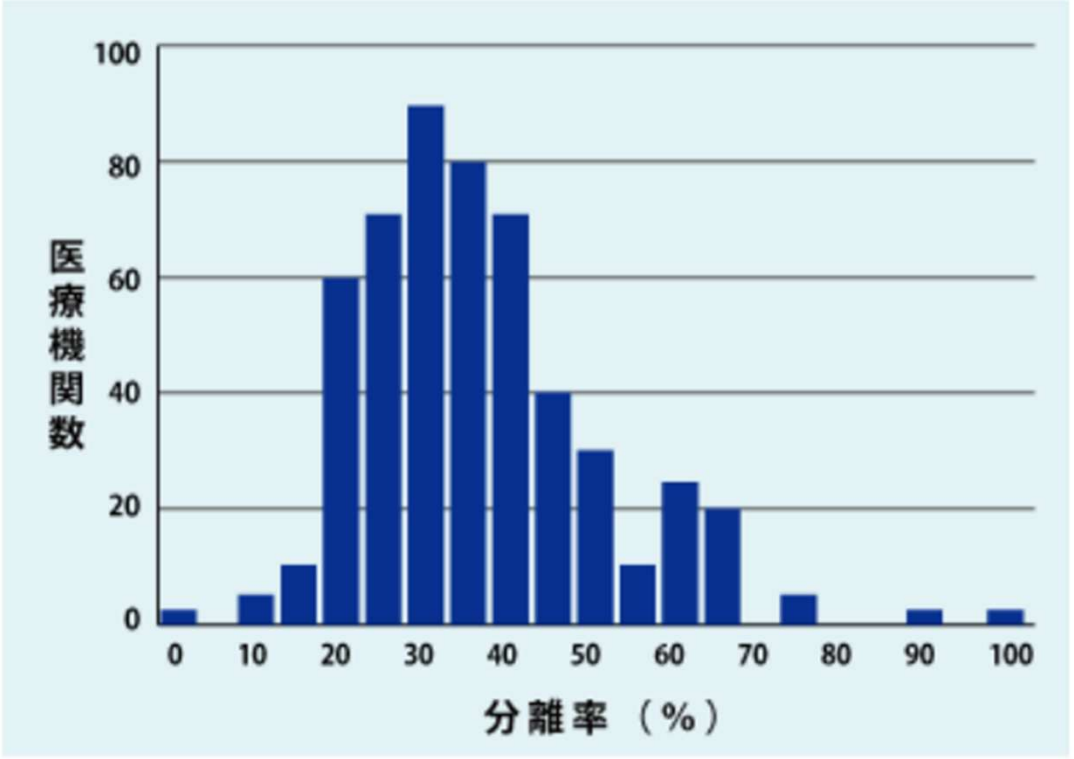
ボックスプロット（箱ひげ図）：



※パーセンタイルとは

値を小さいものから大きいものへと順番に並べ、全体を 100 として何番目であるかを表したもの。
例えば、10 パーセンタイルは、全体を 100 として小さい方から数えて 10 番目の測定値を示している。

ボックスプロット（箱ひげ図）

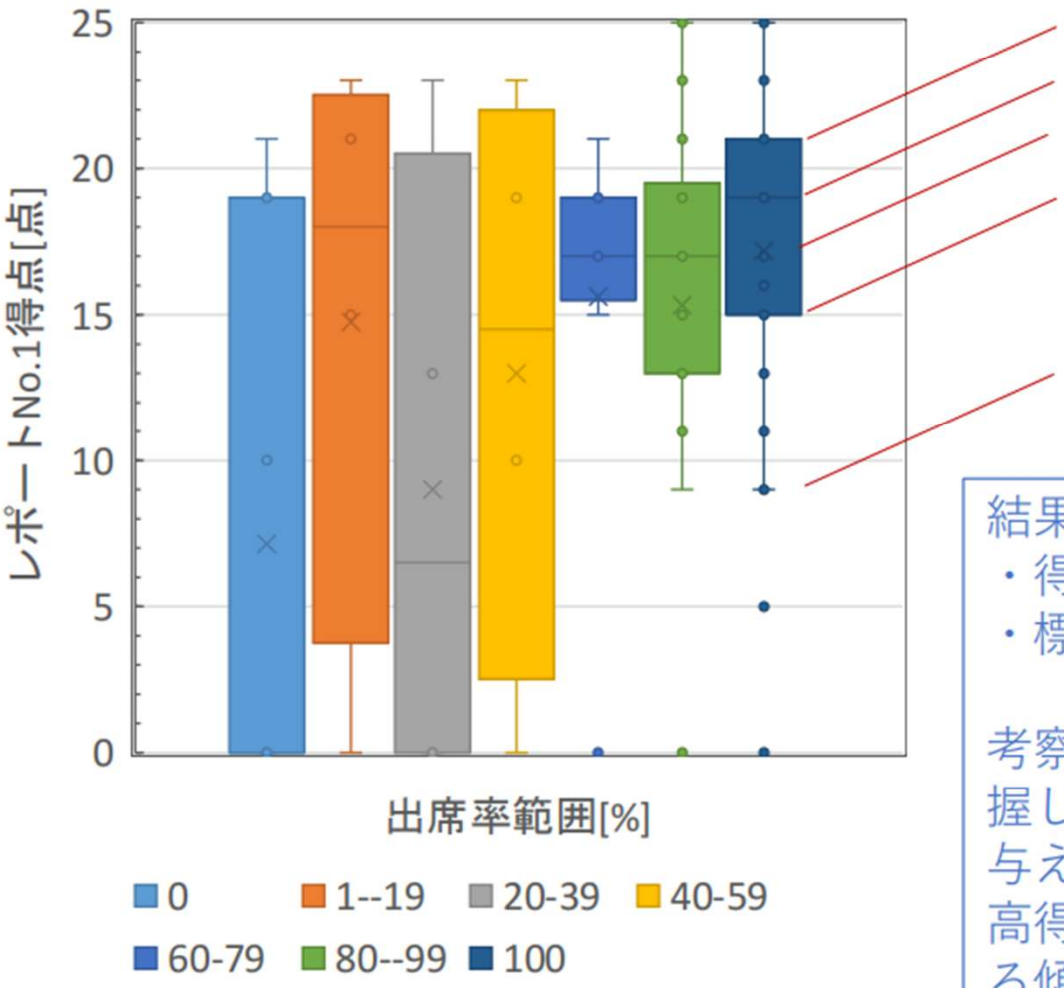


ヒストグラム

ボックスプロットは、ヒストグラムでの分布の様子を簡易的わかりやすく表現できる

厚生労働省　院内感染対策サーベイランス事業（ＪＡＮＩＳ）ＨＰ
https://janis.mhlw.go.jp/datause/index.html

ボックスプロット（箱ひげ図）で比較する



出席率範囲ごとのレポートNo.1得点分布ボックスプロット(2020年度データ)

記述的統計量

第１四分位値(上位２５％の値)

中央値

平均値

第３四分位値（下位２５％の値）

ひげの長さは中央値から四分位までの長さの1.5倍としたり、最大最小としたり様々

結果：出席率の高い学生は低い学生に比べて
・得点が高い傾向にある
・標準偏差が小さい。ばらつきが小さい。

考察：出席したことでレポートの採点条件を把握し、高得点が取れるようなヒントが授業中に与えられたのではないか。もしくはレポートで高得点取れるような人は講義にまじめに出席する傾向が高いのではないか。

意見：講義には出席した方がいい。

統計学の種類

- 記述統計と推測統計に大きく分類できる

記述統計学

・データを整理し、データの特徴(記述統計量)をわかりやすく表す
・データの特徴や傾向を捉える

調べたデータを、全部(母集団)とする。

推測統計学

・標本を使って、その抽出元である母集団の特徴を推測する[2]

全部調べるのは大変だから一部（標本）を調べる。

頻度論的統計

従来の古典的統計学の考え方。手元のデータはたくさん実験したうちの一つであり、事前に立てた仮説が正しいとした場合、そのデータが観測される確率が何％なのかを考える。[2]

TVの視聴率など，600世帯から推定

ベイズ統計[4-5]

知識や経験、新しいデータを柔軟に取り込んで、より正確な分析を目指す。[2]

説明に比重

ベイズ推定

ベイズ推定（Bayesian inference）とは、ベイズ確率の考え方に基づき、観測事象（観測された事実）から、推定したい事柄（それの起因である原因事象）を、確率的な意味で推論することを指す[3]

予測に比重

機械学習へ

[2] 統計学図鑑　栗原伸一・丸山敦史　オーム社
[3] Wikipedia/ベイズ推定
[4] https://to-kei.net/basic-study/bayes/bayes-test-1/
[5] https://www.slideshare.net/matsukenbook/15-59154892

ベイズ推定が必要な場面

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 %
- ウイルスには日本国民の0.01 %が感染している
- この場合，ランダムに選ばれた日本国民が検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？

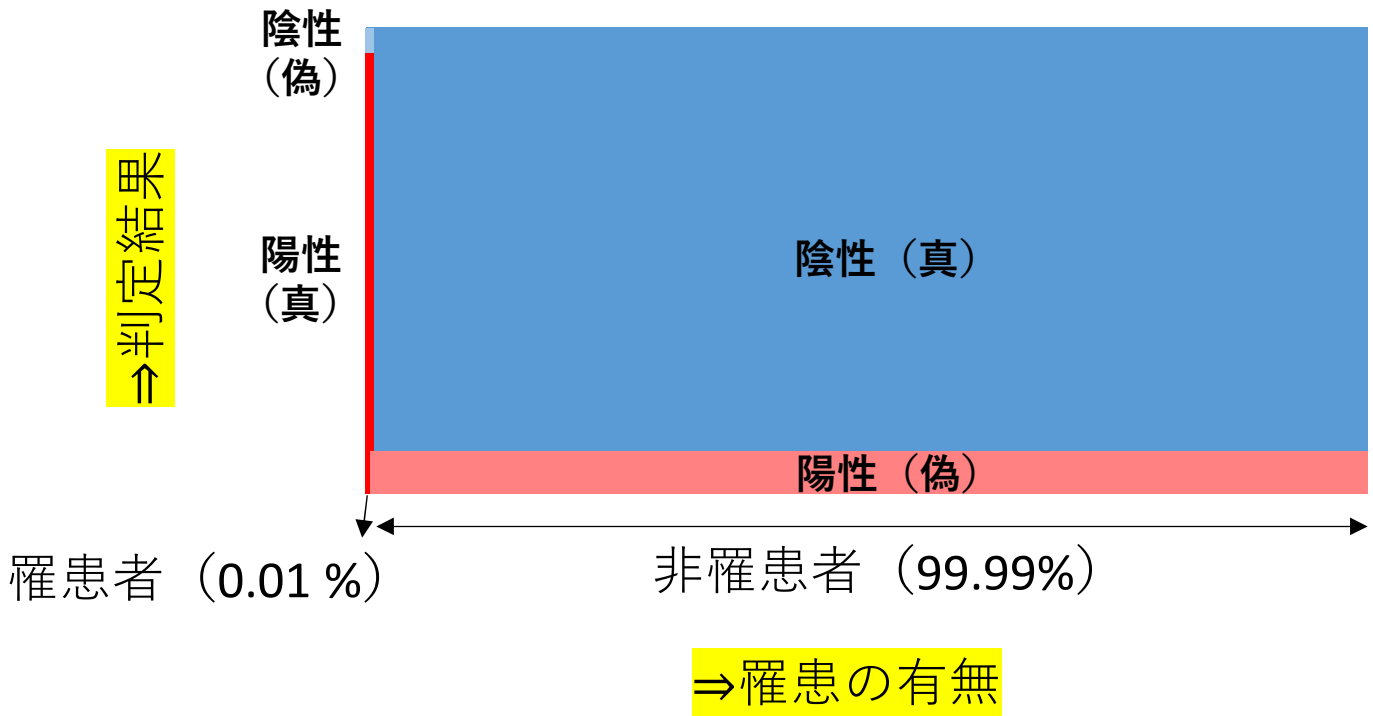
1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. 0.2 %

ベイズ推定が必要な場面

https://bellcurve.jp/statistics/course/6448.html

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 % →5%の確率で陰性者が陽性と判定
- ウイルスには日本国民の0.01 %が感染している
- この場合，検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？⇒**陽性(真)**/{ **陽性(真)**+ **陽性(偽)**}

1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. **0.2 %**



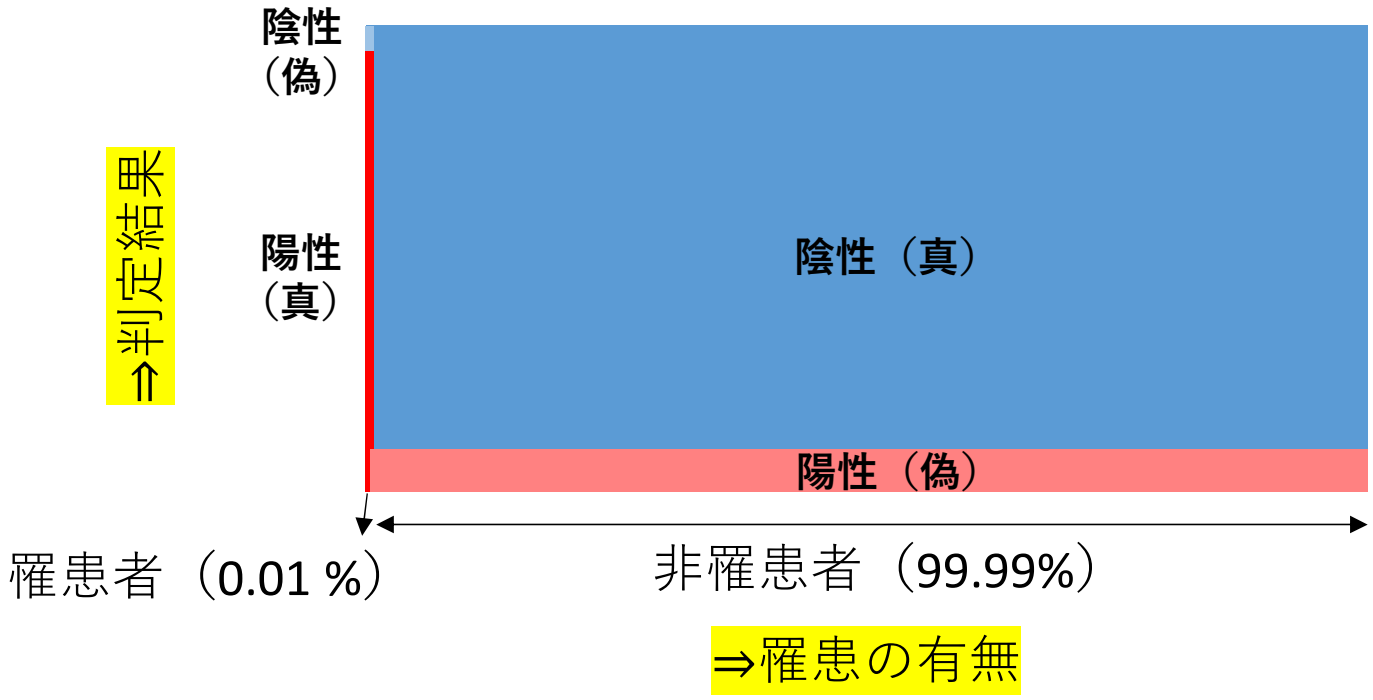
ベイズ推定が必要な場面

https://bellcurve.jp/statistics/course/6448.html

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 % →5%の確率で陰性者が陽性と判定
- ウイルスには日本国民の0.01 %が感染している
- この場合，検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？⇒**陽性(真)**/{ **陽性(真)**+ **陽性(偽)**}

=(0.01*99.9)/{0.01*99.9+99.99*5}

1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. **0.2 %**



グラフを描かずに計算するには？→ベイズ推定

https://bellcurve.jp/statistics/course/6448.html

ベイズ推定

ベイズの定理
Bayes' theorem

$P(B)$を事象Aが起きる前の事象Bが起きる確率（事前確率）、$P(B|A)$を事象Aが起きた後で事象Bが起きる確率（事後確率）とする。このとき$P(B|A)$は下式で表される。これをベイズの定理と言う。

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

- 病気に罹患している確率：$P(B_1) = 0.01\ \%$
- 病気に罹患していない確率：$P(B_2) = 99.99\ \%$
- 実際に罹患している人が検査で陽性となる確率：$P(A|B_1) = 99.9\ \%$
- 実際に罹患していない人が検査で陽性となる確率：$P(A|B_2) = 5\ \%$

一週目

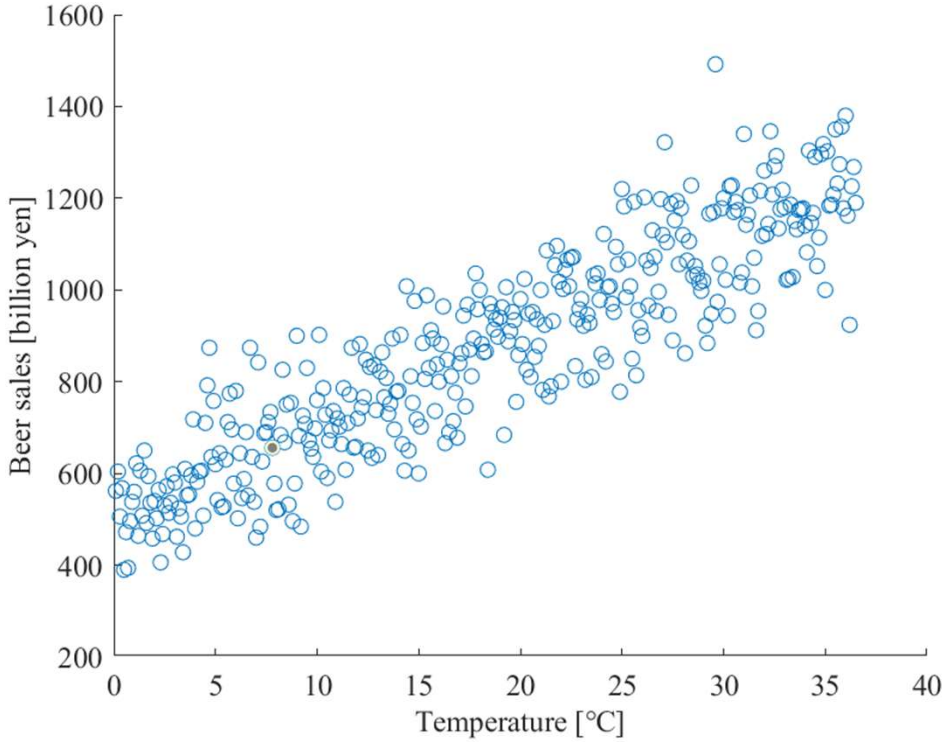
- 統計の復習（CLで見たことがあるはず・・）
  - グラフ作成について
  - 統計学について
  - 2変数データの分析と相関

© 2020 Yuki Kato. All Rights Reserved

気温とビール/アイスの売上-読み込み/プロット

	A	B	C	D	E	F	G	H	I	J	K
1	気温	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	ビール売上	562	604	506	568	390	472	394	496	538	560
3	アイス売上	42.5	55	55.5	64	64.5	47	57.5	34	48.5	69
4											
5											





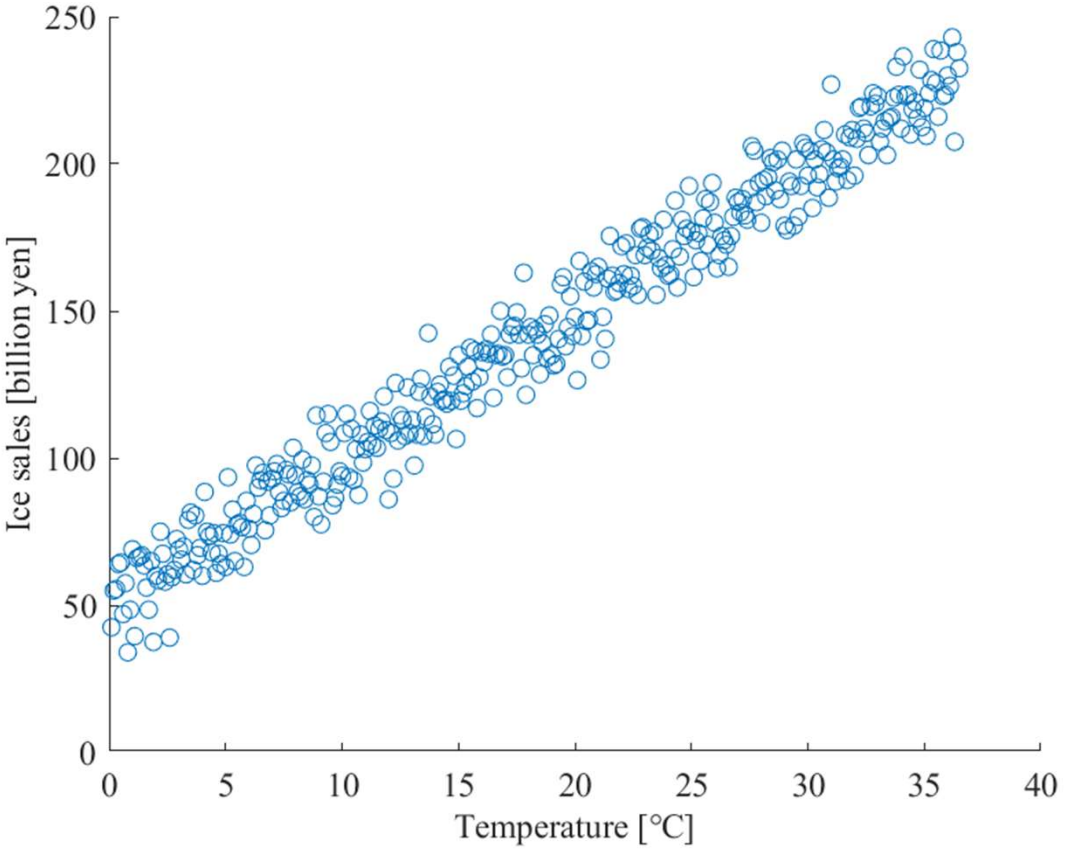


- 横軸気温-縦軸売上でプロット
- 気温が高い日は売り上げも高いことが分かる（相関がある）

※相関：一方が変化すれば他方も変化するように相互に関係しあうこと

気温とアイスの売り上げの関係



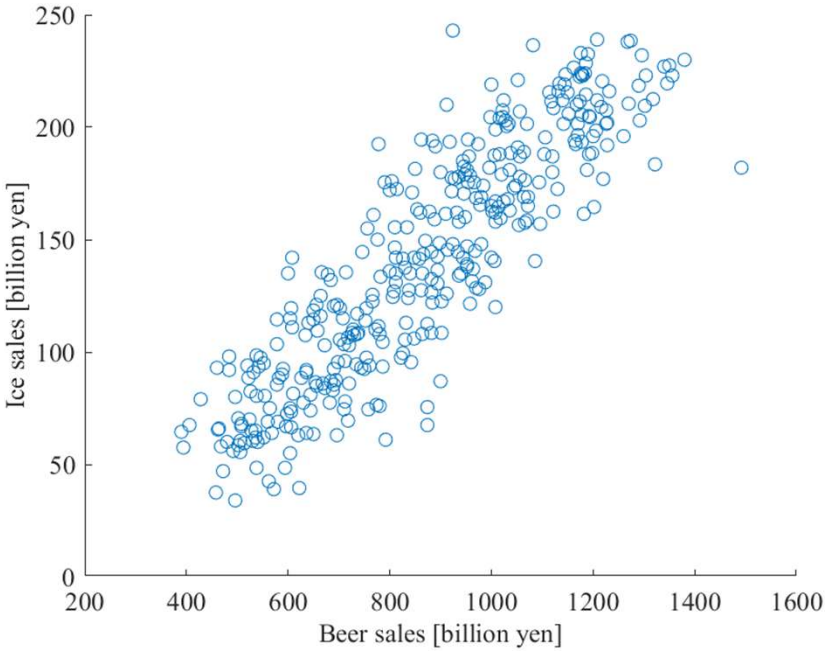




- 横軸気温-縦軸売上でプロット
- 気温が高い日は売り上げも高いことが分かる（相関がある）

ビールとアイスの売り上げの関係







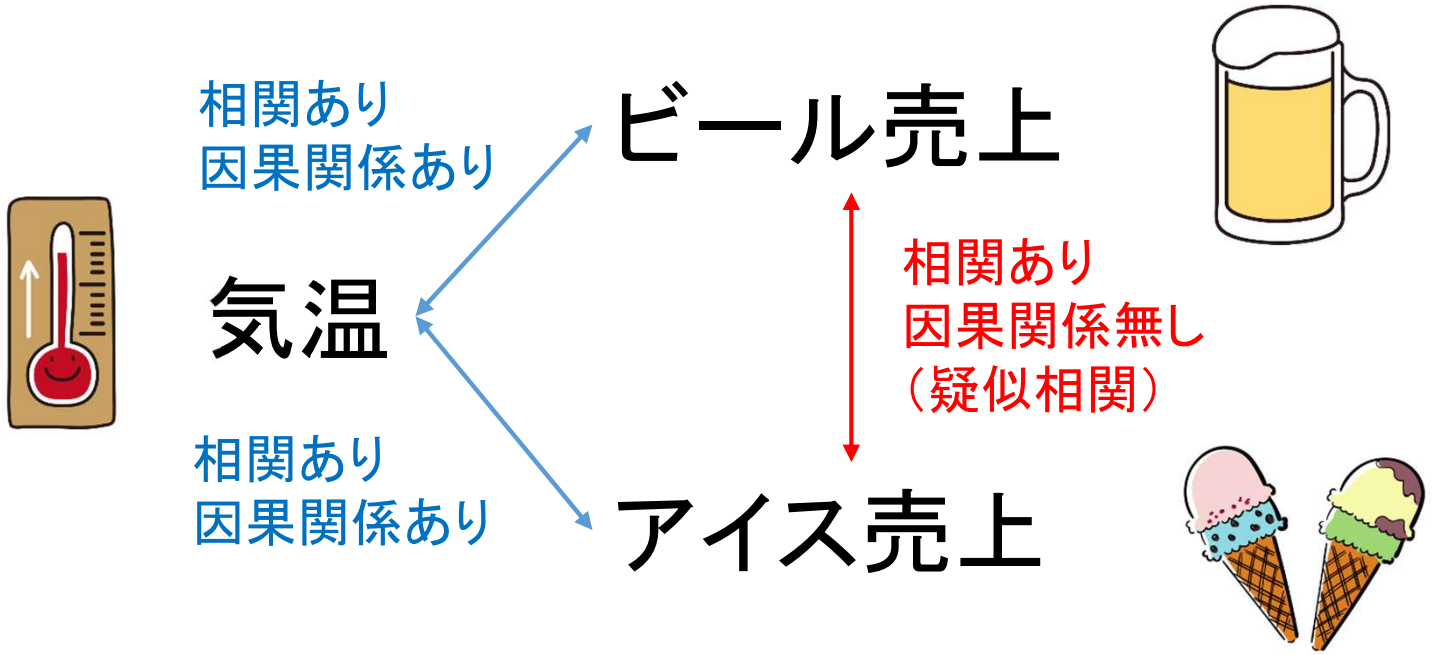
- 横軸ビール売上-縦軸アイス売上でプロット
- ビール売上とアイス売上に相関がある

→疑似相関，因果関係はない

- ビールの売り上げを上げるためにアイスのCMを流してアイスの売上を増加させても意味がない

→因果関係，メカニズムを調べることが重要

ビールとアイスの売り上げの関係



- 横軸ビール売上-縦軸アイス売上でプロット
- ビール売上とアイス売上に相関がある

→疑似相関，因果関係はない

- ビールの売り上げを上げるためにアイスのCMを流してアイスの売上を増加させても意味がない

→因果関係，メカニズムを調べることが重要

統計データ解析
第二回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義について

- 担当：加藤由幹（A423）
- 進め方
  - 基本的には資料を投影して進めます．板書もあります
  - 使用した資料はPDFでKUTLMSにアップロードします
  - ノートPCを使用するので，持参してください
  - 出欠はカードシステムで取ります
- 到達目標と成績評価



到達目標

1. 確率分布の性質を理解し，実験データを正しくグラフに表現し評価できる．
2. 回帰分析やフィルタ処理を用いてデータから知見を抽出できる．
3. 主成分分析や特異値分解を用いて多変量データを解析し，知見を抽出できる．

成績評価の方法・基準

試験は期末試験のみとする．
演習40%，期末試験60%とし，最終成績は得点率で評価する．

※演習（提出課題）は2回を予定しています

- その他

※補講が無い場合の休講は全員出席にします

体調不良や学会等による欠席は認める．最低でも10回以上は出席すること
大学院は研究が主なので，大量の課題等は出さないが，各自自習を怠らないようにすること
数学の基礎知識は各自持っているものとして進めます．（忘れている人は自習復習）

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

予定	修正	授業計画
~~10/6~~	10/9	第1回 統計データの性質とMatlabを用いたデータ解析の基礎
10/9	10/13	第2回 確率分布の性質
10/13	10/20	第3回 グラフと確率分布を用いた実験データの可視化
~~10/16~~	10/23	第4回 さまざまな分布
10/20	10/27	第5回 大数の法則と中心極限定理
10/23	10/30	第6回 時系列データ解析とフィルタ処理
10/27	11/6	第7回 時系列データ解析とフィルタ処理2
10/30	11/10	第8回 画像解析とフィルタ処理
11/6	11/17	第9回 最小二乗法，回帰分析
11/10	11/20	第10回 最尤推定，ベイズ推定
~~11/13~~	11/24	第11回 変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ
11/17	11/27	第12回 スパースモデリング
11/20	12/1	第13回 多変量データ解析の基礎と主成分分析
11/24		第14回 固有値分解，特異値分解
11/27		第15回 期末テスト
12/1	12/2 → (補講3)	第16回 テスト返却と解説

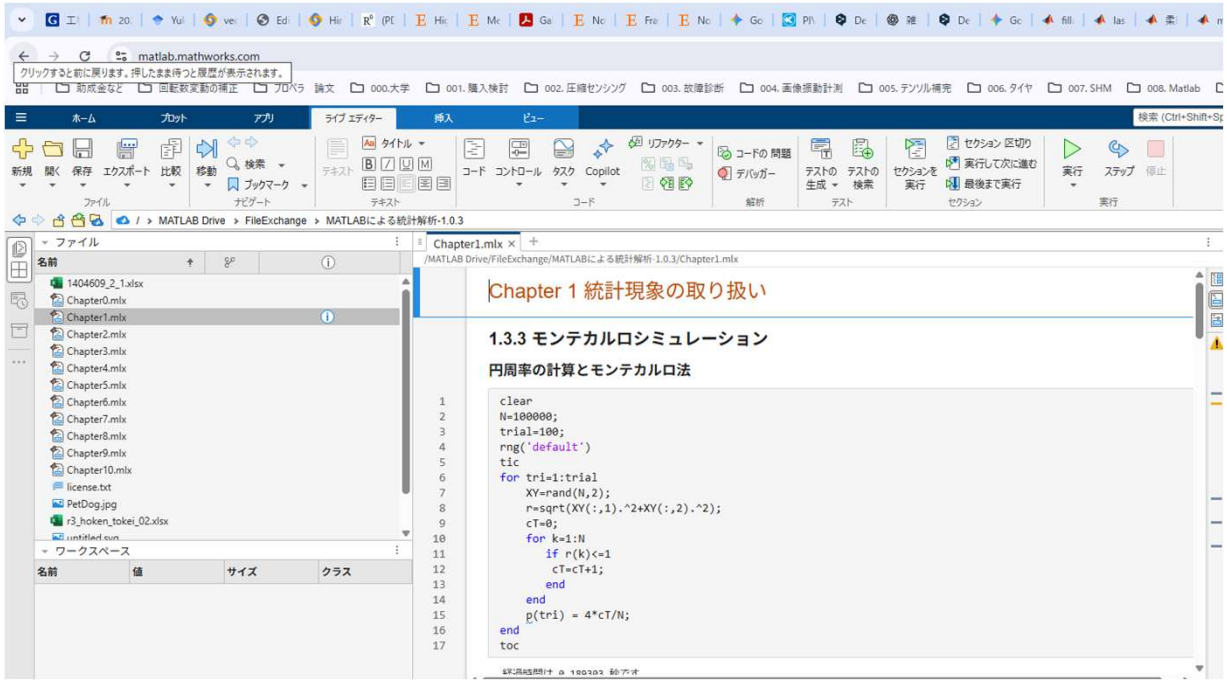
統計解析の基礎
1次元，2変数

統計解析の応用
多変量解析

© 2020 Yuki Kato. All Rights Reserved

当講義におけるプログラミングについて

- 基本的にはMatlabでのプログラミングを推奨するが，PythonやCで同じものを組んでもらっても良い
- 授業で使用するMatlabのソースコードは下記URLから入手できる．ただし，理解を深めるために手打ちでプログラミングすることを推奨する
  - https://jp.mathworks.com/matlabcentral/fileexchange/180435-matlab
  - MATLABによる統計解析で検索



© 2020 Yuki Kato. All Rights Reserved

1.3.4 事象の集合

- 一般に要素の集まりを集合（set）という．ここでは事象の集合が議論の対象となる．また集合の構成要素を元（element）または要素という．x が集合 Aの元であるとき

$$x \in A$$

- と書く．また集合 Aのすべての元が集合 Bの元であるとき

$$A \subseteq B$$

- と書き，A を Bの部分集合（subset）という．とくに $A \neq B$であるとき，A を Bの真部分集合（proper subset）といい

$$A \subset B$$

- と書く．さらに要素を含まないものも考え，{　} または $\emptyset$と書いて，空集合（empty set）と呼ぶ．
- 集合 Aと Bを考えるとき，それら2つの集合（両方の集合に含まれる元があるかもしれないが）の和を和集合（union）といい，$A \cup B$ と書く：

$$A \cup B = \{x | x \in A \text{ または } x \in B\}.$$

- 同様に集合 A, Bの両方に含まれる元から構成される集合を積集合（共通集合，intersection）といい，$A \cap B$ と書く：

$$A \cap B = \{x | x \in A \text{ かつ } x \in B\}.$$

© 2020 Yuki Kato. All Rights Reserved

1.3.5 確率の公理

- 事象全体の中から抽出された事象が作る集合 Ωを標本空間あるいは事象の空間と呼ぶ．またその元を標本点といい ωで表す．さらに以下の性質を持つ Ωの「部分集合の集合」を $\mathcal{A}$と書くことにする．この部分集合の集合 $\mathcal{A}$は次の性質を持つものでなくてはいけない．
  - 性質 1:　$\Omega \in \mathcal{A}$,　　部分集合A以外のすべて
  - 性質 2:　$A \in \mathcal{A}$　ならばその補集合 A^c についても　$A^c \in \mathcal{A}$,
  - 性質 3:　$A_1, A_2, \cdots \in \mathcal{A}$　ならば　$\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$.
- このような部分集合の集合を完全加法族という．
- 部分集合族 $\mathcal{A}$の元である部分集合 Aに関して，次の性質（確率の公理）を満たす $P(A)$を，事象の集合 Aが起きる確率と呼ぶ．

© 2020 Yuki Kato. All Rights Reserved

1.3.5 確率の公理

- 公理 1： $0 \leq P(A) \leq 1,$
- 公理 2： $P(\Omega) = 1,$
- 公理 3： 互いに排他的な事象（$A_i \cap A_j = \emptyset, i \neq j$）ならば

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right)$$

- これはコルモゴロフ（Andrei N. Kolmogorov, 1903-1987）の公理とも呼ばれ，現代確率論の基礎である．
- $\Omega, \mathcal{A}, P$の3つを組として $(\Omega, \mathcal{A}, P)$を確率空間という．
- 一般的に次の加法定理が成り立つ．

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- これから事象 A, Bが排他的であるなら（公理 3）

$$P(A \cup B) = P(A) + P(B)$$

- である．

© 2020 Yuki Kato. All Rights Reserved

大データと少数データの違い

- 統計データに馴染みのない人は，数字に騙されることが多い．コイントスを 10 回行ったときに表が 3 回以下しか出ないというのはままあることだが，トスの回数を 100 回にしたとき表が 30 回以下しか出ないということは「まず」ない．
- 10 回中で表が 3 回以下しか出ない確率： 17.2%
- 100 回中で表が 30 回以下しか出ない確率： 0.0039%
- データ数が少ない場合，偶然により極端な値をとりやすいからである

© 2020 Yuki Kato. All Rights Reserved

1.4.1 条件付確率

- 2つの事象 Aと Bを考える．それぞれの事象が起こる確率を $P(A)$および $P(B)$とする．
- 事象 Bが起こるという条件の下で事象 Aが起こる確率を条件付き確率といい、$P(A|B)$ と書く：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{ただし} \quad P(B) \neq 0.$$

- 逆に，上式で Aと Bとを入れ替えると

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

- である（乗法公式）．
- これから

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \text{ただし} \quad P(A) \neq 0$$

- を得る． ※ベイズの定理

© 2020 Yuki Kato. All Rights Reserved

1.4.2 「サイコロの目」の問題：条件付確率の例

- サイコロの 1 から 6 までの目が出る確率はすべて等しく

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

- であるとする．サイコロを振って偶数の目が出る事象を A, 4 以上の目が出る事象を Bとすれば

$$P(A) = P(2) + P(4) + P(6) = \frac{1}{2}, \quad P(B)$$

$$= P(4) + P(5) + P(6) = \frac{1}{2}$$

- である．それでは，サイコロを振って偶数の目が出た場合，それが 4以上の目である確率はどうであろうか．目が偶数で 4以上（4 か 6）が出る確率は

© 2020 Yuki Kato. All Rights Reserved

1.4.2 「サイコロの目」の問題：条件付確率の例

$$P(A \cap B) = P(4) + P(6) = \frac{1}{3}$$

- である．条件付確率の公式を使えば，サイコロの目が，偶数が出たという条件の下で 4以上である確率は

$$P(B|A) =$$

- となる．一方で $P(A|B)$は，4 以上であるという条件の下でそれが偶数であるという確率であるから

$$P(A|B) =$$

- となる．

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

© 2020 Yuki Kato. All Rights Reserved

1.4.2 「サイコロの目」の問題：条件付確率の例

$$P(A \cap B) = P(4) + P(6) = \frac{1}{3}$$

- である．条件付確率の公式を使えば，サイコロの目が，偶数が出たという条件の下で 4以上である確率は

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{1/2} = \frac{2}{3}$$

- となる．一方で $P(A|B)$は，4 以上であるという条件の下でそれが偶数であるという確率であるから

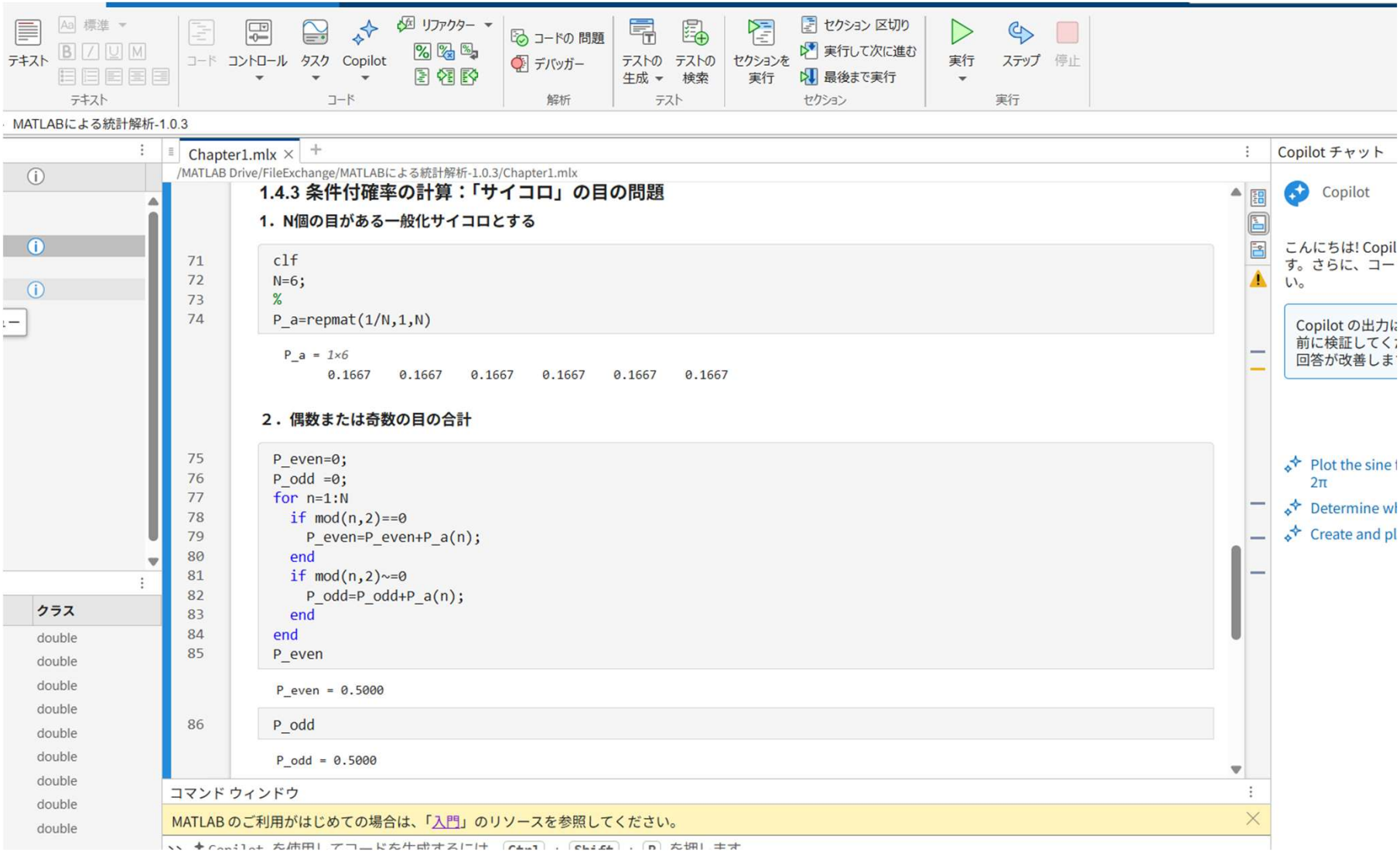
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}$$

- となる．

© 2020 Yuki Kato. All Rights Reserved

Matlabを用いた条件付き確率の計算

■ 資料では，Matlabオンラインを使用します



© 2020 Yuki Kato. All Rights Reserved

ベイズ推定が必要な場面

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 %
- ウイルスには日本国民の0.01 %が感染している
- この場合，ランダムに選ばれた日本国民が検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？

1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. 0.2 %

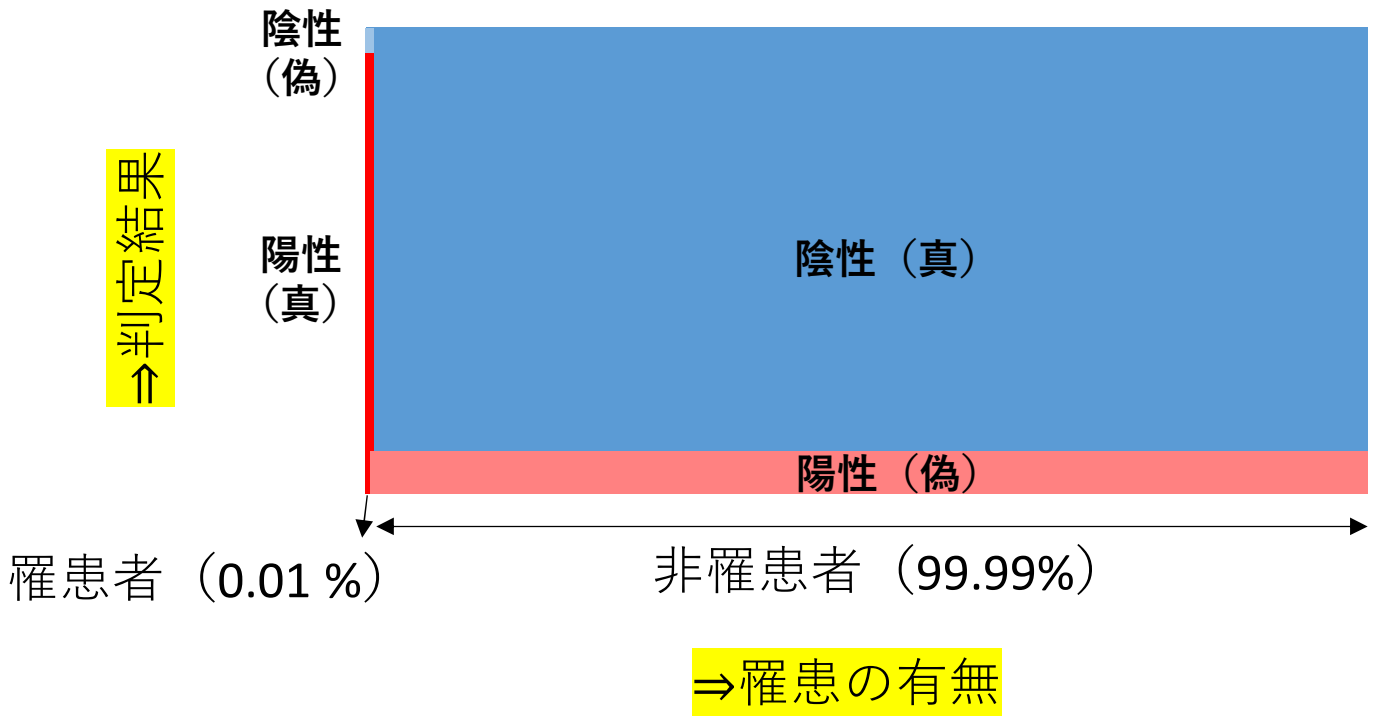
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

ベイズ推定が必要な場面

https://bellcurve.jp/statistics/course/6448.html

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 % →5%の確率で陰性者が陽性と判定
- ウイルスには日本国民の0.01 %が感染している
- この場合，検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？⇒**陽性(真)**/{ **陽性(真)**+ **陽性(偽)**}

1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. **0.2 %**



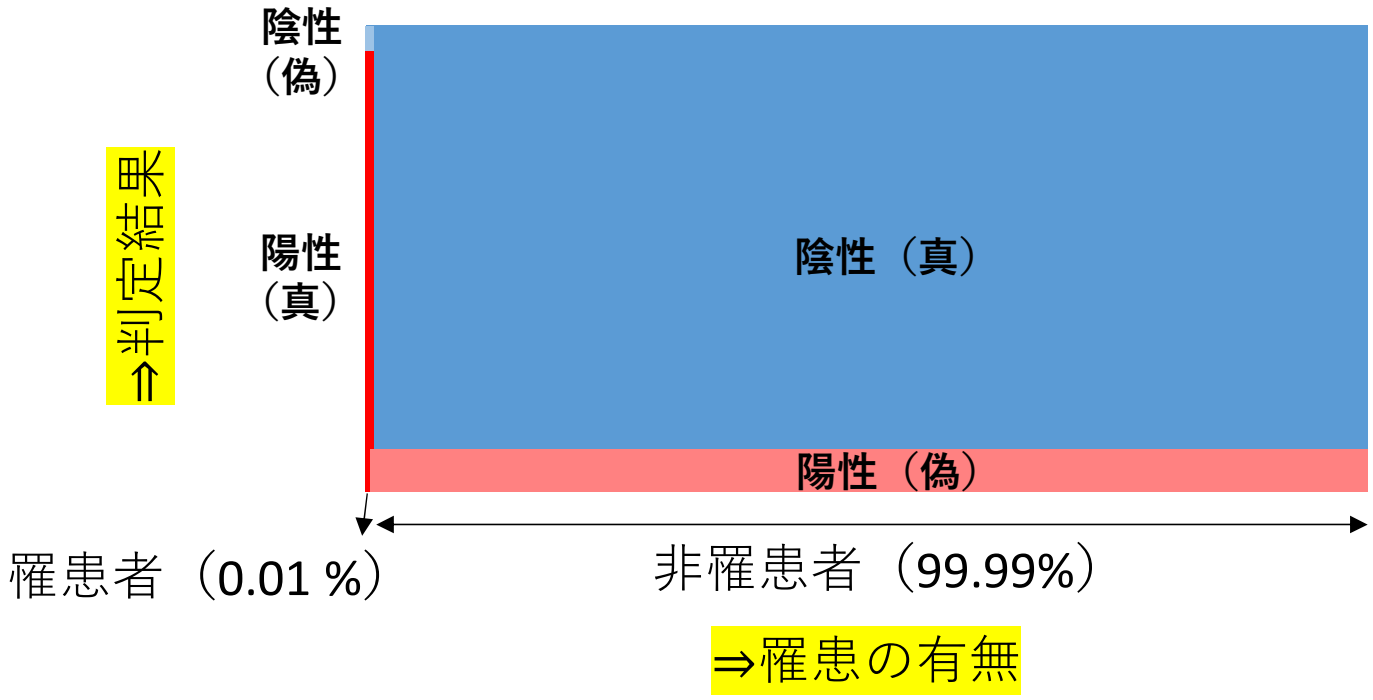
ベイズ推定が必要な場面

https://bellcurve.jp/statistics/course/6448.html

- 検査キットで新型ウイルスの検査を実施
  - 陽性者が陽性と判定される確率：99.9 %
  - 陰性者が陰性と判定される確率：95 % →5%の確率で陰性者が陽性と判定
- ウイルスには日本国民の0.01 %が感染している
- この場合，検査を受けて陽性と判定された人が実際にウイルスに罹患している確率は何％か？⇒**陽性(真)**/{ **陽性(真)**+ **陽性(偽)**}

=(0.01*99.9)/{0.01*99.9+99.99*5}

1. 99.9 %
2. 95 %
3. 10 %
4. 1 %
5. **0.2 %**



グラフを描かずに計算するには？→ベイズ推定

https://bellcurve.jp/statistics/course/6448.html

ベイズ推定

ベイズの定理
Bayes' theorem

$P(B)$を事象Aが起きる前の事象Bが起きる確率（事前確率）、$P(B|A)$を事象Aが起きた後で事象Bが起きる確率（事後確率）とする。このとき$P(B|A)$は下式で表される。これをベイズの定理と言う。

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

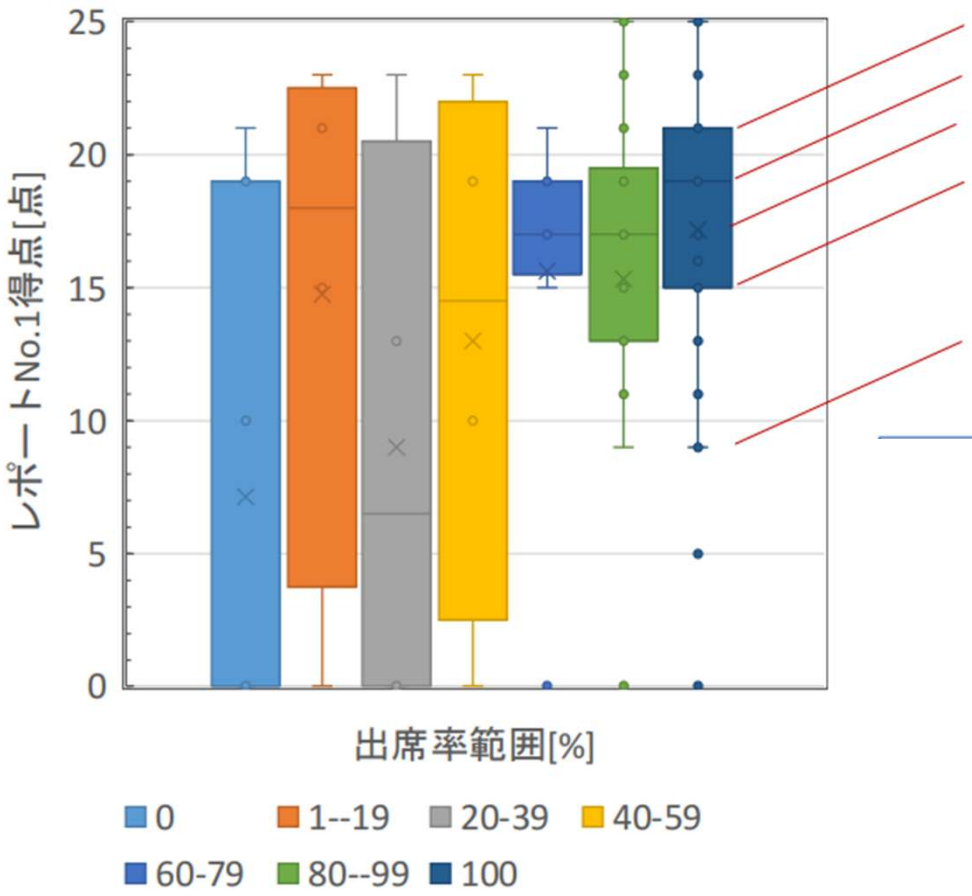
- 病気に罹患している確率：$P(B_1) = 0.01\ \%$
- 病気に罹患していない確率：$P(B_2) = 99.99\ \%$
- 実際に罹患している人が検査で陽性となる確率：$P(A|B_1) = 99.9\ \%$
- 実際に罹患していない人が検査で陽性となる確率：$P(A|B_2) = 5\ \%$

確率変数と分布

- 確率事象に伴って，そこに現れる変数，例えばさいころの目の数を確率変数という
- 確立変数はその事象が現れる確率と対応つけられていなくてはならない
- ここでは，このような確率変数とそれに結び付いた確率の振る舞いをどのように表すかについて，考える

© 2020 Yuki Kato. All Rights Reserved

ボックスプロット



出席率範囲ごとのレポートNo.1得点分布
ボックスプロット(2020年度データ)

記述的統計量

第１四分位値(上位２５％の値)

中央値

平均値

第３四分位値（下位２５％の値）

ひげの長さは中央値から四分位までの長さの1.5倍としたり、最大最小としたり様々

→Matlabでプロットしてみよう

説明

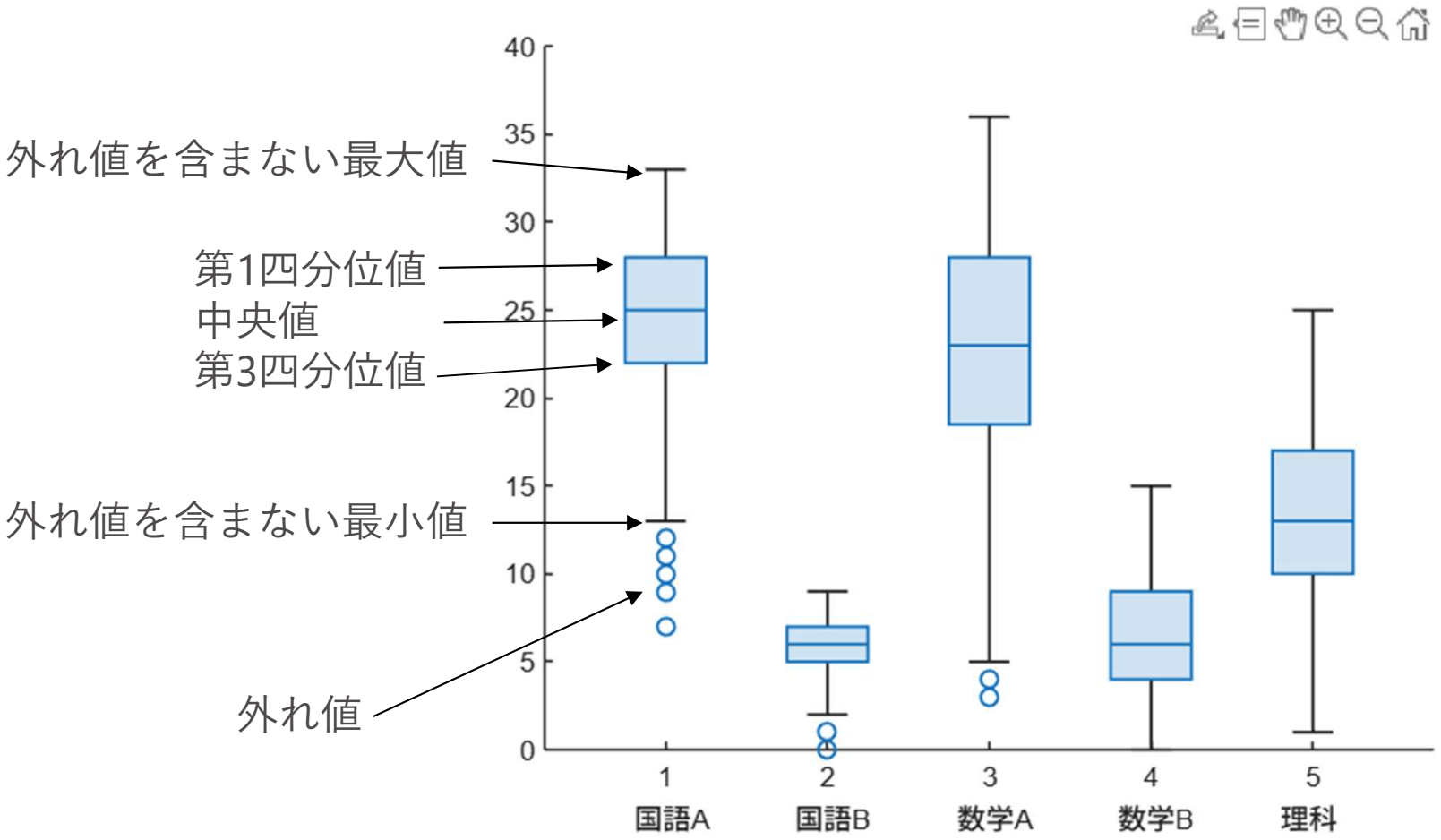
boxchart(ydata) は行列 ydata の各列に対してボックス チャート (箱ひげ図) を作成します。ydata がベクトルの場合、boxchart は単一のボックス チャートを作成します。

各ボックス チャートには、中央値、上位四分位数、下位四分位数、すべての外れ値 (四分位数間範囲を使用して計算)、外れ値でない最小値と最大値の各情報が表示されます。詳細については、ボックス チャート (箱ひげ図)を参照してください。

© 2020 Yuki Kato. All Rights Reserved

Matlabによる箱ひげ図

■ ひげの部分は分野等による．記載するのがベター



説明

boxchart(ydata) は行列 ydata の各列に対してボックス チャート (箱ひげ図) を作成します。ydata がベクトルの場合、boxchart は単一のボックス チャートを作成します。
各ボックス チャートには、中央値、上位四分位数、下位四分位数、すべての外れ値 (四分位数間範囲を使用して計算)、外れ値でない最小値と最大値の各情報が表示されます。詳細については、ボックス チャート (箱ひげ図)を参照してください。

© 2020 Yuki Kato. All Rights Reserved

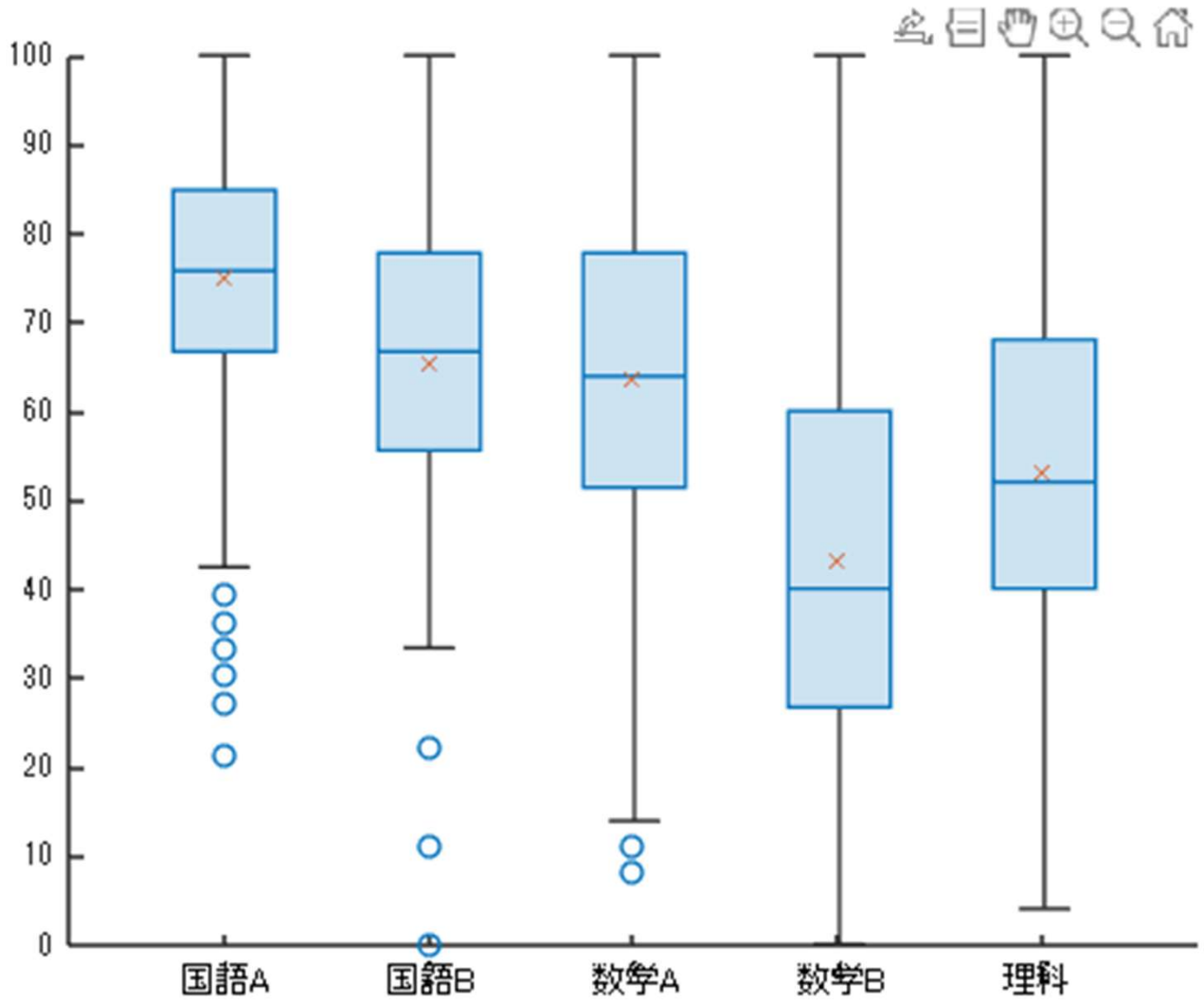
Matlabによる箱ひげ図

- 正答率についても，箱ひげ図を作成しましょう
- 平均値のプロットも追加しましょう

© 2020 Yuki Kato. All Rights Reserved

Matlabによる箱ひげ図

- 正答数についても，箱ひげ図を作成しましょう
- 平均値のプロットも追加しましょう



© 2020 Yuki Kato. All Rights Reserved

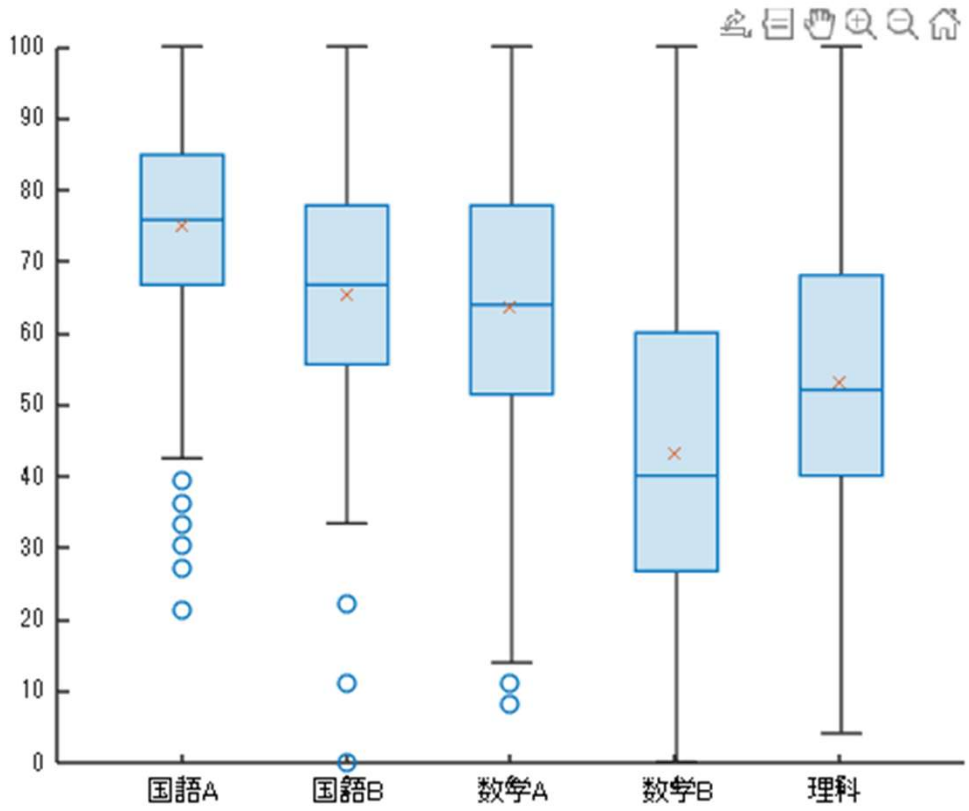
エラーバーのプロット

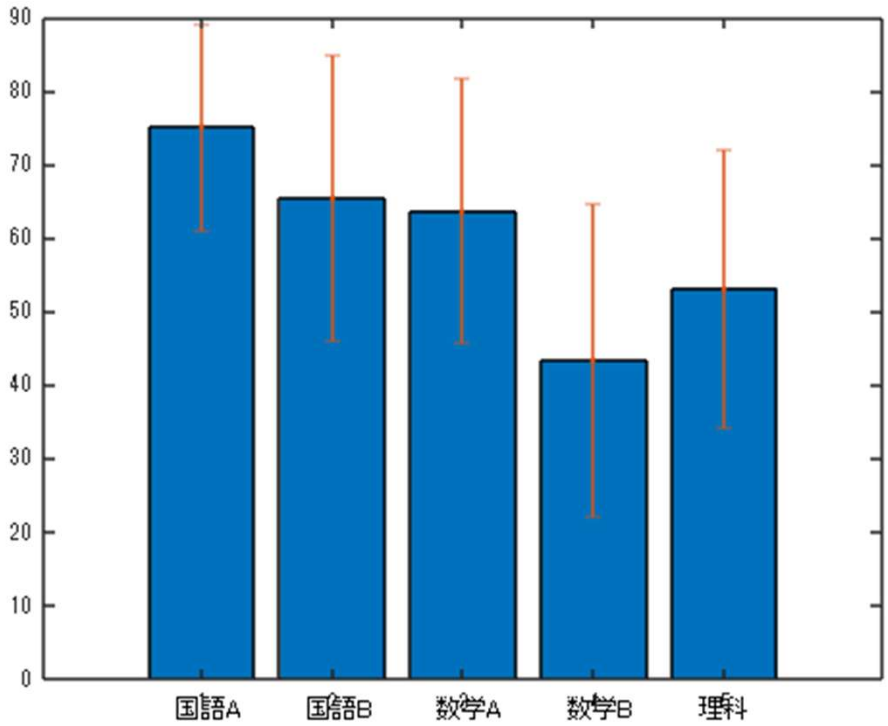
- 正答率について，棒グラフとエラーバー（標準偏差）を用いてプロットしてください

© 2020 Yuki Kato. All Rights Reserved

エラーバーのプロット

■ 正答数について，棒グラフとエラーバーを用いてプロットしてください





© 2020 Yuki Kato. All Rights Reserved

課題

- 3以下で，かつ奇数の目が出る確率

課題

- 3以下で，かつ奇数の目が出る確率
  - 2/6=1/3

```
P_oddAndBE3=0;
for n=1:N
  if mod(n,2)~=0 && n<=3
    P_oddAndBE3=P_oddAndBE3+P_a(n);
  end
end
P_oddAndBE3
```

```
P_oddAndBE3 = 0.3333
```

統計データ解析
第三回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

予定	修正	授業計画	
~~10/6~~	10/9	第1回 統計データの性質とMatlabを用いたデータ解析の基礎	
10/9	10/13	第2回 確率分布の性質	
10/13	10/20	第3回 グラフと確率分布を用いた実験データの可視化	
~~10/16~~	10/23	第4回 さまざまな分布	
10/20	10/27	第5回 大数の法則と中心極限定理	
10/23	10/30	第6回 時系列データ解析とフィルタ処理	
10/27	11/6	第7回 時系列データ解析とフィルタ処理2	
10/30	11/10	第8回 画像解析とフィルタ処理	
11/6	11/17	第9回 最小二乗法，回帰分析	
11/10	11/20	第10回 最尤推定，ベイズ推定	
~~11/13~~	11/24	第11回 変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ	
11/17	11/27	第12回 スパースモデリング	
11/20	12/1	第13回 多変量データ解析の基礎と主成分分析	
11/24		第14回 固有値分解，特異値分解	
11/27		第15回 期末テスト	
12/1	12/2 → (補講3)	第16回 テスト返却と解説	

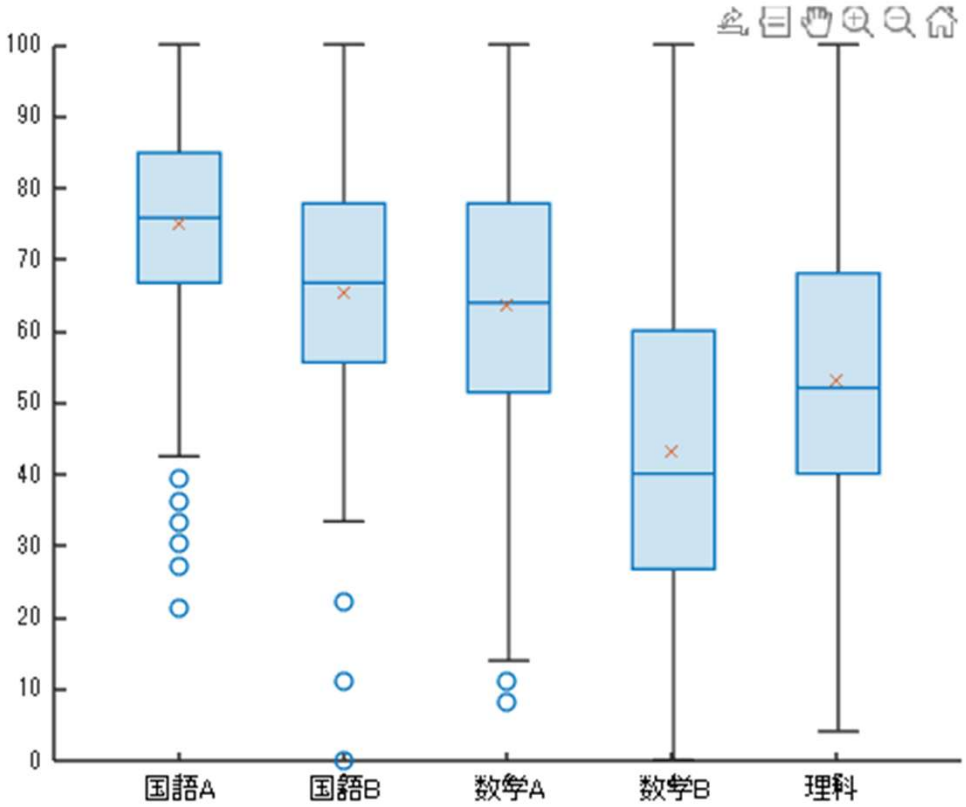
統計解析の基礎
1次元，2変数

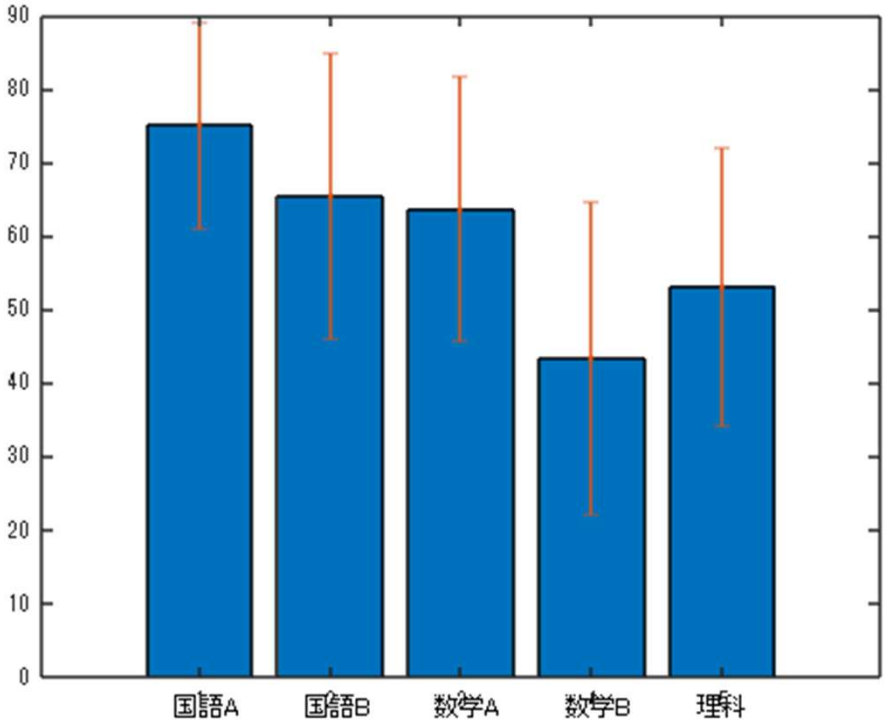
統計解析の応用
多変量解析

© 2020 Yuki Kato. All Rights Reserved

エラーバーのプロット

■ 正答数について，棒グラフとエラーバーを用いてプロットしてください

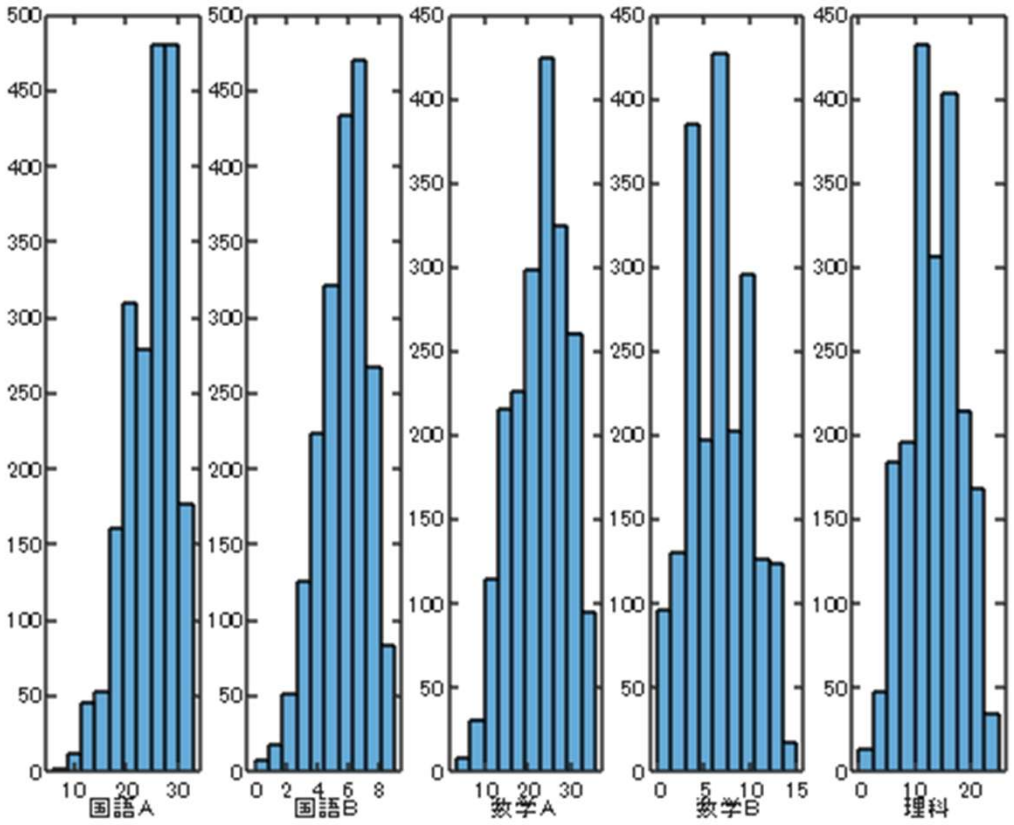




© 2020 Yuki Kato. All Rights Reserved

ヒストグラムの作成

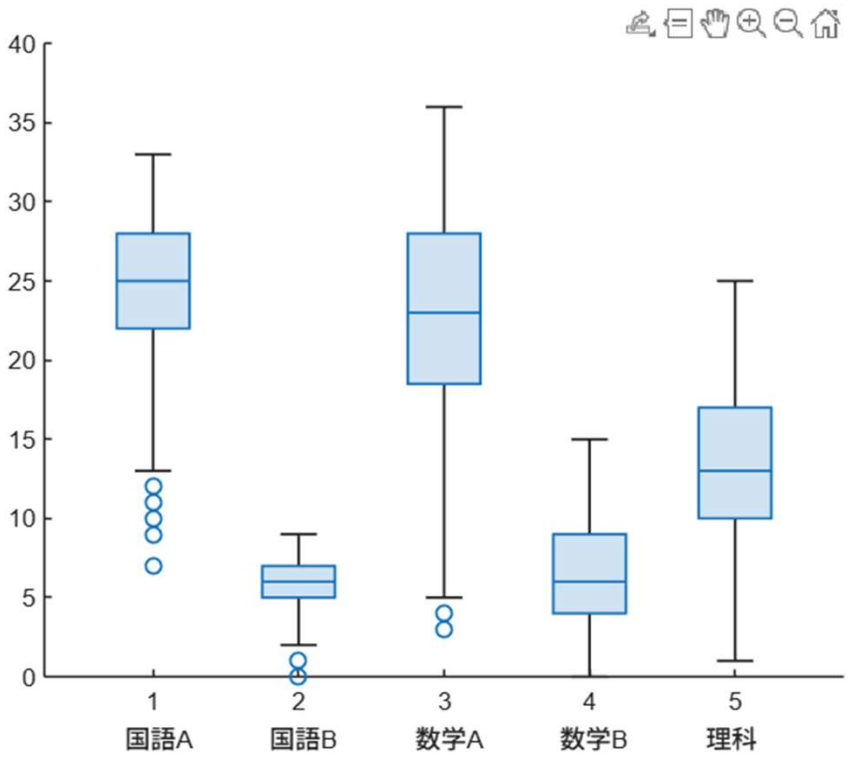
- 標本として得たある変量の値のリストを度数分布表という．
- 変数の領域を互いに重なり合わず接する区間に分け，その区間に分布する度数を棒グラフに書いたものを度数分布図（ヒストグラム）という
- 横軸の正答数を0-40に固定してみよう

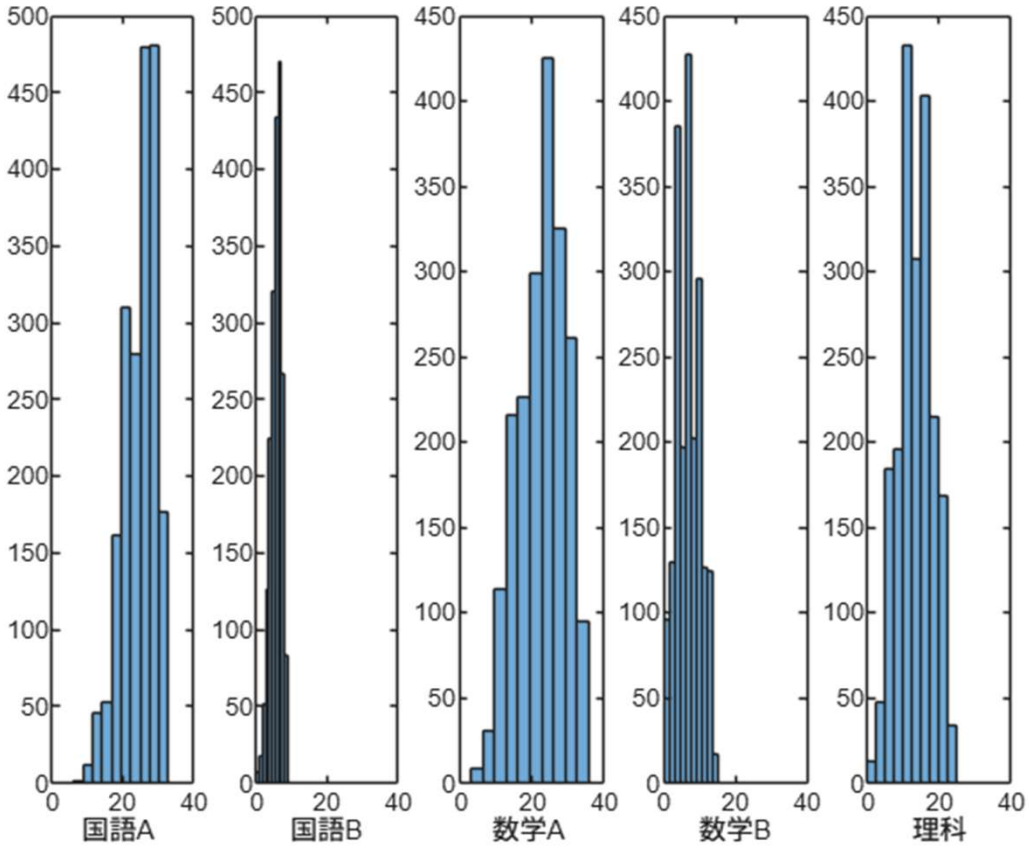


© 2020 Yuki Kato. All Rights Reserved

ボックスプロットの作成とヒストグラム

- 情報量はヒストグラムの方が多いが，ボックスプロットでもおおよその傾向は掴める
- →度数分布を合計で割ると確率関数に変換できる

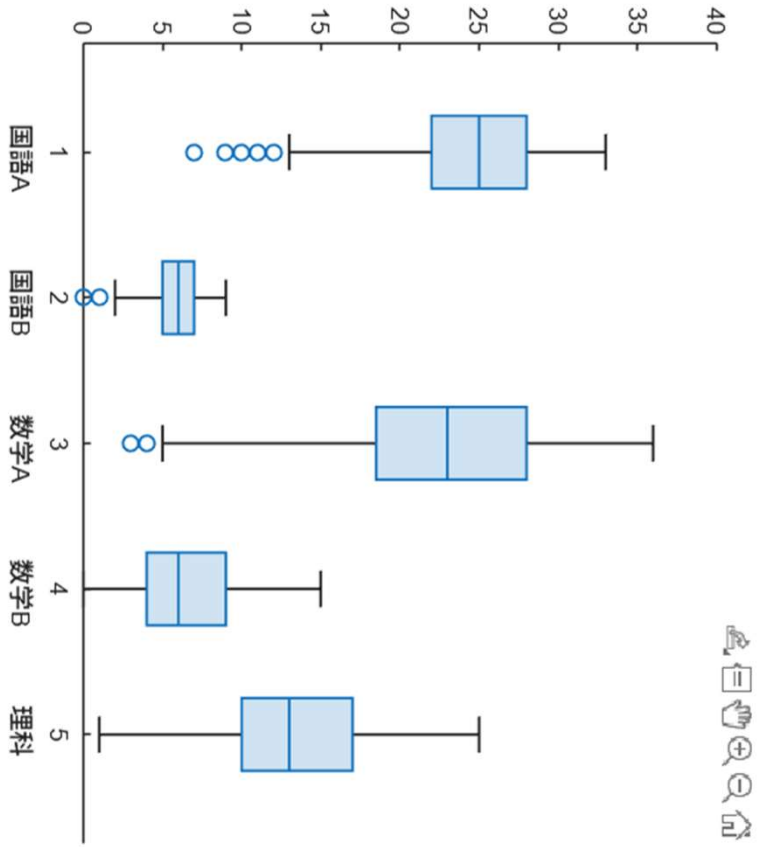


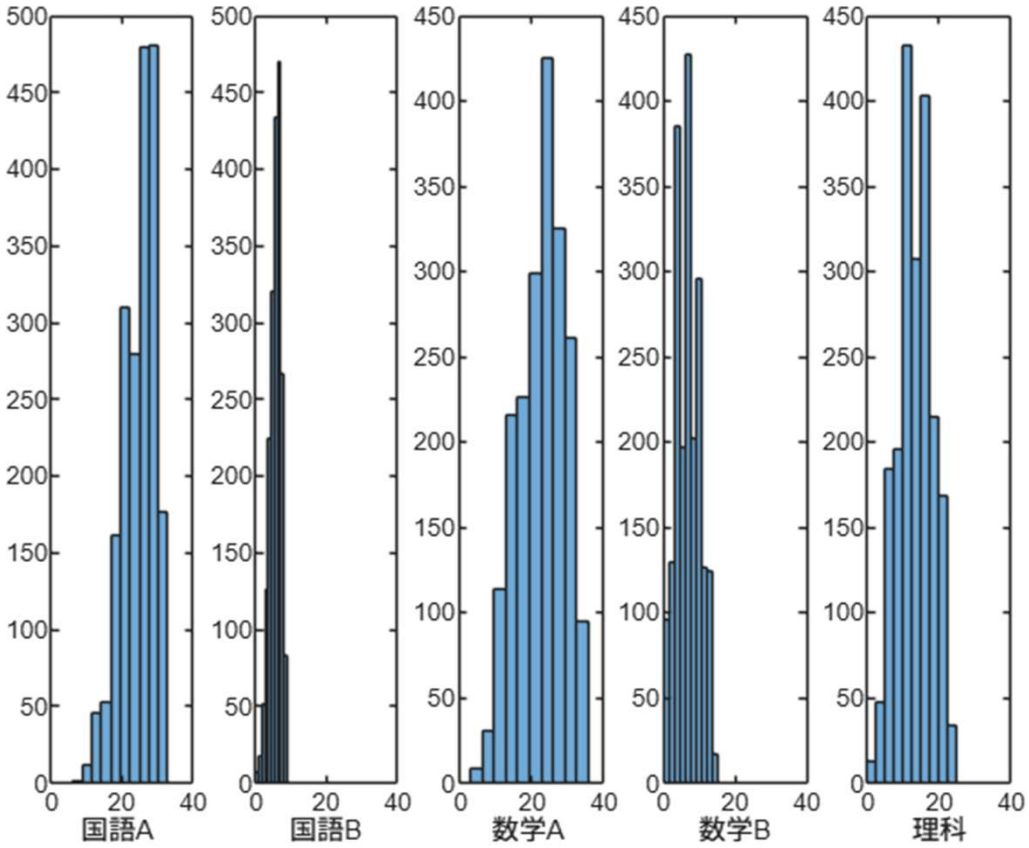


© 2020 Yuki Kato. All Rights Reserved

ボックスプロットの作成とヒストグラム

- 情報量はヒストグラムの方が多いが，ボックスプロットでもおおよその傾向は掴める
- →度数分布を合計で割ると確率関数に変換できる





© 2020 Yuki Kato. All Rights Reserved

確率変数と分布

- 確率事象に伴って，そこに現れる変数，例えばさいころの目の数を確率変数という
- 確立変数はその事象が現れる確率と対応つけられていなくてはならない
- ここでは，このような確率変数とそれに結び付いた確率の振る舞いをどのように表すかについて，考える

© 2020 Yuki Kato. All Rights Reserved

確率変数の定義

- 確率空間 (Ω, A, P) 上のすべての標本点 ω $(\omega \in \Omega)$ について，すべての実数値 x に対して

$$\{\omega : X(\omega) \leq x\} \in A$$

- を満たすようにできる。$X(\omega)$ を確率変数（random variable）という。
- 確率変数 $X(\omega)$ はすべての標本点 ω に実数を対応させる関数である。
- 具体的にみてみよう。

ω：サイコロで1が出るという事象
ω=1-6の全ての標本点について，X(ω)=1-6（100-600）
と当てはめれば，試行結果を具体的な数値に変換できる
このときのX(ω)を確率変数という

© 2020 Yuki Kato. All Rights Reserved

2.2.2 離散型確率変数の場合

- 確率変数のとりうる値が離散的な場合、すなわち $X(\omega)$ が n 個のとびとびの値 $\{v_1, v_2, v_3, \ldots, v_n\}$ をとる場合、確率変数は離散型であるという。
- n は有限のこともあるし、また無限大のこともある。

$$V = \{v_1, v_2, v_3, \ldots, v_n\}。$$

- 離散型確率変数の例：福引の球の色
- 100 名のグループで行ったパーティーの余興に、空くじなしの福引を用意した。5 色の球（赤、黄、青、白、金）の数とその球が出たときの賞品の価格 v_i を表 2.1 に示した。この v_i を確率変数とすることができる。

事象 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5

© 2020 Yuki Kato. All Rights Reserved

2.2.3 連続型確率変数の場合

- 確率変数のとりうる値が連続な場合には確率変数は連続型であるという。
- そのとりうる値の範囲は、無限区間である場合も、あるいは有限区間である場合もある：

$$V = \{x : a < x < b\}$$

- ただし、a, b は正負の無限大のこともある。

- 連続型確率変数の例：ゼンマイ時計の示す時刻
- 時計屋の店先にゼンマイ時計が並んでいる。それら時計の示す時刻の誤差はまちまちで偶然的に（確率的に）決まっているとする。
- もし時計の針が示す時間を、限りなく細かく測ることができれば、時間は連続であるから、時計の示す時刻は連続する値のどれかを示すだろう。人の身長や体重の値の分布についても、その値は連続型確率変数の例である。

© 2020 Yuki Kato. All Rights Reserved

2.3 確率分布

- 確率の散らばり方を確率分布と呼ぶ。度数／（度数の合計）を相対度数という。

- 2.3.1 確率関数と確率密度関数
- (a) 離散分布の確率関数
- 確率変数のとりうる値が離散的な場合

$$V = \{v_1, v_2, v_3, \ldots\}$$

- と書き、$x = v_k$ をとる確率 $f(v_k)$ を確率関数 (probability function) という。離散分布の場合の確率関数の基本的性質は次のようになる。
  - 性質 1： $f(v_k) \geq 0.$
  - 性質 2： $\Sigma_k f(v_k) = 1.$

© 2020 Yuki Kato. All Rights Reserved

2.3 確率分布（2）

- 離散型確率変数の例：福引の球の色
- 福引の例（表 2.1）では、福引の総数が100であるから、 x =100, 200, 300, 500, 800 (赤、黄、青、白、金)のを取る確率 $f(v_k)$ は、それぞれ 0.45, 0.25, 0.15, 0.10, 0.05 となっている。

事象 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5

- 連続型確率変数は下記の通り
  - 性質 1： $f(v_k) \geq 0.$
  - 性質 2： $\Sigma_k f(v_k) = 1.$

性質 1 ： $f(x) \geq 0.$

性質 2 ： $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1.$

確率密度関数

© 2020 Yuki Kato. All Rights Reserved

確率関数VS確率密度関数

比較項目	確率関数（確率質量関数 / PMF）	確率密度関数（PDF）
対象の確率変数	**離散型**（サイコロの目、コインの表裏の回数など）	**連続型**（身長、体重、時間、気温など）
関数の値の意味	その値をとる**「確率そのもの」**	その値の周辺における**「確率の密度（集中度合い）」**
値の範囲	$0 \leq P(X = x) \leq 1$	$f(x) \geq 0$ **（1を超えることもある）**
確率の求め方	関数の値をそのまま読む	**区間で積分する**
全体の合計	$\sum P(X = x) = 1$（全パターンの総和が 1）	$\int_{-\infty}^{\infty} f(x)\,dx = 1$（グラフ全体の面積が 1）
特定の1点での確率	$P(X = a)$ は有限の値をとる	$P(X = a) = 0$（1点だけの確率は常に0）

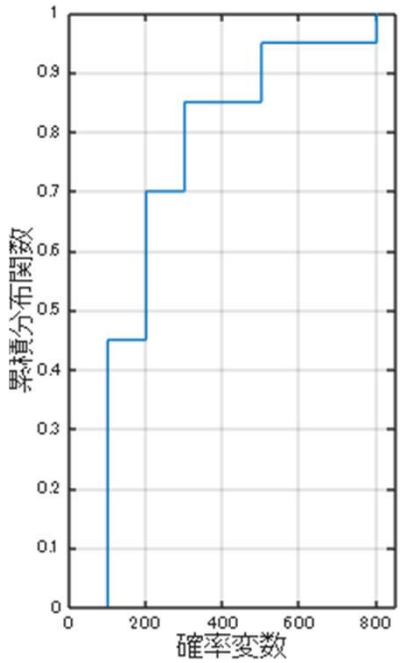
© 2020 Yuki Kato. All Rights Reserved

2.3.2 累積分布関数

- $P(\{\omega : X(\omega) \leq x\})$ は確率変数 $X(\omega)$ が値 x 以下である事象が起きる確率である。
- これを確率変数 X の（累積）分布関数 (cumulative distribution function) という：

$$F_X(x) = P(\{\omega : X(\omega) \leq x\}) \quad \text{ただし} \quad -\infty < x < \infty$$

- 累積分布関数には次のような性質がある。
  1. $a < b$ ならば $F_X(a) \leq F_X(b)$
  2. 右連続、すなわち $\lim_{x \to a+} F_X(x) = F_X(a)$
  3. $\lim_{x \to -\infty} F_X(x) = 0$
  4. $\lim_{x \to +\infty} F_X(x) = 1$



- $F_X(x)$ は、混乱がなければ $F(x)$ と書くこともある。

© 2020 Yuki Kato. All Rights Reserved

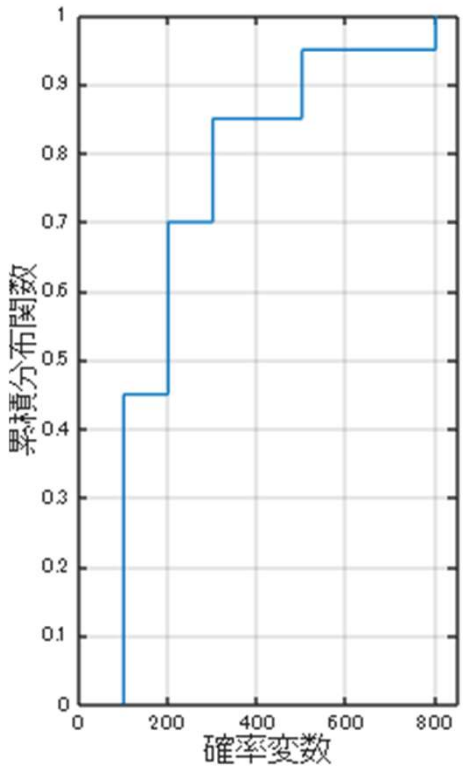
2.3.2 累積分布関数

- 確率関数と累積分布関数との関係は次のとおり
- 実際には累積分布関数を先に決め、そのあとで確率関数を決める方がわかりやすい
- 離散型確率変数の場合
- $$f(v_k) = F(v_k) - F(v_{k-1})$$

- 連続型確率変数の場合

➢ 累積分布関数： $$F(x) = \int_{-\infty}^{\infty} f(v)\,dv$$

➢ 確率密度関数： $$f(x) = \frac{d}{dx} F(x)$$

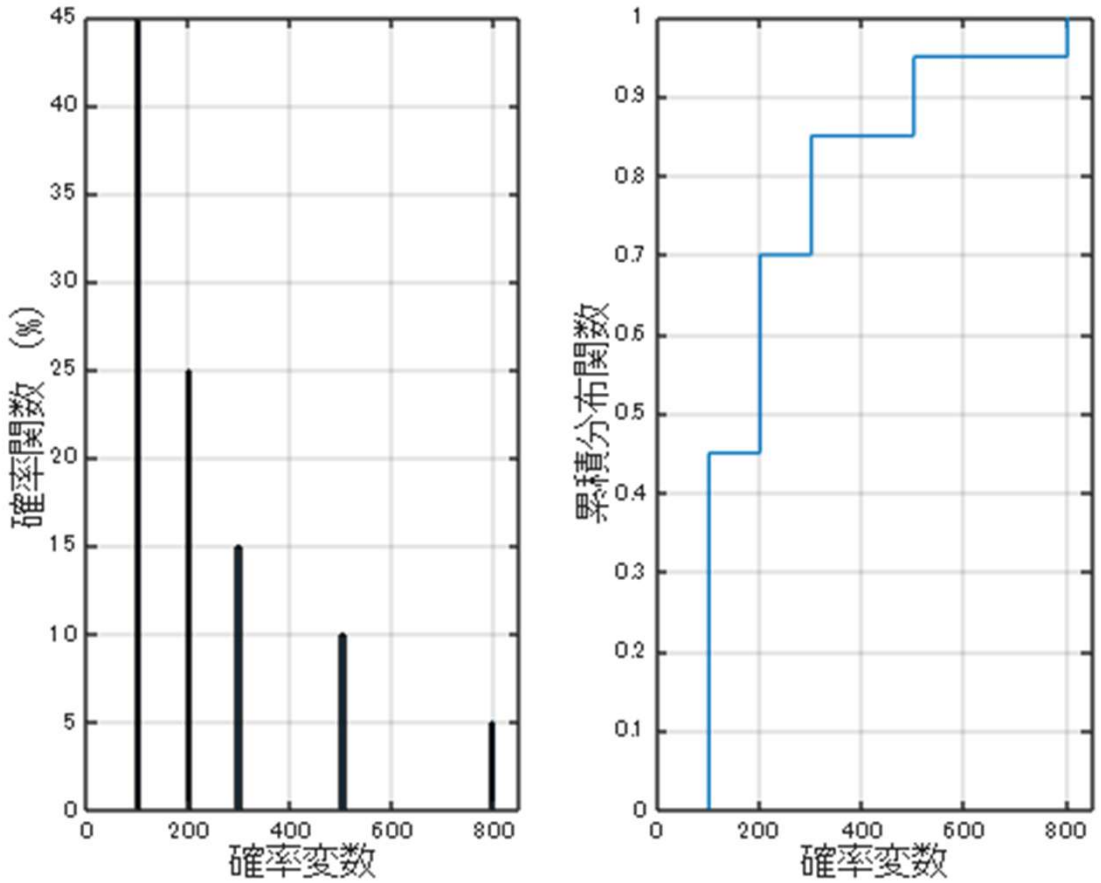


© 2020 Yuki Kato. All Rights Reserved

確率関数と累積分布関数のプロット

■ 右の表に示す確率関数をプロットしましょう

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0



© 2020 Yuki Kato. All Rights Reserved

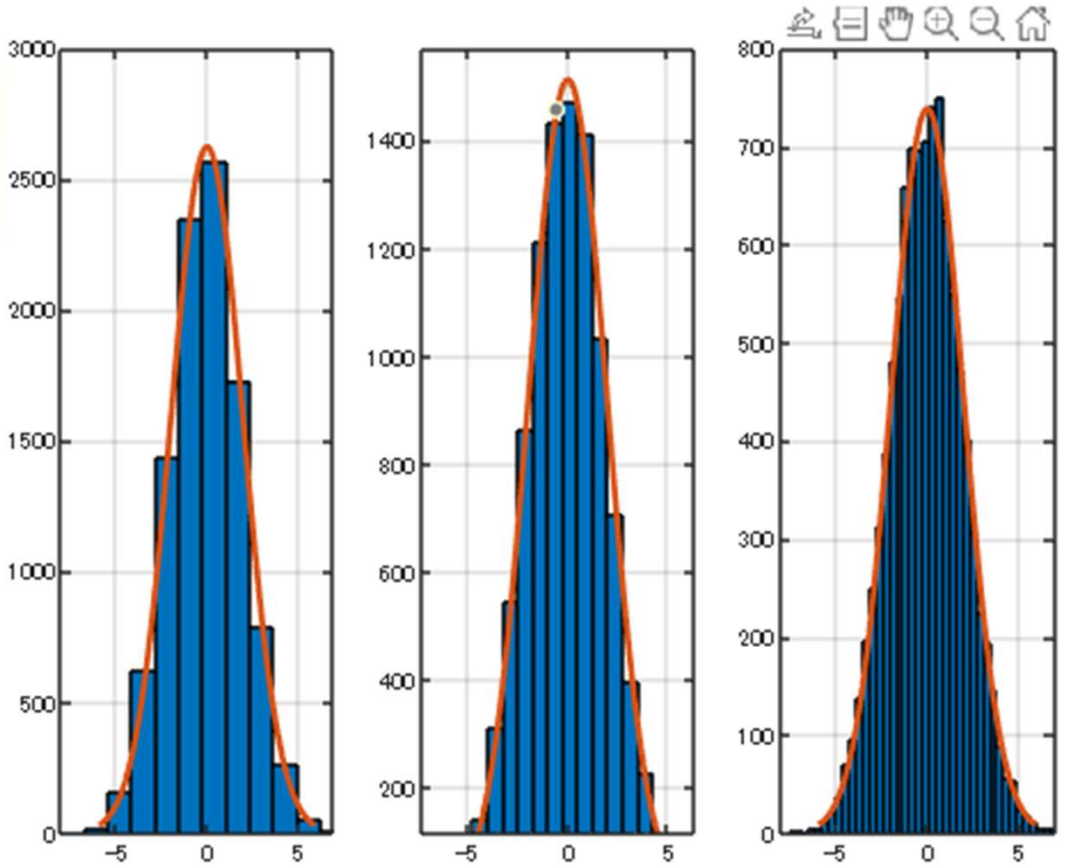
連続型確率変数を有限幅のピンに分けたときの確率変数

- ピンの幅を細かく区切った時の分布の振る舞い及びサンプル数を増やす場合の結果を見る
- 分布結果をサンプル数で割れば確率密度関数が得られる

表 2.3　100 台の時計が正午に示す時刻.

時刻下限	11:53	11:55	11:57	11:59	12:01	12:03
時刻上限	11:55	11:57	11:59	12:01	12:03	12:05
時計の数	1	8	21	39	28	3

時計ごとに時刻にばらつきがあるが,
正解値を平均とした正規分布に
従うものとする



© 2020 Yuki Kato. All Rights Reserved

平均値と分散

- 平均値（mean）と分散（variance）は統計における基本的な概念
- 平均値は与えられた分布の確率（密度）関数を重みとした確率変数の値の算術平均であり、期待値（expected value）とも呼ばれる。
- 分散は、平均値の周りでの分布のばらつきの目安を与える。

2.4.1 平均値および分散の定義

1 確率変数に関する平均値および分散は次のように定義される：　　Σvk (k/N)

$$\text{平均値 } \mu = E[X] = \begin{cases} \sum_k v_k f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} x f(x)\mathrm{d}x & \text{（連続分布）} \end{cases} \tag{2.9a}$$

$$\text{分散 } V[X] = E[\{X - E(X)\}^2] = \begin{cases} \sum_k (v_k - \mu)^2 f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\mathrm{d}x & \text{（連続分布）} \end{cases}$$

分散は σ^2 とも書かれ，分散の平方根 σ は**標準偏差**（standard deviation）という．

© 2020 Yuki Kato. All Rights Reserved

平均値と分散2

2.4.2 分散の性質

1 確率変数の分散には次のような性質がある：

1. $V[X] > 0$.
2. $V[X] = E[X^2] - \mu^2$. (2.10)
3. $V[aX + b] = E[\{(aX + b) - E[aX + b]\}^2] = a^2 V[X]$.

- 平均と分散，標準偏差を求めよ
- （Matlabかエクセルで）

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0

© 2020 Yuki Kato. All Rights Reserved

平均値と分散2

2.4.1　平均値および分散の定義

1 確率変数に関する平均値および分散は次のように定義される：

$$\text{平均値 } \mu = E[X] = \begin{cases} \sum_k v_k f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} x f(x)\mathrm{d}x & \text{（連続分布）} \end{cases} \tag{2.9a}$$

$$\text{分散 } V[X] = E[\{X - E(X)\}^2] = \begin{cases} \sum_k (v_k - \mu)^2 f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\mathrm{d}x & \text{（連続分布）} \end{cases}$$

- 平均と分散，標準偏差を求めよ
- （Matlabかエクセルで）

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0

© 2020 Yuki Kato. All Rights Reserved

平均値と分散2

2.4.1 平均値および分散の定義

1 確率変数に関する平均値および分散は次のように定義される：

$$\text{平均値 } \mu = E[X] = \begin{cases} \sum_k v_k f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} x f(x)\mathrm{d}x & \text{（連続分布）} \end{cases} \tag{2.9a}$$

$$\text{分散 } V[X] = E[\{X - E(X)\}^2] = \begin{cases} \sum_k (v_k - \mu)^2 f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\mathrm{d}x & \text{（連続分布）} \end{cases}$$

- 平均と分散，標準偏差を求めよ
- （分散と標準偏差はMatlabかエクセルで）

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0

© 2020 Yuki Kato. All Rights Reserved

平均値と分散2　こたえ

平均 μ は、「(確率変数の値) × (その確率関数の値)」の合計で求まります。

$$E[X] = \mu = \sum v_i f(v_i)$$

$$\begin{aligned} E[X] &= (100 \times 0.45) + (200 \times 0.25) + (300 \times 0.15) + (500 \times 0.10) + (800 \times 0.05) \\ &= 45 + 50 + 45 + 50 + 40 \\ &= 230 \end{aligned}$$

平均（期待値）は 230 です。

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0

© 2020 Yuki Kato. All Rights Reserved

平均値と分散2　こたえ2

$$V[X] = \sum (v_i - \mu)^2 f(v_i)$$

$\mu = 230$ を用いて計算します。

v_i	$f(v_i)$	$v_i - \mu$	$(v_i - \mu)^2$	$(v_i - \mu)^2 f(v_i)$
100	0.45	-130	16,900	$16,900 \times 0.45 = 7,605$
200	0.25	-30	900	$900 \times 0.25 = 225$
300	0.15	70	4,900	$4,900 \times 0.15 = 735$
500	0.10	270	72,900	$72,900 \times 0.10 = 7,290$
800	0.05	570	324,900	$324,900 \times 0.05 = 16,245$
合計				**32,100**

Google スプレッドシートにエクスポート

$$V[X] = 7,605 + 225 + 735 + 7,290 + 16,245 = 32,100$$

分散は 32,100 です。

標準偏差は179

色 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5
確率 $f(v_i)$	0.45	0.25	0.15	0.10	0.05
累積分布関数 $F(v_i)$	0.45	0.70	0.85	0.95	1.0

© 2020 Yuki Kato. All Rights Reserved

補：母集団全てを計測しない時（標本分散）

- 平均は$\bar{x} = \Sigma x_i / N$で同様の計算だが，標準偏差は

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$ となる．

- 分散計算の基準: 真の分散は、各データ点 xi と母平均 μ の差の二乗和で計算される
- しかし、標本からは母平均 μ が分からないため、代わりに標本平均 $\bar{x}$を使って分散を計算することになる
- $\bar{x}$ は、その標本データセット内で二乗和が最小になるように設計された値である

- したがって、標本平均 $\bar{x}$ を基準として計算された二乗和 $\sum(x_i-\bar{x})^2$ は、真の母平均 μ を基準とした二乗和 $\sum(x_i-\mu)^2$ よりも必ず小さくなる
- この過小評価の偏りを数学的に打ち消し、推定値の期待値が真の値と一致するように補正するために、分母を N−1 に減らす
- これがベッセルの補正であり、N−1 はこの標本データの自由度を表す

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第四回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

統計解析の基礎
1次元，2変数
→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析
→ビッグデータ解析に必須のスキル

© 2020 Yuki Kato. All Rights Reserved

前回の復習：確率変数と確率関数

- 離散型確率変数の例：福引の球の色
- 福引の例（表 2.1）では、福引の総数が100であるから、 $x = v_k = 100, 200, 300, 500, 800$ (赤、黄、青、白、金)のを取る確率 $f(v_k)$ は、それぞれ 0.45, 0.25, 0.15, 0.10, 0.05 となっている。

事象 (ω)	赤	黄	青	白	金
確率変数 $X(\omega) = v_i$	100	200	300	500	800
用意した数	45	25	15	10	5

- 離散/連続型の確率関数は下記の通り
  - 性質 1： $f(v_k) \geq 0.$
  - 性質 2： $\Sigma_k f(v_k) = 1.$

性質 1 ： $f(x) \geq 0.$

性質 2 ： $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1.$

確率密度関数

© 2020 Yuki Kato. All Rights Reserved

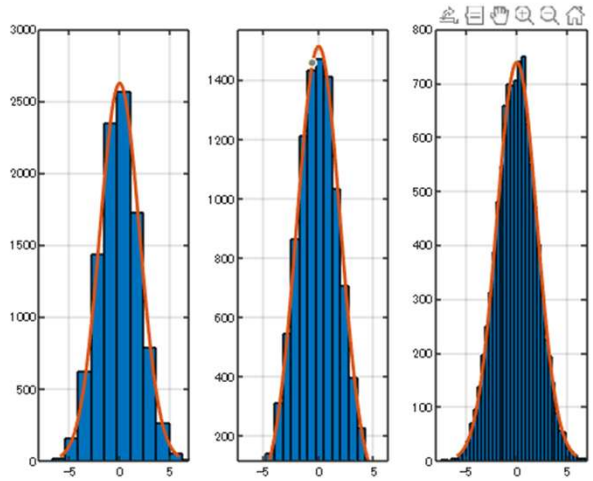
さまざまな分布

- 確率事象において、確率変数の値の分布の様子は様々である
- 確率変数にはある決まった離散的な値をとるものも、あるいは連続的な値をとるものもある
- また値の分布の仕方が一様なものもあれば、ある値を中心として特徴的な様子をみせるものもある
- 分布の仕方は事象の根本的な性質に基づくものである。本章では、それらの分布について考える。

前回プロットした時計の指す時刻についての確率分布
確率分布は無数にあるが，代表的な物とその性質を学んでいこう

表 2.3　100 台の時計が正午に示す時刻.

時刻下限	11:53	11:55	11:57	11:59	12:01	12:03
時刻上限	11:55	11:57	11:59	12:01	12:03	12:05
時計の数	1	8	21	39	28	3



© 2020 Yuki Kato. All Rights Reserved

3.1 離散分布

- 3.1.1 ２点分布
- コイントスのように結果が２種類（表（x=0）か裏（x=1）か）のパターンからなる
- それらが起こる確率 P(x) が P(0)=p、P(1)=q=1−p である試行（実験）をベルヌーイ試行（Bernoulli trial）という
- またこのときの分布を２点分布（two point distribution）という

- 一様乱数からベルヌーイ試行を作るスクリプトを動かそう

© 2020 Yuki Kato. All Rights Reserved

2 点分布

乱数発生器のシード（初期値）を0に
ツイスターアルゴリズム

3.1.1 2点分布（ベルヌーイ試行）

```
figure
rng(0,'twister');
n=100000;
p=0.5;
A= (rand([1,n]) > 1-p);
sum(A,'all')
```

0-1の乱数

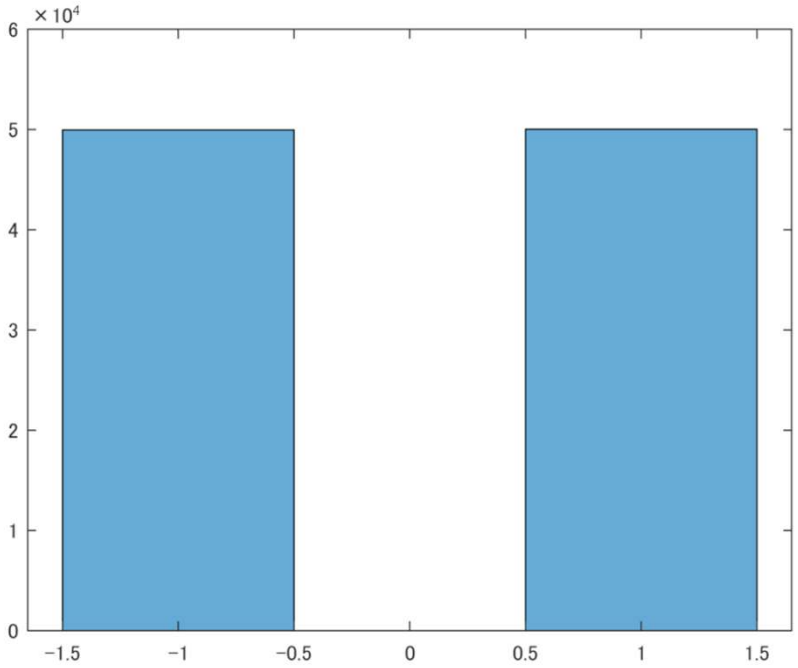
0.5未満で0その他で1
のロジカル値が入る

```
ans = 50036
```

```
n-sum(A,'all')
```

```
ans = 49964
```

```
X=2*A-1;
histogram(X)
hold off
```



Xを成功で1，失敗で-1となる確率変数とする

© 2020 Yuki Kato. All Rights Reserved

3.1.2 2項分布

※Xを成功で1，失敗で-1となる確率変数とする

- ベルヌーイ試行を行ったとき、確率変数 X_i の部分和： $S_n = X_1 + X_2 + \ldots + X_n$ もまた、−n から +n までの整数値をとる確率変数となる。
- n 回のコイントスを行うベルヌーイ試行を繰り返したとき、表が k 回出て、裏が (n − k) 回出る確率関数は
- $f(k) = {}_nC_k\, p^k\, (1-p)^{n-k}, \qquad k = \{0, 1, 2, 3, \ldots, n\}$

n−k 回の失敗が連続して起こる確率

k 回の成功が連続して起こる確率

n 回の試行のうち、成功が k 回発生するパターンの数

- この分布を 2 項分布（binomial distribution）と呼ぶ。
- ${}_nC_k$は 2 項係数 ${}_nC_k = n! / \{k!(n-k)!\}$ である。

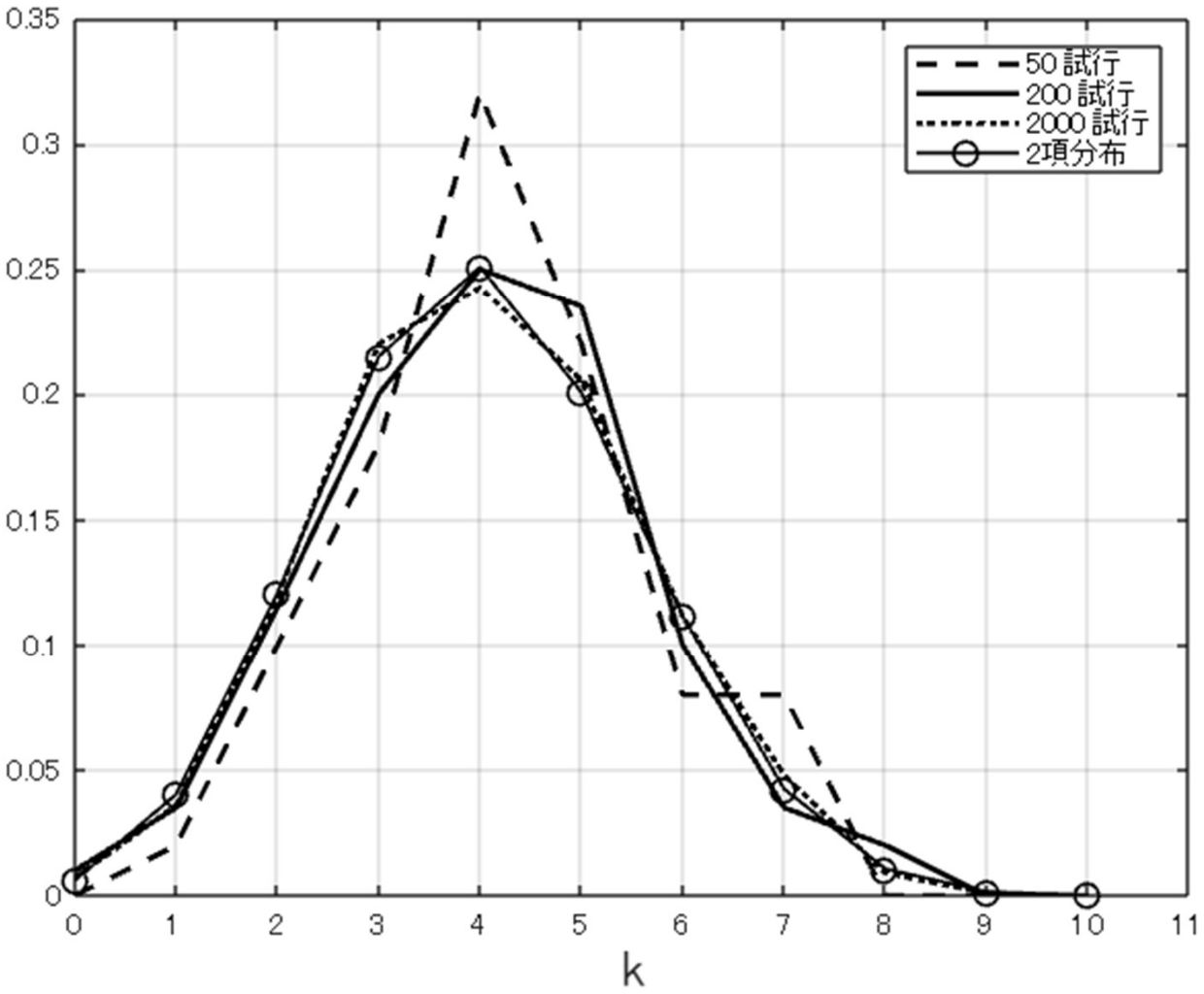
- ベルヌーイ試行を繰り返して 2 項分布を作るスクリプトを実行しよう

© 2020 Yuki Kato. All Rights Reserved

3.1.2 ２項分布　プログラム

■ Nを100回に変えて実行してみよう

2000回施行すればほぼ二項分布と一致

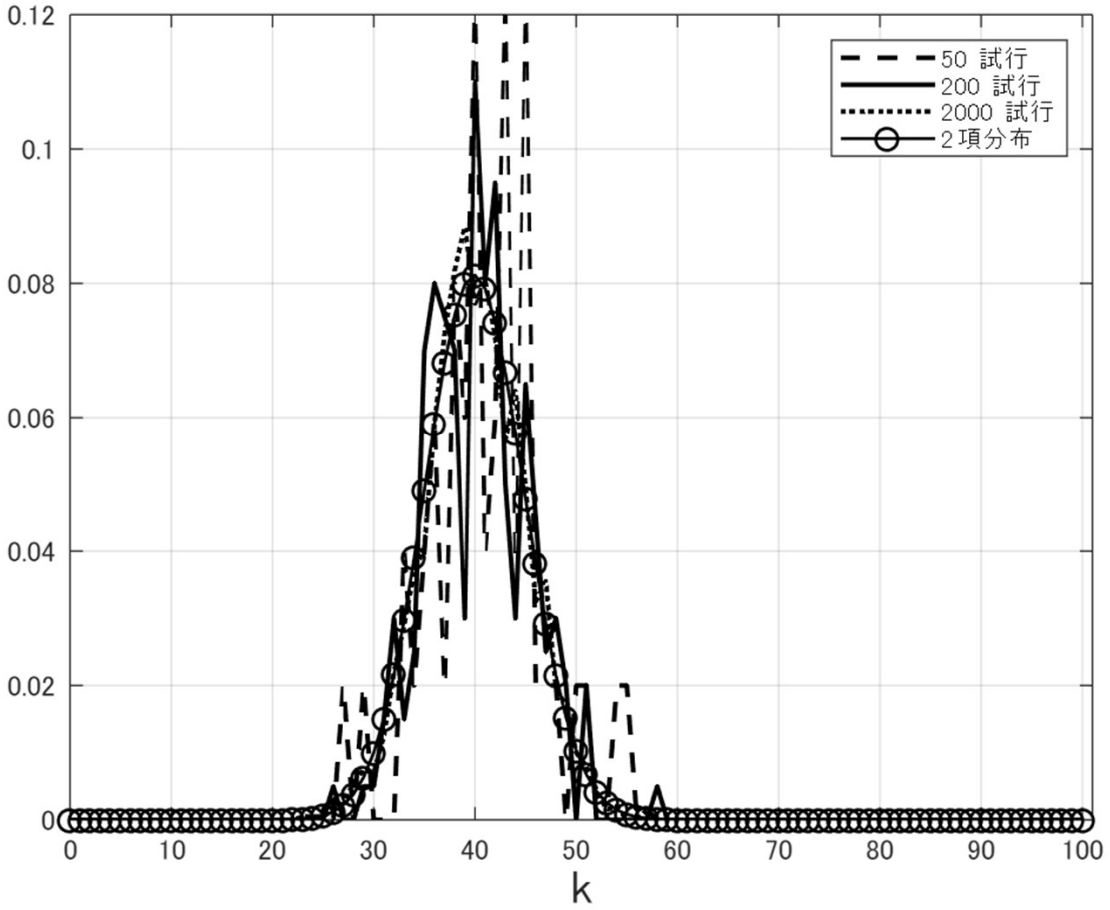


© 2020 Yuki Kato. All Rights Reserved

3.1.2 ２項分布　プログラム

■ Nを100回に変えて実行してみよう

分散が小さく



© 2020 Yuki Kato. All Rights Reserved

3.1.2 2項分布　プログラム

- 2回目の講義で説明した下記問題につながる
- $f(k) = {}_nC_k\, p^k\, (1-p)^{n-k}$

- コイントス：表と裏が1/2
  - 10 回中で表が 3 回以下しか出ない確率：　17.2％
  - 100 回中で表が 30 回以下しか出ない確率：　0.0039％

n=10, p=0.5　　n=100, p=0.5

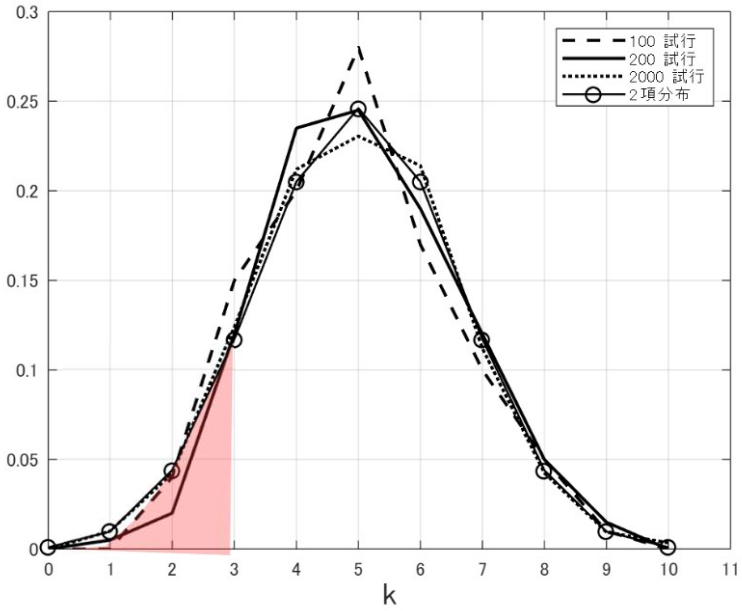
© 2020 Yuki Kato. All Rights Reserved

3.1.2 2項分布　プログラム

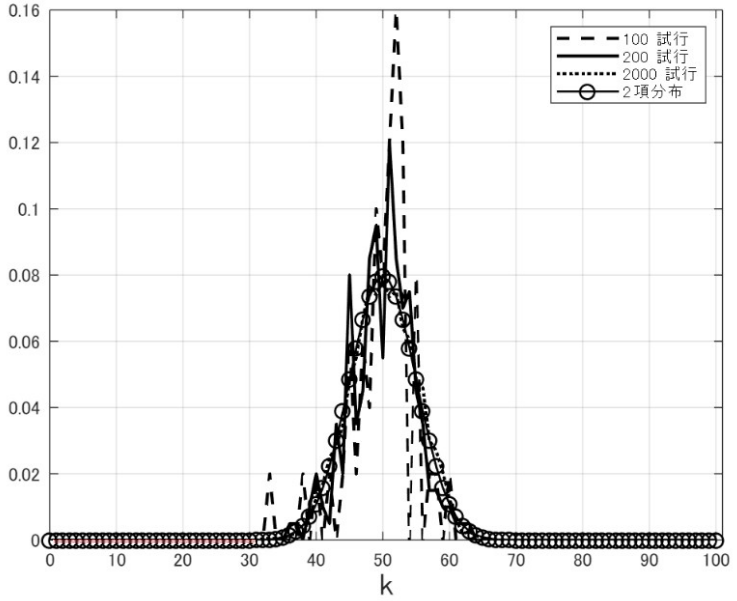
- 2回目の講義で説明した下記問題につながる
- $f(k) = {}_nC_k\, p^k\, (1-p)^{n-k}$

- コイントス：表と裏が1/2
  - 10 回中で表が 3 回以下しか出ない確率：　17.2%
  - 100 回中で表が 30 回以下しか出ない確率：　0.0039%

n=10, p=0.5



n=100, p=0.5



大数の法則

© 2020 Yuki Kato. All Rights Reserved

二項分布の分散　（1）

① 期待値 $E[X_i]$ を求める:

$$\begin{aligned} E[X_i] &= 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) \\ &= 1 \cdot p + 0 \cdot (1 - p) \\ &= p \end{aligned}$$

② $E[X_i^2]$ を求める:

$$\begin{aligned} E[X_i^2] &= 1^2 \cdot P(X_i = 1) + 0^2 \cdot P(X_i = 0) \\ &= 1 \cdot p + 0 \cdot (1 - p) \\ &= p \end{aligned}$$

③ 分散 $V[X_i]$ を求める:

分散の定義式 $V[Y] = E[Y^2] - (E[Y])^2$ を使います。

$$\begin{aligned} V[X_i] &= E[X_i^2] - (E[X_i])^2 \\ &= p - p^2 \\ &= p(1 - p) \end{aligned}$$

© 2020 Yuki Kato. All Rights Reserved

二項分布の分散　（2）

ステップ 2: 独立な確率変数の分散の性質を利用する

確率統計学には、**独立な確率変数の和の分散は、それぞれの分散の和に等しい**という重要な性質があります。

$X_1, X_2, \ldots, X_n$ がすべて独立であるため（これが二項分布の定義条件です）、以下の関係が成り立ちます。

$$V[X] = V[X_1 + X_2 + \cdots + X_n] = V[X_1] + V[X_2] + \cdots + V[X_n]$$

ステップ 3: 二項分布の分散を導出する

ステップ1で求めた $V[X_i] = p(1-p)$ を、上記の和の式に代入します。

$$V[X] = \underbrace{p(1-p) + p(1-p) + \cdots + p(1-p)}_{n個}$$

したがって、二項分布の分散は $p(1-p)$ を n 回足したものであり、以下の式が導出されます。

$$\mathbf{V[X] = np(1-p)}$$

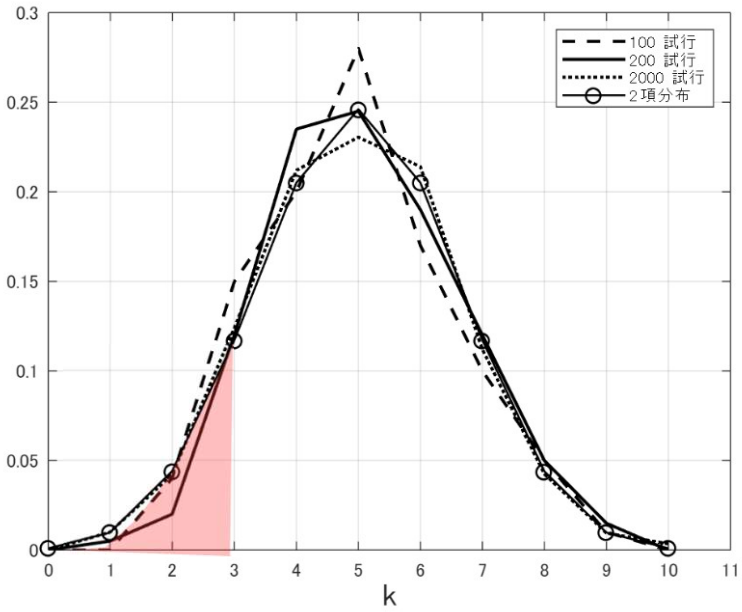
© 2020 Yuki Kato. All Rights Reserved

3.1.2 ２項分布　プログラム

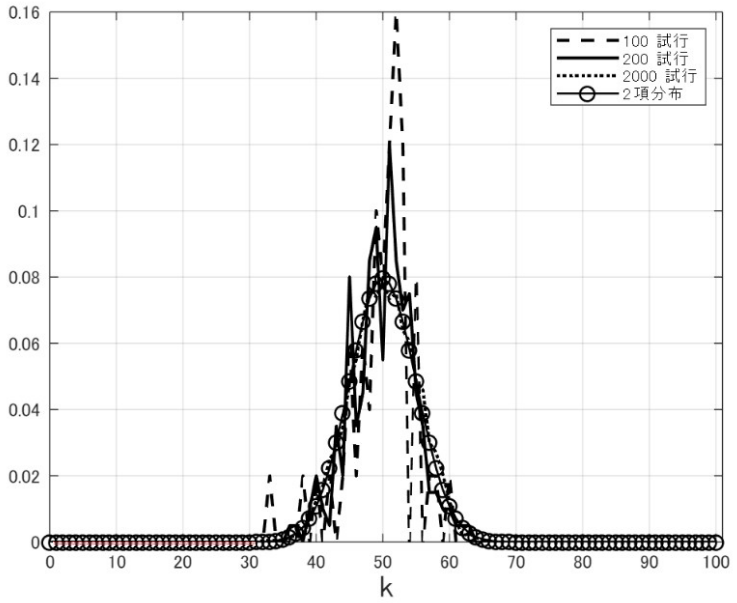
■ 試行回数がn倍，標準偏差は√ n 倍，

$$\mathbf{V}[\mathbf{X}] = \mathbf{np}(\mathbf{1} - \mathbf{p})$$

n=10, p=0.5



n=100, p=0.5



© 2020 Yuki Kato. All Rights Reserved

3.2 連続分布

- 一様分布，指数分布，ガンマ分布など様々な分布がある

- 3.2.4　正規分布
- これからの議論の中で非常に重要な分布となるのは、正規分布（normal distribution）であり、その確率密度関数は次のようになる：

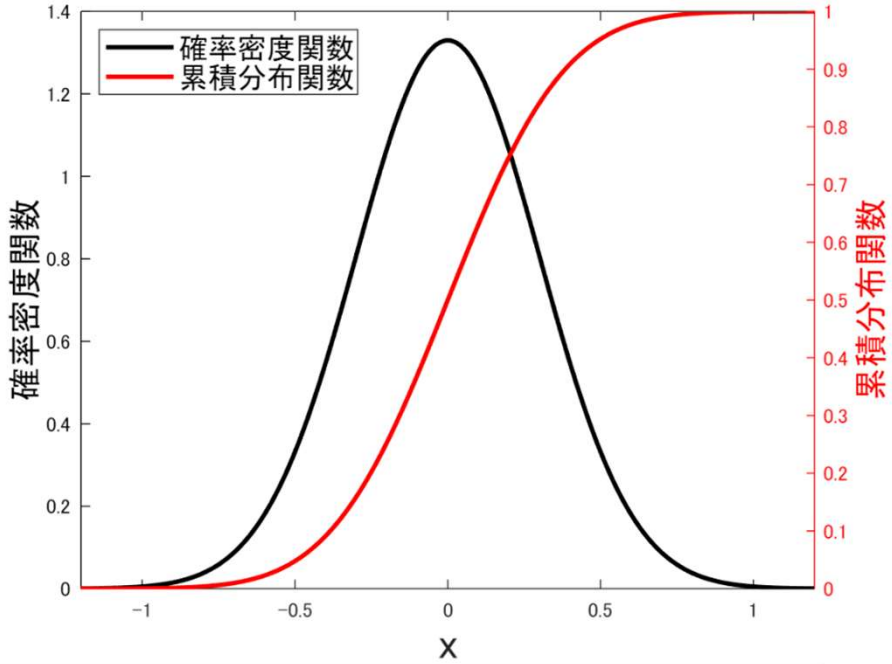
$$f(x) = (1 / \sqrt{(2\pi\sigma^2)}) \exp [-(x - \mu)^2 / (2\sigma^2)].$$

- ただし $-\infty < x < \infty$、$-\infty < \mu < \infty$、$\sigma > 0$。
- 確率密度関数および累積分布関数を求めるためのスクリプトを実行しよう

- μ は分布の平均値、σ^2 は分散（広がりの目安）を与え、この正規分布を $N(\mu, \sigma^2)$ と書く。
- $\mu = 0$、$\sigma^2 = 1$ のとき標準正規分布といい、$N(0, 1)$ と書く。

© 2020 Yuki Kato. All Rights Reserved

正規分布




```
clf
x=-2:0.01:2;
mu=0;
sigma=0.3;
y = normpdf(x,mu,sigma);
z=normcdf(x,mu,sigma);
colororder({'k','r'})
yyaxis left
  plot(x,y,'LineWidth',2);
  ylabel('確率密度関数','FontSize',15)
yyaxis right
  plot(x,z,'LineWidth',2);
  ylabel('累積分布関数','FontSize',15)
xlabel('x','FontSize',20);
legend('確率密度関数','累積分布関数','FontSize',13,'Location','northwest')
xlim([-1.2,1.2])
```

正規分布のPDF
（Probability distribution function）

正規分布のCDF
（Cumulative distribution function）

© 2020 Yuki Kato. All Rights Reserved

正規分布に従うデータ

- 多くの計測値は、平均から離れるほど発生しにくくなるという自然な傾向があるため、正規分布に従うことが多い
- 人の身体的特徴:
  - 身長や体重（特定の年齢・性別の集団内）
  - 手の長さ、足のサイズ、脳の容積など
- 生物の生態:
  - ある種の動物の寿命
  - 植物の葉の長さや実の重さ

- 2. 人的なパフォーマンスや知能指数
- 多くの人が関わるテスト結果や作業効率も、正規分布に近い形になります。
- テストの得点:
  - 知能指数（IQ）や統一された学力試験の得点（平均点が中心になるよう設計されることが多い）
- 作業効率:
  - 特定のタスクを完了するのにかかる時間（非常に速い人も、非常に遅い人も少ない）

- 3. 測定誤差やランダムなノイズ
- 物理学や工学の実験における測定値の誤差は、多くの小さなランダムな要因（環境、機器の微細な振動など）が積み重なって生じるため、正規分布に従うことが知られています。
- 実験の測定誤差:
  - 電子機器のノイズ、天体観測における誤差
- 製品の寸法:
  - 工業製品の製造過程における、わずかな寸法のばらつき

© 2020 Yuki Kato. All Rights Reserved

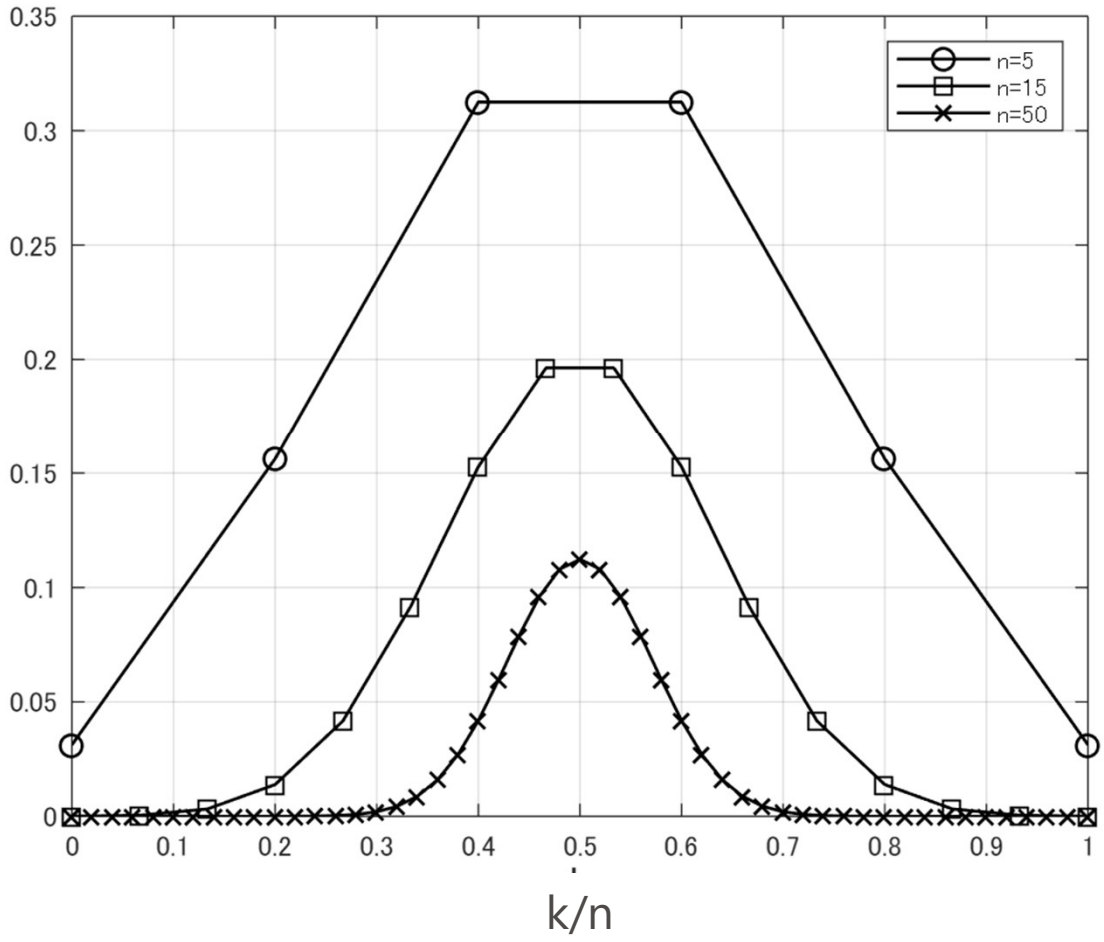
二項分布と正規分布

- コイントスの問題で，5，15，50回の試行を行う場合の二項分布を作成せよ（横軸を0-1で正規化すること）

© 2020 Yuki Kato. All Rights Reserved

二項分布と正規分布

■ コイントスの問題で，5，15，50回の試行を行う場合の二項分布を作成せよ（横軸を0-1で正規化すること）



© 2020 Yuki Kato. All Rights Reserved

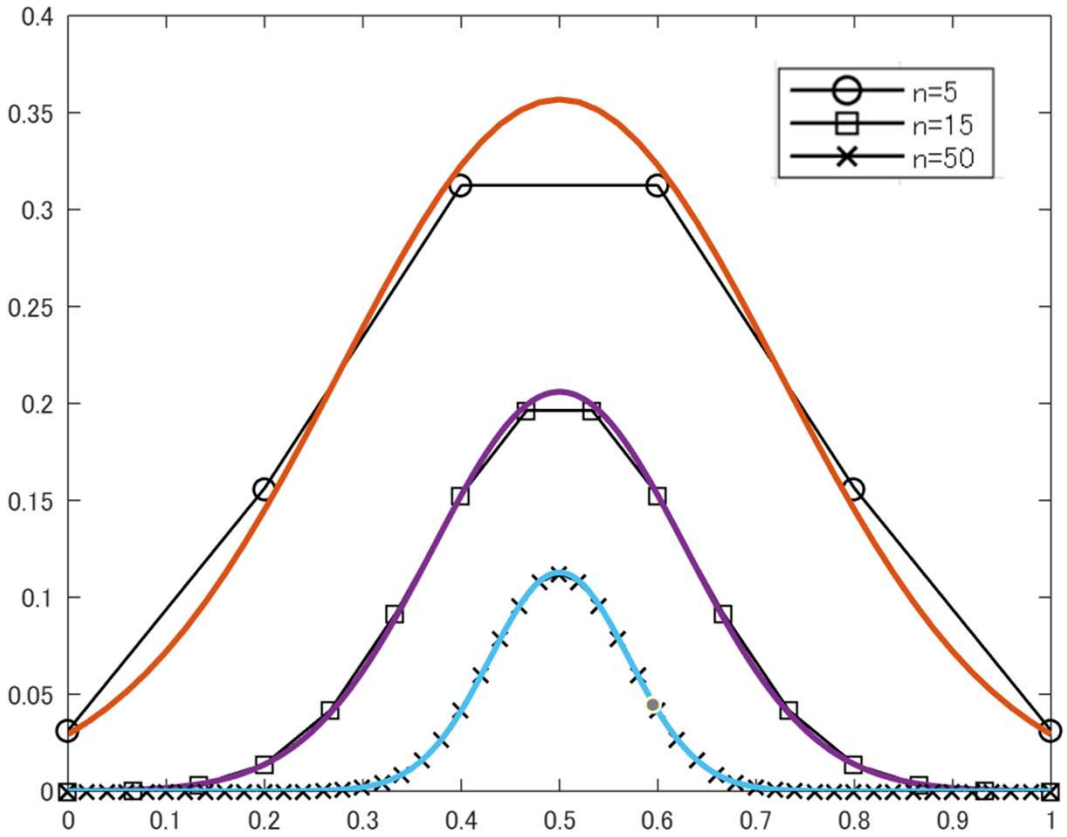
二項分布と正規分布

- 二項分布に正規分布を重ねよ
- 但し，平均はp，標準偏差は$\sqrt{np(1-p)}$を利用すること

© 2020 Yuki Kato. All Rights Reserved

二項分布と正規分布

- 二項分布に正規分布を重ねよ
- 但し，平均はp，標準偏差は$\sqrt{np(1-p)}$を利用すること



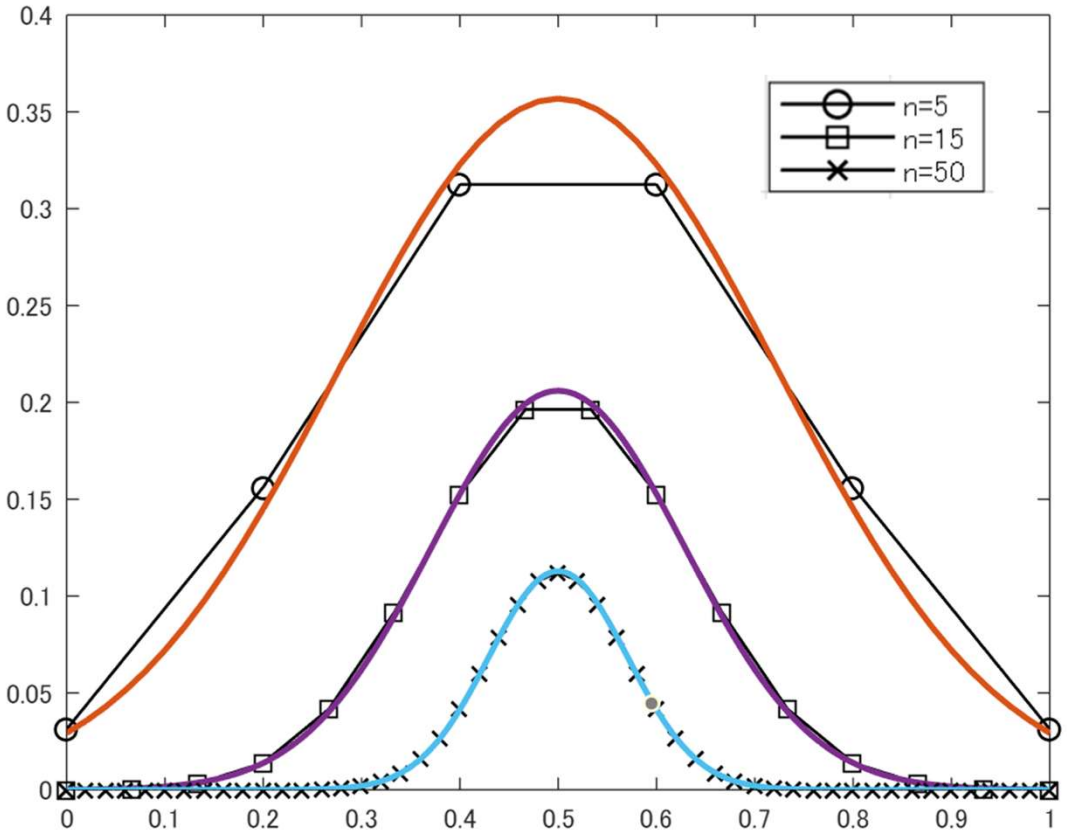
© 2020 Yuki Kato. All Rights Reserved

二項分布と正規分布

- 試行回数が増えると，二項分布は正規分布に近づく

※実は，どんな確率分布であろうと正規分布に収束する

- これを理論的に保証するのが大数の法則と中心極限定理である



© 2020 Yuki Kato. All Rights Reserved

サイコロの例

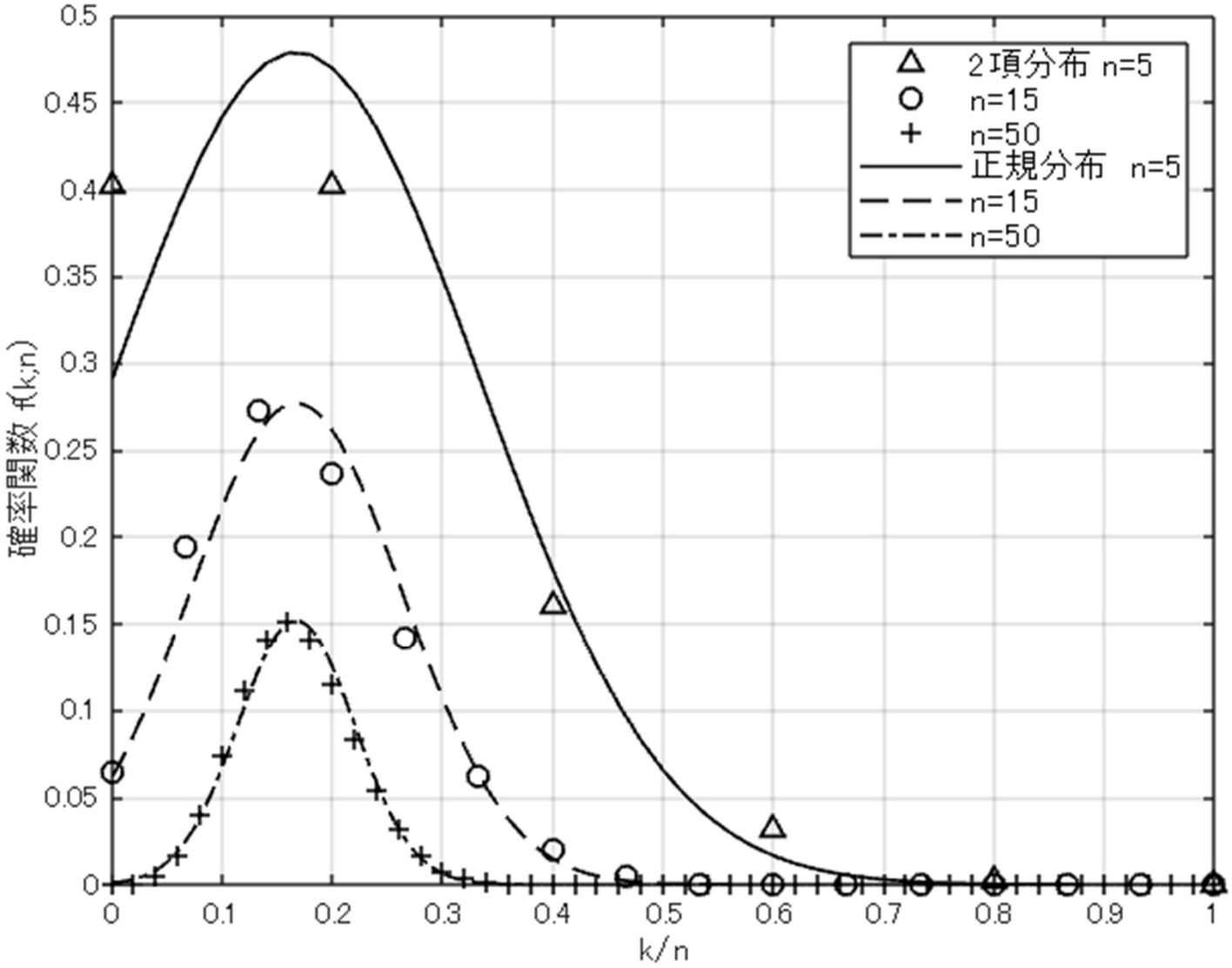
- 元のデータ（1回振ったときの出目）は、1から6までが等しい確率で出る一様分布に従う
- 試行: サイコロを30回振って、その平均値を計算する
- 繰り返し: この「30回の平均値を計算する」という試行を100回、200回、…と繰り返す
- 結果: すると、これらの「平均値の分布」が、元のサイコロの一様分布とは異なり、正規分布（釣鐘型）に近づいていく

- ※このサイコロを何度振っても、個々の出目が出る確率が正規分布に変わることはない点に注意
- 要は，母集団の分布は変わらないが，標本平均の分布が正規分布に近づいていくという事

⇒標本平均の話を正規分布を元に進めることができる

© 2020 Yuki Kato. All Rights Reserved

サイコロの例2



© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第五回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

統計解析の基礎
1次元，2変数
→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析
→ビッグデータ解析に必須のスキル

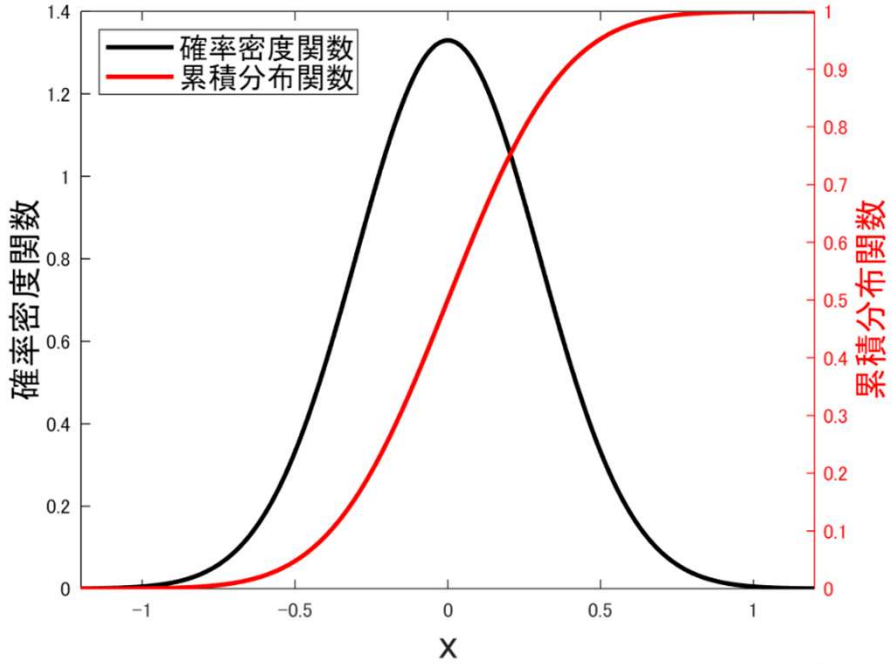
© 2020 Yuki Kato. All Rights Reserved

はじめに

- LMSから参考データとプログラムをダウンロードしておいてください

© 2020 Yuki Kato. All Rights Reserved

前回の復習：正規分布




```
clf
x=-2:0.01:2;
mu=0;
sigma=0.3;
y = normpdf(x,mu,sigma);
z=normcdf(x,mu,sigma);
colororder({'k','r'})
yyaxis left
  plot(x,y,'LineWidth',2);
  ylabel('確率密度関数','FontSize',15)
yyaxis right
  plot(x,z,'LineWidth',2);
  ylabel('累積分布関数','FontSize',15)
xlabel('x','FontSize',20);
legend('確率密度関数','累積分布関数','FontSize',13,'Location','northwest')
xlim([-1.2,1.2])
```

正規分布のPDF
（Probability distribution function）

正規分布のCDF
（Cumulative distribution function）

© 2020 Yuki Kato. All Rights Reserved

前回の復習：正規分布に従うデータ

- 多くの計測値は、平均から離れるほど発生しにくくなるという自然な傾向があるため、正規分布に従うことが多い
- 人の身体的特徴:
  - 身長や体重（特定の年齢・性別の集団内）
  - 手の長さ、足のサイズ、脳の容積など
- 生物の生態:
  - ある種の動物の寿命
  - 植物の葉の長さや実の重さ

- 2. 人的なパフォーマンスや知能指数
- 多くの人が関わるテスト結果や作業効率も、正規分布に近い形になります。
- テストの得点:
  - 知能指数（IQ）や統一された学力試験の得点（平均点が中心になるよう設計されることが多い）
- 作業効率:
  - 特定のタスクを完了するのにかかる時間（非常に速い人も、非常に遅い人も少ない）

- 3. 測定誤差やランダムなノイズ
- 物理学や工学の実験における測定値の誤差は、多くの小さなランダムな要因（環境、機器の微細な振動など）が積み重なって生じるため、正規分布に従うことが知られています。
- 実験の測定誤差:
  - 電子機器のノイズ、天体観測における誤差
- 製品の寸法:
  - 工業製品の製造過程における、わずかな寸法のばらつき

© 2020 Yuki Kato. All Rights Reserved

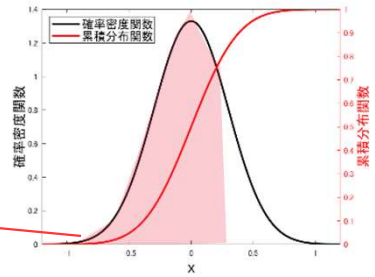
応用例：高校男子生徒の身長調査

- 全国の高等学校の 3 年男子生徒についての身長調査では、平均身長 170 cm、標準偏差 6 cm であった。
- この分布が正規分布であるとしたら、身長 165～175 cm の人の数は全体の何％を占めるか？

- この分布は $N(170, 6^2)$ である。身長が 165～175 cm である生徒の割合は MATLAB を用いれば
  - 59％

```
cdf('Normal',175,170,6)-cdf('Normal',165,170,6)

ans = 0.5953
```



- それでは，160cm以下の男子高校生の割合は？
- 130cm以下の男子高校生の割合は？
- 2m15cm以上の男子高校生の割合は？

© 2020 Yuki Kato. All Rights Reserved

応用例：高校男子生徒の身長調査

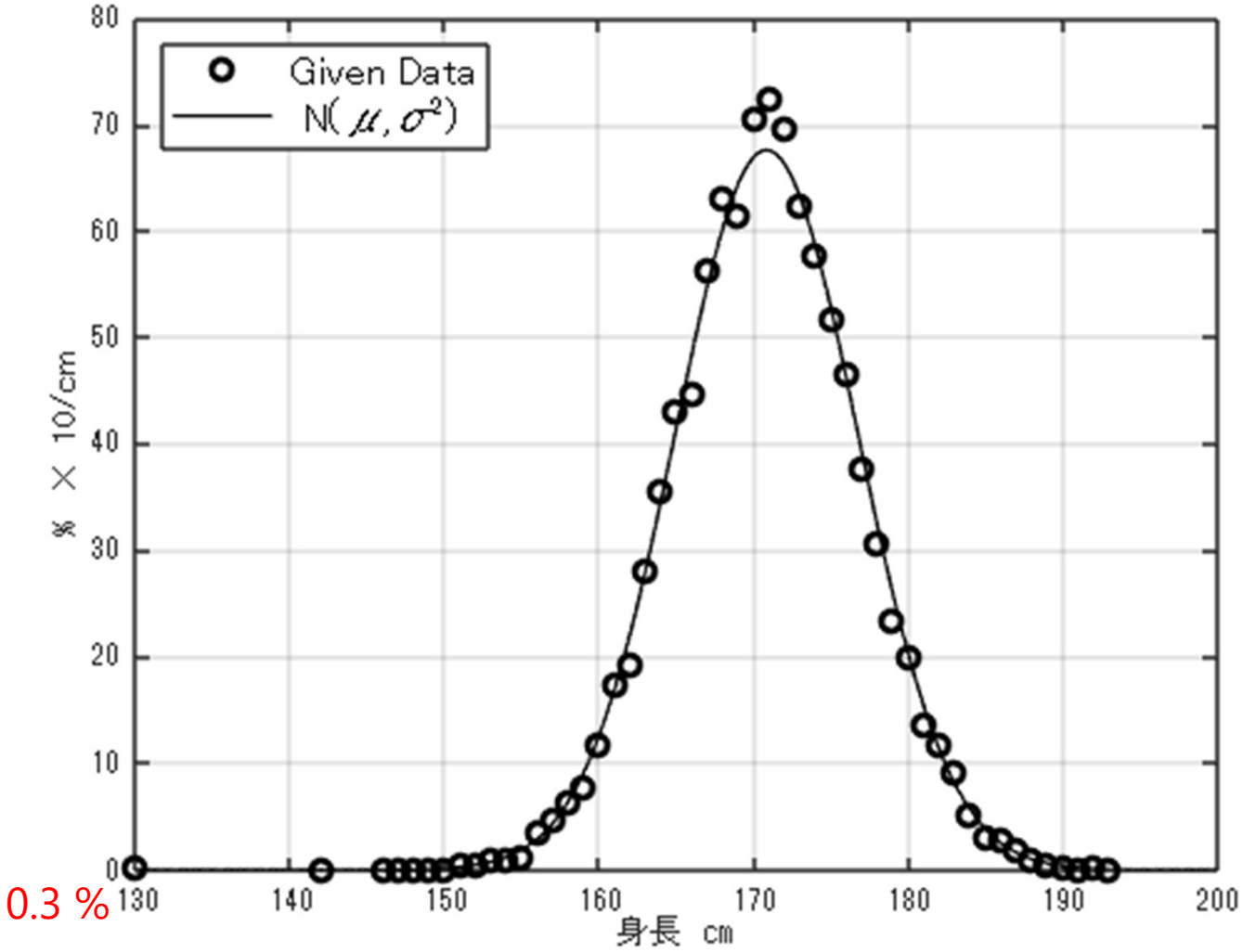
- それでは，160cm以下の男子高校生の割合は？→4.8 %
- 130cm以下の男子高校生の割合は？　→$1.3*10^{-9}$ %
- 2m15cm以上の男子高校生の割合は？　→$3.2*10^{-12}$ %
  - 1学年に全国で100万人⇒$3.2*10^{-6}$ %？→数万年観測しても存在せず？
  - 実際には過去数十年で数人存在する

- なぜ異なる？
  - 年代、地域、栄養状態、競技選抜（バスケットボール選手のような選抜集団）、さらには医学的要因（下垂体性巨人症や特定の遺伝的症候群など）が別々の分布を作る
  - これらが混ざると、単一の正規分布より右裾が厚くなり、215 cm超のような極端値の頻度が理論より増える（外れ値）
- どうする？
  - 「正規分布は近似」と割り切り、極端値を扱うときはサブグループごと（性別・地域・年代・選抜集団など）に別々の分布を当てる
  - 混合分布（Mixture of Gaussians）や極値理論（上位 α%だけを一般化パレート分布で近似など）を検討する
  - 経験分布（実測の分位点）で評価する

© 2020 Yuki Kato. All Rights Reserved

応用例：高校男子生徒の身長調査

■ 実際のデータを重ねてプロット



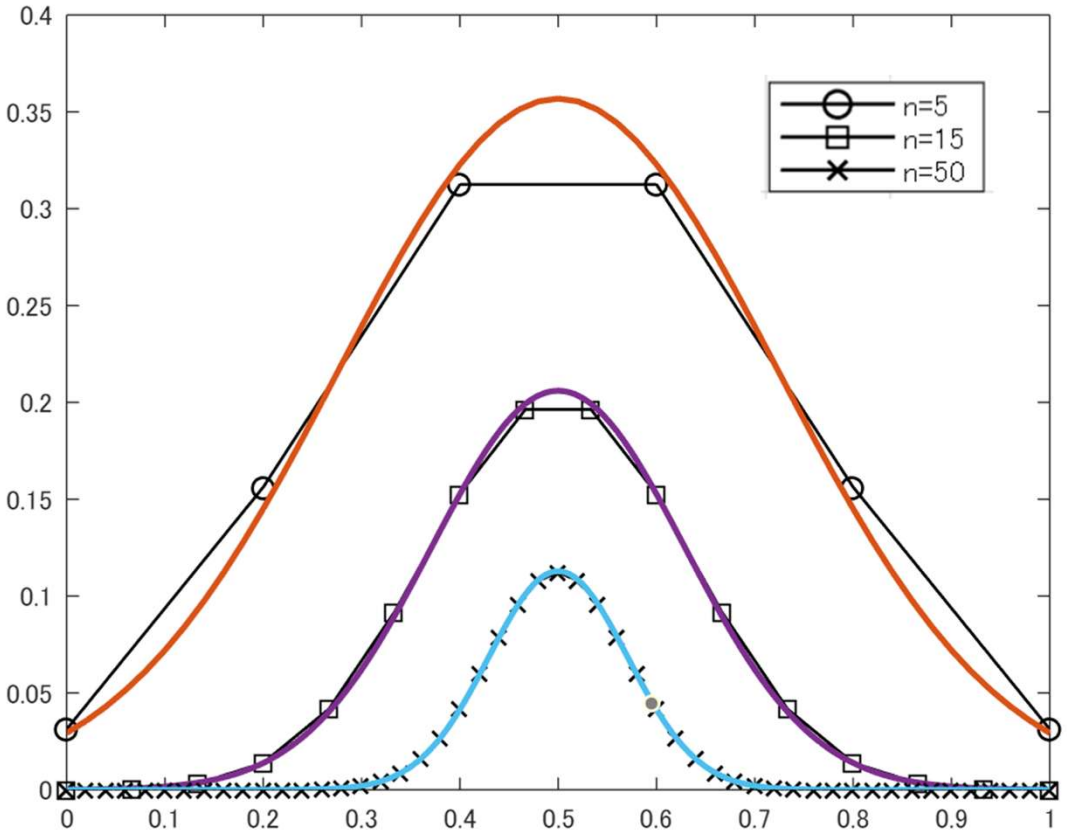
© 2020 Yuki Kato. All Rights Reserved

前回の復習：二項分布と正規分布

- 試行回数が増えると，二項分布は正規分布に近づく

※実は，どんな確率分布であろうと正規分布に収束する

- これを理論的に保証するのが大数の法則と中心極限定理である



© 2020 Yuki Kato. All Rights Reserved

大数の法則とは？

※証明は教科書参照

- 試行回数を増やせば増やすほど、標本から得られる平均値や相対度数が、理論上の真の値に近づいていくことを保証する定理
- 試行回数 n が無限大に近づくとき、標本平均（観測されたデータの平均）*Xn* が、母平均（理論上の期待値）μ に収束する

$$\bar{X}_n = (X_1 + X_2 + \ldots X_n)/n \qquad \lim_{n\to\infty} \bar{X}_n = \mu$$

- 直感的な例（コイン投げ）：
  - 公平なコインを10回投げたとき、表が出る相対度数（例：7/10 = 0.7）は、理論値 (0.5) と大きくズレることがある
  - しかし、コインを1万回、10万回と投げる回数を増やしていくと、表の出る相対度数は限りなく 0.5 に近づく
- 大数の法則は、「確率」という抽象的な概念を、現実の「相対度数」という観察結果に基づいて推定することの正当性を与える
- 統計学で標本調査の結果から母集団の平均を推定できるのは、この法則があるため

© 2020 Yuki Kato. All Rights Reserved

中心極限定理とは？

※証明は教科書参照

- 標本平均の「分布の形状」に関する定理
- 元のデータの分布がどんな形であっても、多数の標本平均を集めると、その分布が正規分布に近づくことを示す
- 母集団の分布が正規分布でなくても、その母集団から十分大きなサイズ n の標本を独立に繰り返し抽出するとき、標本平均 $\bar{X}$ の分布は、標本サイズ n が大きいほど正規分布（釣鐘型の分布）に近づく
- 収束する正規分布のパラメータ:
  - 平均: 母平均 μ
  - 分散: σ^2 /n(母分散 σ^2 を標本サイズ n で割ったもの)
  - 確率変数Zを下記で定義すると，その分布は標準正規分布に近づき

$$\bar{X}_n = \left(X_1 + X_2 + \ldots X_n\right) / n$$

$$Z = \sqrt{n}\left(X - \mu\right) / \sigma$$

$$\lim_{n\to\infty} P\left(Z \le z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-y^2 / 2\right) dy$$

=N(0, 1)

© 2020 Yuki Kato. All Rights Reserved

中心極限定理とは？ 2

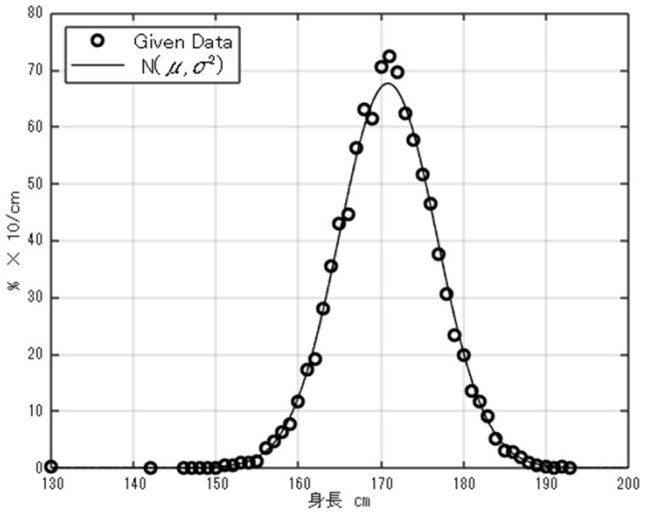
- 直感的な例（非正規分布からのサンプリング）：
  - サイコロの目（一様分布）のように、元のデータが均等な分布である
  - このサイコロを30個振り、その平均値 $\bar{X}$ を記録する試行を1000回繰り返すと、記録された平均値 $\bar{X}$ のヒストグラムは、元のサイコロの分布とは異なり、きれいな釣鐘型（正規分布）になる
- 中心極限定理は、統計的推測のほとんどの根拠となっている
- 多くの統計的検定（t検定など）は、データが正規分布に従うことを前提としている
- 元のデータが正規分布でなくても、中心極限定理のおかげで、標本平均を扱う際には正規分布の理論を適用できる
- 本平均から母平均を推定する際の信頼区間を計算できるのも、標本平均のばらつきが正規分布として捉えられるから

© 2020 Yuki Kato. All Rights Reserved

注意点

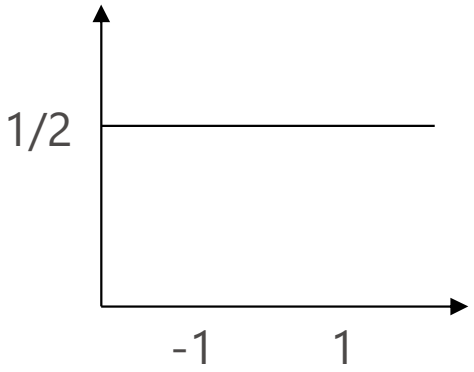
- 母集団の性質が正規分布に従うか？という問題と
- 標本を取り出した際の平均値の分布が正規分布に従う（中心極限定理）という問題は異なることに注意すること

身長

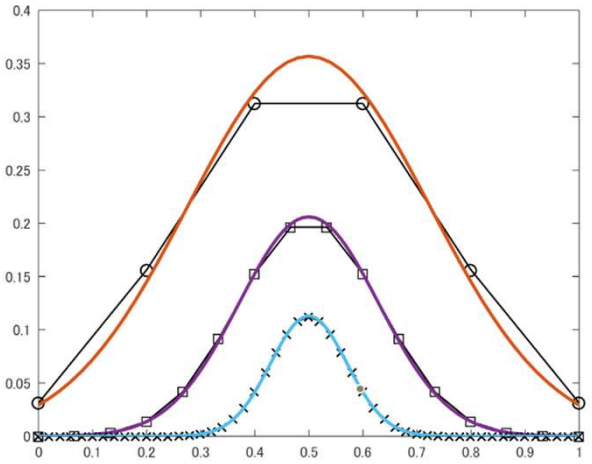


コイントス

母集団の性質
（確率）



標本平均の性質



© 2020 Yuki Kato. All Rights Reserved

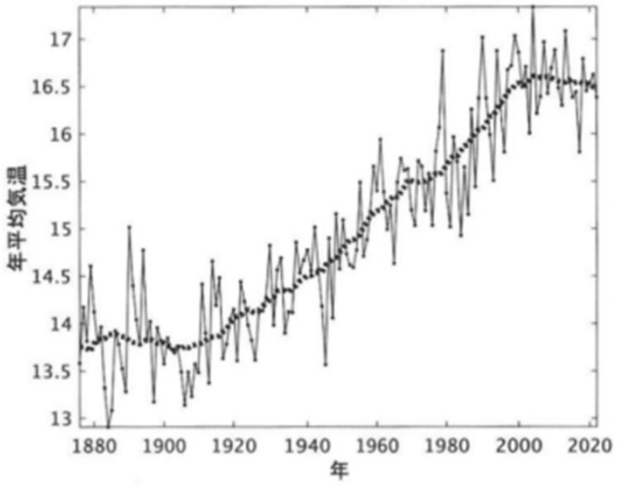
統計学について

- ここまでの部分で，純粋な統計学の話は終わり
- 以降はこの統計学を活かして，いかにデータ解析（時系列信号解析，画像解析，多変量解析，機械学習など）に役立てていくかという応用面の話を進めていく

© 2020 Yuki Kato. All Rights Reserved

時系列データとは？

- ある観測量の時間的な変化を記録したデータを時系列データという。
- このような時間順序がついたデータの現在までの情報をもとに、今後の動き（時期変動）を予測したり
- それに基づいて判断・制御を行いたいという問題は実用上非常に多くある
- このような解析のための統計的な手法、時系列解析について扱う。



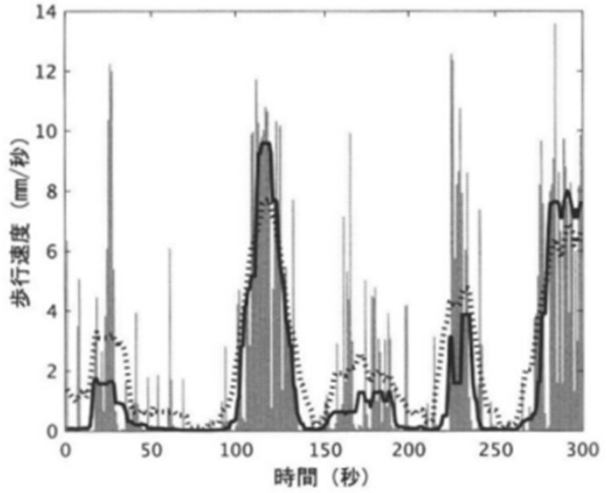


図 10.1　（上）東京の年平均気温（点線は前後 10 年の 21 年移動平均値）と（下）ショウジョウバエの歩行速度（点線と実線はそれぞれ前後 10 秒の 21 秒移動平均値と移動中央値）.

© 2020 Yuki Kato. All Rights Reserved

時系列データとは？

- 10.1.1　時系列
- ある観測量の時間的な変化を記録したものを時系列という。一般には、
- 観測した時刻の組 $\{t_n\} = t_1, t_2, \ldots, t_n$ についての観測値の組 $\{x_n\} = x(t_1), x(t_2), \ldots, x(t_n)$ の形をとる。

- x_n が 1 つの数値（スカラー値）の場合は単変量時系列といい、多数の値の組の場合は多変量時系列
- 観測時刻 t_n が離散的な値のみをとる場合（毎秒/時/日/週/月/年のような定期的観測など）を離散時間時系列
- 連続的な値をとる場合（何かのイベントが起きたときにその時刻とともに記録される場合など）を連続時間時系列

© 2020 Yuki Kato. All Rights Reserved

機械工学で良く用いられるデータ解析

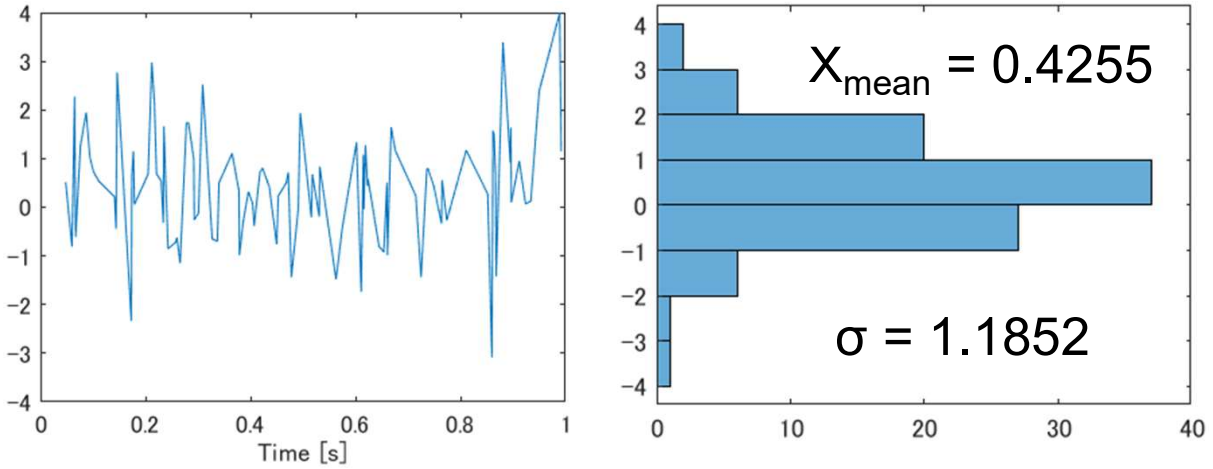
- 単変量時系列データ解析
  - 平均と標準偏差
  - ヒストグラム
  - リサンプリング
  - 差分と積分
  - 移動平均と移動中央値（外れ値フィルタ）
  - 離散フーリエ変換（DFT）と高速フーリエ変換（FFT）
  - 周波数フィルタ
  - ウェーブレット変換
- 多変量時系列データ解析（2変量）
  - 相互相関関数
  - コヒーレンス

© 2020 Yuki Kato. All Rights Reserved

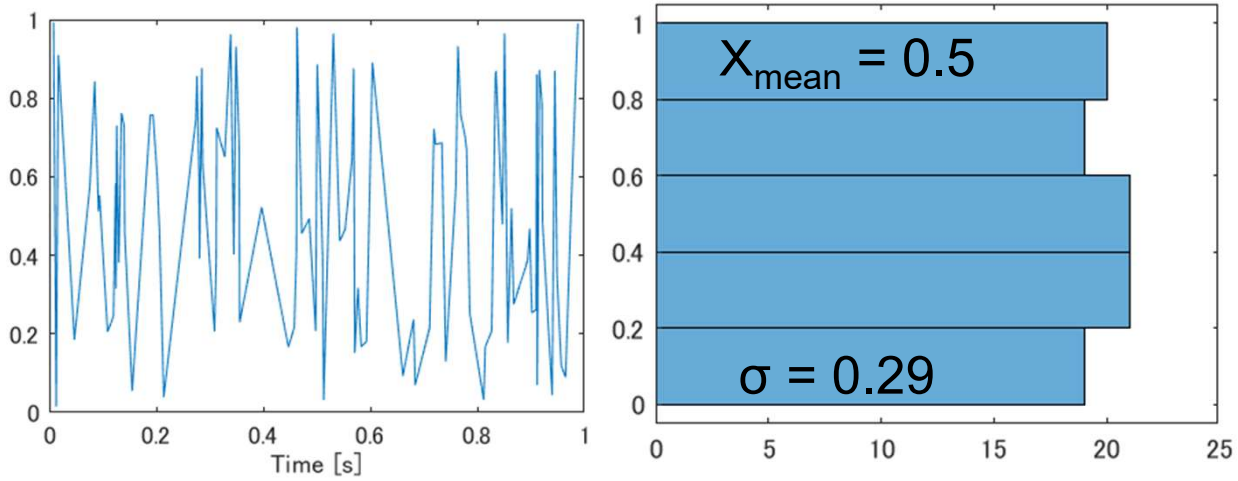
平均と標準偏差，ヒストグラム　ランダム

- 典型的時系列データの平均や標準偏差，ヒストグラム
- 1msの時間刻みで1秒のデータ(n=1000)．そのうちm=100をサンプリング

ランダム（ガウス）



ランダム（一様乱数）

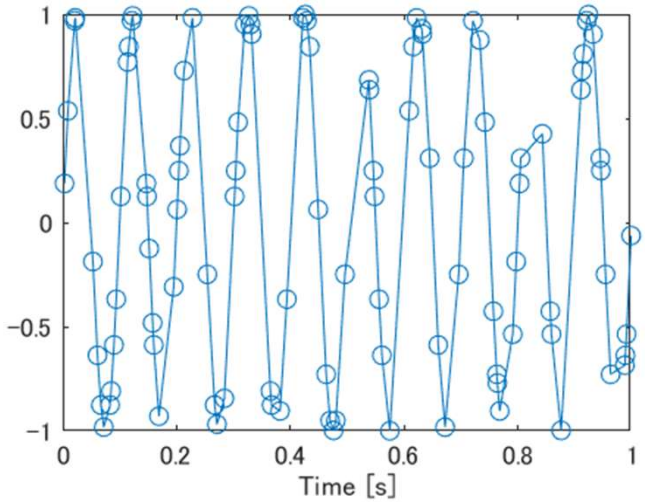


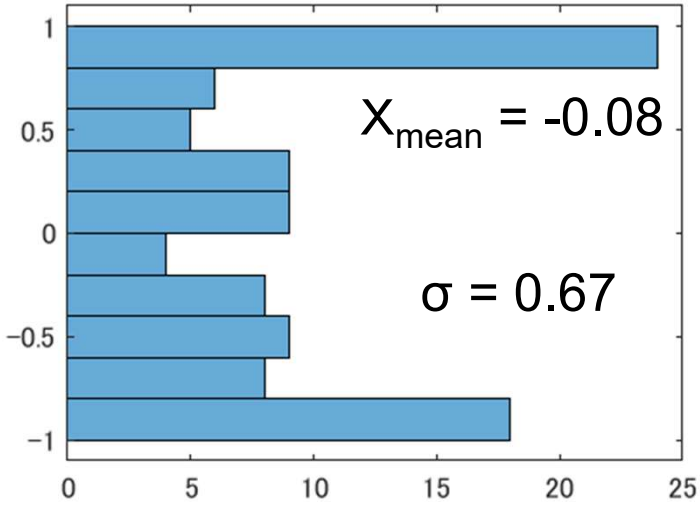
© 2020 Yuki Kato. All Rights Reserved

平均と標準偏差，ヒストグラム　振動10Hz

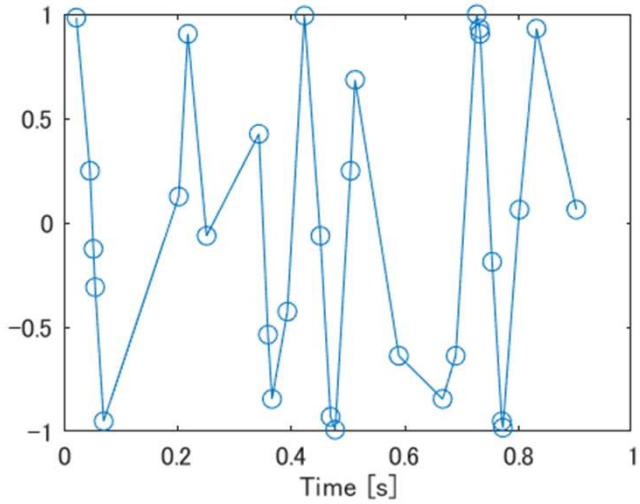
■ 典型的時系列データの平均や標準偏差，ヒストグラム

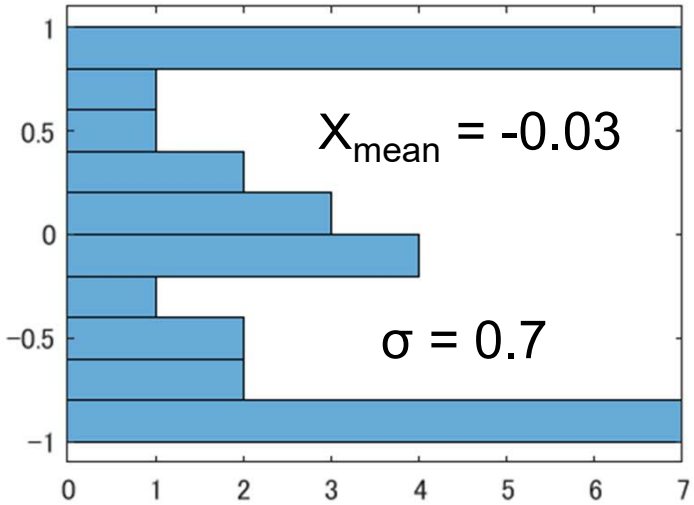
100サンプル





30サンプル





■ 等間隔でないと時系列解析はやりづらい・・・

© 2020 Yuki Kato. All Rights Reserved

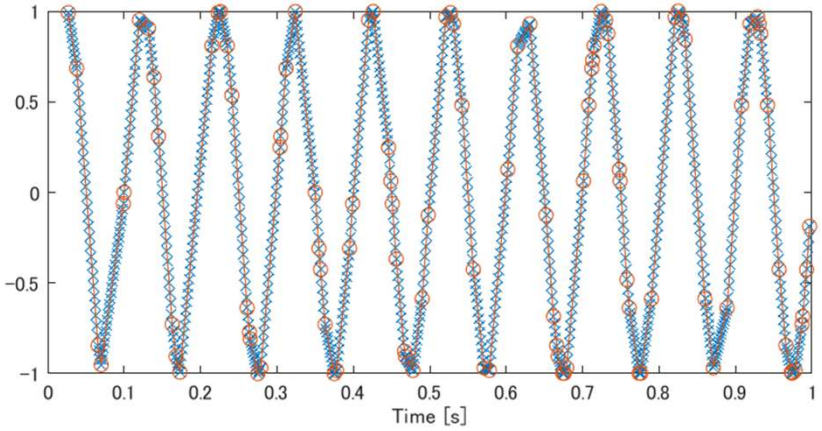
リサンプリング

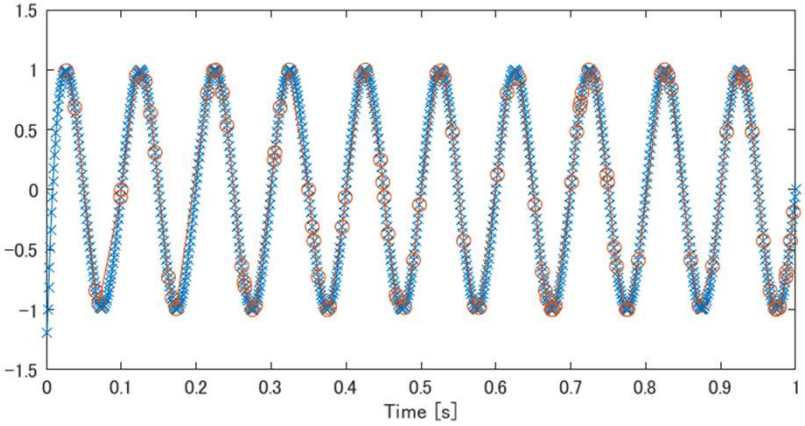
X：内挿
○：サンプリングしたデータ

線形内挿

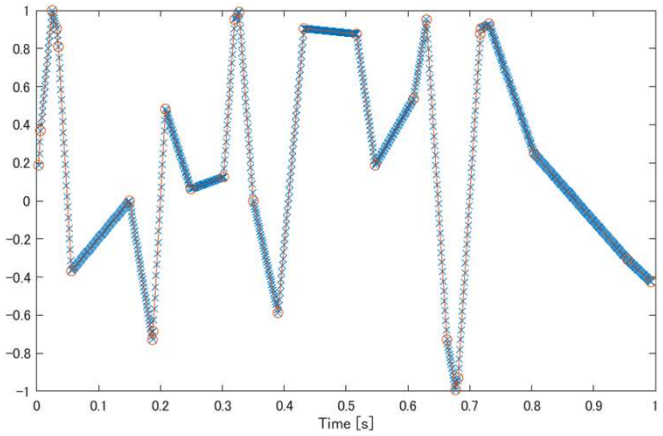
スプライン内挿

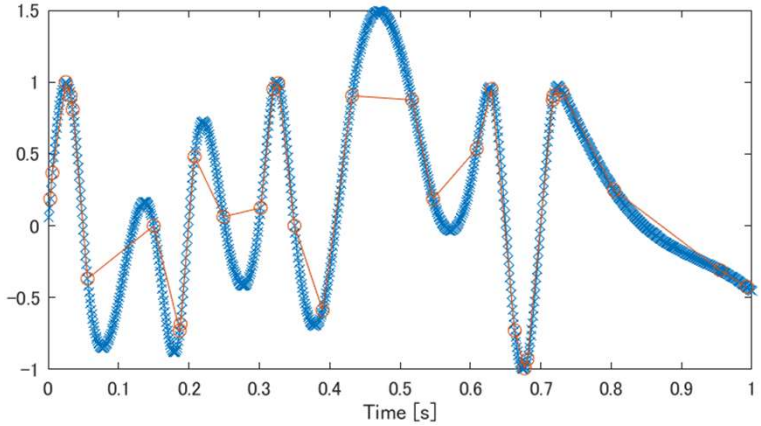
100サンプル





30サンプル





- ギャップが大きい場合は自己回帰モデル
- サンプルレートが不足の場合は圧縮センシング等の特殊な推定アルゴリズムが必要

© 2020 Yuki Kato. All Rights Reserved

以降の注意

- これより、とくに説明がない限り等間隔離散時間の単変量時系列

- $\{x_t\} = x_1, x_2, \ldots, x_n \qquad (t_i = t_{i-1} + \Delta t)$

- を対象とする

© 2020 Yuki Kato. All Rights Reserved

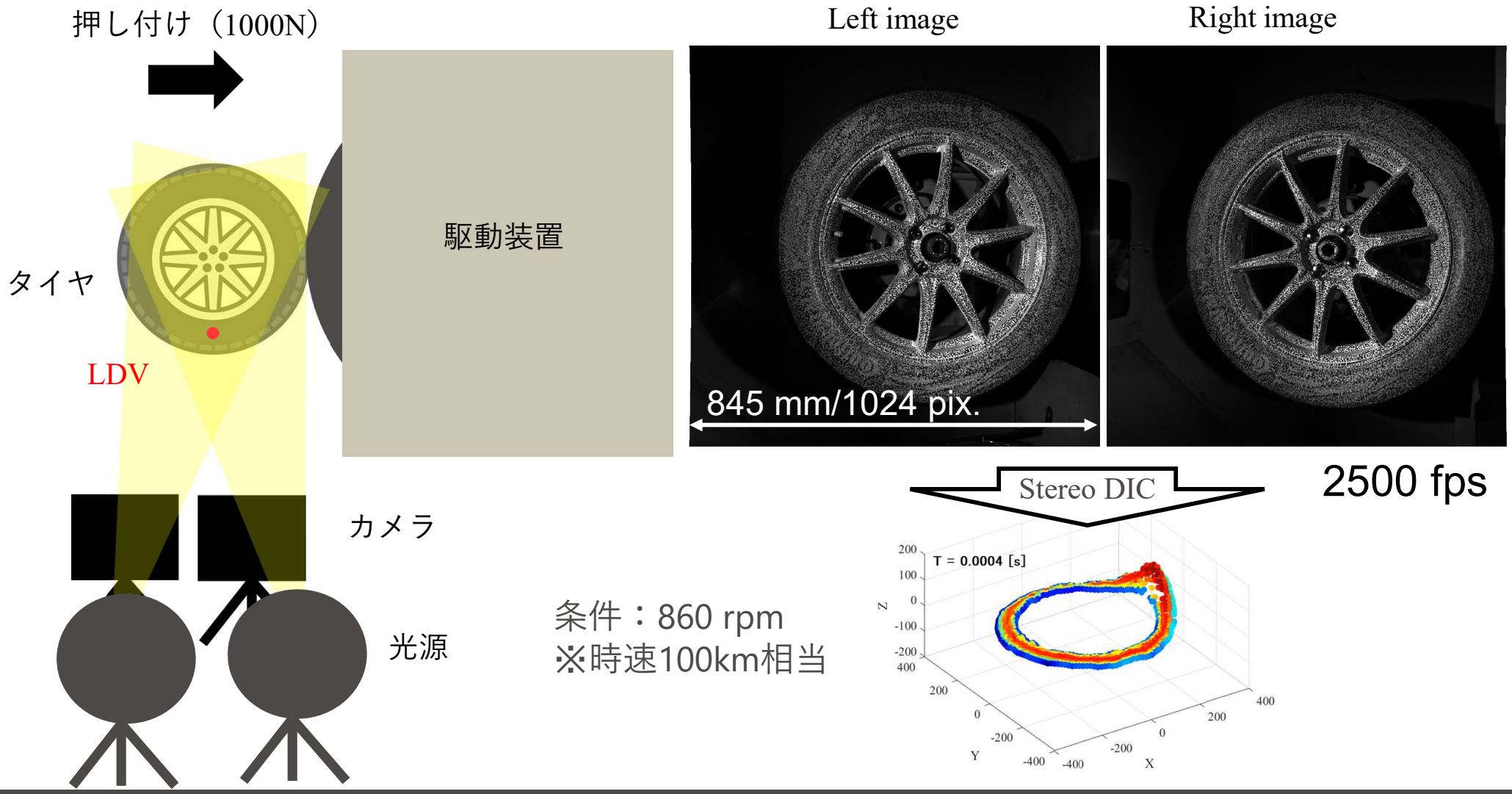
機械工学で良く用いられるデータ解析

- 単変量時系列データ解析
  - 平均と標準偏差
  - ヒストグラム
  - リサンプリング
  - 差分と積算(微積分)
  - 移動平均と移動中央値（外れ値フィルタ）
  - 離散フーリエ変換（DFT）と高速フーリエ変換（FFT）
  - 周波数フィルタ
  - ウェーブレット変換
- 多変量時系列データ解析（2変量）
  - 相互相関関数
  - コヒーレンス

© 2020 Yuki Kato. All Rights Reserved

今回用いるデータ

- タイヤの振動をLDVで計測
- 変位データをHSカメラを用いたDICにより計測

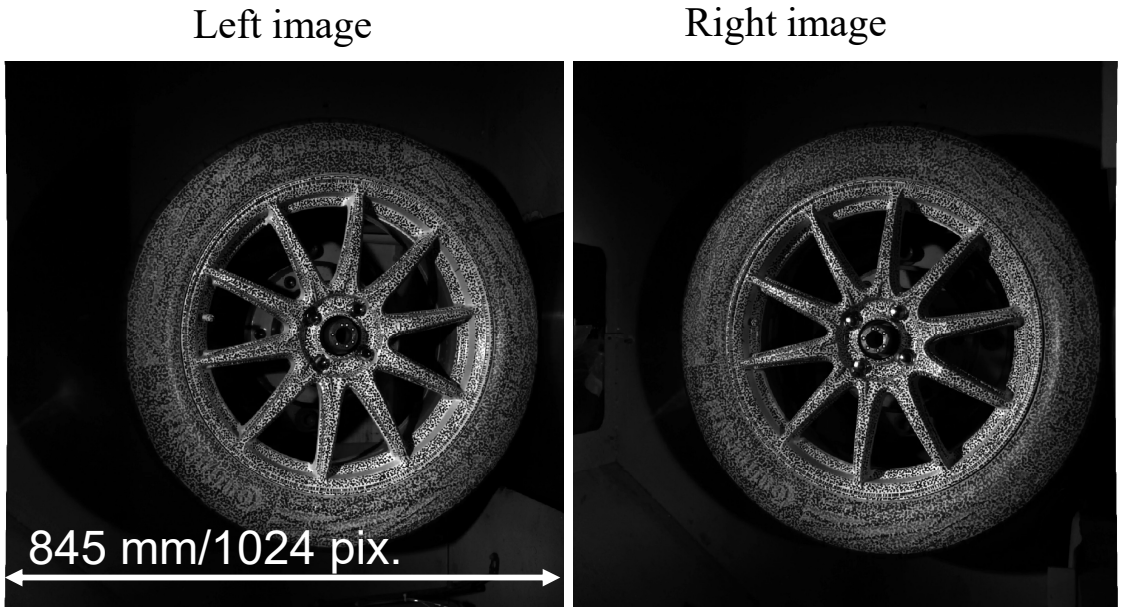


© 2020 Yuki Kato. All Rights Reserved

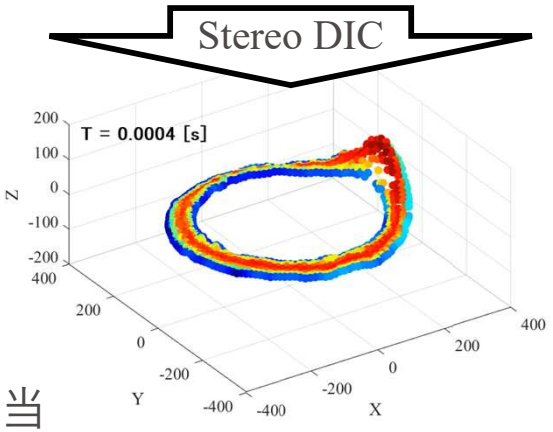
今回用いるデータ

- タイヤの振動をLDVで計測
- 変位データをHSカメラを用いたDICにより計測





2500 fps



条件：860 rpm
※時速100km相当

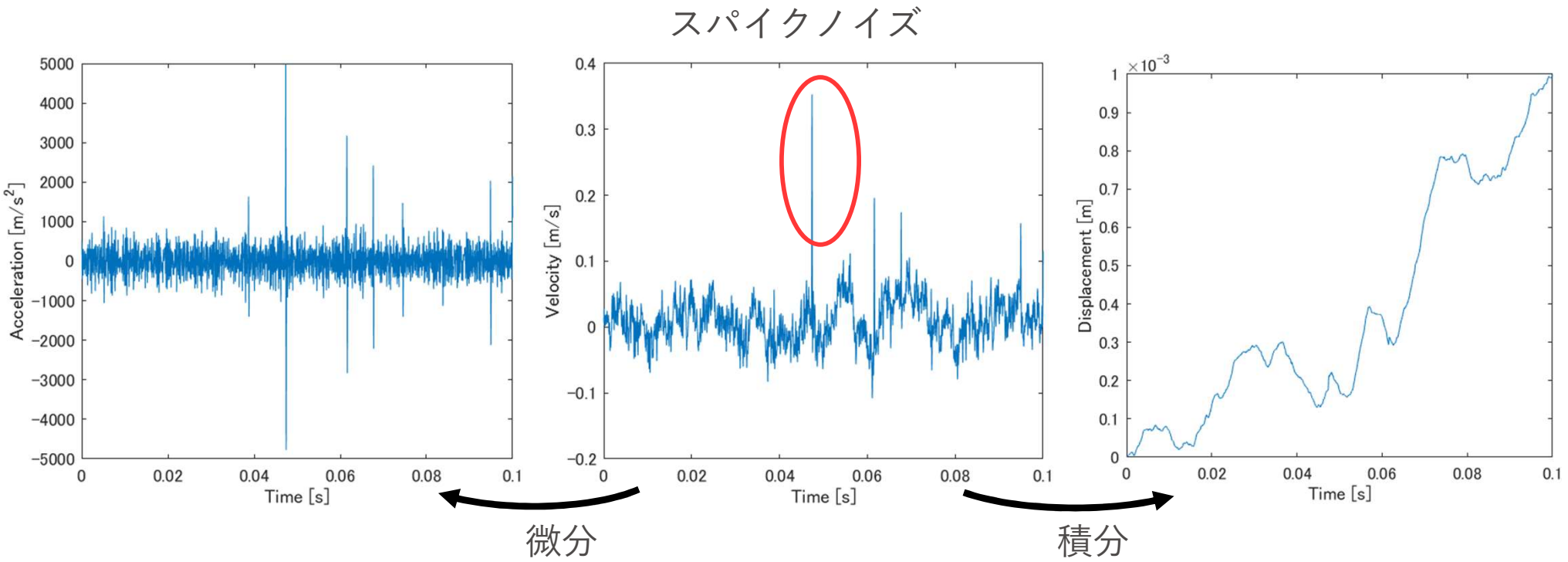
© 2020 Yuki Kato. All Rights Reserved

生データと差分と積算

■ 差分⇒微分　　積算⇒積分

$$a = \frac{dv}{dt}, \quad a_i = \frac{v_{i+1} - v_i}{\Delta t} \qquad x = \int v dt, \quad x_n = \sum_{i=1}^{n} v_i \Delta t$$

■ 計測した速度の差分から加速度，積算量から変位を求めることができる



© 2020 Yuki Kato. All Rights Reserved

移動平均と移動中央値

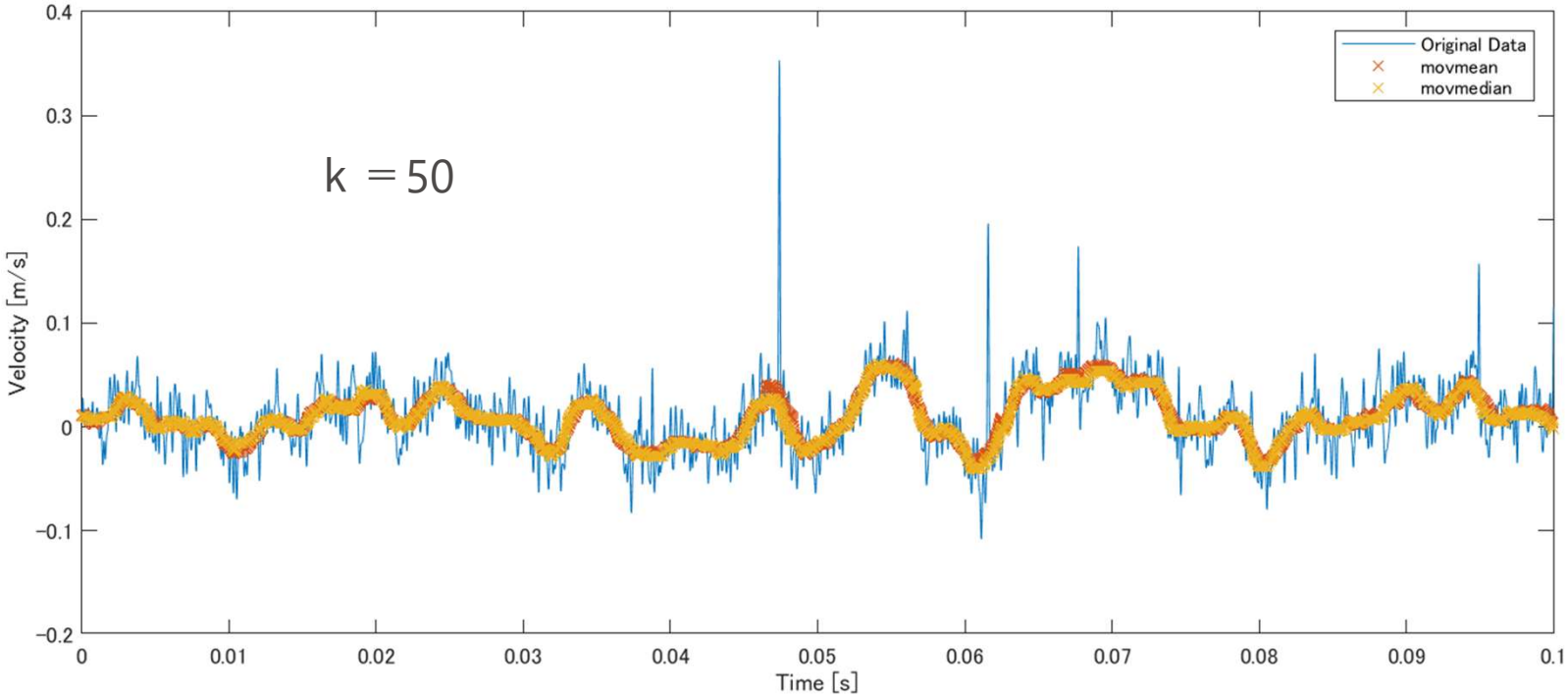
■ 移動平均値

$$z_t = \frac{x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}}{2k+1}$$

移動中央値

$$z_t = median\left(x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}\right)$$

■ 計測したデータの高周波ノイズを減算できるが，変動を過小評価してしまう・・・分散も小さくなる

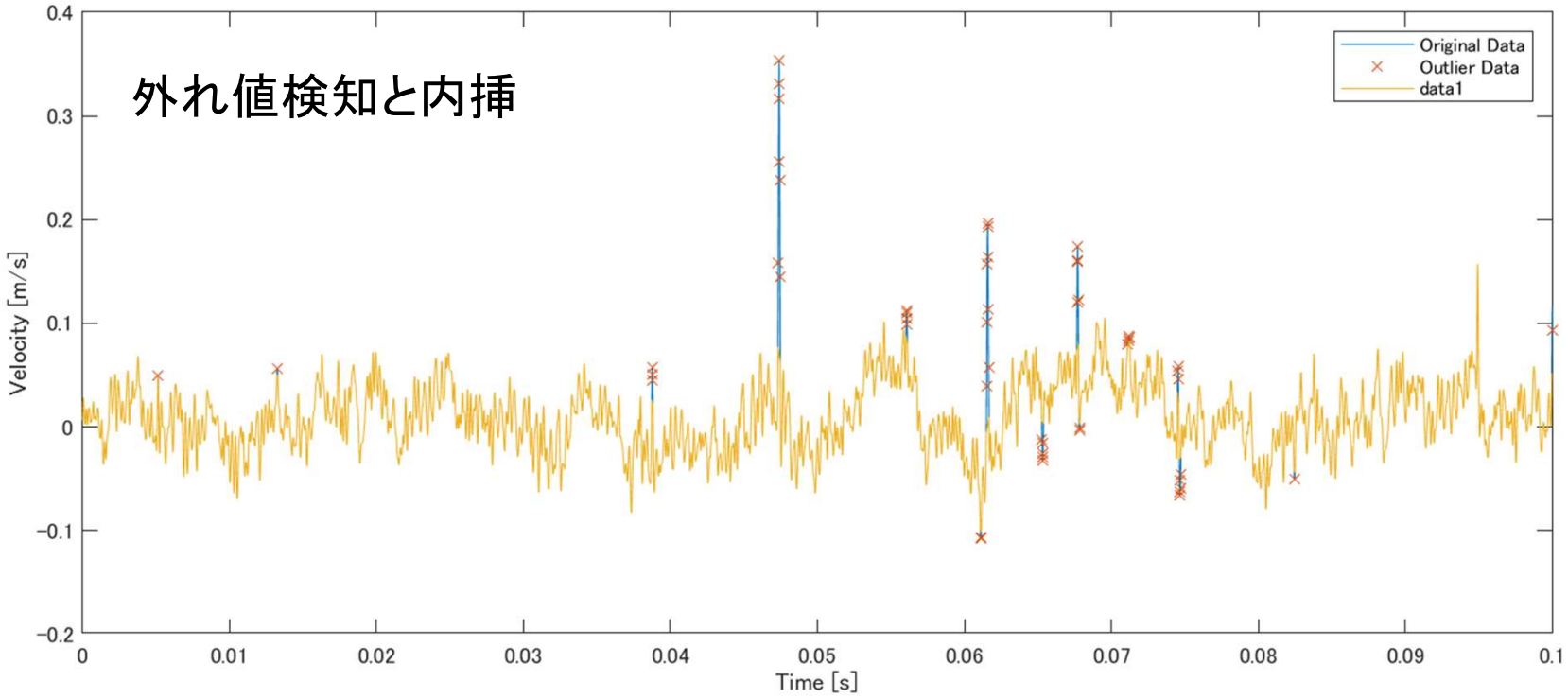


© 2020 Yuki Kato. All Rights Reserved

移動中央値を用いた外れ値検知と補完

■ 移動中央値と標準偏差を計算し，その3倍以上離れているデータを線形補完する

$$z_t = median\left(x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}\right)$$



© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第六回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

✓課題1出題

✓課題1〆

✓(課題2出題)

✓('課題2〆)

統計解析の基礎
1次元，2変数

→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析

→ビッグデータ解析に必須のスキル

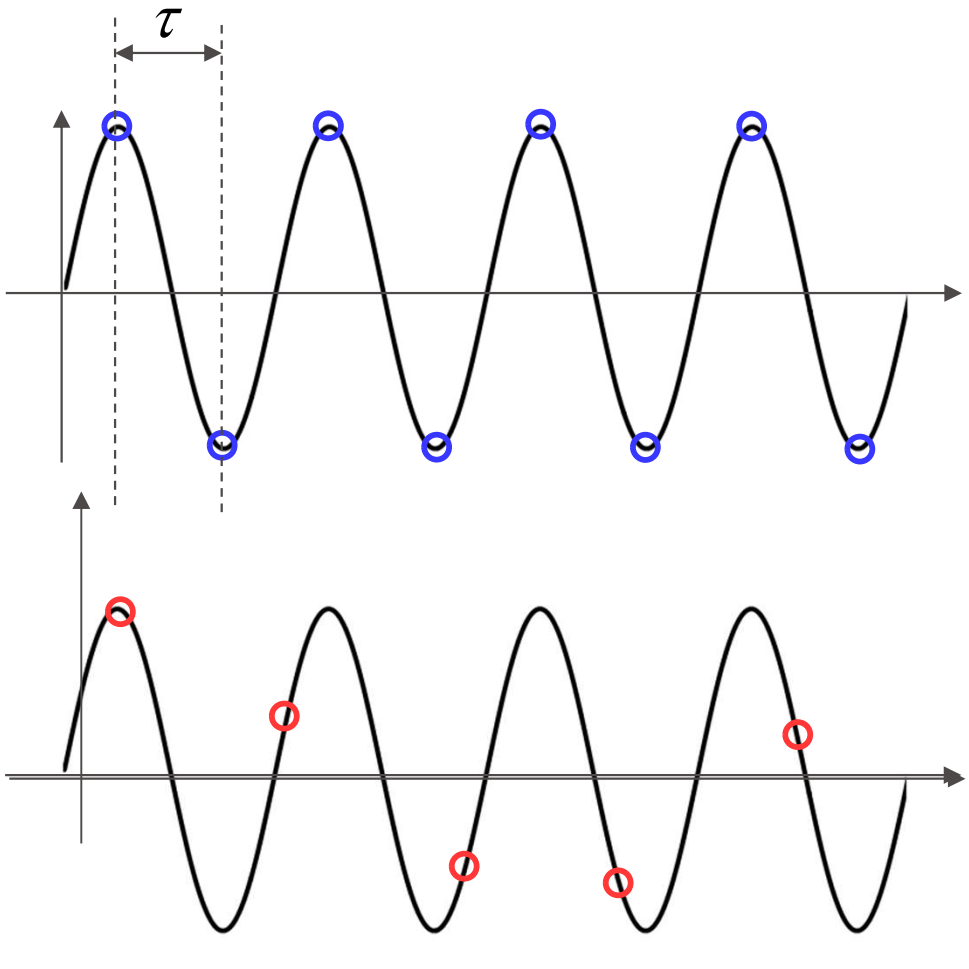
© 2020 Yuki Kato. All Rights Reserved

補：計測周波数とエイリアシング

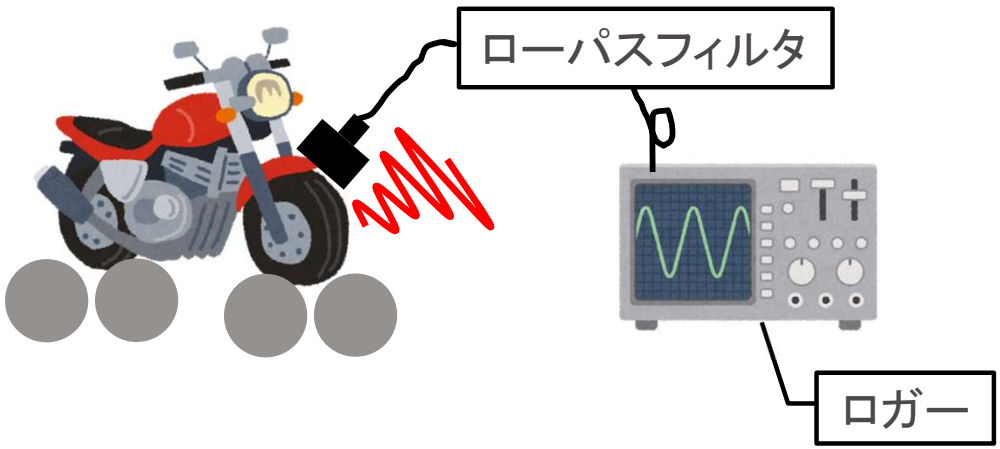
- 計測周波数は対象物の振動の2-10倍程度に設定
- エイリアシングを防ぐために、計測周波数以上の信号をカットするローパスフィルタが必要（計測機に付属）

最大周波数：
（ナイキスト周波数）

$$f_{\max} = \frac{1}{2\tau}$$



エイリアシング



© 2020 Yuki Kato. All Rights Reserved

機械工学で良く用いられるデータ解析

- 単変量時系列データ解析
  - 平均と標準偏差
  - ヒストグラム
  - リサンプリング
  - 差分と積算(微積分)
  - 移動平均と移動中央値（外れ値フィルタ）
  - 離散フーリエ変換（DFT）と高速フーリエ変換（FFT）
  - 周波数フィルタ
- 多変量時系列データ解析（2変量）
  - 相互相関関数
  - コヒーレンス

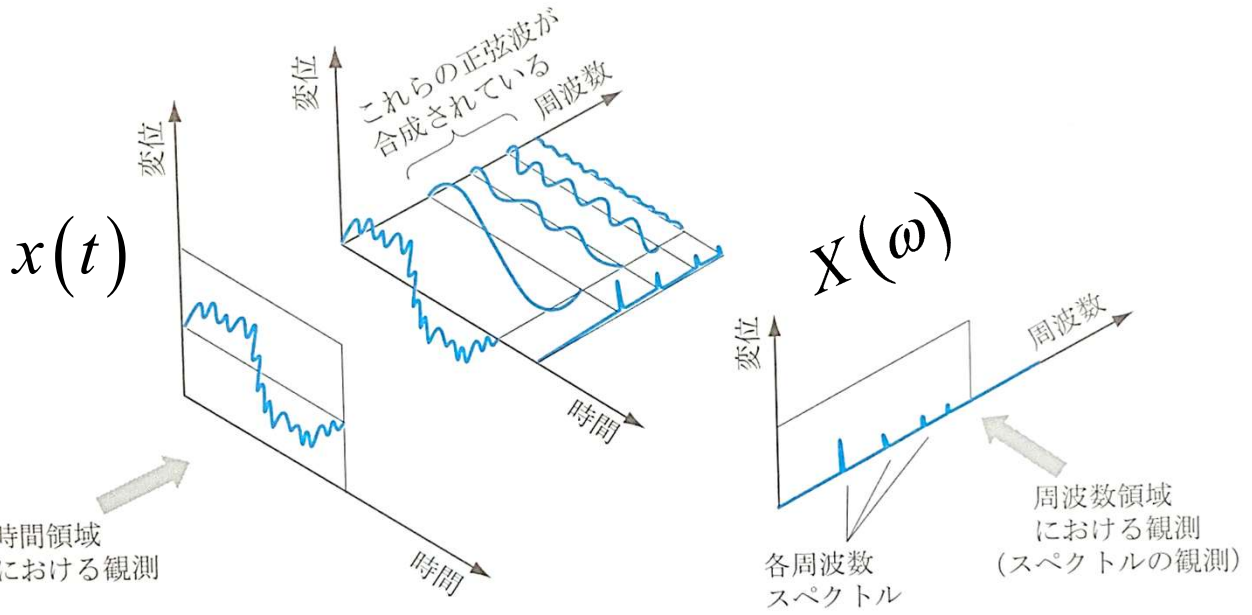
© 2020 Yuki Kato. All Rights Reserved

離散フーリエ変換を用いた周波数分析方法

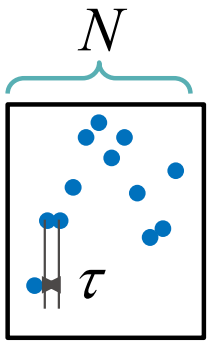
- 振動波形を時間tの関数x(t)とすると、下式で示すフーリエ変換を用いることで時間関数x(t)は周波数関数X(ω)と変換される

$$X(\omega)=\int_{-\infty}^{\infty}x(t)e^{-i\omega t}dt \qquad x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}X(\omega)e^{i\omega t}d\omega$$

- ここで、ωは角振動数である。右式は逆フーリエ変換



離散フーリエ変換



$$X(kf_0)=\sum_{n=0}^{N-1}x(n\tau)e^{-i2\pi kn/N}$$

$$x(n\tau)=\frac{1}{N}\sum_{k=0}^{N-1}X(kf_0)e^{i2\pi kn/N}$$

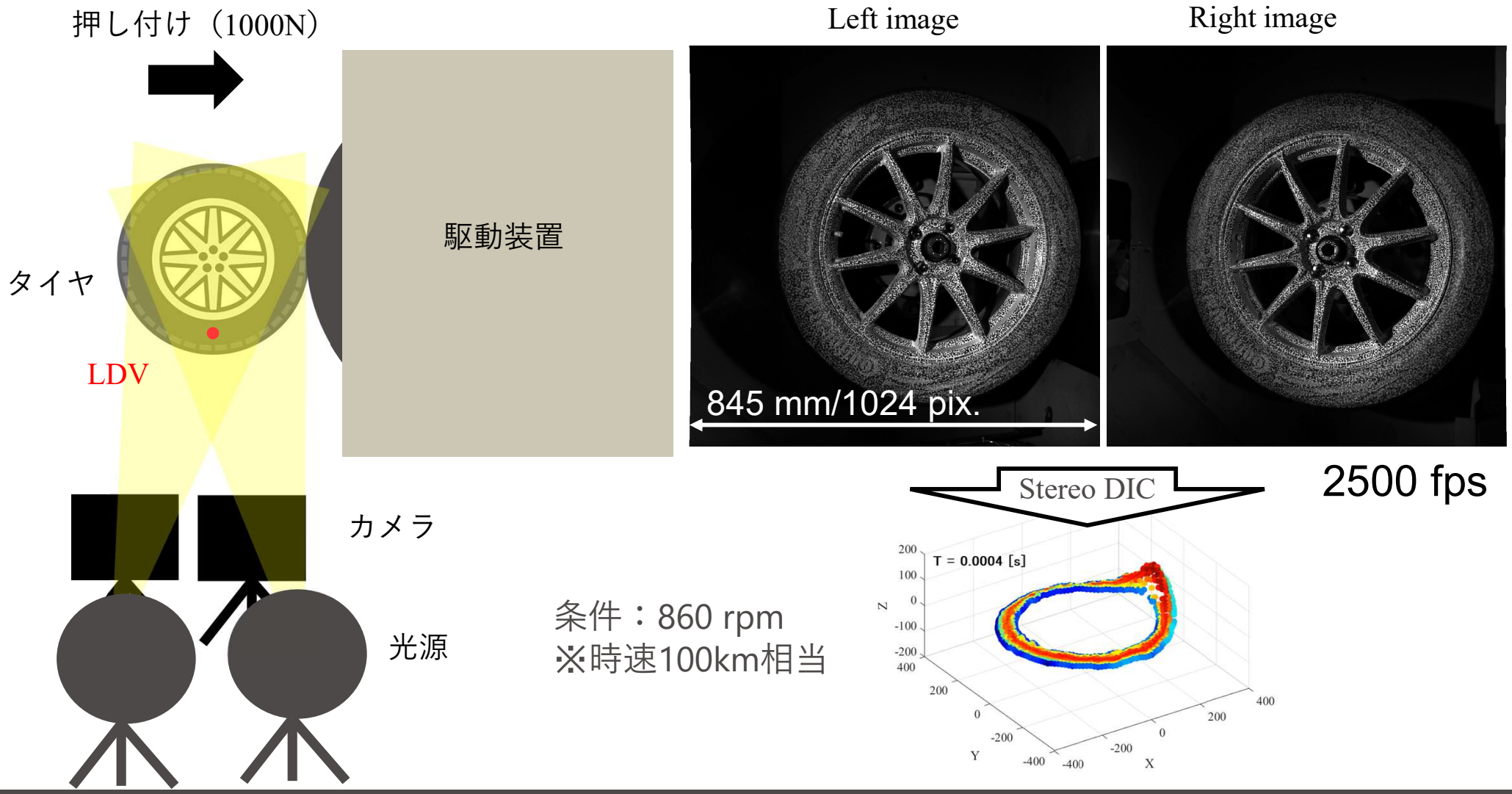
$$f_0=1/n\tau$$

FFTを用いると速いが，MatlabはDFTも高速化してある

© 2020 Yuki Kato. All Rights Reserved

今回用いるデータ

- タイヤの振動をLDVで計測
- 変位データをHSカメラを用いたDICにより計測



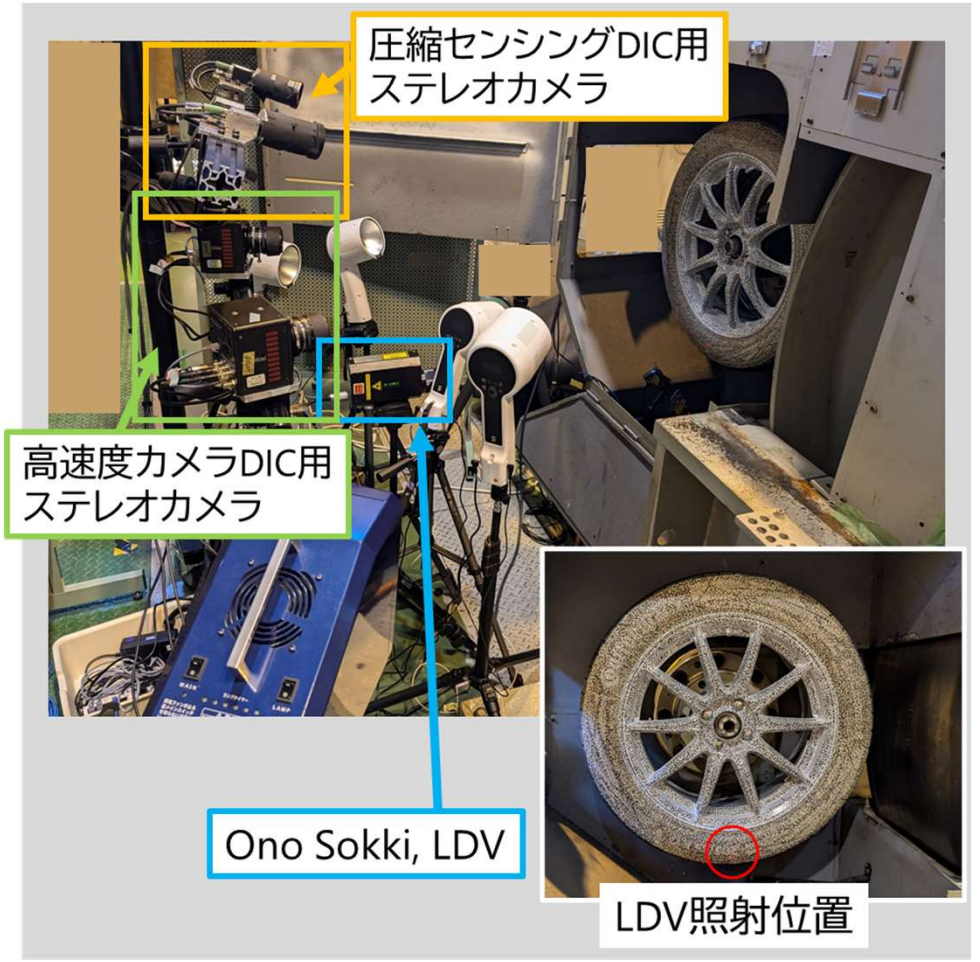
© 2020 Yuki Kato. All Rights Reserved

今回用いるデータ

- タイヤの振動をLDVで計測
- 変位データをHSカメラを用いたDICにより計測



条件：860 rpm
※時速100km相当



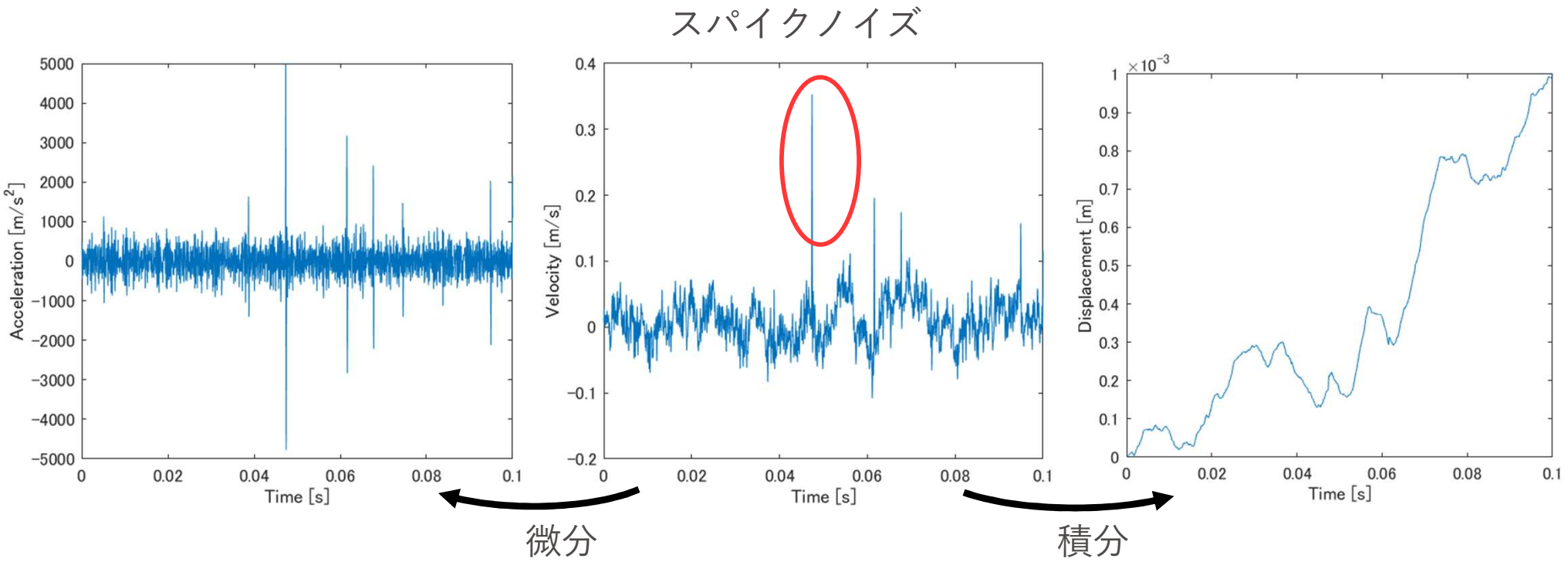
© 2020 Yuki Kato. All Rights Reserved

生データと差分と積算

■ 差分⇒微分　　積算⇒積分

$$a = \frac{dv}{dt}, \quad a_i = \frac{v_{i+1} - v_i}{\Delta t} \qquad x = \int v dt, \quad x_n = \sum_{i=1}^{n} v_i \Delta t$$

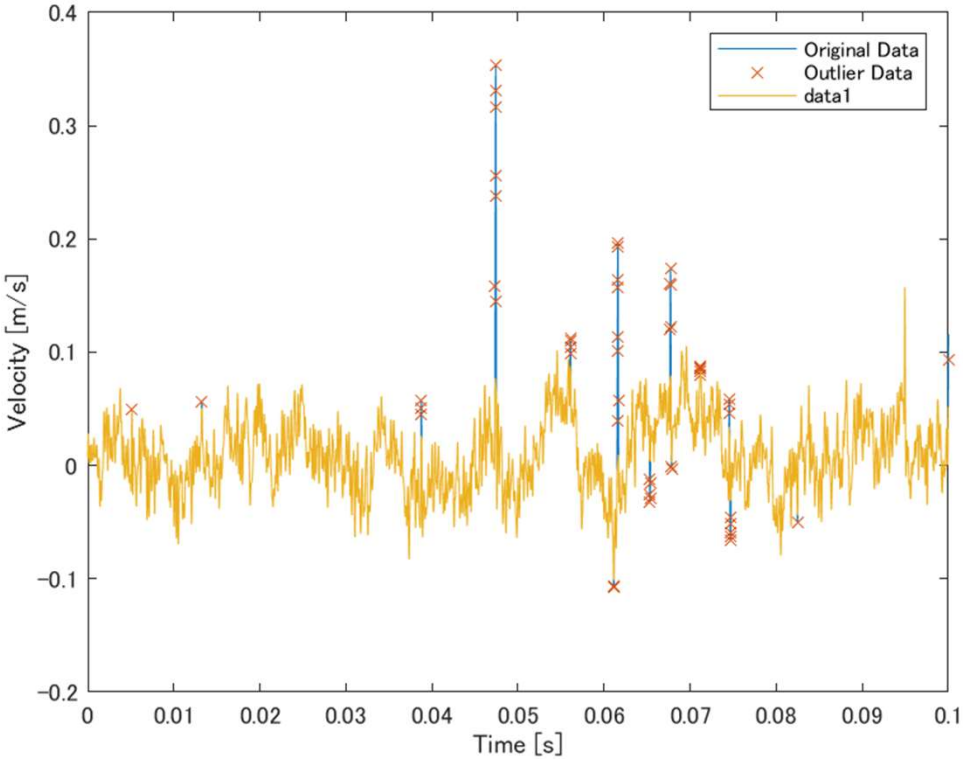
■ 計測した速度の差分から加速度，積算量から変位を求めることができる

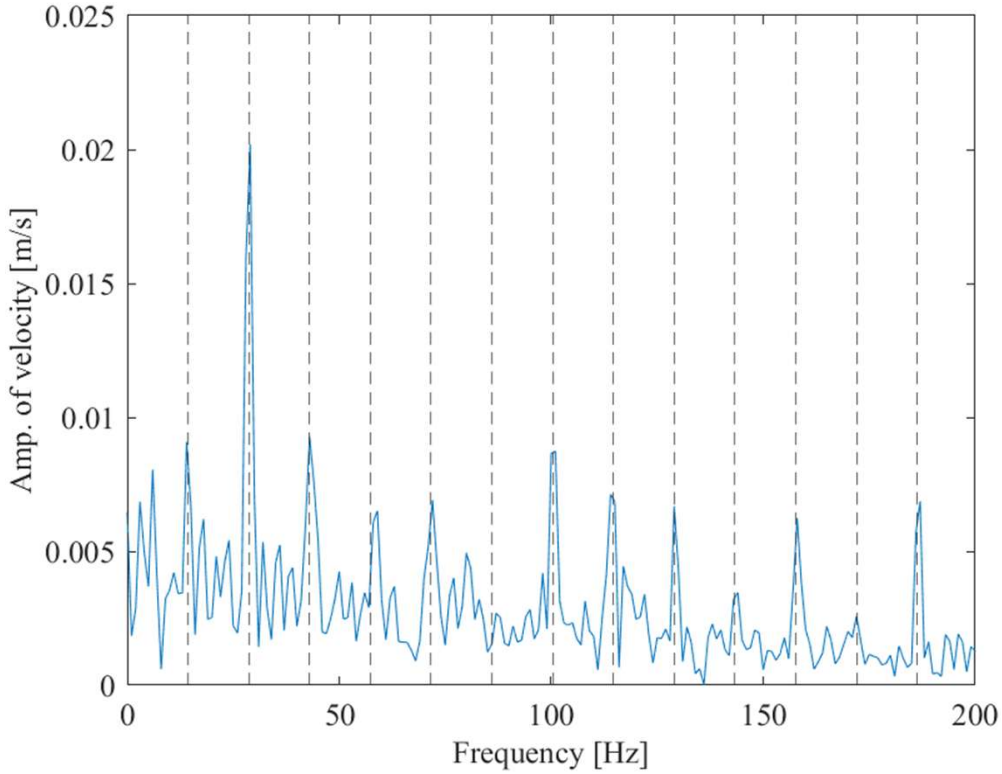


© 2020 Yuki Kato. All Rights Reserved

速度データのDFT

- 回転周波数の次数成分でピークが出現
- 縦軸は対数表示にする場合も多い

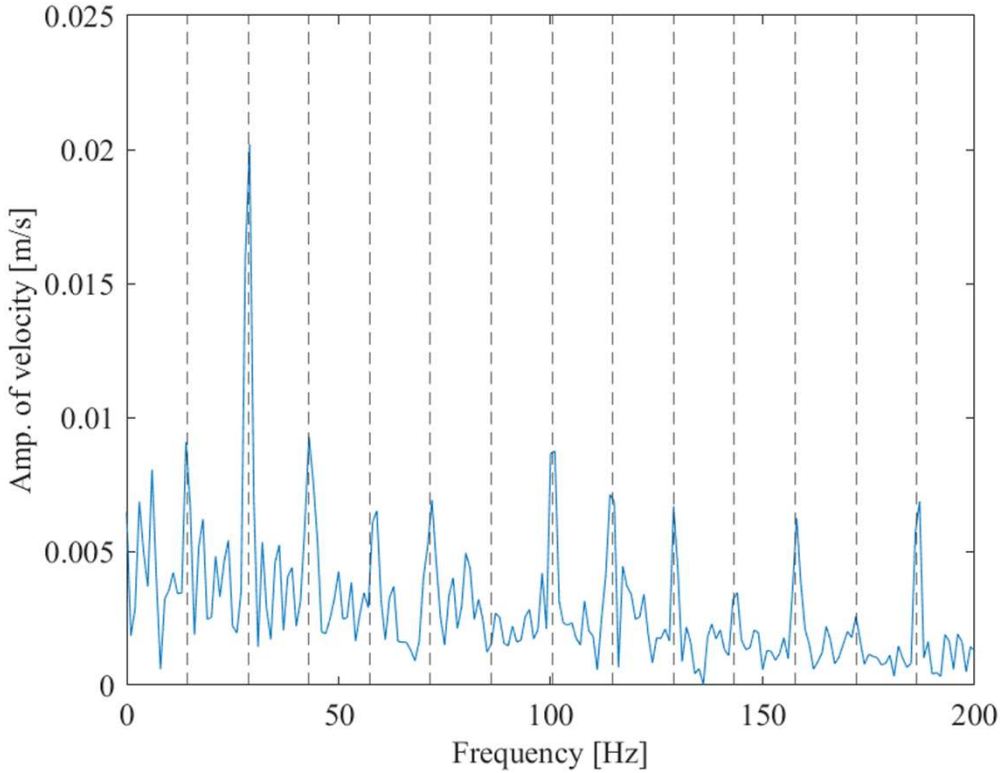


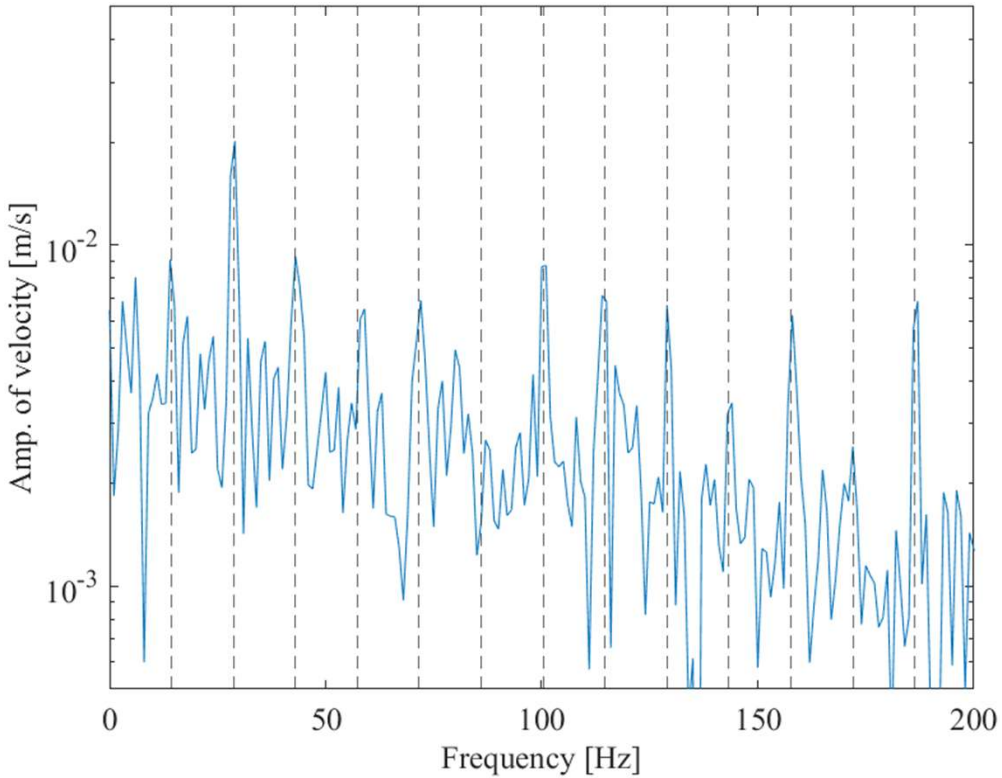


© 2020 Yuki Kato. All Rights Reserved

速度データのDFT

- 回転周波数の次数成分でピークが出現
- 縦軸は対数表示にする場合も多い





© 2020 Yuki Kato. All Rights Reserved

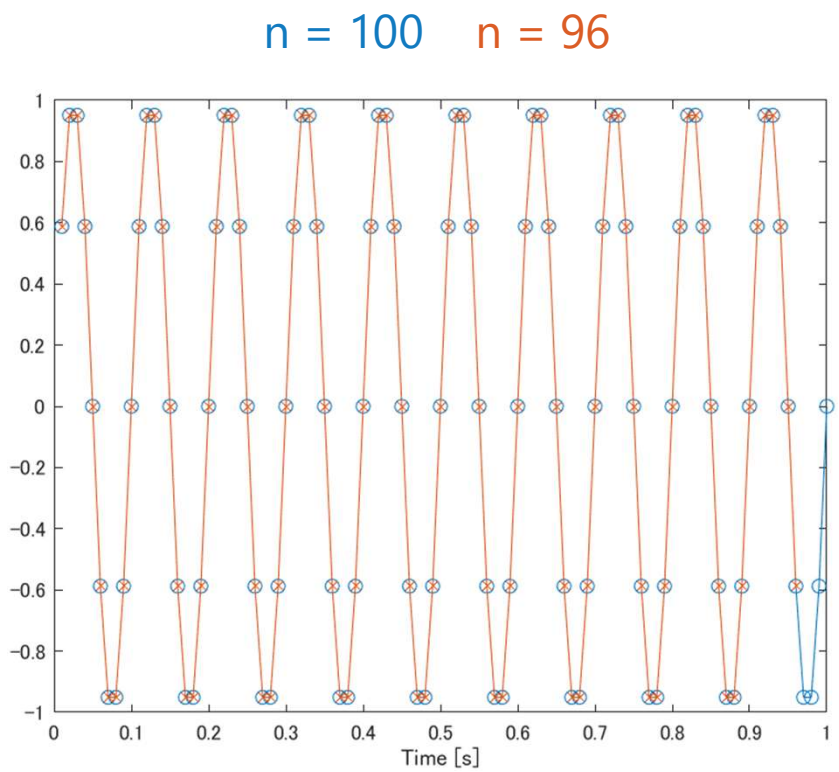
窓関数とスペクトル漏れ1

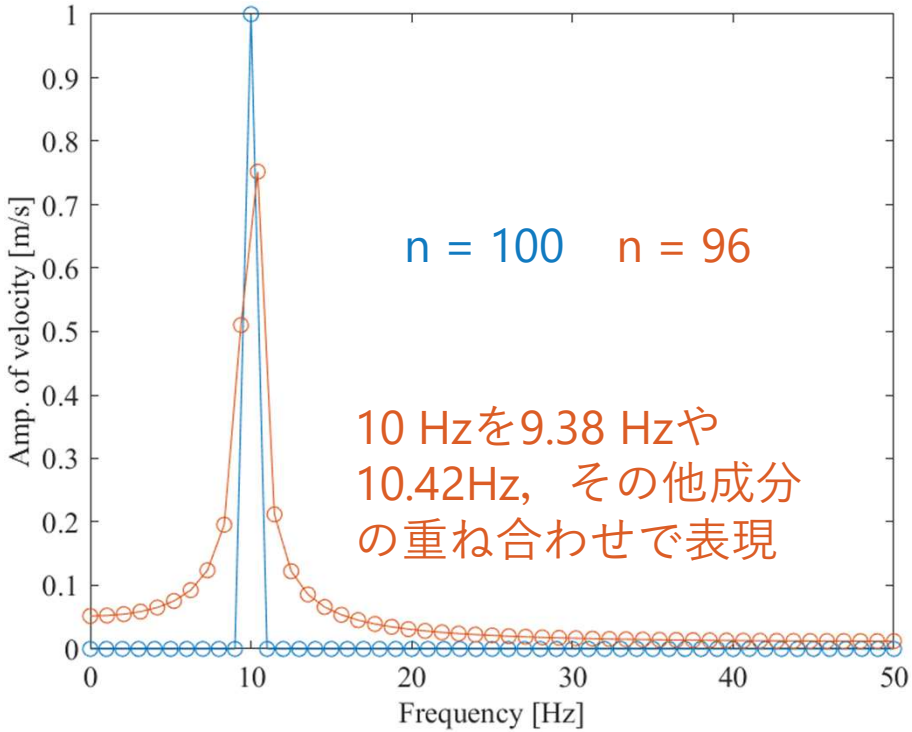
- DFTは、入力信号が時間的に無限に繰り返されていると仮定して計算する
- しかし、実際には有限な時間区間（フレーム）の信号を切り出して解析するため、次の問題が生じる
- 不連続性:
  - 切り出した信号の始点と終点の値が異なると、信号全体として大きな不連続点が生じる
- 高調波成分の発生:
  - この不連続な信号は、本来の周波数成分に加え、多くの不必要な高周波成分（サイドローブ）をスペクトルに発生させる。これがスペクトル漏れ
- 問題点:
  - スペクトル漏れにより、本来の信号のピーク（主ローブ）が広がり、隣接する周波数成分が隠されたり、誤って検出されたりする

© 2020 Yuki Kato. All Rights Reserved

窓関数とスペクトル漏れ2

- スペクトル漏れの例
- N=96では始まりと終わりのデータが不連続



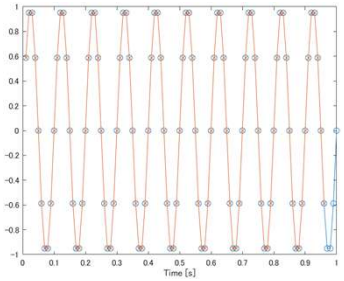


© 2020 Yuki Kato. All Rights Reserved

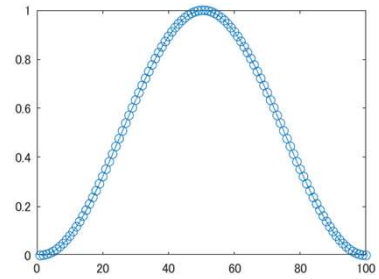
窓関数とスペクトル漏れ3

- 窓関数をかけて，不連続性をなくす
- 窓関数 w[n] は、切り出す信号 x[n] に時間領域で乗算する

$$x_{窓}[n] = x[n] \cdot w[n]$$

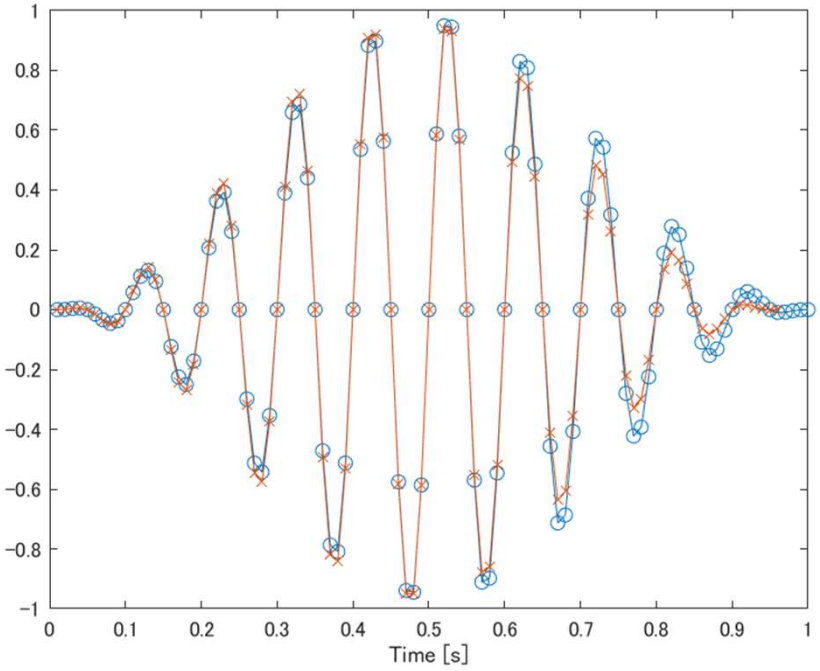


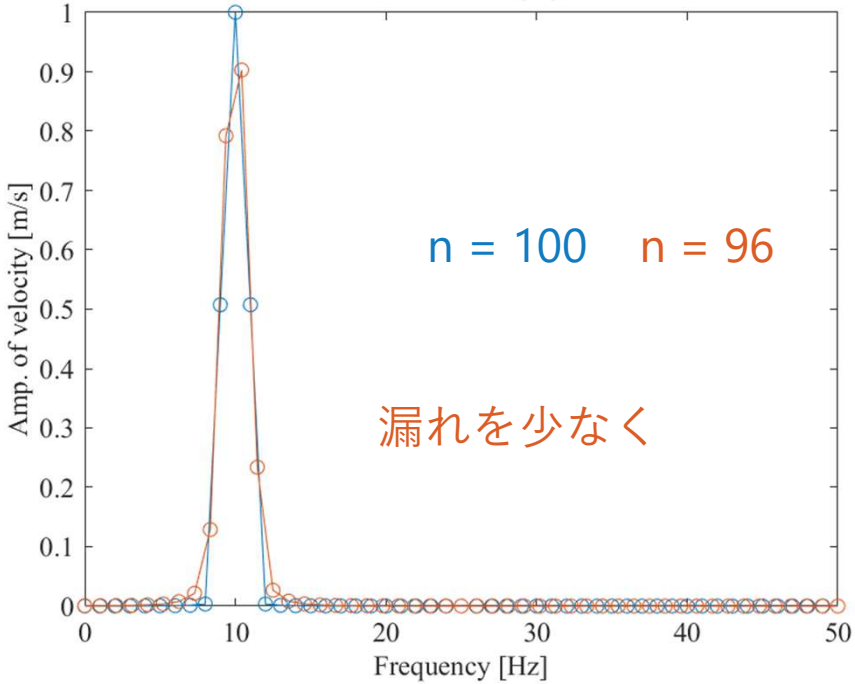
×



$$w(x) = 0.5 - 0.5\cos 2\pi x$$

n = 100　n = 96



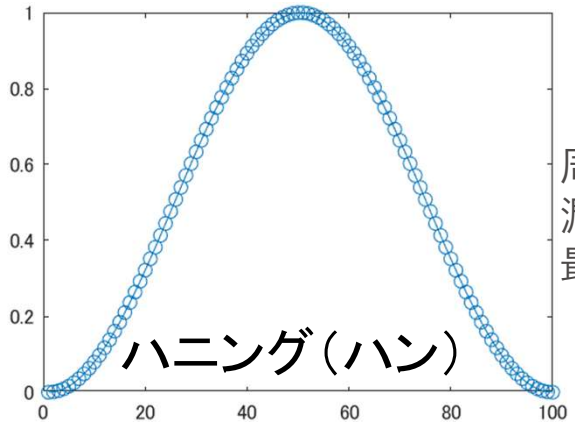


※振幅補正を忘れない事

© 2020 Yuki Kato. All Rights Reserved

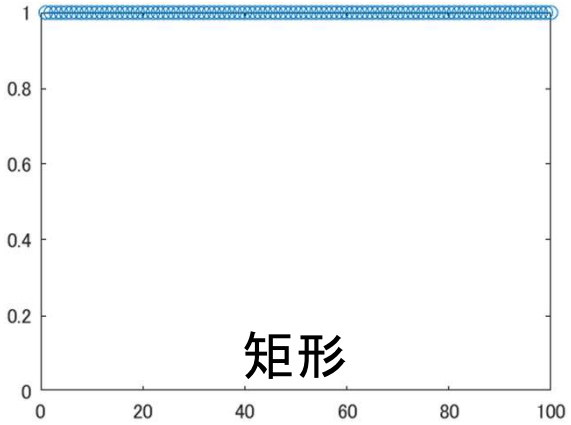
窓関数とスペクトル漏れ4

■ 窓関数の種類様々あるが，よく使われるのはハニング（ハン），矩形，ブラックマン，フラットトップ



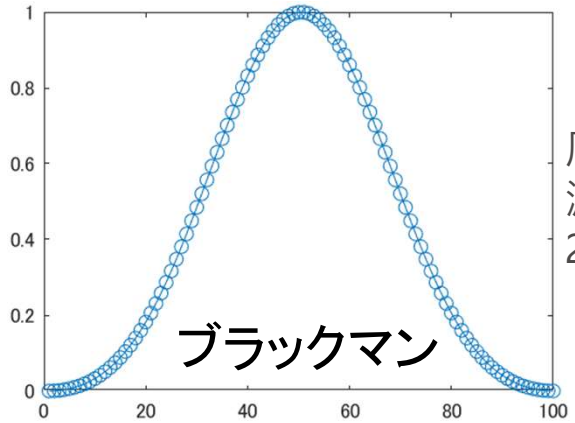
周波数特性が不明な時
漏れが少ない
最もよく使用される

$$w(x) = 0.5 - 0.5\cos 2\pi x$$

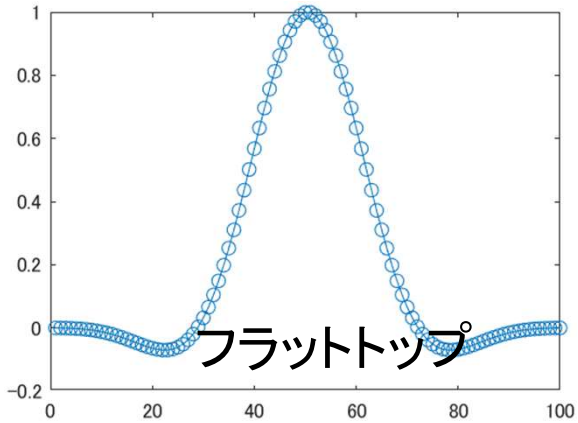


周波数特性が既知で
データ長を制御可能

$$w(x) = 1$$



周波数特性が不明な時
漏れが少ない
2-3番目に使用される



振幅を定量的に評価
裾が広がる

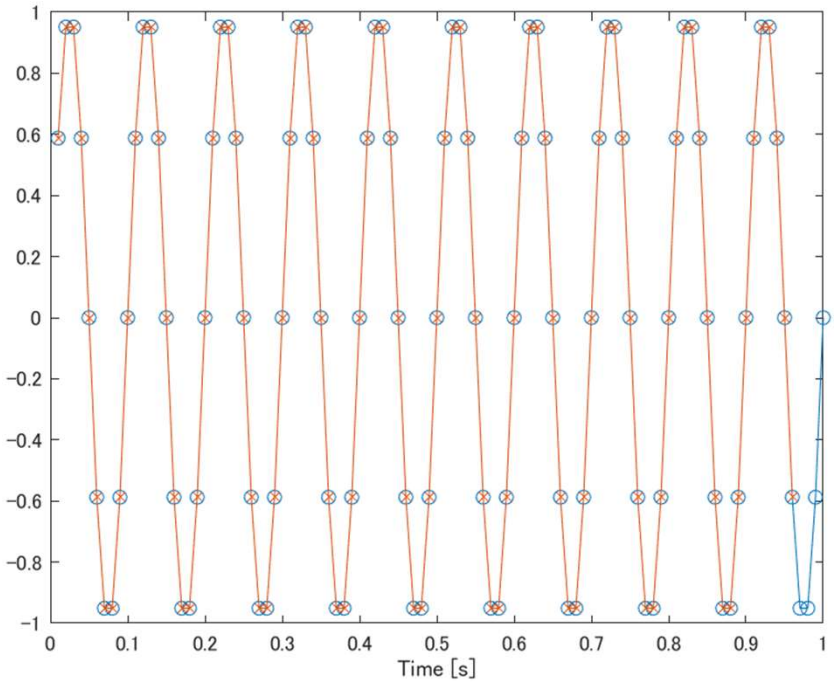
$$w(x) = 0.42 - 0.5\cos 2\pi x + 0.08\cos 4\pi x$$

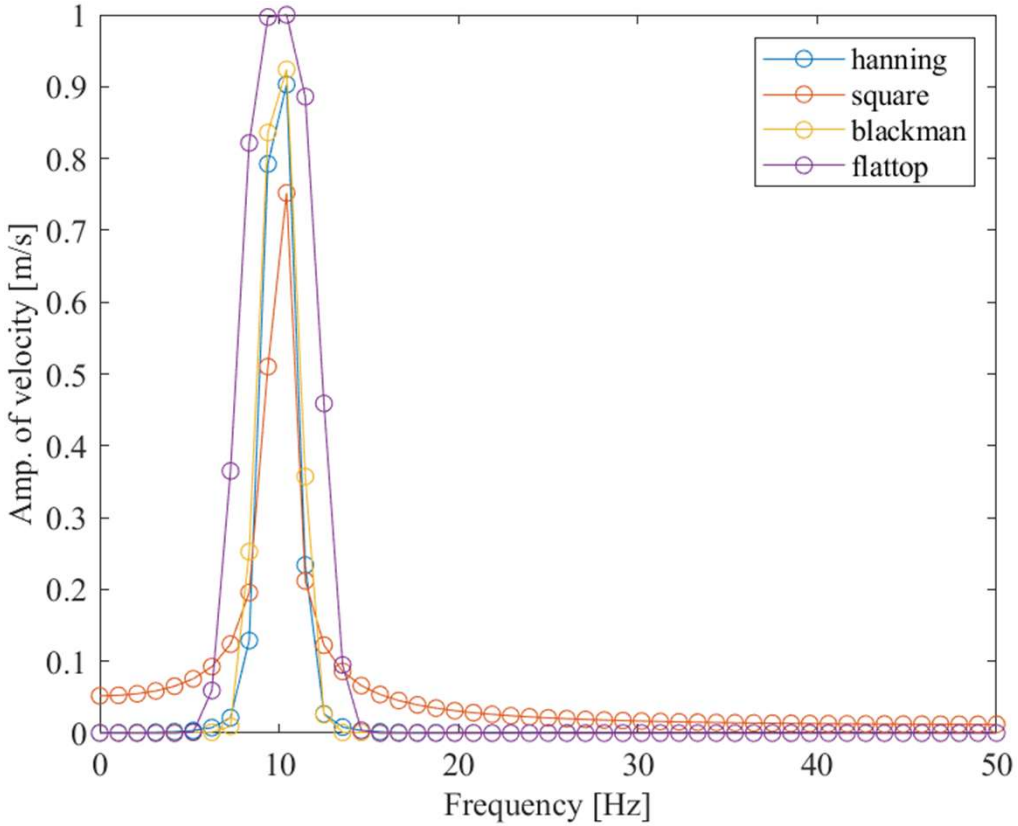
$$w(x) = 1 - 1.93\cos 2\pi x + 1.29\cos 4\pi x - 0.388\cos 6\pi x + 0.032\cos 8\pi x$$

© 2020 Yuki Kato. All Rights Reserved

問題

- 下記模擬波形について，ブラックマン，フラットトップ窓を用いて解析を実施すること





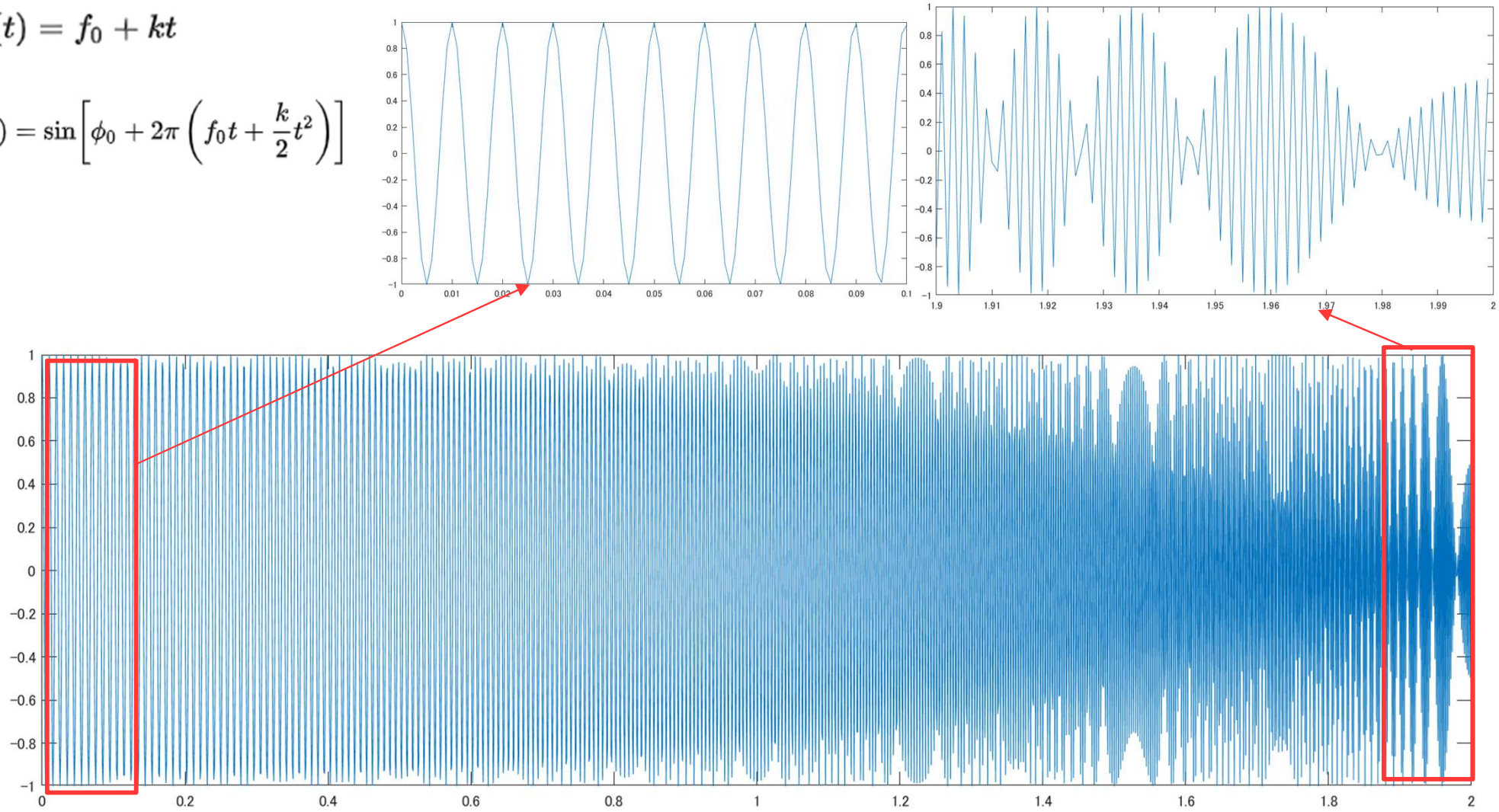
© 2020 Yuki Kato. All Rights Reserved

STFT（Short Time Fourier Transform）

- 短時間周波数分析を実施すると，周波数変化を可視化したり，スペクトルの平均化が出来る

$$f(t) = f_0 + kt$$

$$x(t) = \sin\left[\phi_0 + 2\pi\left(f_0 t + \frac{k}{2}t^2\right)\right]$$

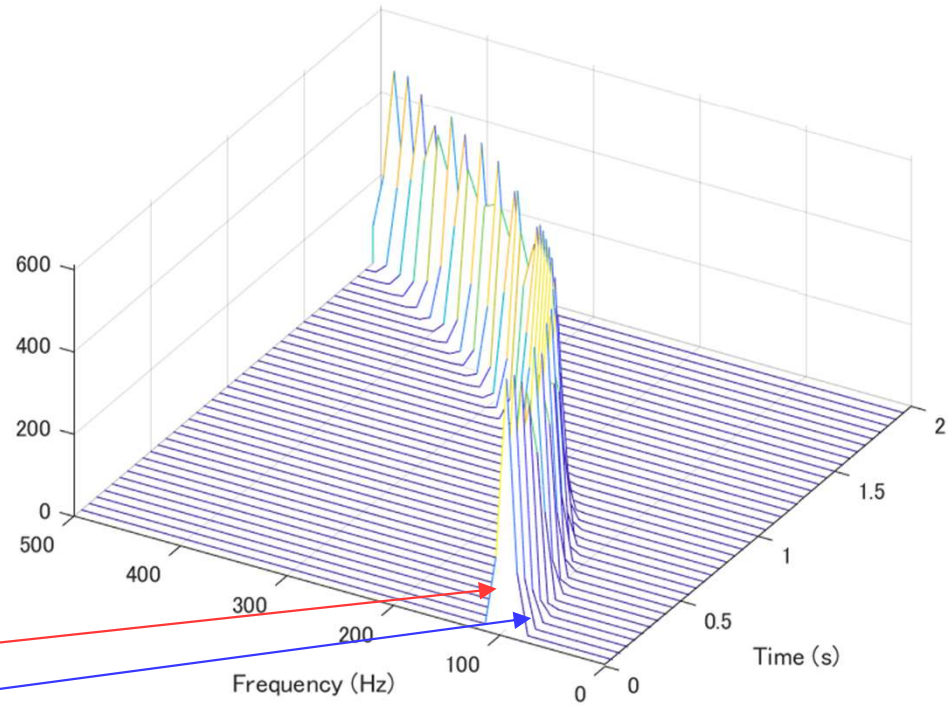


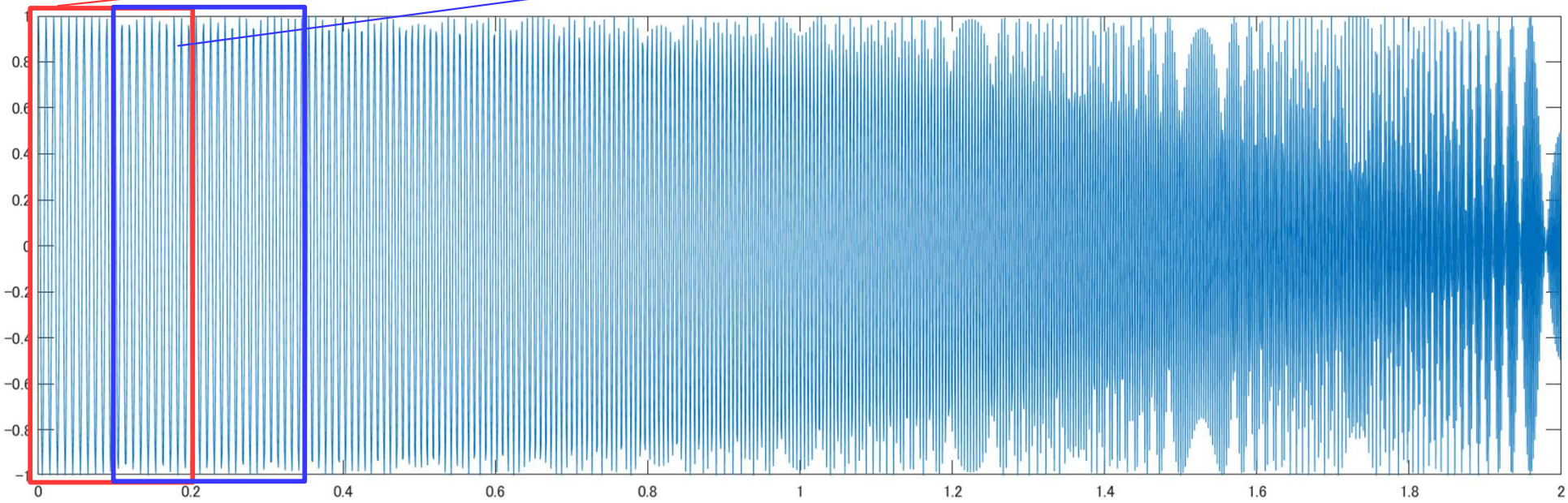
© 2020 Yuki Kato. All Rights Reserved

STFT（Short Time Fourier Transform）

$$f(t) = f_0 + kt$$

- 窓を動かしつつ，FFTをかけていく
- 周波数の変化をプロットるすることができる



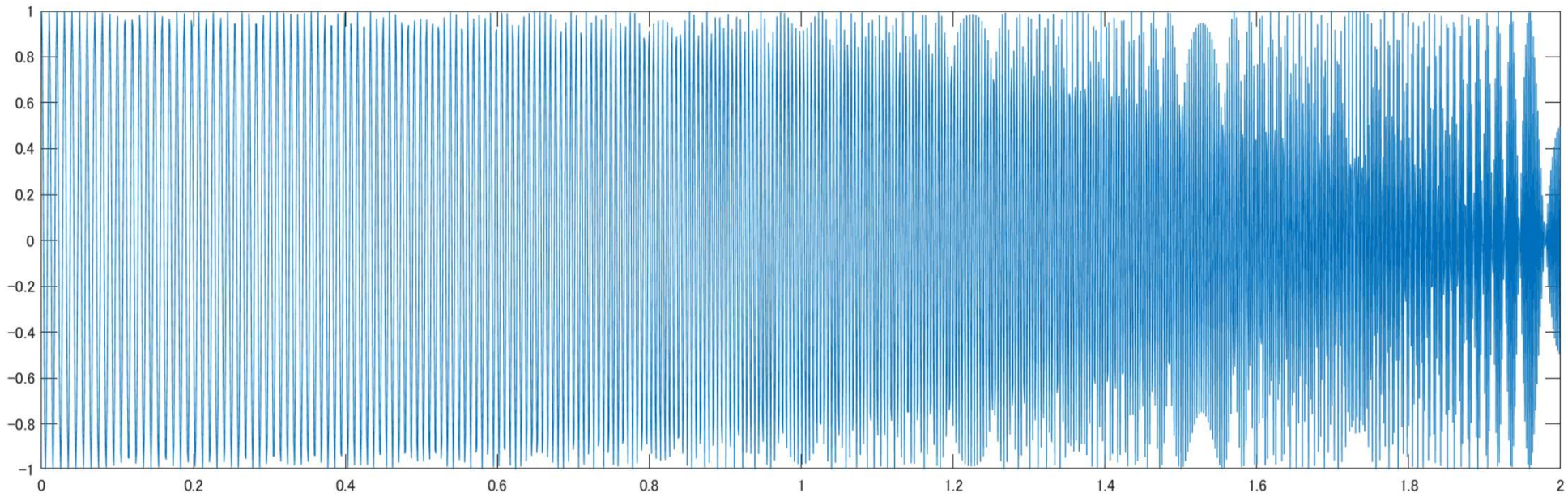


© 2020 Yuki Kato. All Rights Reserved

STFT（Short Time Fourier Transform）

$$f(t) = f_0 + kt$$

- 窓関数をフラットトップにして，オーバーラップを0でプロットしてみよう

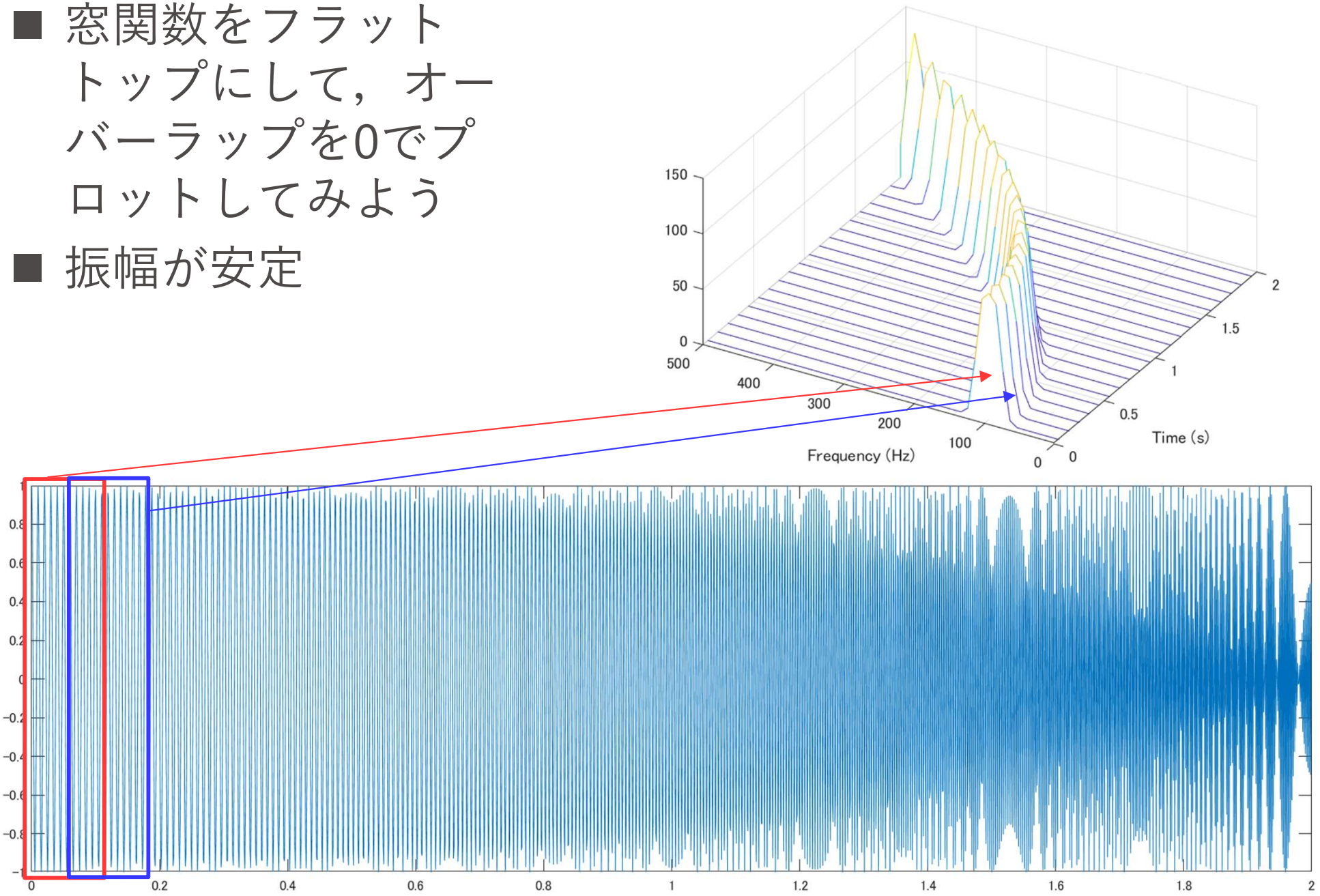


© 2020 Yuki Kato. All Rights Reserved

STFT（Short Time Fourier Transform）

$$f(t) = f_0 + kt$$

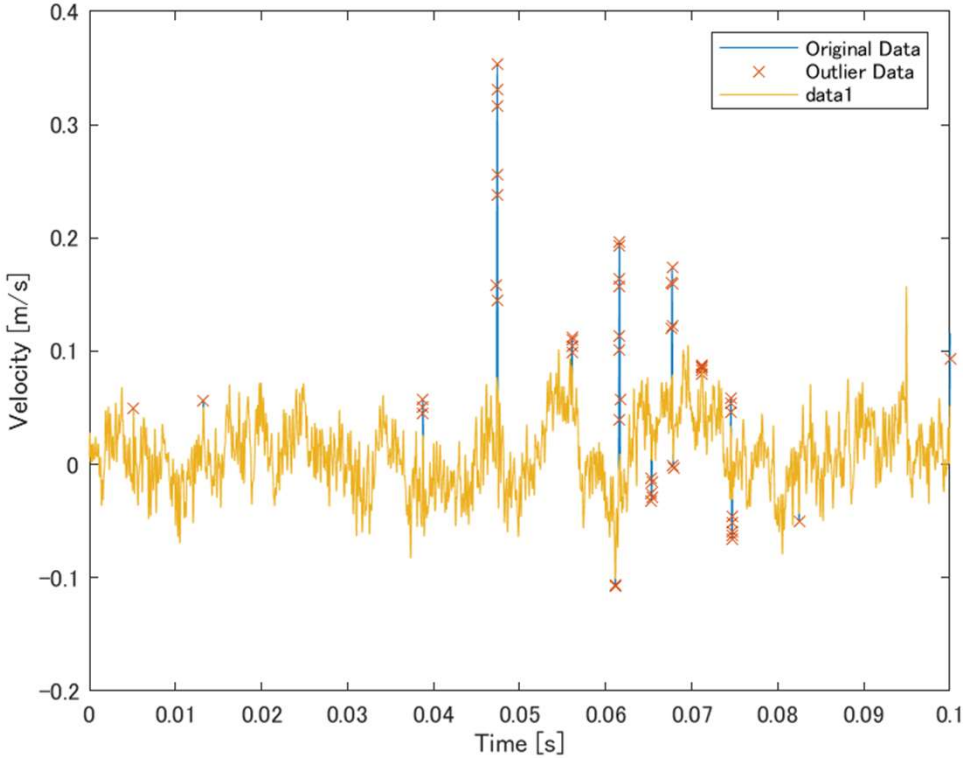
- 窓関数をフラットトップにして，オーバーラップを0でプロットしてみよう
- 振幅が安定

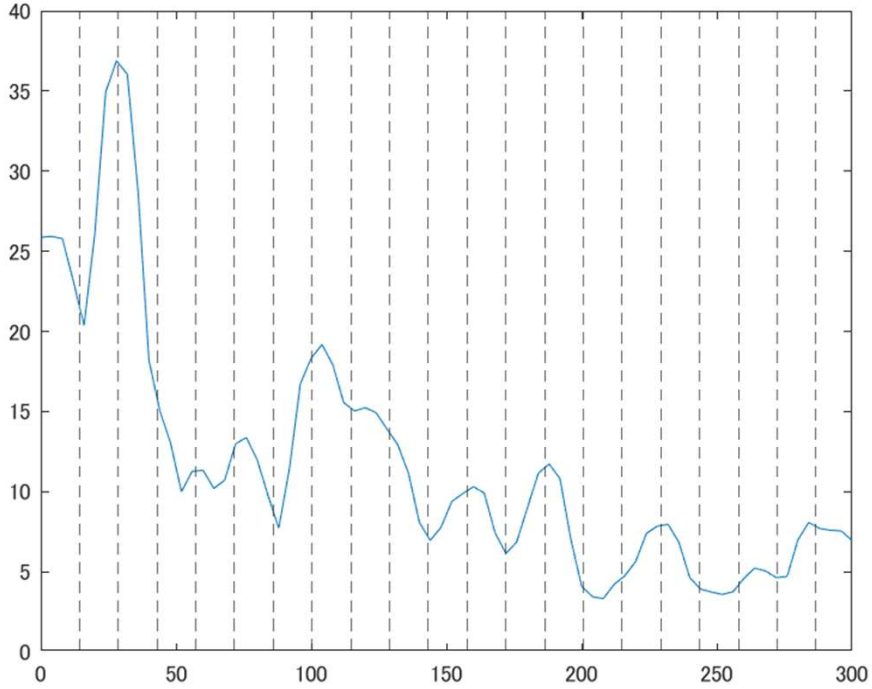


© 2020 Yuki Kato. All Rights Reserved

問題

- LDVの計測波形についてSTFTの結果をプロットせよ
- また，窓長を0.25 s, オーバーラップ: 0.125 sとしてSTFTを実施し，平均スペクトルを計算せよ



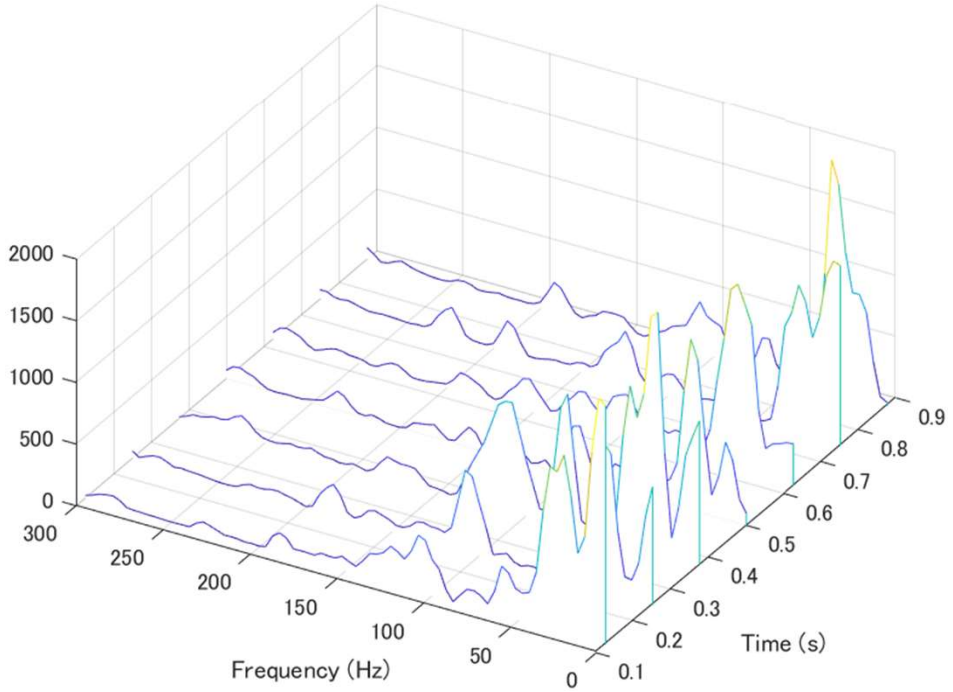


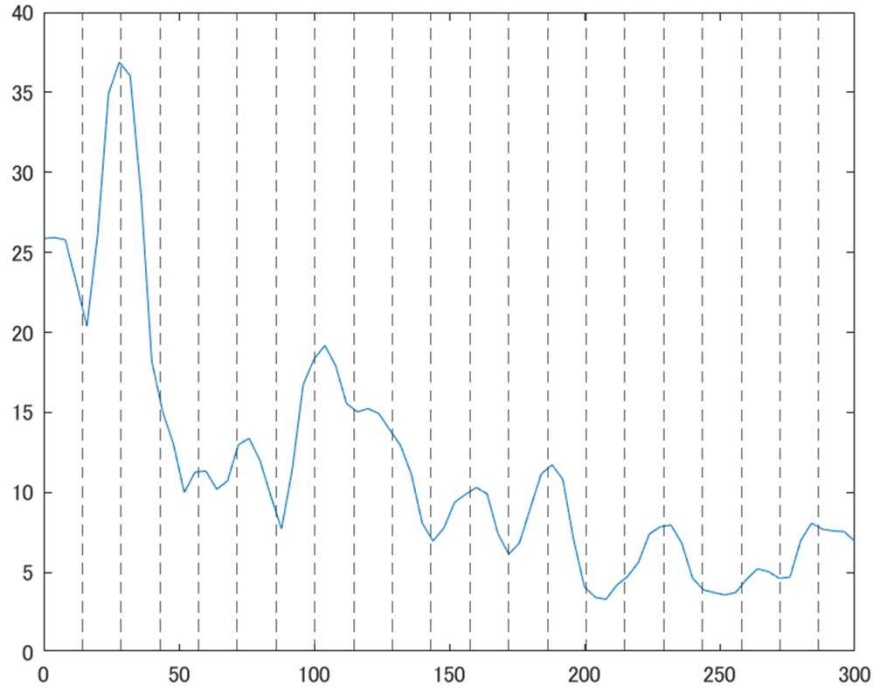
© 2020 Yuki Kato. All Rights Reserved

問題

- 不安定？
- 回転二次の成分は安定している
- 平均特性の評価には，より長いサンプリング時間が必要

それぞれの周波数における標準偏差を評価するのも有用





© 2020 Yuki Kato. All Rights Reserved

機械工学で良く用いられるデータ解析

- 単変量時系列データ解析
  - 平均と標準偏差
  - ヒストグラム
  - リサンプリング
  - 差分と積算(微積分)
  - 移動平均と移動中央値（外れ値フィルタ）
  - 離散フーリエ変換（DFT）と高速フーリエ変換（FFT）
  - 周波数フィルタ
- 多変量時系列データ解析（2変量）
  - 相互相関関数
  - コヒーレンス

© 2020 Yuki Kato. All Rights Reserved

周波数フィルタとは？

- 信号処理で、特定の周波数帯域の成分を通過させたり、阻止したりするためのシステムまたはアルゴリズム
- 電気回路、音響学、画像処理、時系列解析など、信号を扱うあらゆる分野で用いられる基本的なツール
- 基本的な役割
  - フィルタの主な役割は、信号に含まれる不要な成分を取り除き、必要な成分を抽出すること
  - ノイズの除去: 高周波数ノイズ（ヒスノイズなど）や電源周波数ノイズ（ハムノイズなど）を除去し、信号の品質を向上させる
  - 信号の分離: 複合信号から特定の情報（例：ラジオ放送の特定のチャンネル、音声の低音成分）を取り出す
  - 帯域制限: 信号の周波数範囲を特定の目的に合わせて制限する（例：通信帯域の確保）

© 2020 Yuki Kato. All Rights Reserved

周波数フィルタの種類

■ アナログフィルタ

- 抵抗（R）、インダクタ（L）、キャパシタ（C）といった受動部品や、オペアンプなどの能動部品を用いて電気回路として構成される
- ロガーにはエイリアシングを防ぐために，計測周波数の1/2以下の成分のみを通すアナログフィルタがほぼ必ず実装されている

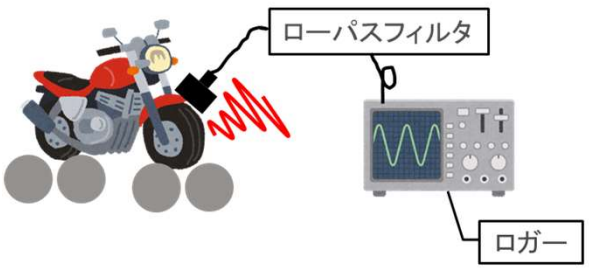
■ デジタルフィルタ

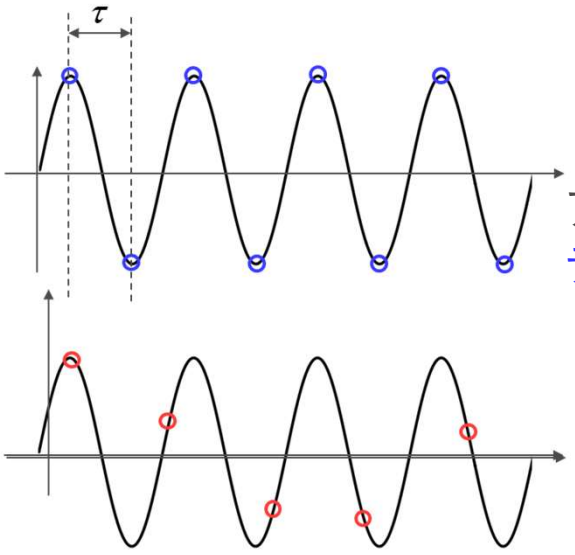
- コンピュータやDSP（デジタル信号処理プロセッサ）上で、サンプリングされた離散信号に対して数学的な計算（アルゴリズム）を用いて実現される

最大周波数:
（ナイキスト周波数）

$$f_{\max} = \frac{1}{2\tau}$$

エイリアシング



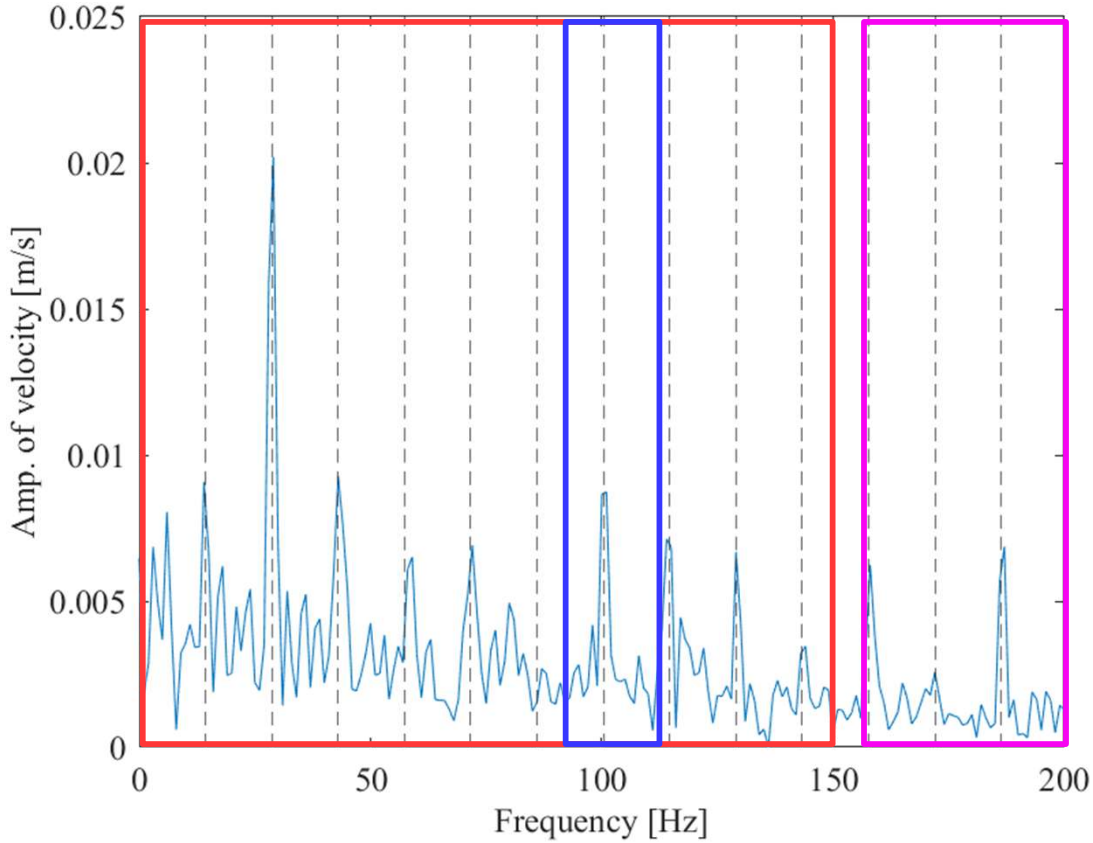


アナログフィルタをかける
ディジタルフィルタをかける

© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（周波数帯）

- ローパスフィルタ **(Low-Pass Filter, LPF)**
- ハイパスフィルタ **(High-Pass Filter, HPF)**
- バンドパスフィルタ **(Band-Pass Filter, BPF)**
- バンドストップフィルタ **(Band-Stop Filter, BSF) /** ノッチフィルタ **(Notch Filter)**⇒青以外



© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（周波数帯）

- ローパスフィルタ (Low-Pass Filter, LPF)
  - 機能: 低い周波数成分を通過させ、高い周波数成分を阻止します。
  - 用途: ノイズ（通常、高周波数成分）の除去、信号の平滑化
- ハイパスフィルタ (High-Pass Filter, HPF)
  - 機能: 高い周波数成分を通過させ、低い周波数成分を阻止します。
  - 用途: 信号の直流成分（バイアス）の除去、低周波数ノイズ（ハムノイズなど）の除去、エッジ強調（画像処理）
- バンドパスフィルタ (Band-Pass Filter, BPF)
  - 機能: 特定の周波数帯域だけを通過させ、それより低い周波数と高い周波数の両方を阻止します。
  - 用途: 特定の通信チャンネルの選択、特定の音階や周波数の抽出。
- バンドストップフィルタ (Band-Stop Filter, BSF) / ノッチフィルタ (Notch Filter)
  - 機能: 特定の周波数帯域だけを阻止し、それ以外の周波数を通過させます。
  - 用途: 狭い周波数範囲の不要なノイズ（例：商用電源の50Hz/60Hzのハムノイズ）をピンポイントで除去する。

© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（アルゴリズム別）

■ デジタルフィルタ

- **FIRフィルタ (Finite Impulse Response):** 過去の入力値のみを用いて計算する。常に安定で、線形位相特性（信号の波形を歪ませない）を容易に実現できる。

$$y[n] = \sum_{k=0}^{M-1} b_k x[n-k]$$

  - M: フィルタの**タップ数**または**次数**（フィルタの長さ）
  - b_k: フィルタ係数（重み）

- **IIRフィルタ (Infinite Impulse Response):** 過去の入力値と過去の出力値の両方を用いて計算する。少ない計算量で急峻な特性を得やすいが、安定性の確保や線形位相の実現が複雑になる。

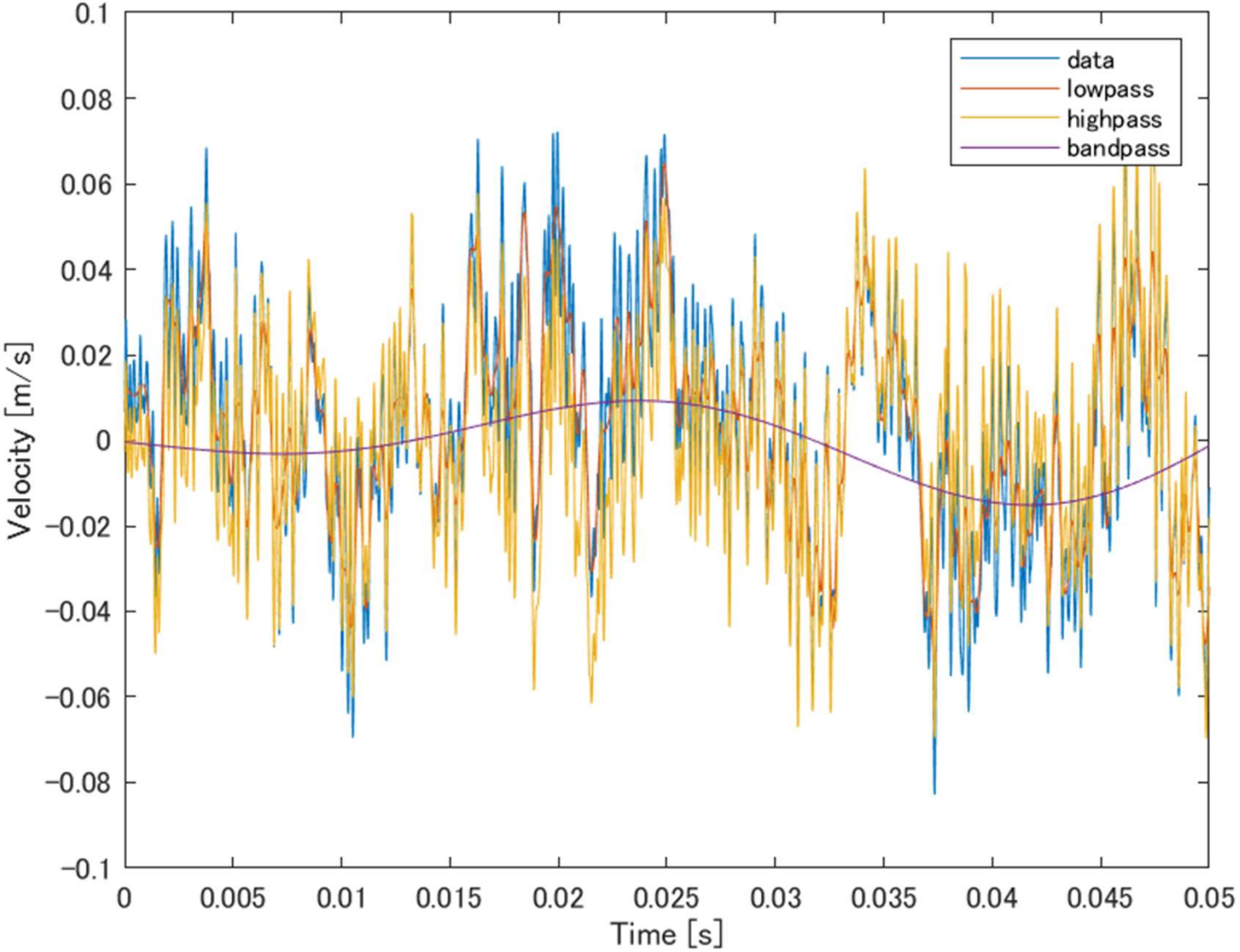
$$y[n] = \sum_{k=0}^{M} b_k x[n-k] - \sum_{j=1}^{N} a_j y[n-j]$$

  - b_k: 入力側のフィルタ係数
  - a_j: 出力側のフィルタ係数（フィードバック係数）

© 2020 Yuki Kato. All Rights Reserved

Matlabでの実装

■ それぞれのフィルタ結果は以下の通り



© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第七回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．
補講があれば，また連絡します

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

✓ 課題①，②出題

✓ 課題①②

統計解析の基礎
1次元，2変数

→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析

→ビッグデータ解析に必須のスキル

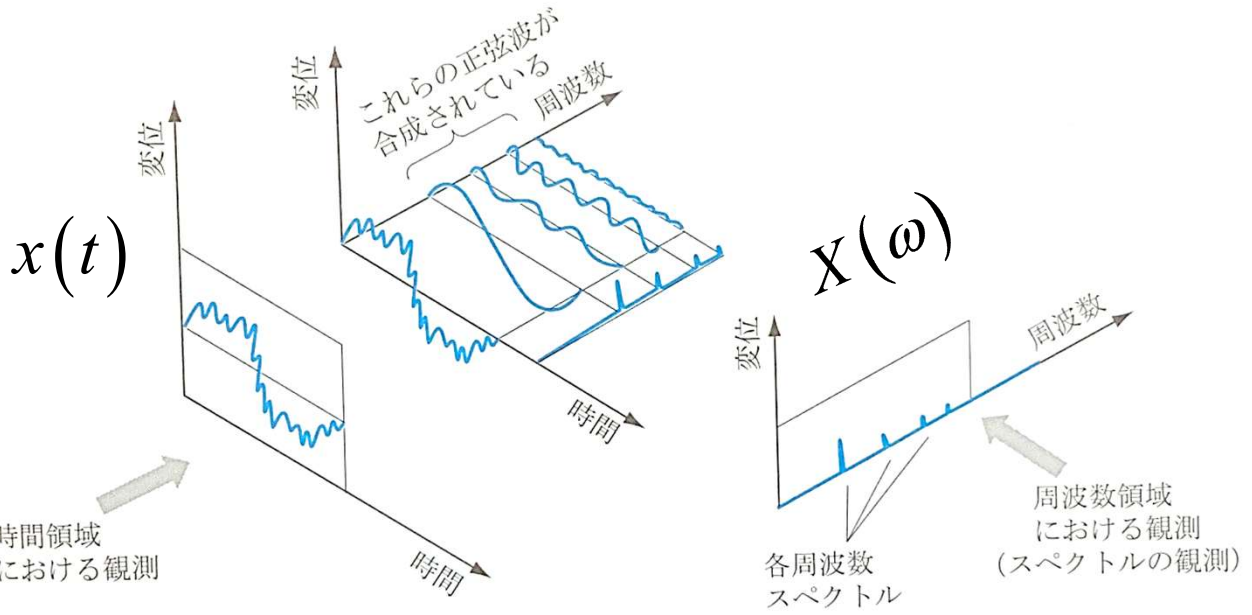
© 2020 Yuki Kato. All Rights Reserved

補：離散フーリエ変換を用いた周波数分析方法

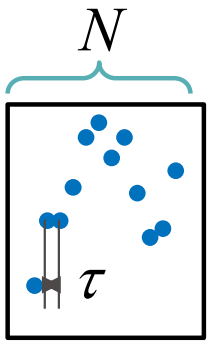
- 振動波形を時間tの関数x(t)とすると、下式で示すフーリエ変換を用いることで時間関数x(t)は周波数関数X(ω)と変換される

$$X(\omega)=\int_{-\infty}^{\infty}x(t)e^{-i\omega t}dt \qquad x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}X(\omega)e^{i\omega t}d\omega$$

- ここで、ωは角振動数である。右式は逆フーリエ変換



離散フーリエ変換



$$X(kf_0)=\sum_{n=0}^{N-1}x(n\tau)e^{-i2\pi kn/N}$$

$$x(n\tau)=\frac{1}{N}\sum_{k=0}^{N-1}X(kf_0)e^{i2\pi kn/N}$$

$$f_0=1/n\tau$$

FFTを用いると速いが，MatlabはDFTも高速化してある

© 2020 Yuki Kato. All Rights Reserved

補：平均値と分散，標準偏差

- 平均値（mean）と分散（variance）は統計における基本的な概念
- 平均値は与えられた分布の確率（密度）関数を重みとした確率変数の値の算術平均であり、期待値（expected value）とも呼ばれる。
- 分散は、平均値の周りでの分布のばらつきの目安を与える。

2.4.1　平均値および分散の定義

1 確率変数に関する平均値および分散は次のように定義される：　　Σvk (k/N)

$$\text{平均値 } \mu = E[X] = \begin{cases} \sum_k v_k f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} x f(x)\mathrm{d}x & \text{（連続分布）} \end{cases} \tag{2.9a}$$

$$\text{分散 } V[X] = E[\{X - E(X)\}^2] = \begin{cases} \sum_k (v_k - \mu)^2 f(v_k) & \text{（離散分布）} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\mathrm{d}x & \text{（連続分布）} \end{cases}$$

分散は σ^2 とも書かれ，分散の平方根 σ は**標準偏差**（standard deviation）という．

© 2020 Yuki Kato. All Rights Reserved

補：母集団全てを計測しない時（標本分散）

- 平均は$\bar{x} = \Sigma x_i / N$で同様の計算だが，標準偏差は

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$ となる．

- 分散計算の基準: 真の分散は、各データ点 xi と母平均 μ の差の二乗和で計算される
- しかし、標本からは母平均 μ が分からないため、代わりに標本平均 $\bar{x}$を使って分散を計算することになる
- $\bar{x}$ は、その標本データセット内で二乗和が最小になるように設計された値である

- したがって、標本平均 $\bar{x}$ を基準として計算された二乗和 $\sum(x_i-\bar{x})^2$ は、真の母平均 μ を基準とした二乗和 $\sum(x_i-\mu)^2$ よりも必ず小さくなる
- この過小評価の偏りを数学的に打ち消し、推定値の期待値が真の値と一致するように補正するために、分母を N−1 に減らす
- これがベッセルの補正であり、N−1 はこの標本データの自由度を表す

© 2020 Yuki Kato. All Rights Reserved

周波数フィルタの種類

■ アナログフィルタ

- 抵抗（R）、インダクタ（L）、キャパシタ（C）といった受動部品や、オペアンプなどの能動部品を用いて電気回路として構成される
- ロガーにはエイリアシングを防ぐために，計測周波数の1/2以下の成分のみを通すアナログフィルタがほぼ必ず実装されている

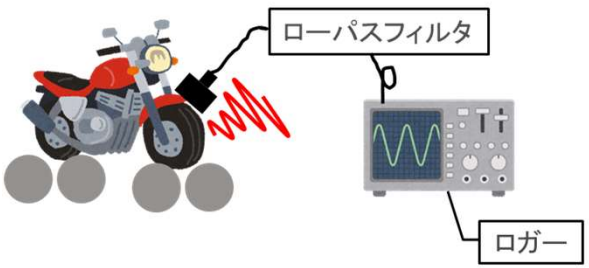
■ デジタルフィルタ

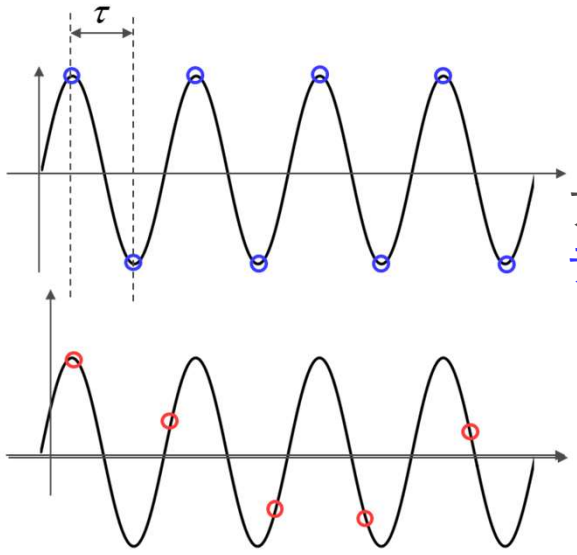
- コンピュータやDSP（デジタル信号処理プロセッサ）上で、サンプリングされた離散信号に対して数学的な計算（アルゴリズム）を用いて実現される

最大周波数:
（ナイキスト周波数）

$$f_{\max} = \frac{1}{2\tau}$$

エイリアシング



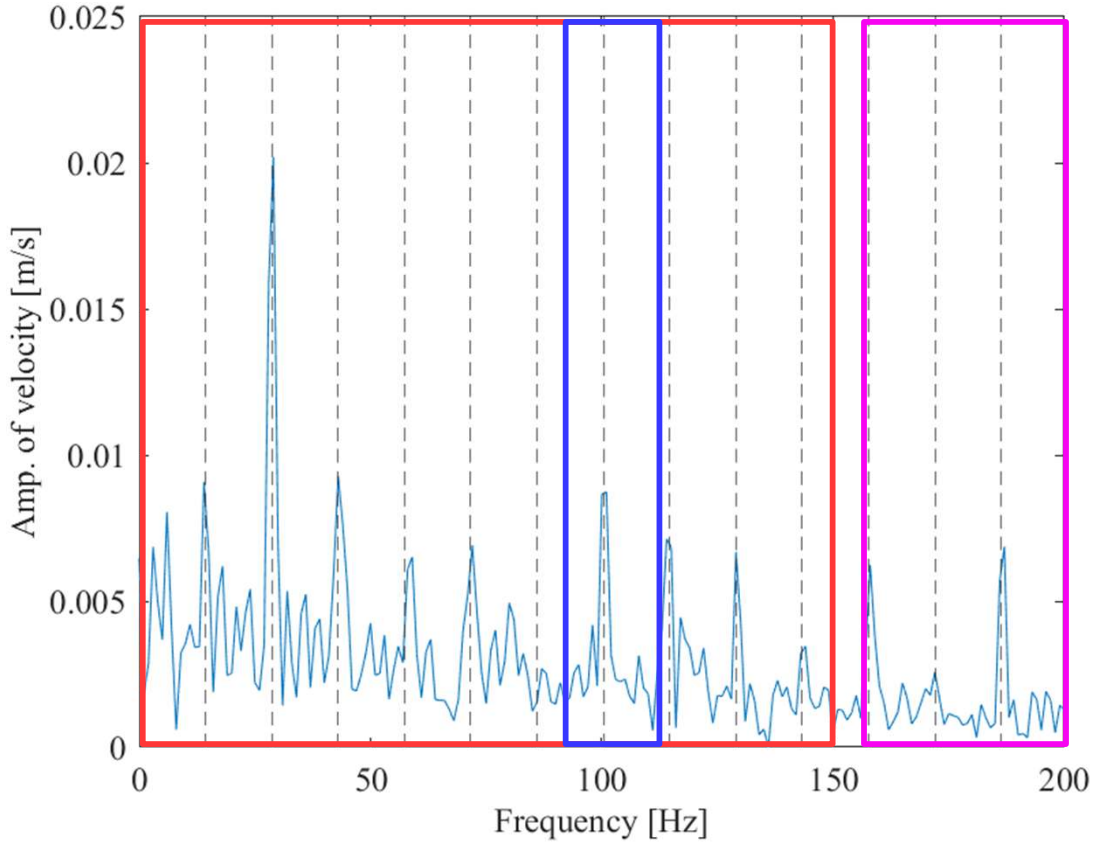


アナログフィルタをかける
ディジタルフィルタをかける

© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（周波数帯）

- ローパスフィルタ **(Low-Pass Filter, LPF)**
- ハイパスフィルタ **(High-Pass Filter, HPF)**
- バンドパスフィルタ **(Band-Pass Filter, BPF)**
- バンドストップフィルタ **(Band-Stop Filter, BSF) / ノッチフィルタ (Notch Filter)**⇒青以外



© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（周波数帯）

- ローパスフィルタ (Low-Pass Filter, LPF)
  - 機能: 低い周波数成分を通過させ、高い周波数成分を阻止します。
  - 用途: ノイズ（通常、高周波数成分）の除去、信号の平滑化
- ハイパスフィルタ (High-Pass Filter, HPF)
  - 機能: 高い周波数成分を通過させ、低い周波数成分を阻止します。
  - 用途: 信号の直流成分（バイアス）の除去、低周波数ノイズ（ハムノイズなど）の除去、エッジ強調（画像処理）
- バンドパスフィルタ (Band-Pass Filter, BPF)
  - 機能: 特定の周波数帯域だけを通過させ、それより低い周波数と高い周波数の両方を阻止します。
  - 用途: 特定の通信チャンネルの選択、特定の音階や周波数の抽出。
- バンドストップフィルタ (Band-Stop Filter, BSF) / ノッチフィルタ (Notch Filter)
  - 機能: 特定の周波数帯域だけを阻止し、それ以外の周波数を通過させます。
  - 用途: 狭い周波数範囲の不要なノイズ（例：商用電源の50Hz/60Hzのハムノイズ）をピンポイントで除去する。

© 2020 Yuki Kato. All Rights Reserved

ディジタルフィルタの種類（アルゴリズム別）

■ デジタルフィルタ

- **FIRフィルタ (Finite Impulse Response):** 過去の入力値のみを用いて計算する。常に安定で、線形位相特性（信号の波形を歪ませない）を容易に実現できる。

$$y[n] = \sum_{k=0}^{M-1} b_k x[n-k]$$

  - M: フィルタの**タップ数**または**次数**（フィルタの長さ）
  - b_k: フィルタ係数（重み）

- **IIRフィルタ (Infinite Impulse Response):** 過去の入力値と過去の出力値の両方を用いて計算する。少ない計算量で急峻な特性を得やすいが、安定性の確保や線形位相の実現が複雑になる。

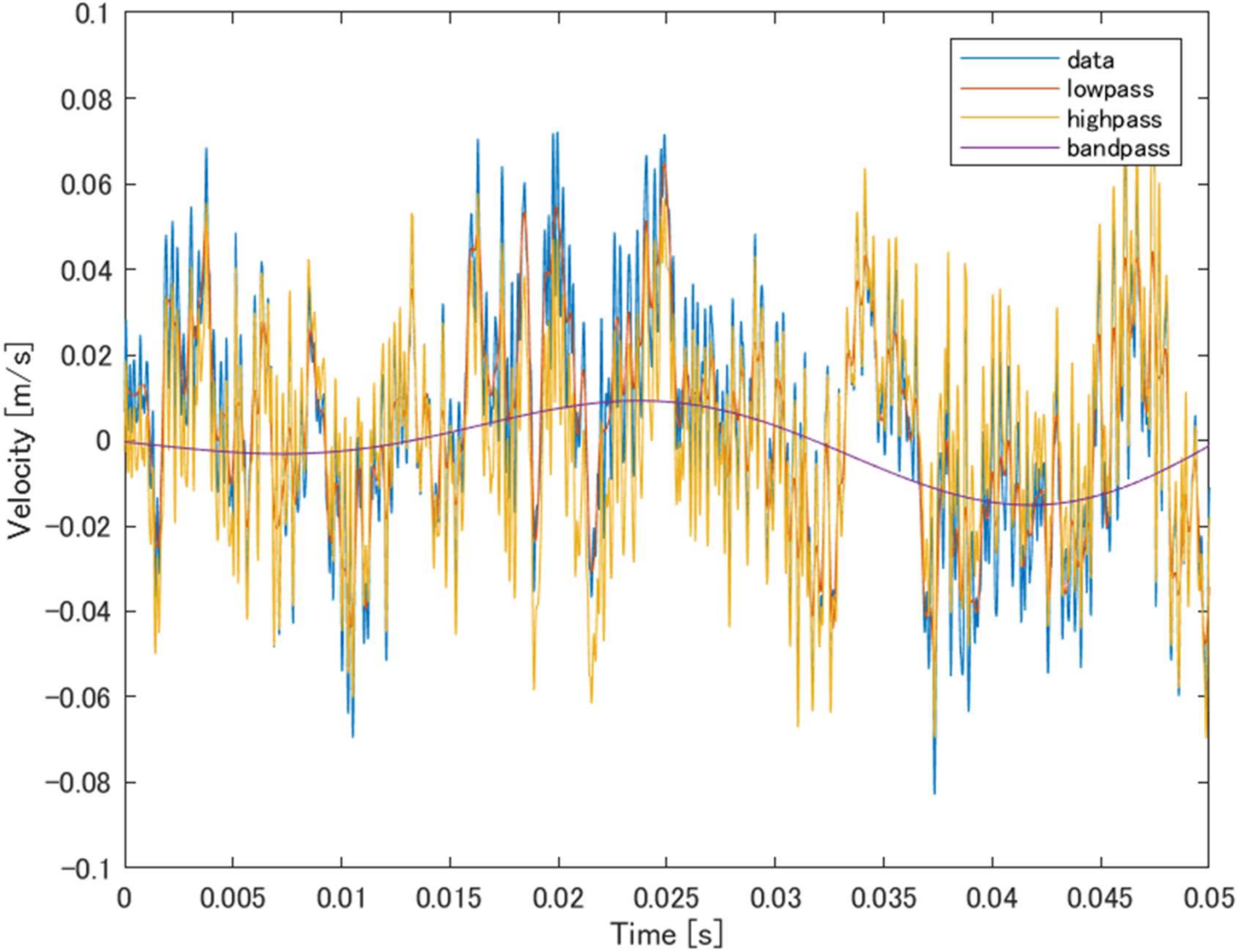
$$y[n] = \sum_{k=0}^{M} b_k x[n-k] - \sum_{j=1}^{N} a_j y[n-j]$$

  - b_k: 入力側のフィルタ係数
  - a_j: 出力側のフィルタ係数（フィードバック係数）

© 2020 Yuki Kato. All Rights Reserved

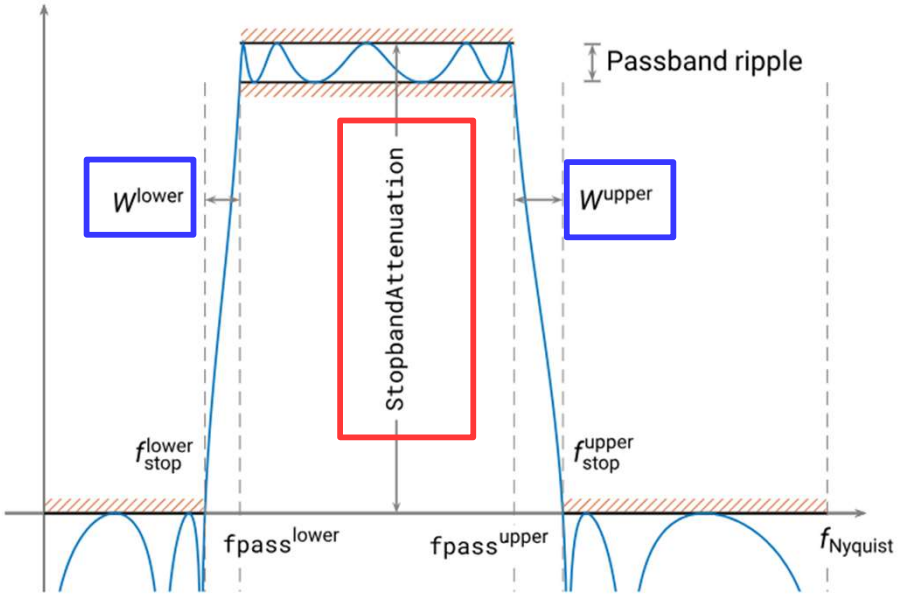
Matlabでの実装

■ それぞれのフィルタ結果は以下の通り



© 2020 Yuki Kato. All Rights Reserved

フィルタ設計



規定値60dB
規定値[0.85 0.85]

遷移帯の幅をコントロールするには、2 要素ベクトル $[s^{lower},s^{upper}]$、またはスカラーとして Steepness を指定できます。

Steepness をベクトルとして指定した場合、関数は以下を実行します。

- 遷移幅下限を以下で計算します。

  $W^{lower} = (1 - s^{lower}) \times \text{fpass}^{lower}$.

  - Steepness の最初の要素が 0.5 に等しい場合、遷移幅は、fpass^{lower} の 50% です。
  - Steepness の最初の要素が 1 に近づくにつれて、遷移幅は最小値の fpass^{lower} の 1% に達するまで徐々に狭くなります。
- 遷移幅上限を以下で計算します。

  $W^{upper} = (1 - s^{upper}) \times (f_{Nyquist} - \text{fpass}^{upper})$.

  - Steepness の 2 番目の要素が 0.5 に等しい場合、遷移幅は、$(f_{Nyquist} - \text{fpass}^{upper})$. の 50% です。
  - Steepness の 2 番目の要素が 1 に近づくにつれて、遷移幅は最小値の $(f_{Nyquist} - \text{fpass}^{upper})$ の 1% に達するまで徐々に狭くなります。

Steepness をスカラーとして指定した場合、関数は以下を実行します。

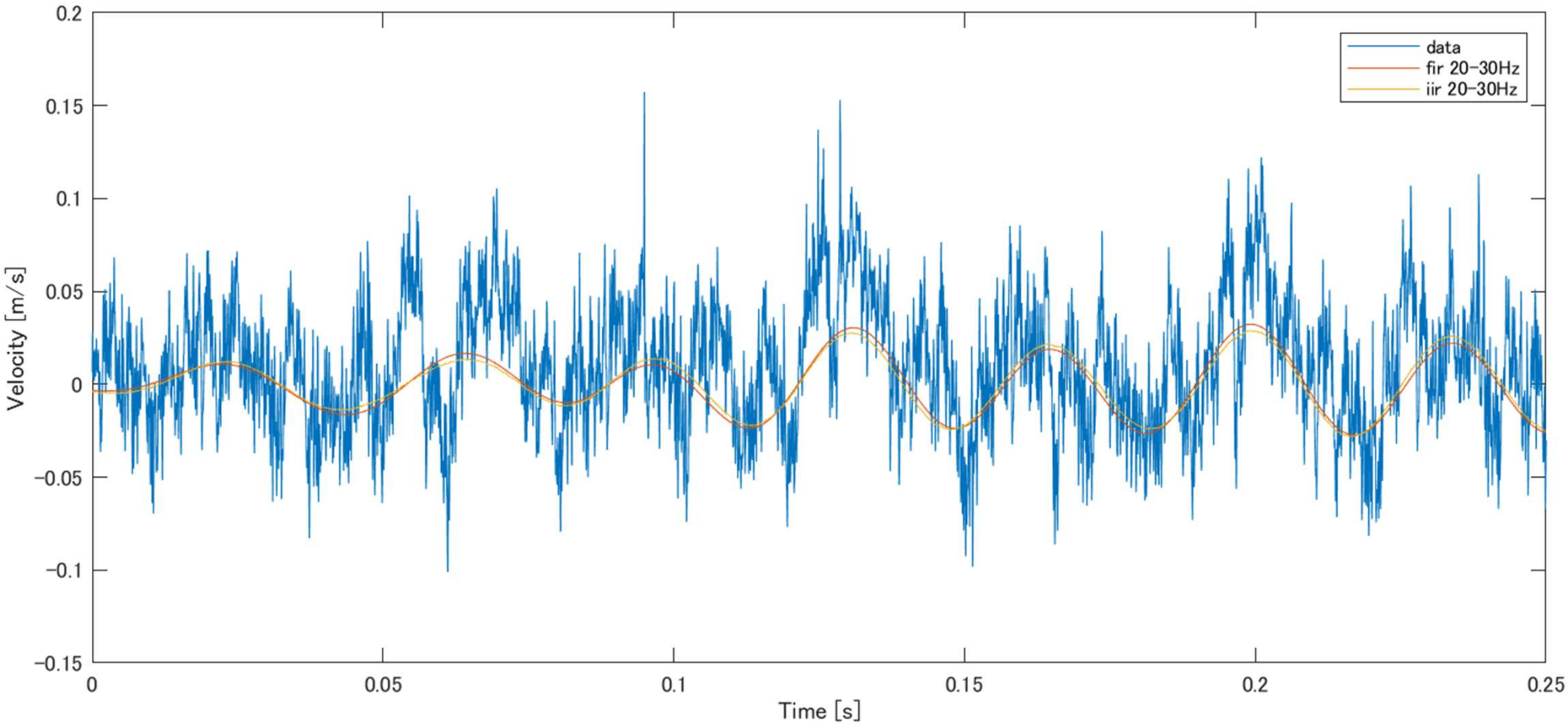
- 指定したスカラー値を使用して、W^{lower} と W^{upper} を計算します。
- W^{lower} と W^{upper} の小さい方に等しい遷移幅下限と遷移幅上限をもつ対称フィルターを設計します。

Steepness の既定値は [0.85 0.85] です。

© 2020 Yuki Kato. All Rights Reserved

FIRフィルタとIIRフィルタ

- 制御に使う/位相情報も積極的に使う場合等の用途を抜きにすれば，IIRフィルタで十分
- 但し，不安定になることがあるので，結果は必ず確認



© 2020 Yuki Kato. All Rights Reserved

演習

- LDVのデータに+1のオフセットを加えて，20-30 Hzのバンドパスフィルタをかけよ
- なお，Steepnessは20, Wは0.7とし，IIRとFIRで比較せよ

遷移帯の幅をコントロールするには、2 要素ベクトル $[s^{lower},s^{upper}]$、またはスカラーとして `Steepness` を指定できます。

`Steepness` をベクトルとして指定した場合、関数は以下を実行します。

- 遷移幅下限を以下で計算します。

  $W^{lower} = (1 - s^{lower}) \times \mathtt{fpass}^{lower}$.

  - `Steepness` の最初の要素が 0.5 に等しい場合、遷移幅は、$\mathtt{fpass}^{lower}$ の 50% です。
  - `Steepness` の最初の要素が 1 に近づくにつれて、遷移幅は最小値の $\mathtt{fpass}^{lower}$ の 1% に達するまで徐々に狭くなります。
- 遷移幅上限を以下で計算します。

  $W^{upper} = (1 - s^{upper}) \times (f_{Nyquist} - \mathtt{fpass}^{upper})$.

  - `Steepness` の 2 番目の要素が 0.5 に等しい場合、遷移幅は、$(f_{Nyquist} - \mathtt{fpass}^{upper})$. の 50% です。
  - `Steepness` の 2 番目の要素が 1 に近づくにつれて、遷移幅は最小値の $(f_{Nyquist} - \mathtt{fpass}^{upper})$ の 1% に達するまで徐々に狭くなります。

`Steepness` をスカラーとして指定した場合、関数は以下を実行します。

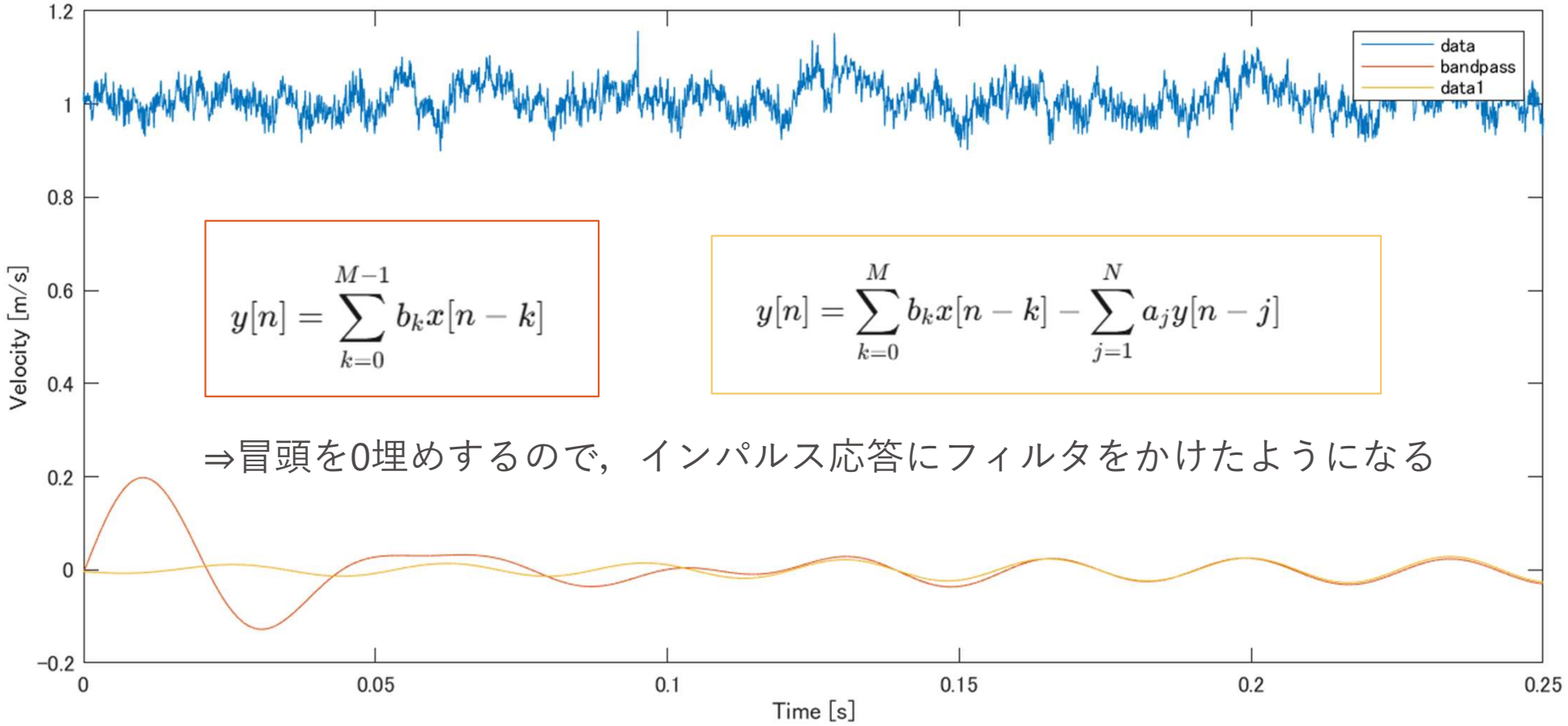
- 指定したスカラー値を使用して、W^{lower} と W^{upper} を計算します。
- W^{lower} と W^{upper} の小さい方に等しい遷移幅下限と遷移幅上限をもつ対称フィルターを設計します。

`Steepness` の既定値は `[0.85 0.85]` です。

© 2020 Yuki Kato. All Rights Reserved

演習

- オフセット成分はどちらのフィルタも除去できている
- しかし，FIRでは過去の入力のみを見てフィルタをかけるので，冒頭の波形が不適
- FIRを使いたい場合は平均値を減算すること



$$y[n] = \sum_{k=0}^{M-1} b_k x[n-k]$$

$$y[n] = \sum_{k=0}^{M} b_k x[n-k] - \sum_{j=1}^{N} a_j y[n-j]$$

⇒冒頭を0埋めするので，インパルス応答にフィルタをかけたようになる

© 2020 Yuki Kato. All Rights Reserved

機械工学で良く用いられるデータ解析

- 単変量時系列データ解析
  - 平均と標準偏差
  - ヒストグラム
  - リサンプリング
  - 差分と積分
  - 移動平均と移動中央値（外れ値フィルタ）
  - 離散フーリエ変換（DFT）と高速フーリエ変換（FFT）
  - 周波数フィルタ
  - ウェーブレット変換
- 多変量時系列データ解析（2変量）
  - 相互相関関数
  - コヒーレンス

© 2020 Yuki Kato. All Rights Reserved

2変量時系列データ

- 等間隔離散時間の単変量時系列
- $\{x_t\} = x_1, x_2, \ldots, x_n \qquad (t_i = t_{i-1} + \Delta t)$
- 等間隔離散時間の2変量時系列
- $\{x_t\} = x_1, x_2, \ldots, x_n \qquad (t_i = t_{i-1} + \Delta t)$
- $\{y_t\} = y_1, y_2, \ldots, y_n \qquad (t_i = t_{i-1} + \Delta t)$

- 2変量時系列の分析では,
  - データ同士に共通の周期性があるか？
  - データ間に時系列のズレがあるか/あればどのていどか？

  という点が重要になる
- 相互相関関数やコヒーレンスを学ぼう

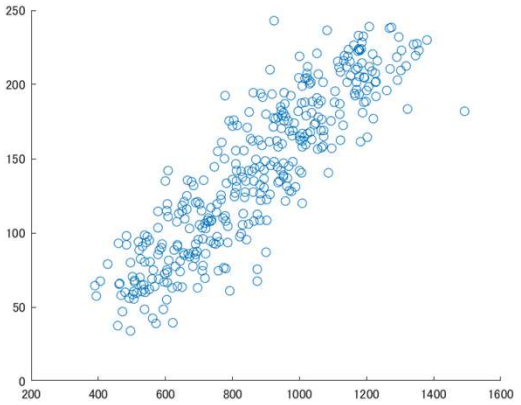
© 2020 Yuki Kato. All Rights Reserved

相関係数

- 相互相関関数を学ぶ前に，似た概念である相関関数を復習する
- 相関係数は，2つの変数間に存在する線形的な関係の強さと方向を示す統計的な指標であり，以下の式で計算する

$$r = \frac{\text{共分散}(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $\bar{x}, \bar{y}$: 各変数の標本平均
- σ_x, σ_y: 各変数の標本標準偏差
- 分子：**共分散**。2つの変数のズレの積の平均。



- 0-0.3　：ほぼ無関係　　0.3-0.5　：非常に弱い相関
- 0.5-0.7　：相関がある　　0.7-0.9　：強い相関がある
- 0.9以上　：非常に強い相関　（負の値も同様）

※分野により異なるので要チェック

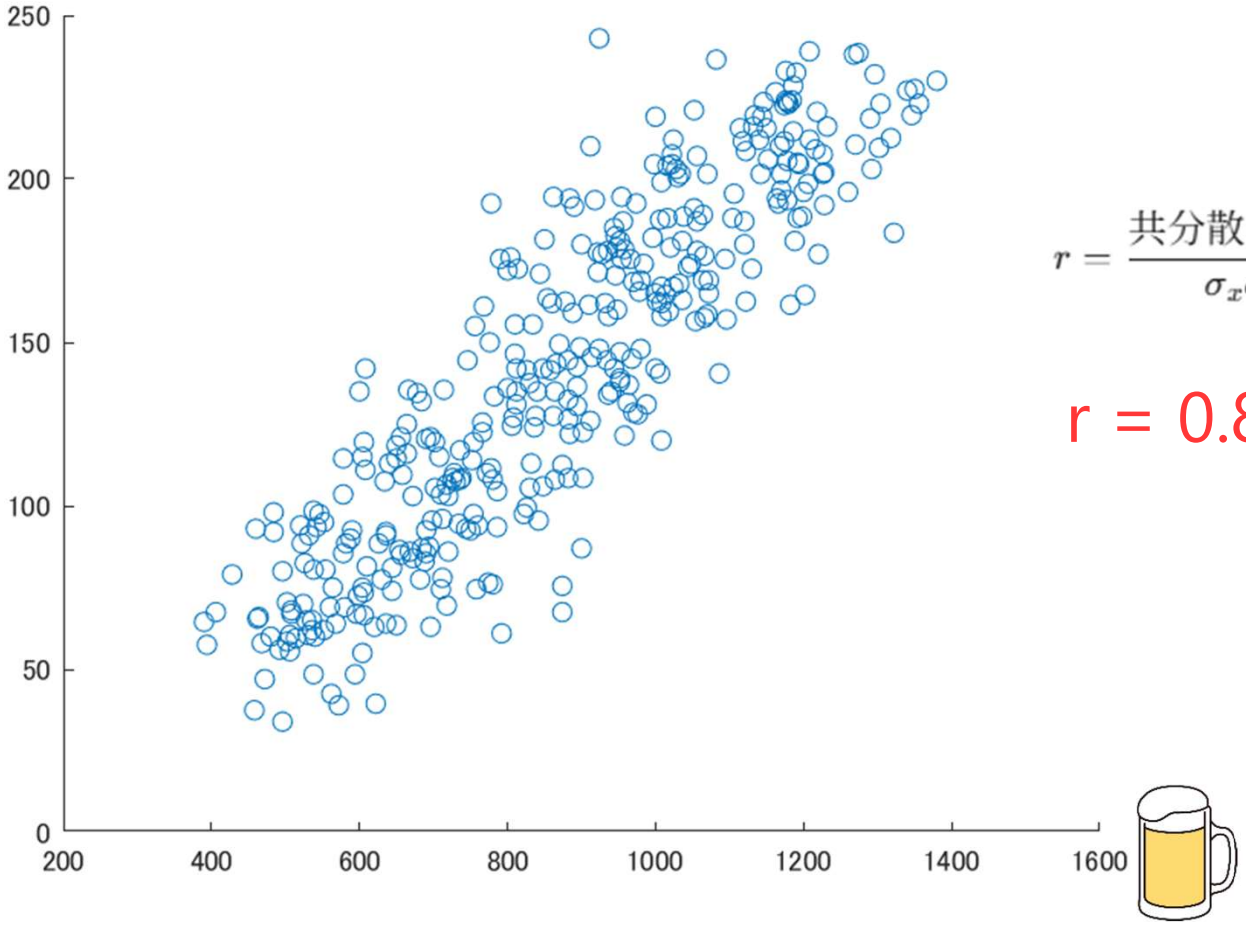
© 2020 Yuki Kato. All Rights Reserved

相関係数を計算してみよう

■ 強い正の相関がある

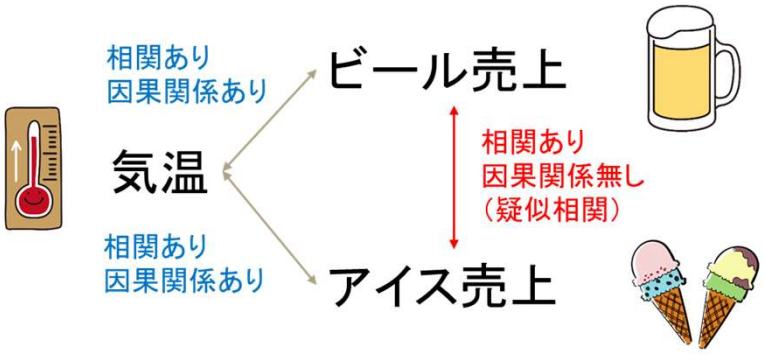
横軸ビール売上-縦軸アイス売上でプロット
ビール売上とアイス売上に相関がある
→疑似相関，因果関係はない
ビールの売り上げを上げるためにアイスのCMを流してアイスの売上を増加させても意味がない
→因果関係，メカニズムを調べることが重要





$$r = \frac{\text{共分散}(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r = 0.88



© 2020 Yuki Kato. All Rights Reserved

相互相関関数

- 2つの時系列データ x(t) と y(t) があるとき，相互相関関数は以下のように定義される

$$r_{xy}(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau}(x_t - \bar{x})(y_{t+\tau} - \bar{y})$$

- τ は時間差（ラグ）であり、 y が x に対して τ だけ先行しているか、遅延しているかを示します
- 各データの標準偏差で正規化すると，

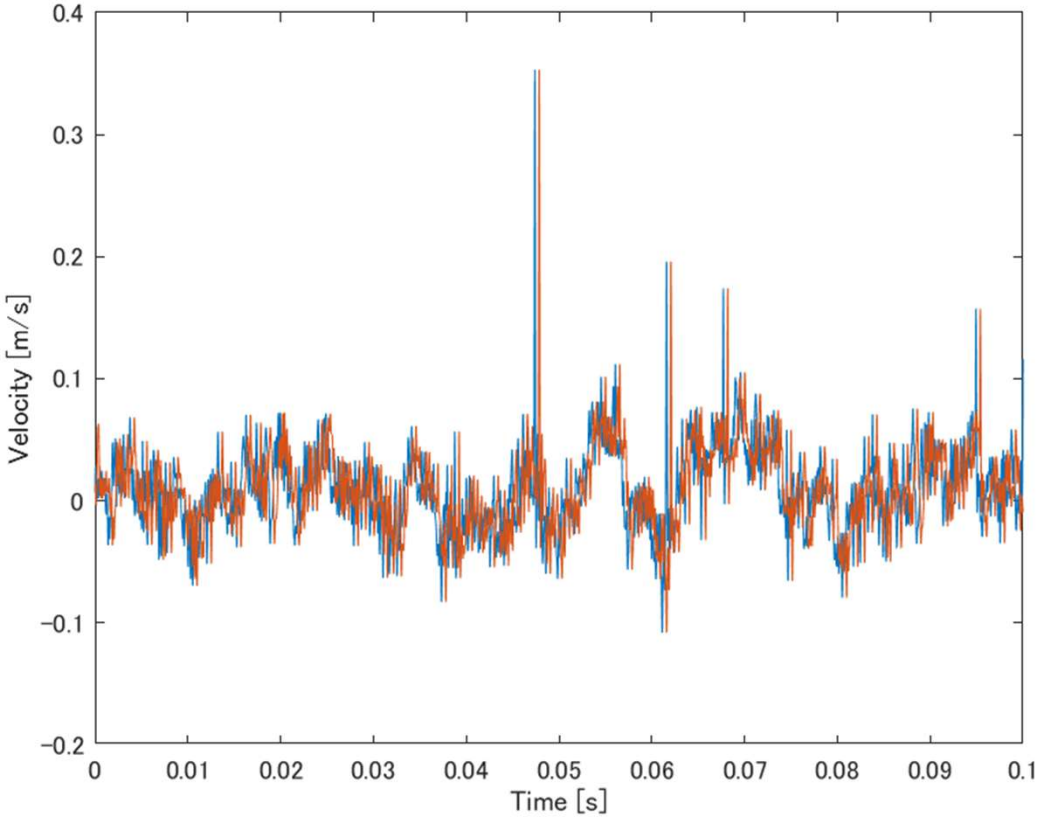
$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_x \sigma_y}$$

要は，時刻をずらしていき相関係数を順次計算する

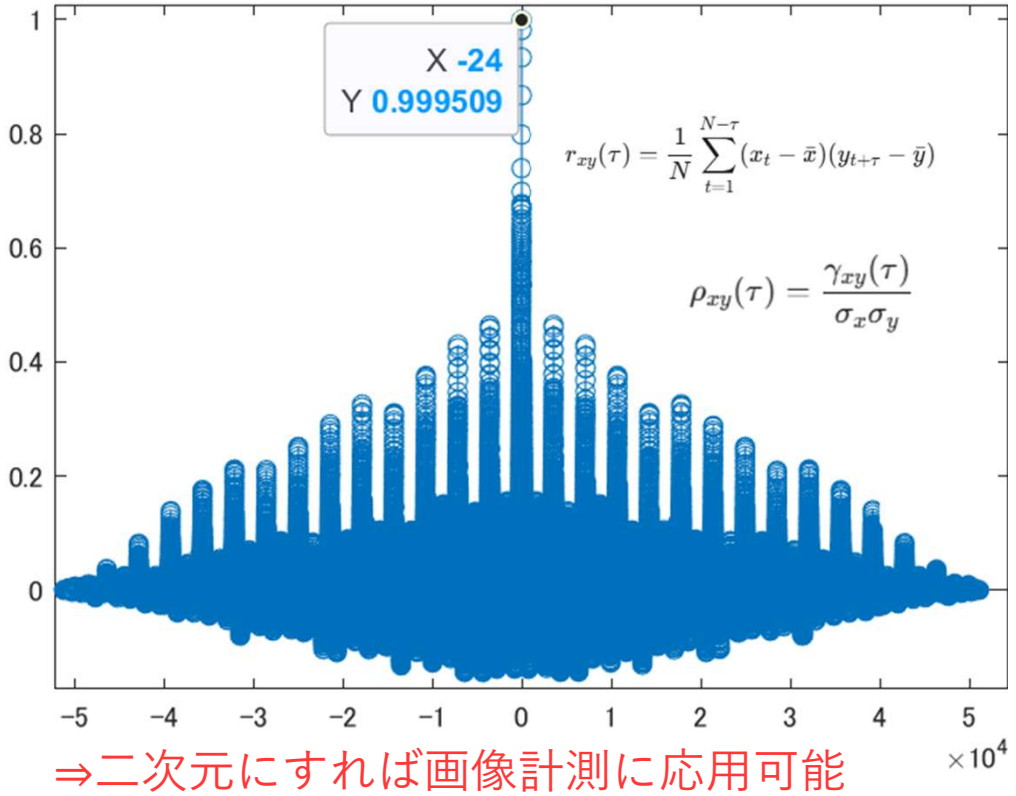
© 2020 Yuki Kato. All Rights Reserved

LDVデータでの検証

- 二つのデータに遅れがある.
- 二台のロガーで同じ信号を記録した場合などに発生
- 遅れが分かれば同期計測できる
- また，二つの物理量の遅れと相関係数を知れば，ダイナミクスの評価ができる



データ2が1に対して24遅れている.
（24/fs = 0.47 ms）



⇒二次元にすれば画像計測に応用可能

© 2020 Yuki Kato. All Rights Reserved

相関まとめ

- 相関係数で変数間の線形な相関を分析できる
- 相互相関関数を使えば，二つの変数間に発生する遅れ量，そしてその際の相関係数を同時に分析ができる
- 但し，実現象では様々な周波数成分があり，それぞれの周波数成分ごとにどれだけ線形な相関があるのか分析したい場合がある
- その際に活躍するのが，コヒーレンスである
  - コヒーレンスが 0.9である周波数 f は、その周期の振動成分が X と Y で非常に強く連動していることを示す
  - **ノイズの除去:** ある周波数成分がノイズやランダムな変動によって発生しているのか、それとも本当に両信号間で共通のメカニズムによって連動しているのかを判断するのに役立ちます。

© 2020 Yuki Kato. All Rights Reserved

コヒーレンス

■ コヒーレンスは、2つの信号の相互スペクトル $G_{xy}(f)$ と、それぞれの信号のパワースペクトル $G_{xx}(f)$、$G_{yy}(f)$ を用いて定義される

$$\text{コヒーレンス二乗 } \gamma_{xy}^2(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)}$$

■ 相互スペクトル：

- 相互スペクトル $G_{xy}(f)$ は、2つの信号 X(t) と Y(t) の相互相関関数 $\gamma_{xy}(\tau)$ をフーリエ変換したもの
- 二つの信号の複素振幅の掛け算でも計算できる

■ パワースペクトル

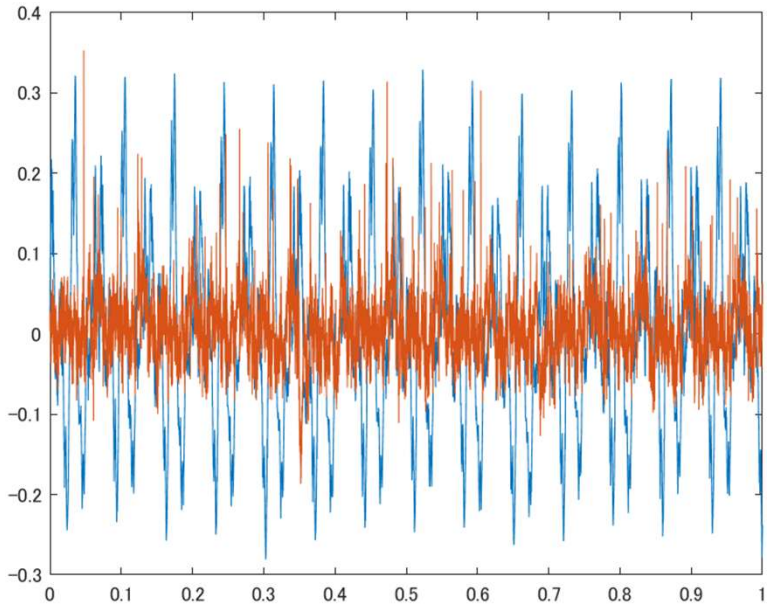
- T振幅の二乗

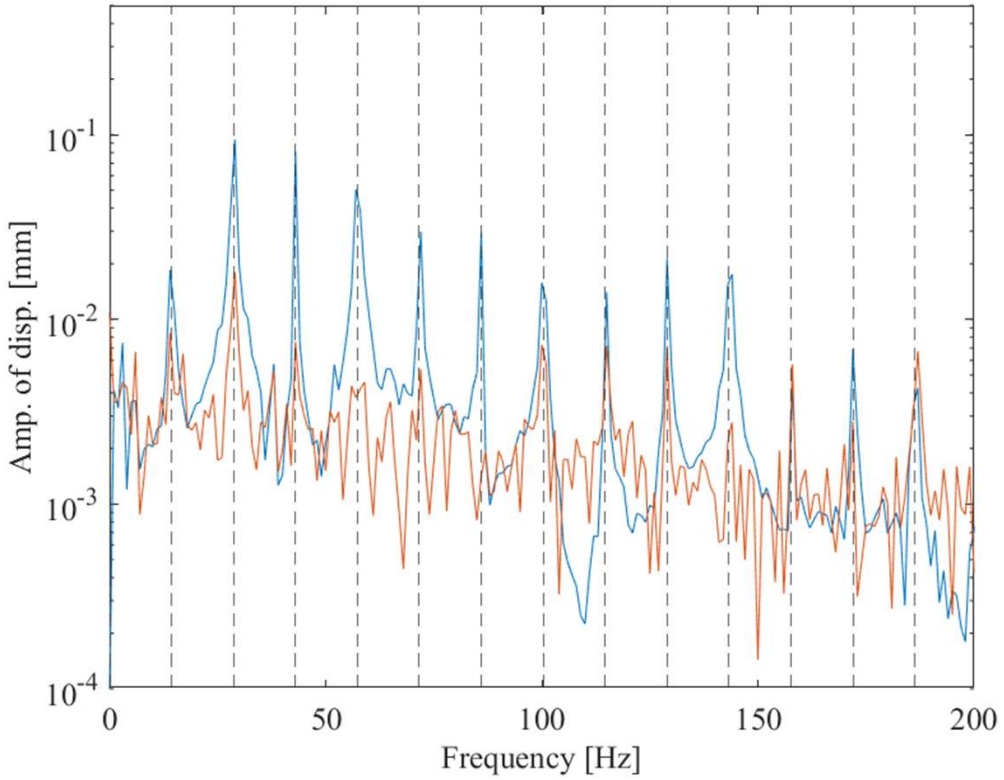
© 2020 Yuki Kato. All Rights Reserved

演習

- LDVによる計測データと，DICによる計測データのコヒーレンスを計算しよう
  - LDVにより求めた変位について，低い周波数は信頼できない
  - アップサンプリングする場合は，元の計測周波数以上の情報は評価できない

$$\text{コヒーレンス二乗 } \gamma_{xy}^2(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)}$$

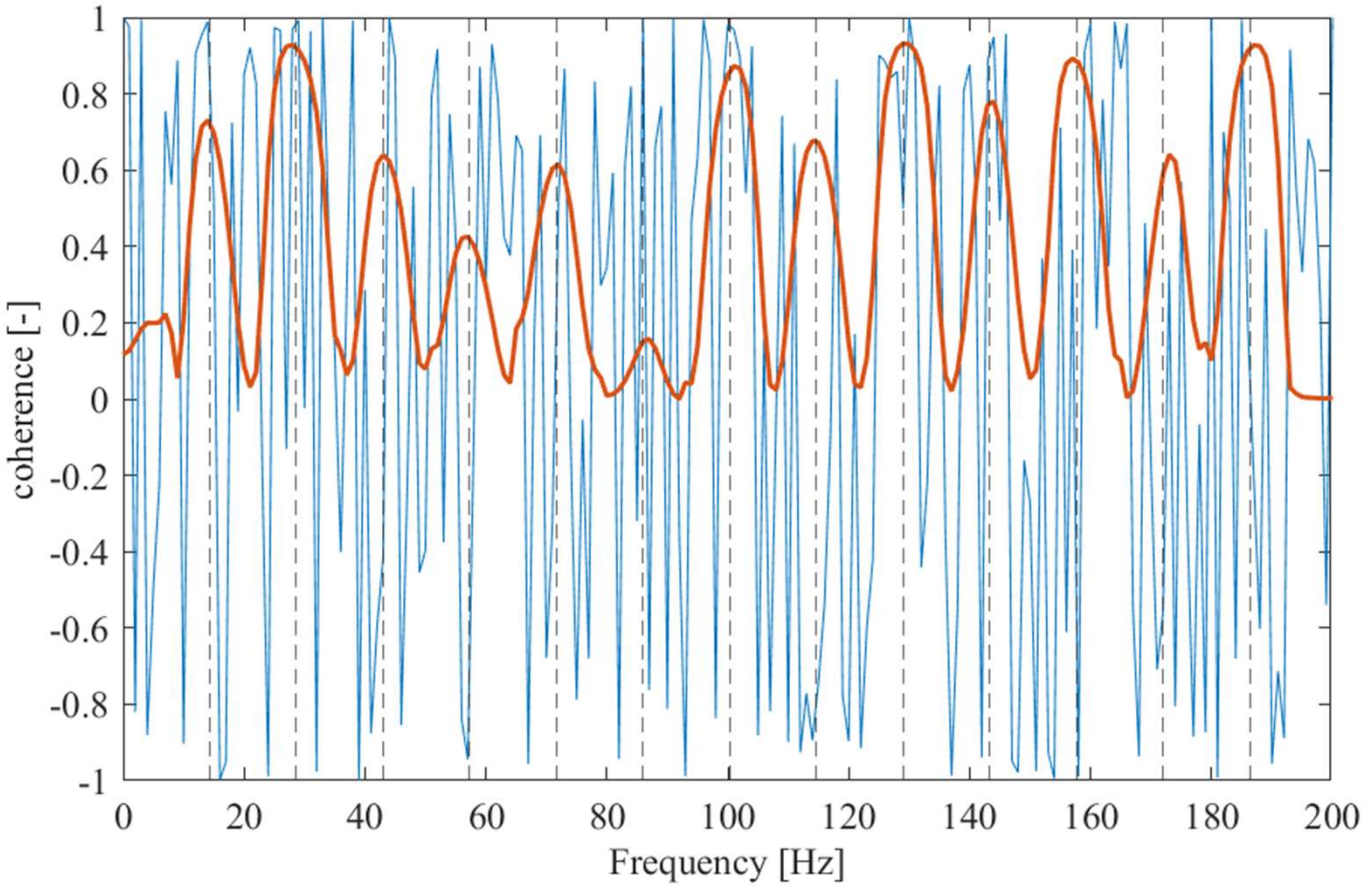




© 2020 Yuki Kato. All Rights Reserved

演習

- 平均化なしでは良く分からない
- 平均化すると，回転周波数とその次数成分ではコヒーレンスが高く，相関があると言える．
- どの程度が相関があると言えるかは・・・分野による



© 2020 Yuki Kato. All Rights Reserved

振動工学界隈

モード解析（モーダルテスト）では、構造物に加える**励起力**（入力）と、それによって生じる**応答**（出力）を測定し、その比率である**周波数応答関数（FRF）**を求めます。

コヒーレンス $\gamma^2(f)$ は、このFRFの品質を評価するために使われます。

$$\gamma^2(f) = \frac{|\text{相互スペクトル}(f)|^2}{\text{入力パワースペクトル}(f) \cdot \text{出力パワースペクトル}(f)}$$

- $\gamma^2(f) \approx 1$: 入力と出力の間に**完全な線形関係**があり、FRFが信頼できることを示します。
- $\gamma^2(f) \ll 1$: 入力と出力が**無関係**であるか、**ノイズ**、**非線形性**、または**不十分な平均化**によってFRFが信頼できないことを示します。

コヒーレンス $\gamma^2(f)$ の値	信頼性/解釈
$0.9 \sim 1.0$	**極めて良好**。非常に信頼性の高いFRFであり、線形かつ低ノイズの理想的な測定を示します。
$0.8 \sim 0.9$	**良好**。一般的にモード抽出（パラメータ同定）を行うのに十分な品質と見なされます。
0.7 以下	**要注意**。このレベルを下回ると、FRFがノイズや非線形性の影響を強く受けている可能性が高く、モードパラメータの抽出が困難になる、または誤った結果を導くリスクが高まります。

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．
補講があれば，また連絡します

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

✓ 課題①，②出題

✓ 課題①②

統計解析の基礎
1次元，2変数

→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析

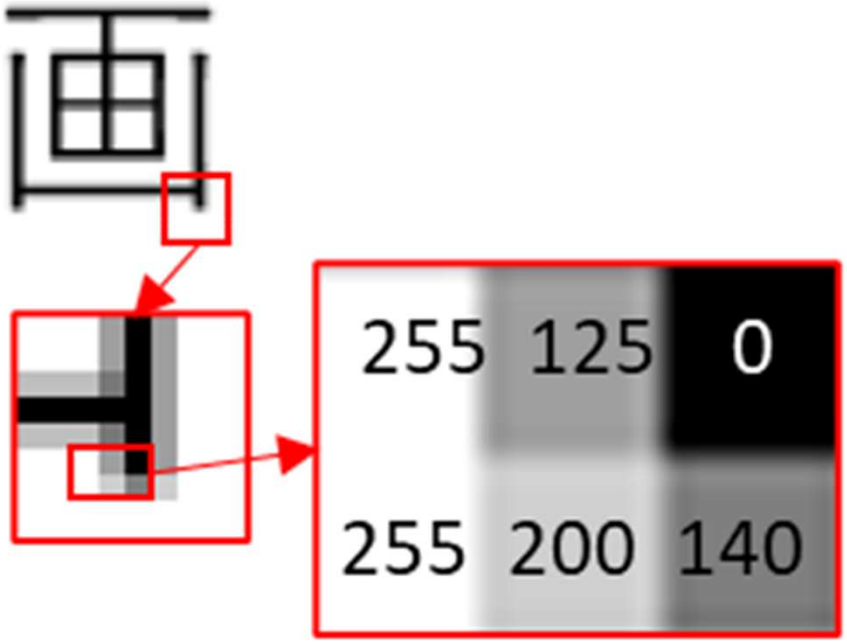
→ビッグデータ解析に必須のスキル

© 2020 Yuki Kato. All Rights Reserved

デジタル画像とは？

- カメラやスキャナなどの装置により，2次元平面座標に量子化されたデータ
- 重要な概念：解像度，諧調

現実世界の連続値である輝度をサンプリングし，0-255で量子化（8bit）
空間的に連続的な輝度変化を特定のグリッドで解像する　　階調
⇒解像度：1000×1000 pix＝100万画素　　白黒は1bit
⇒300 dpi（dot per inch）1インチ(25.4mm)あたり300pix　　ハイエンドだと12bitも



※カラー画像だと，RGBで三つの値を持つ

© 2020 Yuki Kato. All Rights Reserved

画像のフォーマット

- 様々な拡張子で保存される
- **A. ラスター形式 (Raster Graphics / ビットマップ形式)**
  - 画像をピクセル（画素）の格子状の集まりとして表現する形式
  - デジカメ写真やWeb画像など、**色やグラデーションが複雑な画像**に適しています。
  - **特徴:** 拡大すると画像が粗くなる（ジャギーが発生する）。
  - **主な形式:** JPEG, PNG, GIF, TIFF, BMP
- **B. ベクター形式 (Vector Graphics)**
  - 画像を点、線、円などの幾何学的図形と、それらの関係を記述する数式（ベクトル）で表現する形式です。ロゴやイラスト、フォントなど、**拡大・縮小しても劣化しない画像**に適しています。
  - **特徴:** 拡大・縮小しても劣化しない。複雑な写真の保存には不向き。
  - **主な形式:** SVG, AI, EPS

© 2020 Yuki Kato. All Rights Reserved

画像のフォーマット-ラスター形式

■ A. ラスター形式 (Raster Graphics / ビットマップ形式)

フォーマット	拡張子	圧縮方式	特徴と主な用途
JPEG (Joint Photographic Experts Group)	`.jpg`, `.jpeg`	**非可逆圧縮** (Lossy)	**最も一般的**な写真・Web画像形式。圧縮率が高くファイルサイズが小さいが、保存のたびに画質が劣化する。透明度には非対応。
PNG (Portable Network Graphics)	`.png`	**可逆圧縮** (Lossless)	**透明度（透過）**をサポートするWeb画像形式。グラデーションの少ないWeb画像、ロゴ、アイコン、スクリーンショットに適している。JPEGよりファイルサイズが大きい。
GIF (Graphics Interchange Format)	`.gif`	**可逆圧縮** (Lossless)	256色以下の画像を扱う。**アニメーション（複数の画像を連続再生）**をサポートする唯一の形式。色数が少ないロゴや簡単なイラスト向け。
TIFF (Tagged Image File Format)	`.tif`, `.tiff`	可逆または非可逆	**高品質な保存、印刷、DTP向け**。非圧縮または可逆圧縮で画像情報を完全に保持できるため、印刷業界や医療分野で広く使われる。ファイルサイズは極めて大きい。
BMP (Bitmap)	`.bmp`	非圧縮または可逆	圧縮をほとんど行わない、Windowsで使われる基本的な形式。ファイルサイズが非常に大きいため、通常は使われない。

容量が小さい
情報を保全

画像読み込み
が高速

© 2020 Yuki Kato. All Rights Reserved

画像処理

- 古典的な画像処理（理屈に基づく手法）
  - フィルタリング
  - 色空間変換: RGB、HSV、Labなどの異なる色表現の間で変換を行う。
  - 幾何学的変換: 画像を回転、拡大・縮小、平行移動、透視変換など
  - エッジ検出: 画像の輝度変化が大きい場所（輪郭）を抽出する（例：Sobel、Cannyフィルタ）。
  - 特徴量抽出（伝統的）: SIFT、SURF、HOGなどのアルゴリズムを用いて、画像の局所的な特徴を抽出する。
- 統計的パターン認識手法
  - 特徴量（SIFT, HOGなど）を入力として、統計モデルやシンプルな機械学習モデルを用いて、物体を識別・分類する手法．SVM,PCAなど
- 機械学習やディープラーニングによる手法
  - 現代の画像処理の主流であり、特徴量の抽出から分類・認識までを学習モデルやニューラルネットワークが行う

機械学習（最尤推定，ベイズ統計，深層学習等）を用いた機械学習による画像分類や画像生成（オートエンコーダ）などを紹介予定

© 2020 Yuki Kato. All Rights Reserved

フィルタ処理

- 平滑化
  - ガウスフィルタ
  - メディアンフィルタ
  - ・・・
- 先鋭化
  - 微分フィルタ（輪郭抽出等）
    - https://jp.mathworks.com/discovery/edge-detection.html
- その他：周波数フィルタ
  - https://jp.mathworks.com/help/images/design-linear-filters-in-the-frequency-domain.html

© 2020 Yuki Kato. All Rights Reserved

ガウスフィルタ

- ガウス関数を参照した重みづけをすることで，平滑化を行うフィルタ
- 下は3×3のフィルタサイズの例

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

ガウス分布関数

156	159	158	155	158	156	159	158	157
160	255	157	158	157	159	158	158	158
156	159	158	155	158	156	159	158	157
160	154	157	158	157	159	158	158	158
156	153	155	159	159	155	156	155	155
155	155	155	157	156	159	152	158	156
156	153	157	156	153	155	154	155	157
159	159	156	158	156	159	157	161	162
158	155	158	154	156	160	162	155	159

×

1/16	2/16	1/16
2/16	4/16	1/16
1/16	2/16	1/16

=

158	167	159	156	157	157	159	158	157
167	218	165	157	157	158	158	158	158
158	166	159	156	157	157	158	158	157
159	155	157	158	157	158	158	158	158
156	154	155	158	158	156	156	156	156
155	155	155	157	156	157	154	157	156
156	154	156	156	154	155	154	156	157
159	158	157	157	156	158	158	160	161
158	156	157	155	157	160	160	157	159

© 2020 Yuki Kato. All Rights Reserved

ガウスフィルタ

■ 演習

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

ガウス分布関数

Original Image (Left) Vs. Gaussian Filtered Image (Right)



© 2020 Yuki Kato. All Rights Reserved

メディアンフィルタ

- 中央値を計算し埋め込むことで，主にノイズ除去を行うフィルタ
- 下は3×3のフィルタサイズの例

ノイズ除去にはガウスより効果的

156	159	158	155	158	156	159	158	157
160	255	157	158	157	159	158	158	158
156	159	158	155	158	156	159	158	157
160	154	157	158	157	159	158	158	158
156	153	155	159	159	155	156	155	155
155	155	155	157	156	159	152	158	156
156	153	157	156	153	155	154	155	157
159	159	156	158	156	159	157	161	162
158	155	158	154	156	160	162	155	159

中央値=158

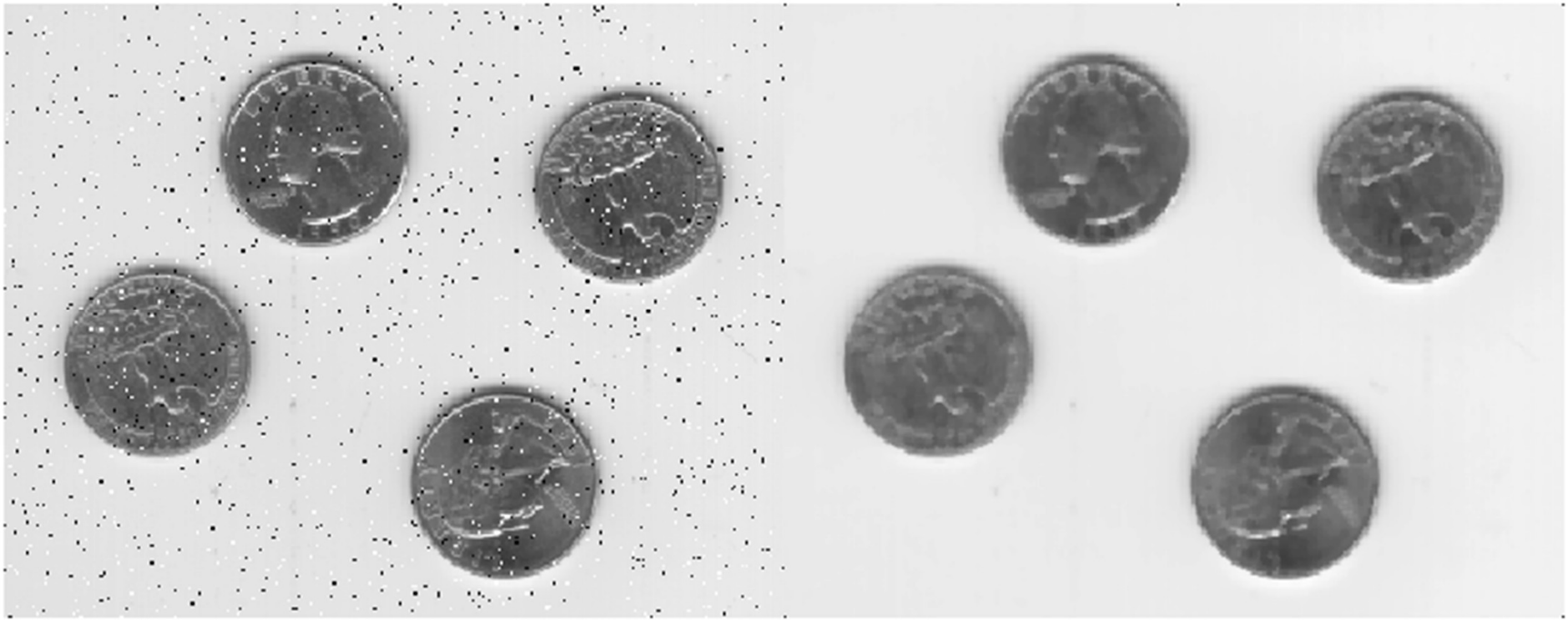
0	157	157	157	156	157	158	158	158
156	158	158	158	157	158	158	158	158
156	158	158	157	158	158	158	158	158
154	156	157	158	158	158	158	158	158
154	155	155	157	158	157	158	156	158
153	155	155	156	156	155	155	155	156
155	156	156	156	156	156	157	157	157
155	157	156	156	156	156	157	157	157
155	157	157	157	158	159	160	159	160

"zeros" (既定)	イメージを 0 でパディングします。
"symmetric"	境界でイメージを対称的に拡張します。

© 2020 Yuki Kato. All Rights Reserved

メディアンフィルタ

■ 演習



© 2020 Yuki Kato. All Rights Reserved

画像計測

- 最近の工学分野では，画像を用いた物理量の計測がよく用いられている
- ここでは流体計測に用いられるPIV(Particle Image Velocimetry)と，構造物の振動計測に用いられるDIC(Digital Image Correlation)を解説する

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．
補講があれば，また連絡します

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布，大数の法則と中心極限定理
- 第５回　時系列データ解析とフィルタ処理
- 第６回　画像解析とフィルタ処理
- 第７回　最小二乗法
- 第８回　ベイズ推定
- 第９回　最尤推定
- 第１０回　オートエンコーダと変分オートエンコーダ
- 第１１回　多変量データ解析の基礎と主成分分析
- 第１２回　固有値分解
- 第１３回　特異値分解
- 第１４回　スパースモデリング
- 第１５回　期末テスト
- 第１６回　返却

⇓統計を使ってデータを解析

✓ 課題①，②出題

✓ 課題①②

統計解析の基礎
1次元，2変数

→データ解析/論文作成に必須のスキル

統計解析の応用
多変量解析

→ビッグデータ解析に必須のスキル

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第八回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布
- 第５回　大数の法則と中心極限定理
- 第６回　時系列データ解析とフィルタ処理1
- 第７回　時系列データ解析とフィルタ処理2，画像解析　✓課題①，②出題

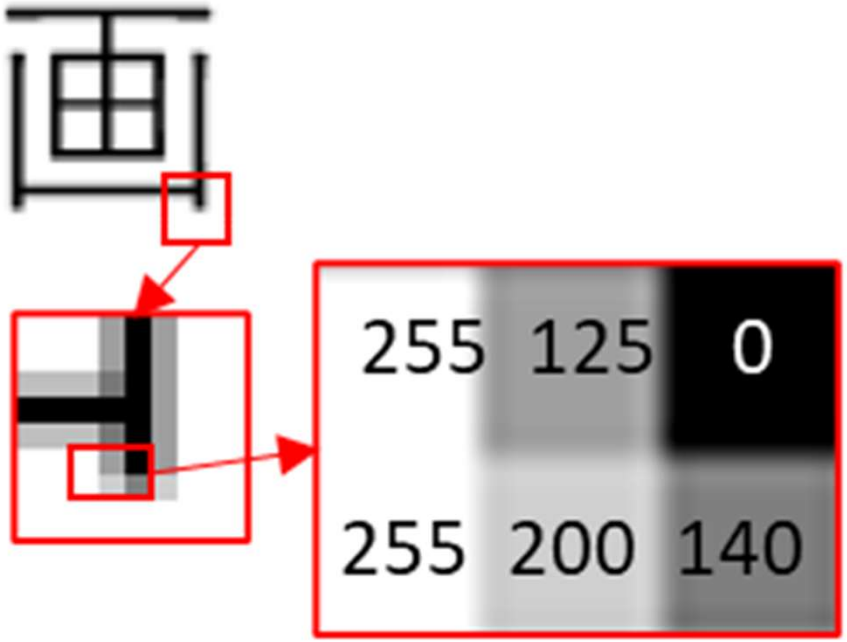
10/28 ■ 第８回　画像解析
10/31 ■ 第９回　最尤推定，最小二乗法
11/4 ■ 第１０回　ベイズ推定
11/14 ■ 第１１回　オートエンコーダと変分オートエンコーダ
11/18 ■ 第１２回　多変量データ解析の基礎と主成分分析　✓11/14: 課題①②〆，両方とも，講義開始時に集めます
11/21 ■ 第１３回　固有値分解，特異値分解　答案の配布と，課題の解説をどこかで実施予定
11/25 ■ 第１４回　スパースモデリング
11/28 ■ 第１５回　期末テスト
12/2 ■ 第１６回　返却　⇒返却のみ，補講扱いですが，全員出席とするので任意参加

© 2020 Yuki Kato. All Rights Reserved

デジタル画像とは？

- カメラやスキャナなどの装置により，2次元平面座標に量子化されたデータ
- 重要な概念：解像度，諧調

現実世界の連続値である輝度をサンプリングし，0-255で量子化（8bit）
空間的に連続的な輝度変化を特定のグリッドで解像する　　階調
⇒解像度：1000×1000 pix＝100万画素　　白黒は1bit
⇒300 dpi（dot per inch）1インチ(25.4mm)あたり300pix　　ハイエンドだと12bitも



※カラー画像だと，RGBで三つの値を持つ

© 2020 Yuki Kato. All Rights Reserved

画像のフォーマット

- 様々な拡張子で保存される
- **A. ラスター形式 (Raster Graphics / ビットマップ形式)**
  - 画像をピクセル（画素）の格子状の集まりとして表現する形式
  - デジカメ写真やWeb画像など、**色やグラデーションが複雑な画像**に適しています。
  - **特徴:** 拡大すると画像が粗くなる（ジャギーが発生する）。
  - **主な形式:** JPEG, PNG, GIF, TIFF, BMP
- **B. ベクター形式 (Vector Graphics)**
  - 画像を点、線、円などの幾何学的図形と、それらの関係を記述する数式（ベクトル）で表現する形式です。ロゴやイラスト、フォントなど、**拡大・縮小しても劣化しない画像**に適しています。
  - **特徴:** 拡大・縮小しても劣化しない。複雑な写真の保存には不向き。
  - **主な形式:** SVG, AI, EPS

© 2020 Yuki Kato. All Rights Reserved

画像のフォーマット-ラスター形式

■ A. ラスター形式 (Raster Graphics / ビットマップ形式)

フォーマット	拡張子	圧縮方式	特徴と主な用途
JPEG (Joint Photographic Experts Group)	`.jpg`, `.jpeg`	**非可逆圧縮** (Lossy)	**最も一般的**な写真・Web画像形式。圧縮率が高くファイルサイズが小さいが、保存のたびに画質が劣化する。透明度には非対応。
PNG (Portable Network Graphics)	`.png`	**可逆圧縮** (Lossless)	**透明度（透過）**をサポートするWeb画像形式。グラデーションの少ないWeb画像、ロゴ、アイコン、スクリーンショットに適している。JPEGよりファイルサイズが大きい。
GIF (Graphics Interchange Format)	`.gif`	**可逆圧縮** (Lossless)	256色以下の画像を扱う。**アニメーション（複数の画像を連続再生）**をサポートする唯一の形式。色数が少ないロゴや簡単なイラスト向け。
TIFF (Tagged Image File Format)	`.tif`, `.tiff`	可逆または非可逆	**高品質な保存、印刷、DTP向け**。非圧縮または可逆圧縮で画像情報を完全に保持できるため、印刷業界や医療分野で広く使われる。ファイルサイズは極めて大きい。
BMP (Bitmap)	`.bmp`	非圧縮または可逆	圧縮をほとんど行わない、Windowsで使われる基本的な形式。ファイルサイズが非常に大きいため、通常は使われない。

容量が小さい
情報を保全

画像読み込み
が高速

© 2020 Yuki Kato. All Rights Reserved

画像処理

身近なものだと，人物写真加工も最近は深層学習が多い

■ 古典的な画像処理（理屈に基づく手法）

- フィルタリング
- 色空間変換: RGB、HSV、Labなどの異なる色表現の間で変換を行う。
- 幾何学的変換: 画像を回転、拡大・縮小、平行移動、透視変換など
- エッジ検出: 画像の輝度変化が大きい場所（輪郭）を抽出する（例：Sobel、Cannyフィルタ）。
- 特徴量抽出（伝統的）: SIFT、SURF、HOGなどのアルゴリズムを用いて、画像の局所的な特徴を抽出する。

■ 統計的パターン認識手法

- 特徴量（SIFT, HOGなど）を入力として、統計モデルやシンプルな機械学習モデルを用いて、物体を識別・分類する手法．SVM,PCAなど

■ 機械学習やディープラーニングによる手法

- 現代の画像処理の主流であり、特徴量の抽出から分類・認識までを学習モデルやニューラルネットワークが行う

機械学習（最尤推定，ベイズ統計，深層学習等）を用いた機械学習による画像分類や画像生成（オートエンコーダ）などを紹介予定

© 2020 Yuki Kato. All Rights Reserved

フィルタ処理

- 平滑化
  - ガウスフィルタ
  - メディアンフィルタ
  - ・・・
- 先鋭化
  - 微分フィルタ（輪郭抽出等）
    - https://jp.mathworks.com/discovery/edge-detection.html
- その他：周波数フィルタ
  - https://jp.mathworks.com/help/images/design-linear-filters-in-the-frequency-domain.html

© 2020 Yuki Kato. All Rights Reserved

ガウスフィルタ

- ガウス関数を参照した重みづけをすることで，平滑化を行うフィルタ
- 下は3×3のフィルタサイズの例

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

ガウス分布関数

156	159	158	155	158	156	159	158	157
160	255	157	158	157	159	158	158	158
156	159	158	155	158	156	159	158	157
160	154	157	158	157	159	158	158	158
156	153	155	159	159	155	156	155	155
155	155	155	157	156	159	152	158	156
156	153	157	156	153	155	154	155	157
159	159	156	158	156	159	157	161	162
158	155	158	154	156	160	162	155	159

×

1/16	2/16	1/16
2/16	4/16	1/16
1/16	2/16	1/16

=

158	167	159	156	157	157	159	158	157
167	218	165	157	157	158	158	158	158
158	166	159	156	157	157	158	158	157
159	155	157	158	157	158	158	158	158
156	154	155	158	158	156	156	156	156
155	155	155	157	156	157	154	157	156
156	154	156	156	154	155	154	156	157
159	158	157	157	156	158	158	160	161
158	156	157	155	157	160	160	157	159

© 2020 Yuki Kato. All Rights Reserved

ガウスフィルタ

■ 演習

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

ガウス分布関数

Original Image (Left) Vs. Gaussian Filtered Image (Right)



© 2020 Yuki Kato. All Rights Reserved

メディアンフィルタ

- 中央値を計算し埋め込むことで，主にノイズ除去を行うフィルタ
- 下は3×3のフィルタサイズの例

ノイズ除去にはガウスより効果的

156	159	158	155	158	156	159	158	157
160	255	157	158	157	159	158	158	158
156	159	158	155	158	156	159	158	157
160	154	157	158	157	159	158	158	158
156	153	155	159	159	155	156	155	155
155	155	155	157	156	159	152	158	156
156	153	157	156	153	155	154	155	157
159	159	156	158	156	159	157	161	162
158	155	158	154	156	160	162	155	159

中央値=158

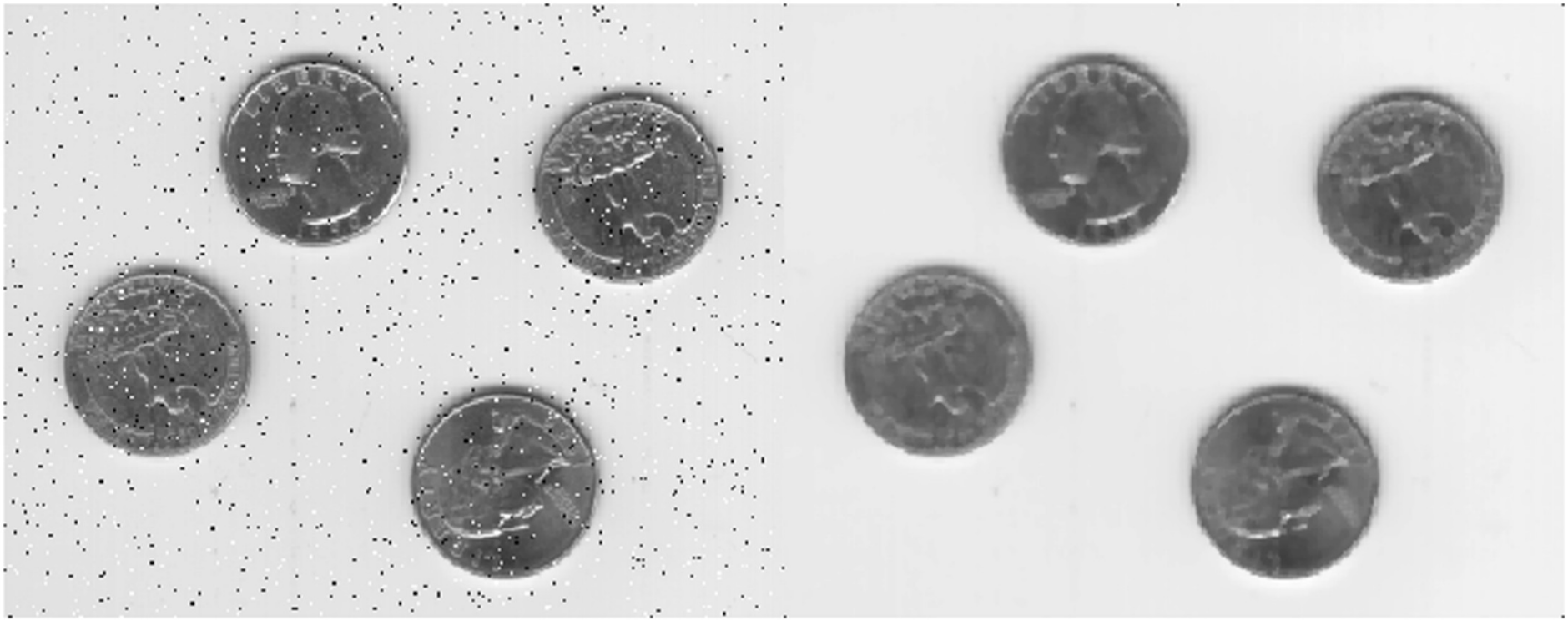
0	157	157	157	156	157	158	158	158
156	158	158	158	157	158	158	158	158
156	158	158	157	158	158	158	158	158
154	156	157	158	158	158	158	158	158
154	155	155	157	158	157	158	156	158
153	155	155	156	156	155	155	155	156
155	156	156	156	156	156	157	157	157
155	157	156	156	156	156	157	157	157
155	157	157	157	158	159	160	159	160

"zeros" (既定)	イメージを 0 でパディングします。
"symmetric"	境界でイメージを対称的に拡張します。

© 2020 Yuki Kato. All Rights Reserved

メディアンフィルタ

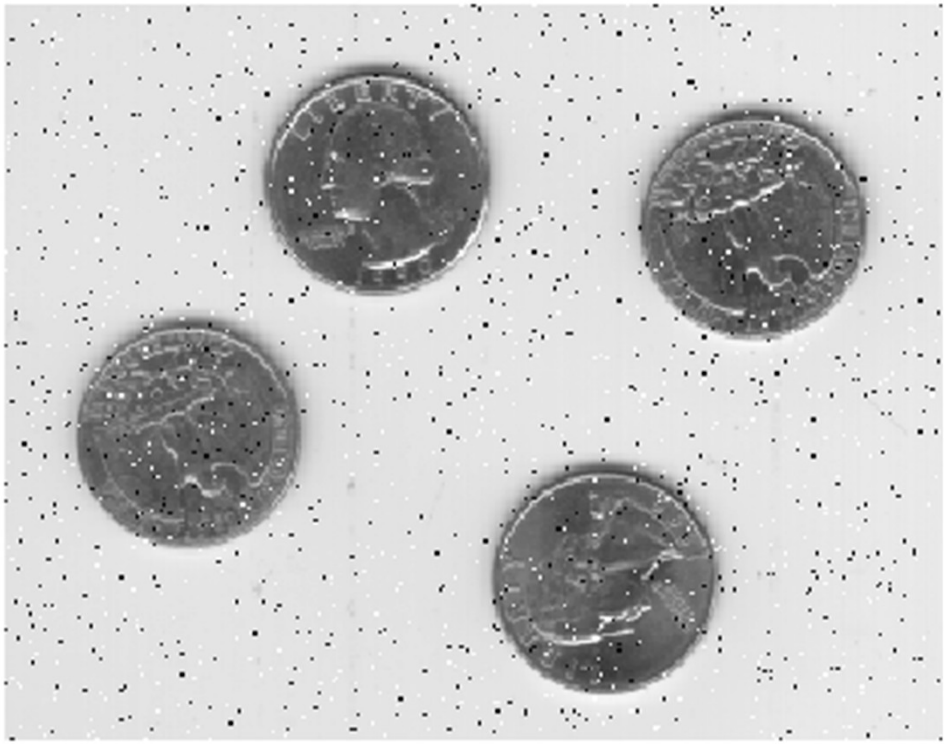
■ 演習



© 2020 Yuki Kato. All Rights Reserved

同じ画像にガウシアンフィルタをかけてみよう

■ 演習



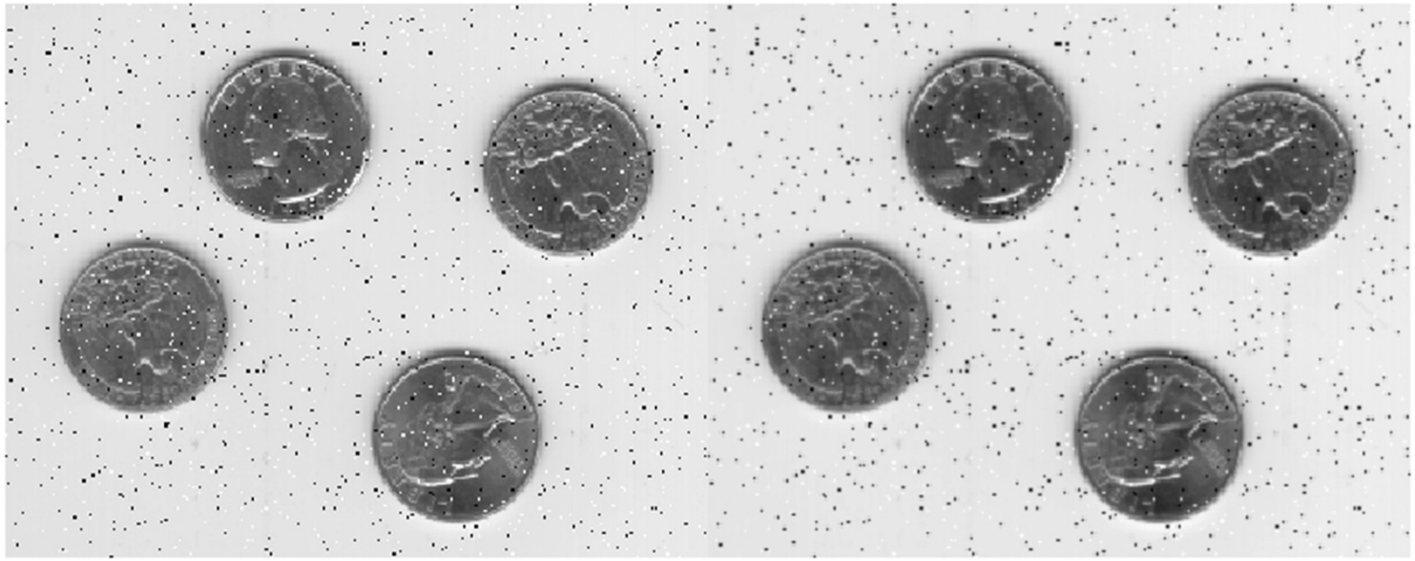
?

© 2020 Yuki Kato. All Rights Reserved

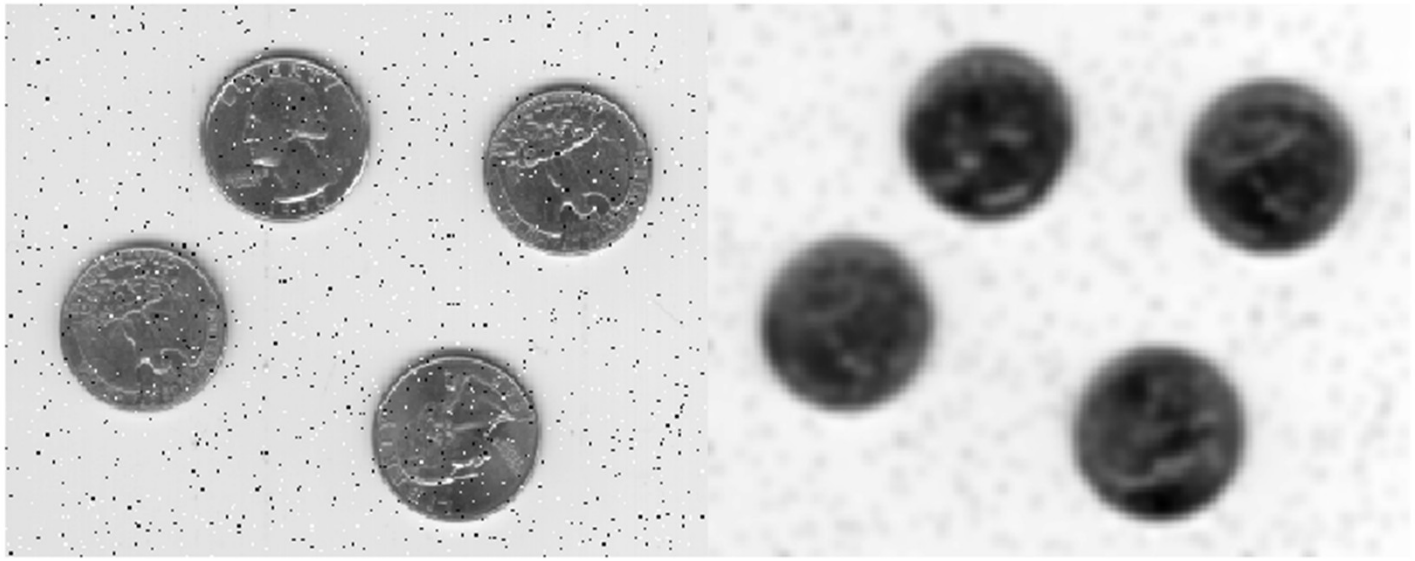
同じ画像にガウシアンフィルタをかけてみよう

■ 演習

σ=0.5



σ=2.5



© 2020 Yuki Kato. All Rights Reserved

移動平均と移動中央値

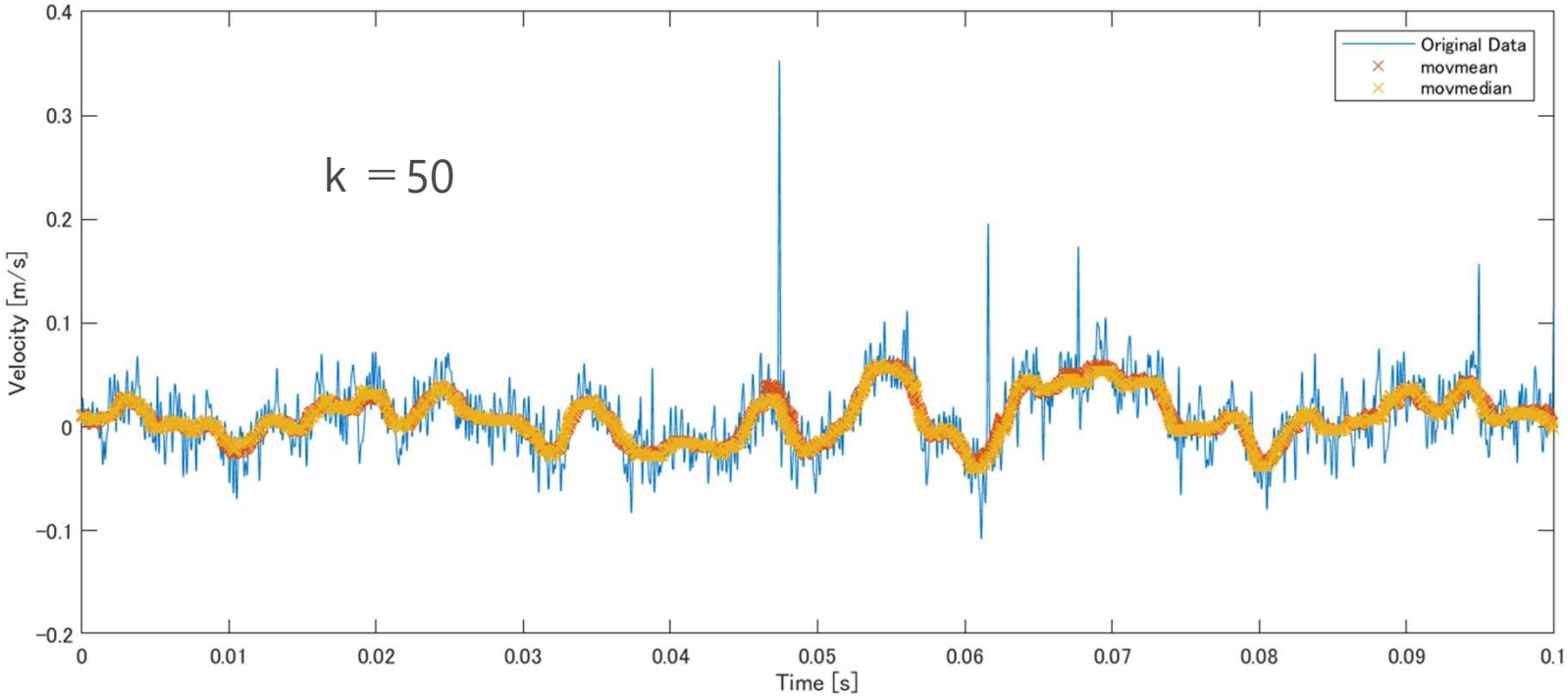
■ 移動平均値

$$z_t = \frac{x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}}{2k+1}$$

移動中央値

$$z_t = median\left(x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}\right)$$

■ 計測したデータの高周波ノイズを減算できるが，変動を過小評価してしまう・・・分散も小さくなる

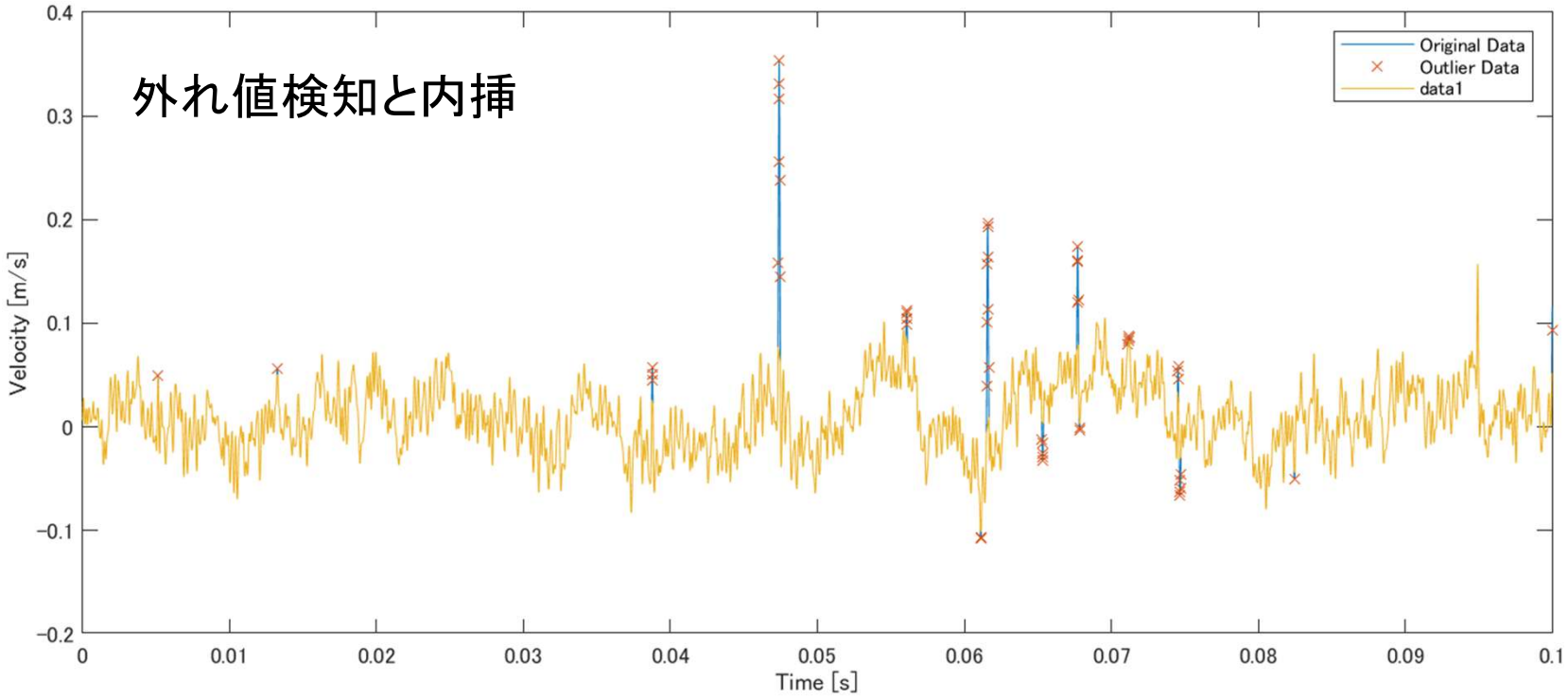


© 2020 Yuki Kato. All Rights Reserved

移動中央値を用いた外れ値検知と補完

■ 移動中央値と標準偏差を計算し，その3倍以上離れているデータを線形補完する

$$z_t = median\left(x_{t-k} + x_{t-k+1} \ldots x_{t+k-1} + x_{t+k}\right)$$



© 2020 Yuki Kato. All Rights Reserved

画像計測

- 最近の工学分野では，画像を用いた物理量の計測がよく用いられている
- ここでは，
  - ✓ 流体計測に用いられるPIV(Particle Image Velocimetry)
  - ✓ 構造物の振動計測に用いられるDIC(Digital Image Correlation)

  を解説する

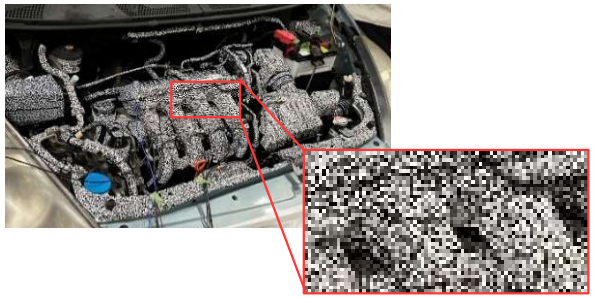
© 2020 Yuki Kato. All Rights Reserved

画像を用いた振動（変位）計測法

✓ 画像振動計測→詳細なモードやひずみ/応力を計測可能

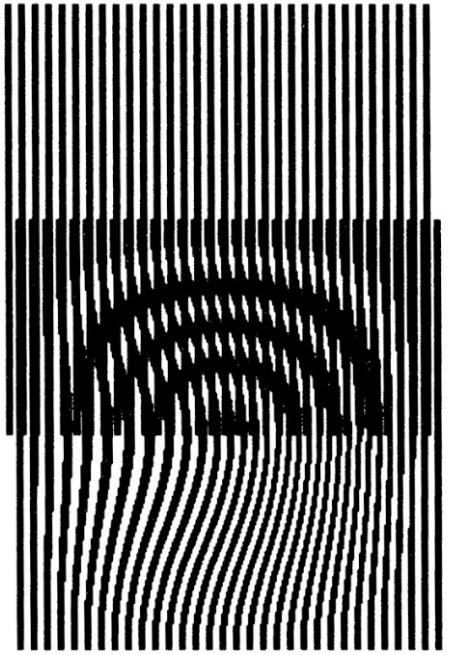
■ デジタル画像相関法（DIC）

- 撮影した画像のドットパターンを追跡し、変位/ひずみを計測
- 面全体の変形・振動を高分解能で取得可能. 2D-DICと3D-DIC（ステレオカメラ使用）がある. 曲面の計測が可能



■ サンプリングモアレ法

- 格子パターンを利用し、振動による変位を干渉縞の変化として可視化
- サブミクロン級の変位検出が可能. 特に微小振動に有効. 平面が得意



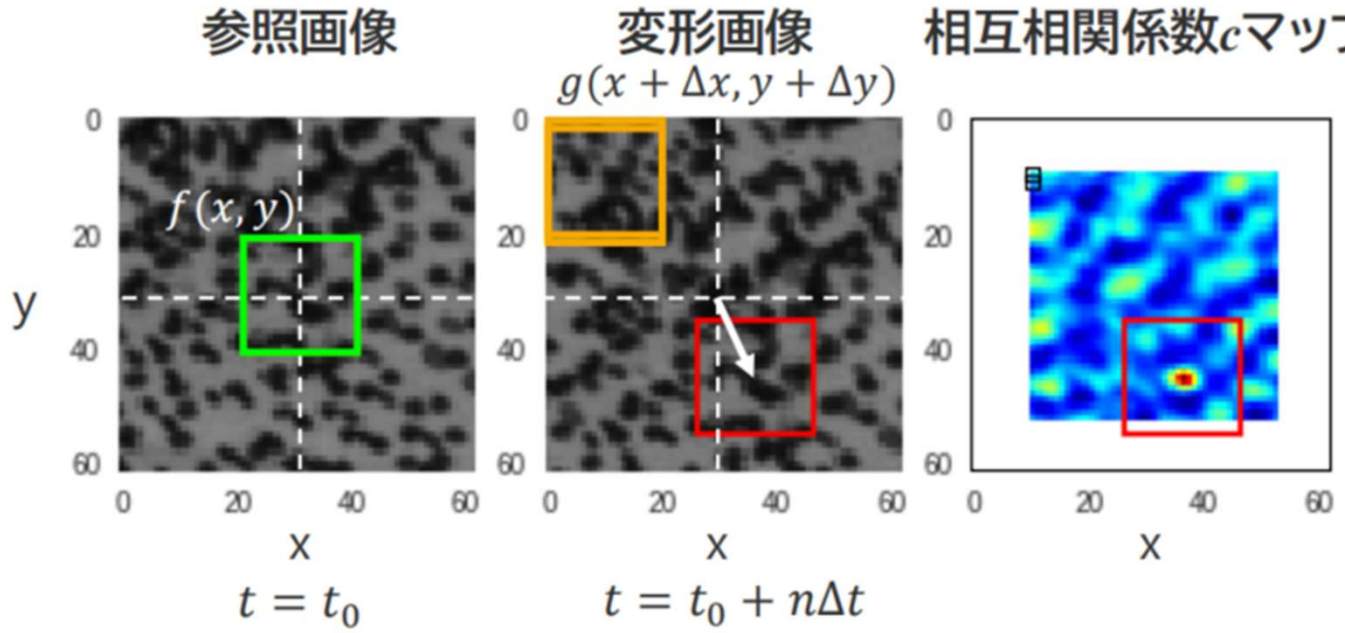
■ その他（ホログラフィ干渉法, スペックル散乱・・）

Ref: JSME機械工学辞典, モアレ法

© 2020 Yuki Kato. All Rights Reserved

DICのメカニズム

- ランダムドットを追跡．1/10~1/50pix程度の分解能
- 商用ソフト：アラミス，VIC-DIC，Dipp-strain



※フリーだと，
2D: Ncorr
3D: Duo-DIC

- 変形画像中のファセット$g(x,y)$を走査し、参照画像中の格子点上のファセット$f(x,y)$との相互相関係数C分布を算出
- 相互相関係数Cの最大位置を取得

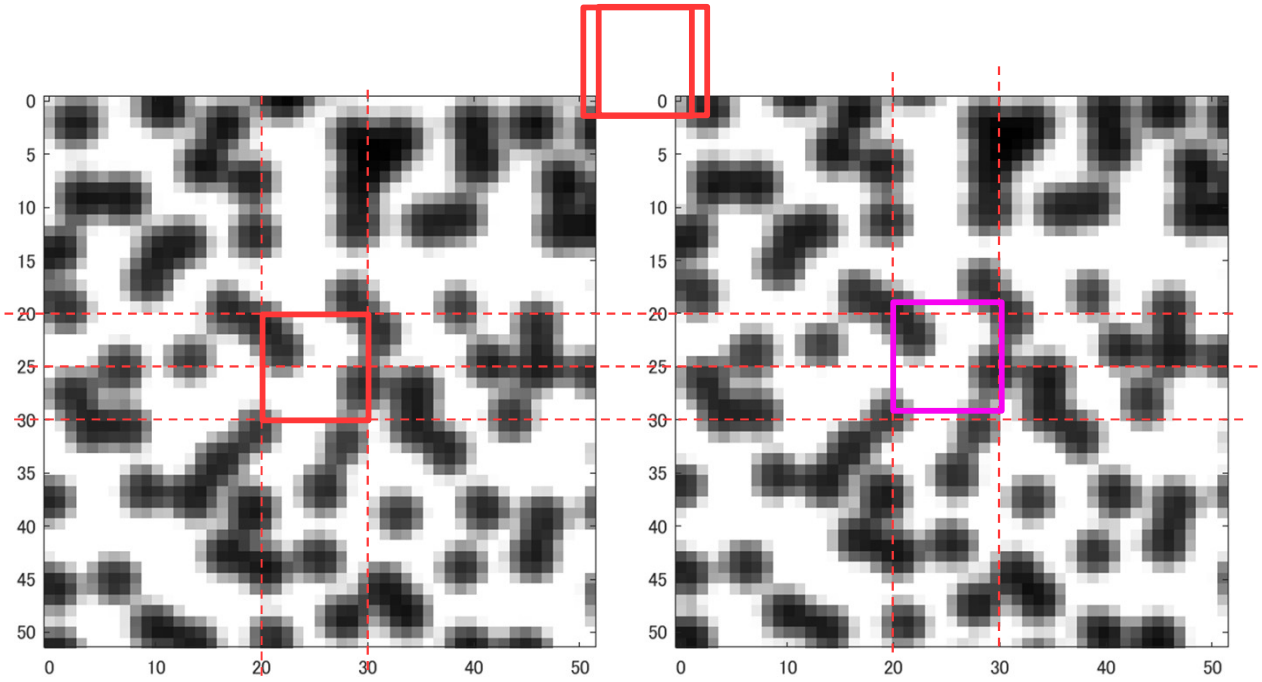
相互相関関数　$$c(\Delta x, \Delta y) = \frac{\langle f(x,y), g(x+\Delta x, y+\Delta y)\rangle}{|f(x,y)| \cdot |g(x,y)|}$$

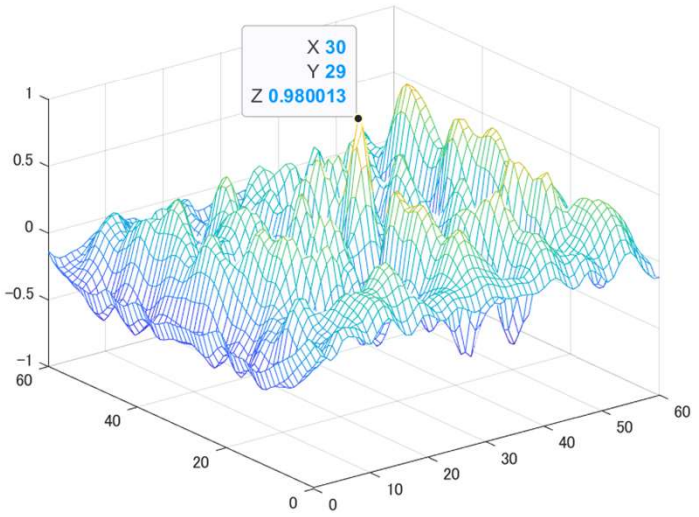
M. A. Sutton., et al., (2009).

© 2020 Yuki Kato. All Rights Reserved

DIC-演習-相互相関関数

■ 下記の画像について，相互相関関数により変位量を計算せよ





相互相関関数　$$c(\Delta x, \Delta y) = \frac{\langle f(x,y), g(x+\Delta x, y+\Delta y)\rangle}{|f(x,y)| \cdot |g(x,y)|}$$

※最近はZNCC（**ゼロ平均正規化相互相関**）を使うことが多い

© 2020 Yuki Kato. All Rights Reserved

サブピクセル処理

- ピクセル単位ではなく，サブピクセル単位で追跡
- 下記簡単な補完では最大で1/10の精度が出る（経験則）

ガウス分布

$$P_{sub} = i - \frac{1}{2}\frac{\ln R_{i+1} - \ln R_{i-1}}{\ln R_{i+1} - 2\ln R_i + \ln R_{i-1}}$$

2 次曲線

$$P_{sub} = i - \frac{1}{2}\frac{R_{i+1} - R_{i-1}}{R_{i+1} - 2R_i + R_{i-1}}$$

重心

$$P_{sub} = \frac{(i+1)R_{i+1} + iR_i + (i-1)R_{i-1}}{R_{i+1} + R_i + R_{i-1}}$$

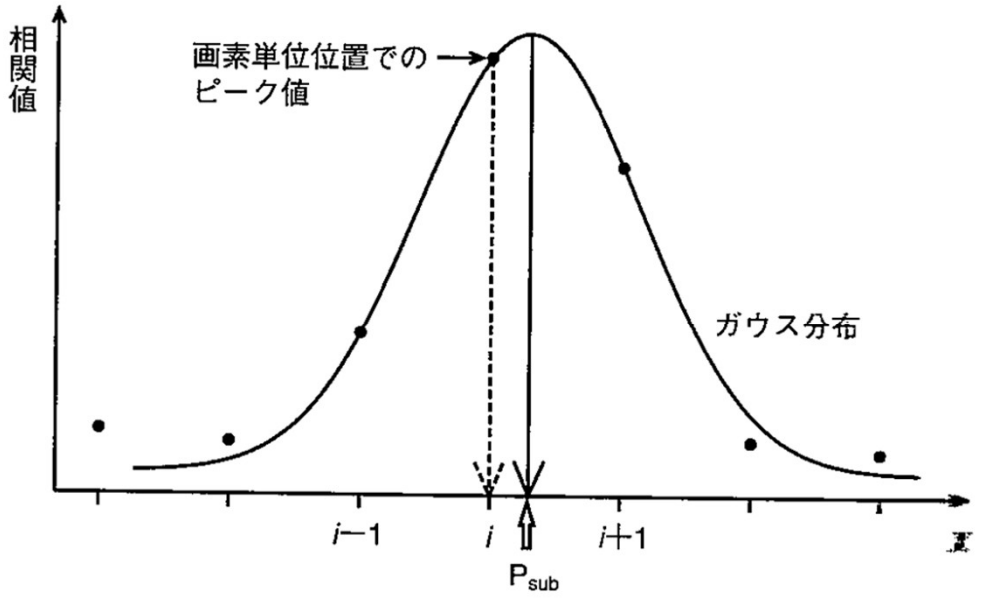


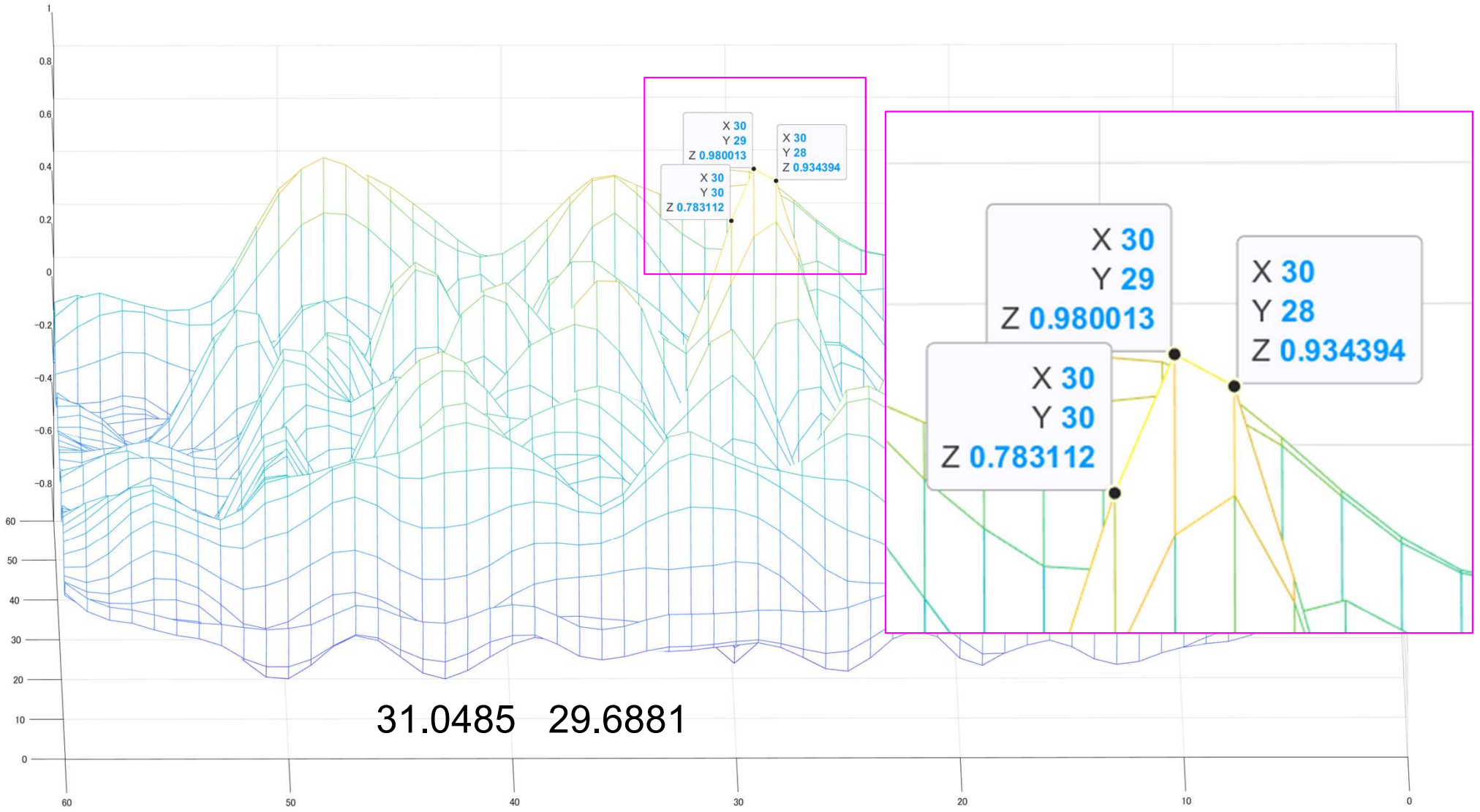
図 4.7　サブピクセル精度（ガウス分布の場合）

Ref: PIVハンドブック

© 2020 Yuki Kato. All Rights Reserved

サブピクセル処理

- ピクセル単位ではなく，サブピクセル単位で追跡
- 下記簡単な補完では最大で1/10の精度が出る（経験則）



© 2020 Yuki Kato. All Rights Reserved

演習

■ ガウス分布と2次曲線による補完について，式変形して導出しましょう．

R_{i-1}, R_i, R_{i+1}
ax^2+bx+c
exp(-x^2/σ^2)

ガウス分布 $$P_{sub} = i - \frac{1}{2}\frac{\ln R_{i+1} - \ln R_{i-1}}{\ln R_{i+1} - 2\ln R_i + \ln R_{i-1}}$$

2 次曲線 $$P_{sub} = i - \frac{1}{2}\frac{R_{i+1} - R_{i-1}}{R_{i+1} - 2R_i + R_{i-1}}$$

重心 $$P_{sub} = \frac{(i+1)R_{i+1} + iR_i + (i-1)R_{i-1}}{R_{i+1} + R_i + R_{i-1}}$$

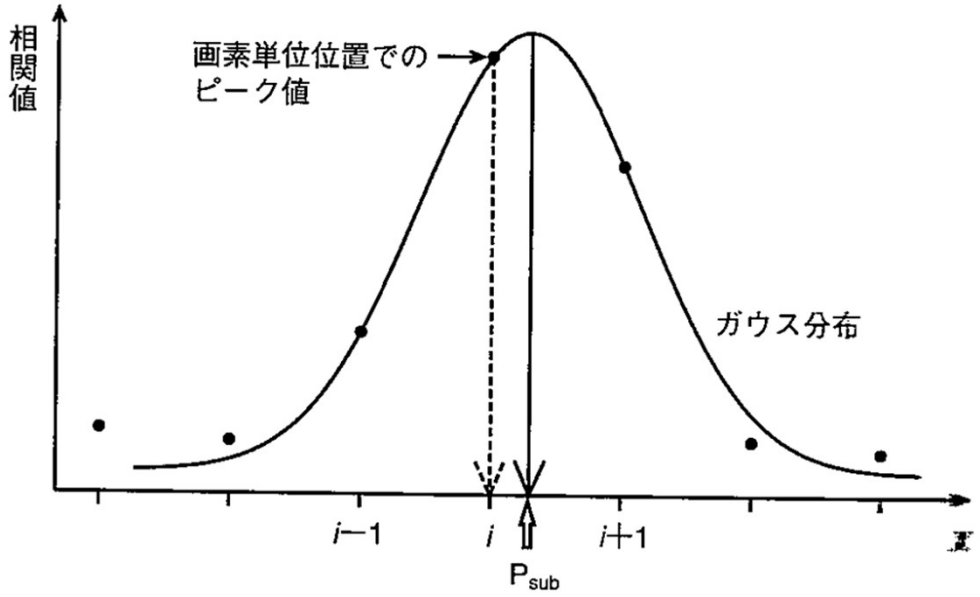


図 4.7　サブピクセル精度（ガウス分布の場合）

Ref: PIVハンドブック

© 2020 Yuki Kato. All Rights Reserved

2次曲線

ガウス分布 $P_{sub} = i - \frac{1}{2}\frac{\ln R_{i+1} - \ln R_{i-1}}{\ln R_{i+1} - 2\ln R_i + \ln R_{i-1}}$

2 次曲線 $P_{sub} = i - \frac{1}{2}\frac{R_{i+1} - R_{i-1}}{R_{i+1} - 2R_i + R_{i-1}}$

モデル

ピーク近傍の相関を**直接二次式**で近似：

$$R(i) = ai^2 + bi + c.$$

同様に局所座標 $k \in \{-1, 0, +1\}$ を使って

$$R(k) = ak^2 + bk + c, \qquad R_{-1} = R(-1),\ R_0 = R(0),\ R_{+1} = R(+1).$$

3 点から係数を求める

$$\begin{aligned} R(-1) &= a - b + c = R_{-1}, \\ R(0) &= c = R_0, \\ R(+1) &= a + b + c = R_{+1}. \end{aligned}$$

したがって

$$a = \frac{R_{-1} - 2R_0 + R_{+1}}{2}, \qquad b = \frac{R_{+1} - R_{-1}}{2}, \qquad c = R_0.$$

頂点（ピーク）位置

放物線 $R(k) = ak^2 + bk + c$ の極大は

$$k^{\backslash *} = -\frac{b}{2a} = -\frac{(R_{+1} - R_{-1})/2}{(R_{-1} - 2R_0 + R_{+1})} = \frac{R_{-1} - R_{+1}}{2\,(R_{-1} - 2R_0 + R_{+1})}.$$

よって

$$\boxed{\delta_{\mathrm{Quad}} = \frac{R_{-1} - R_{+1}}{2\big(R_{-1} - 2R_0 + R_{+1}\big)}}$$

© 2020 Yuki Kato. All Rights Reserved

ガウス分布

ガウス分布　$P_{sub} = i - \frac{1}{2}\frac{\ln R_{i+1} - \ln R_{i-1}}{\ln R_{i+1} - 2\ln R_i + \ln R_{i-1}}$

2 次曲線　$P_{sub} = i - \frac{1}{2}\frac{R_{i+1} - R_{i-1}}{R_{i+1} - 2R_i + R_{i-1}}$

モデル

ピーク近傍の相関を**（対数を取ると二次）**なガウス形とみなす：

$$R(i) \;=\; A\exp\!\Big(-\frac{(i-x^{\backslash *})^2}{2\sigma^2}\Big), \qquad A>0.$$

両辺の対数を取ると

$$\ell(i) \;:=\; \ln R(i) \;=\; \ln A - \frac{(i-x^{\backslash *})^2}{2\sigma^2} \;=\; \underbrace{-\frac{1}{2\sigma^2}}_{\alpha} i^2 + \underbrace{\frac{x^{\backslash *}}{\sigma^2}}_{\beta} i + \underbrace{\ln A - \frac{x^{\backslash *2}}{2\sigma^2}}_{\gamma}.$$

すなわち $\ell(i)$ は i の**二次式** $\alpha i^2 + \beta i + \gamma$ です。

整数ピーク周りの 3 点

$$\ell_{-1} := \ln R_{-1} = \ell(i_0 - 1), \quad \ell_0 := \ln R_0 = \ell(i_0), \quad \ell_{+1} := \ln R_{+1} = \ell(i_0 + 1)$$

を、原点を i_0 に合わせた局所座標 $k \in \{-1, 0, +1\}$ で表すと

$$\ell(k) = \alpha k^2 + \beta k + \gamma.$$

© 2020 Yuki Kato. All Rights Reserved

ガウス分布2

ガウス分布　$P_{\text{sub}} = i - \frac{1}{2}\frac{\ln R_{i+1} - \ln R_{i-1}}{\ln R_{i+1} - 2\ln R_i + \ln R_{i-1}}$

2 次曲線　$P_{\text{sub}} = i - \frac{1}{2}\frac{R_{i+1} - R_{i-1}}{R_{i+1} - 2R_i + R_{i-1}}$

すなわち $\ell(i)$ は i の**二次式** $\alpha i^2 + \beta i + \gamma$ です。

整数ピーク周りの 3 点

$$\ell_{-1} := \ln R_{-1} = \ell(i_0 - 1), \quad \ell_0 := \ln R_0 = \ell(i_0), \quad \ell_{+1} := \ln R_{+1} = \ell(i_0 + 1)$$

を、原点を i_0 に合わせた局所座標 $k \in \{-1, 0, +1\}$ で表すと

$$\ell(k) = \alpha k^2 + \beta k + \gamma.$$

3 点から係数を求める

$$\begin{aligned} \ell(-1) &= \alpha - \beta + \gamma = \ell_{-1}, \\ \ell(0) &= \gamma = \ell_0, \\ \ell(+1) &= \alpha + \beta + \gamma = \ell_{+1}. \end{aligned}$$

よって

$$\alpha = \frac{\ell_{-1} - 2\ell_0 + \ell_{+1}}{2}, \qquad \beta = \frac{\ell_{+1} - \ell_{-1}}{2}, \qquad \gamma = \ell_0.$$

頂点（ピーク）位置

二次式 $\ell(k) = \alpha k^2 + \beta k + \gamma$ の極大は

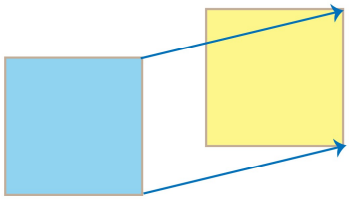
$$k^{\backslash *} = -\frac{\beta}{2\alpha} = -\frac{(\ell_{+1} - \ell_{-1})/2}{(\ell_{-1} - 2\ell_0 + \ell_{+1})} = \frac{\ell_{-1} - \ell_{+1}}{2(\ell_{-1} - 2\ell_0 + \ell_{+1})}.$$

これは局所座標でのサブピクセル量なので $\delta = k^{\backslash *}$。したがって

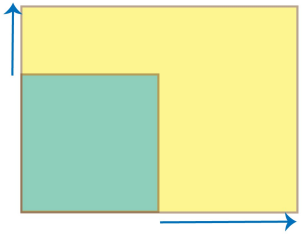
$$\boxed{\delta_{\text{Gauss}} = \frac{\ln R_{-1} - \ln R_{+1}}{2\big(\ln R_{-1} - 2\ln R_0 + \ln R_{+1}\big)}}$$

© 2020 Yuki Kato. All Rights Reserved

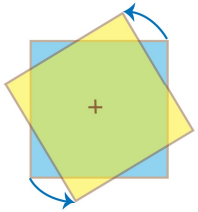
アフィン変換を用いたサブピクセル処理



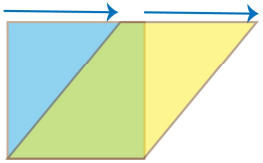
平行移動



拡大縮小 (スケーリング)



回転



シアー (せん断) or スキュー

画像変換名	アフィン行列
平行移動(Translation)	$\begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}$
回転 (Rotation)	$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$
スケーリング (Scaling)	$\begin{pmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{pmatrix}$
シアー (Shear)	$\begin{pmatrix} 1 & sh_x & 0 \\ sh_y & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

変換後の座標　拡大・縮小 回転 剪断変形　変換前の座標　平行移動

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

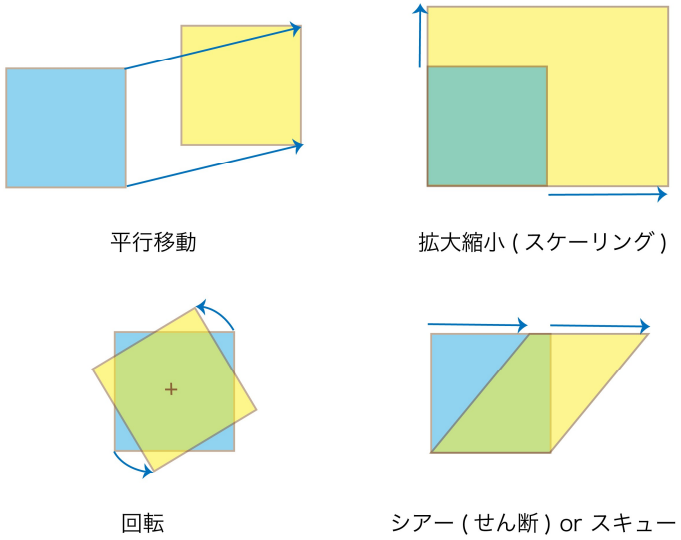
- 上記6つの未知数を最適化により決定

Ref:https://cvml-expertguide.net/terms/affine-transformation/

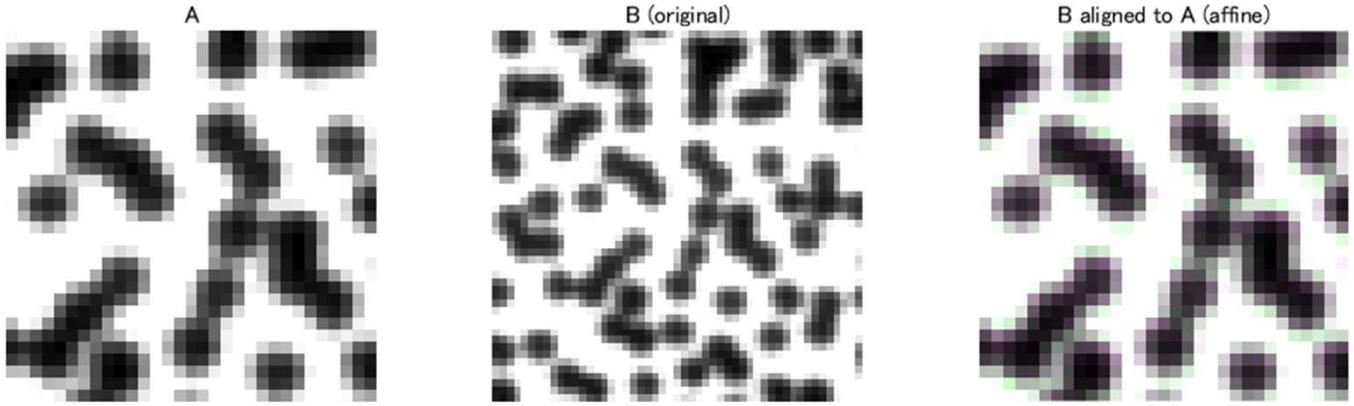
© 2020 Yuki Kato. All Rights Reserved

アフィン変換を用いたサブピクセル処理

■ 精度の高い追跡が可能（PIVにはあまり使われない）



XCorr 初期推定 (B->A): dx=0.003, dy=-1.324
--- Affine params (approx) ---
Translation: tx=-0.0337 px, ty=1.3282 px
Rotation : 179.9603 deg
Scale : sx=-1.001700, sy=1.002288
Shear : 179.9735 deg



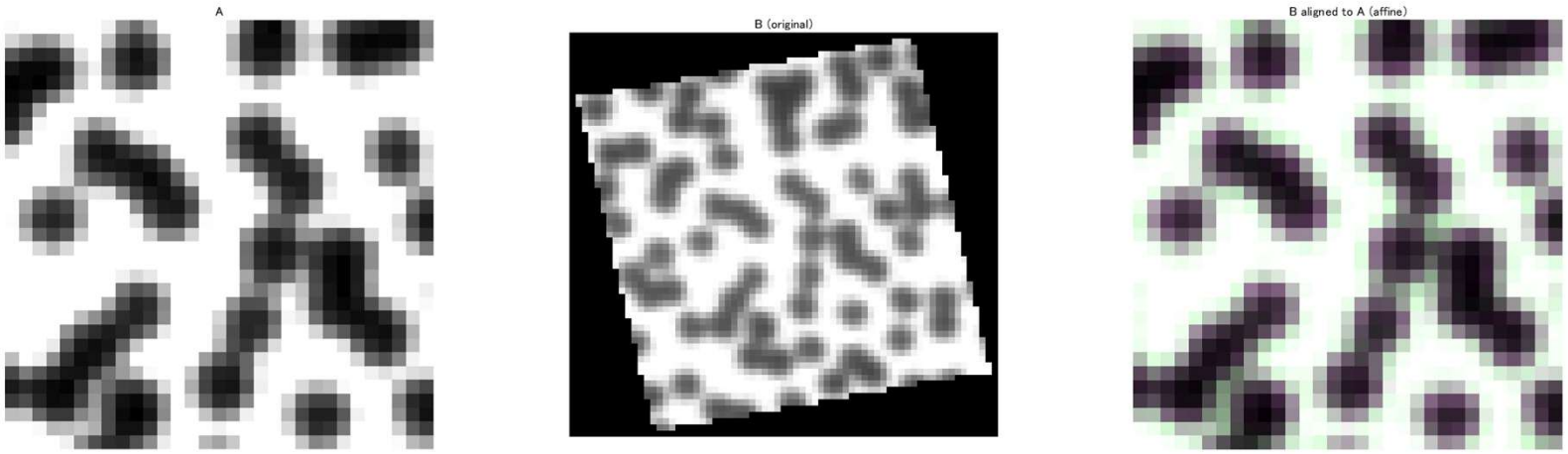
Ref:https://cvml-expertguide.net/terms/affine-transformation/

© 2020 Yuki Kato. All Rights Reserved

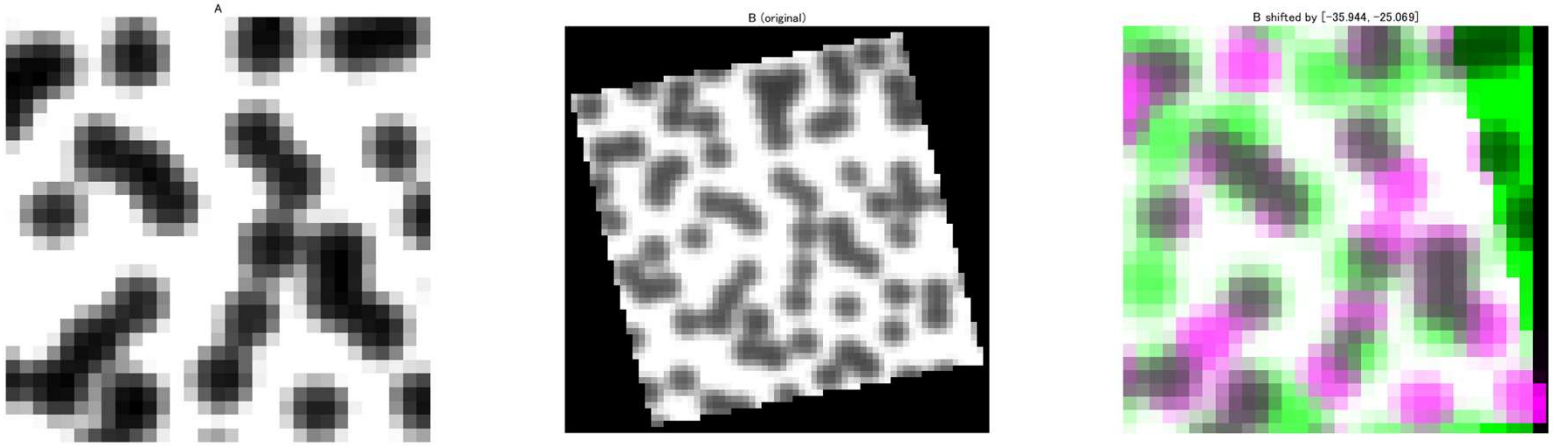
大変形の例

■ 変形や回転，拡大を含む場合も，精度の高い追跡が可能

アフィン変換DIC



通常DIC



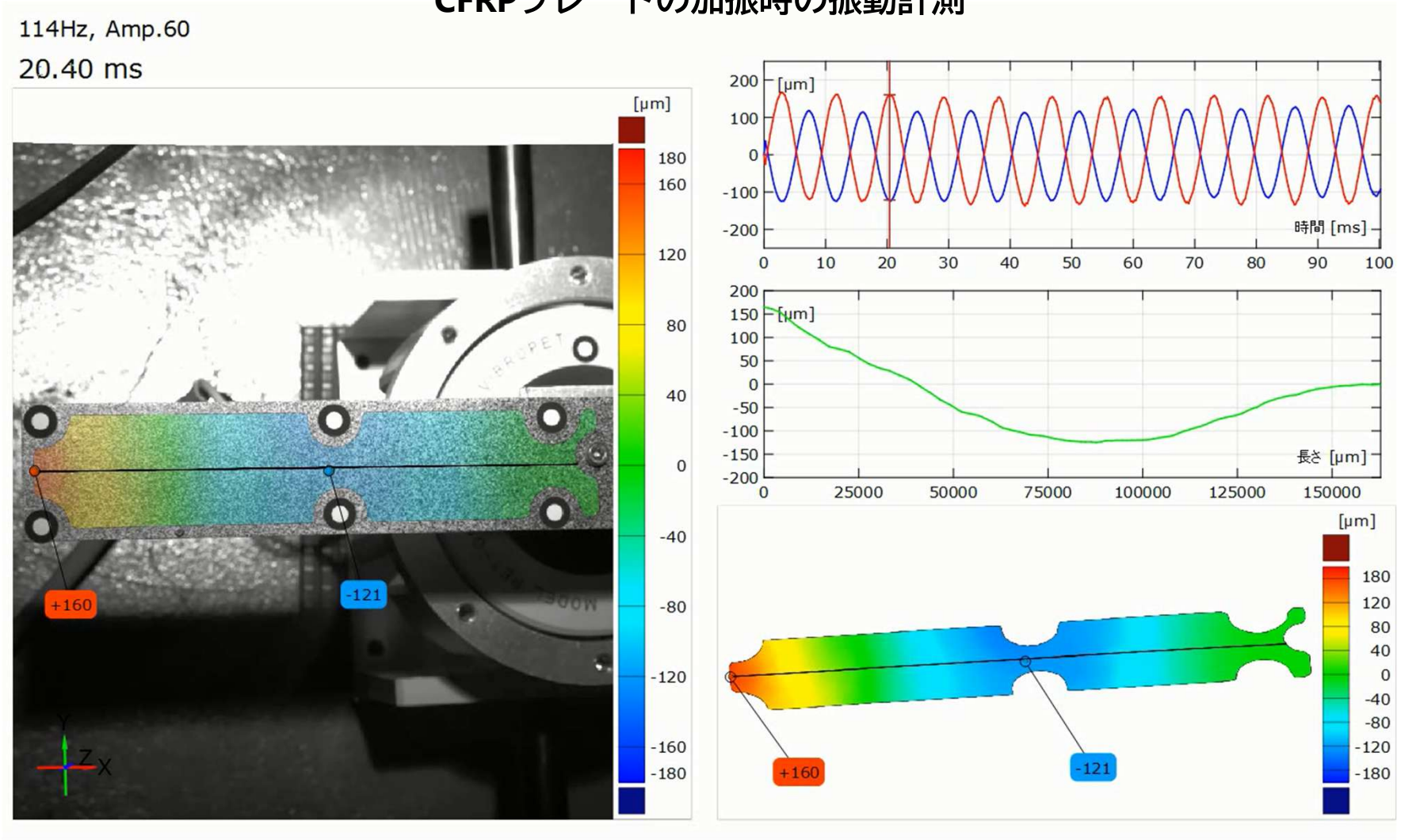
※Ncorr はZNCC+アフィン変換

© 2020 Yuki Kato. All Rights Reserved

DICによる計測例

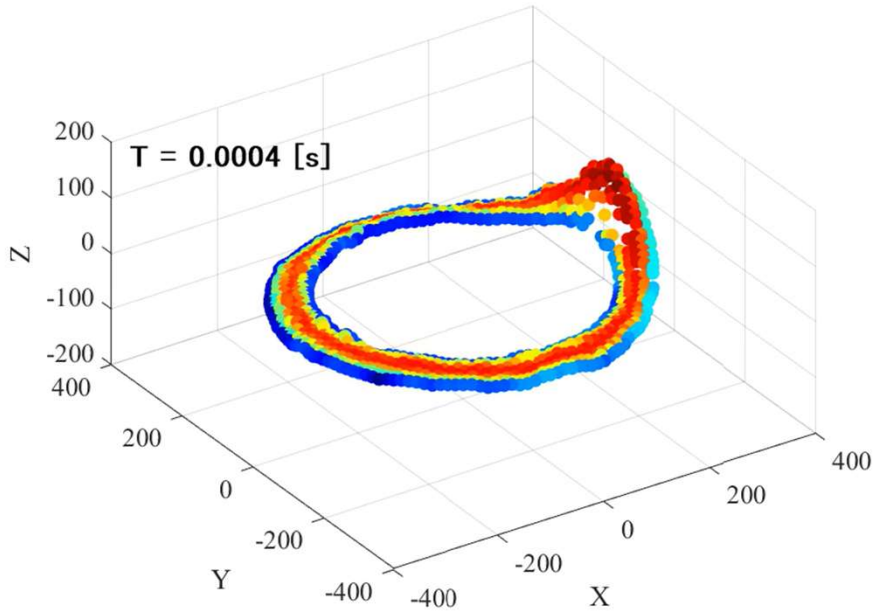
- ステレオカメラを用いれば，三次元変位を計測可能
- ひずみや応力の評価も可能→疲労試験などにも応用

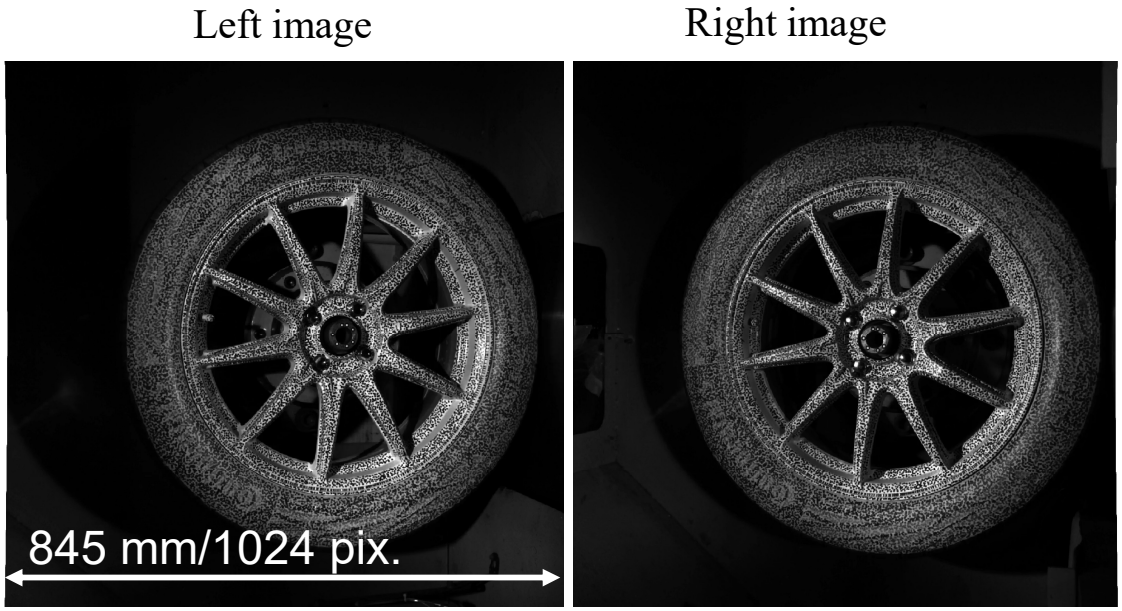
CFRPプレートの加振時の振動計測



© 2020 Yuki Kato. All Rights Reserved

タイヤ計測





2500 fps

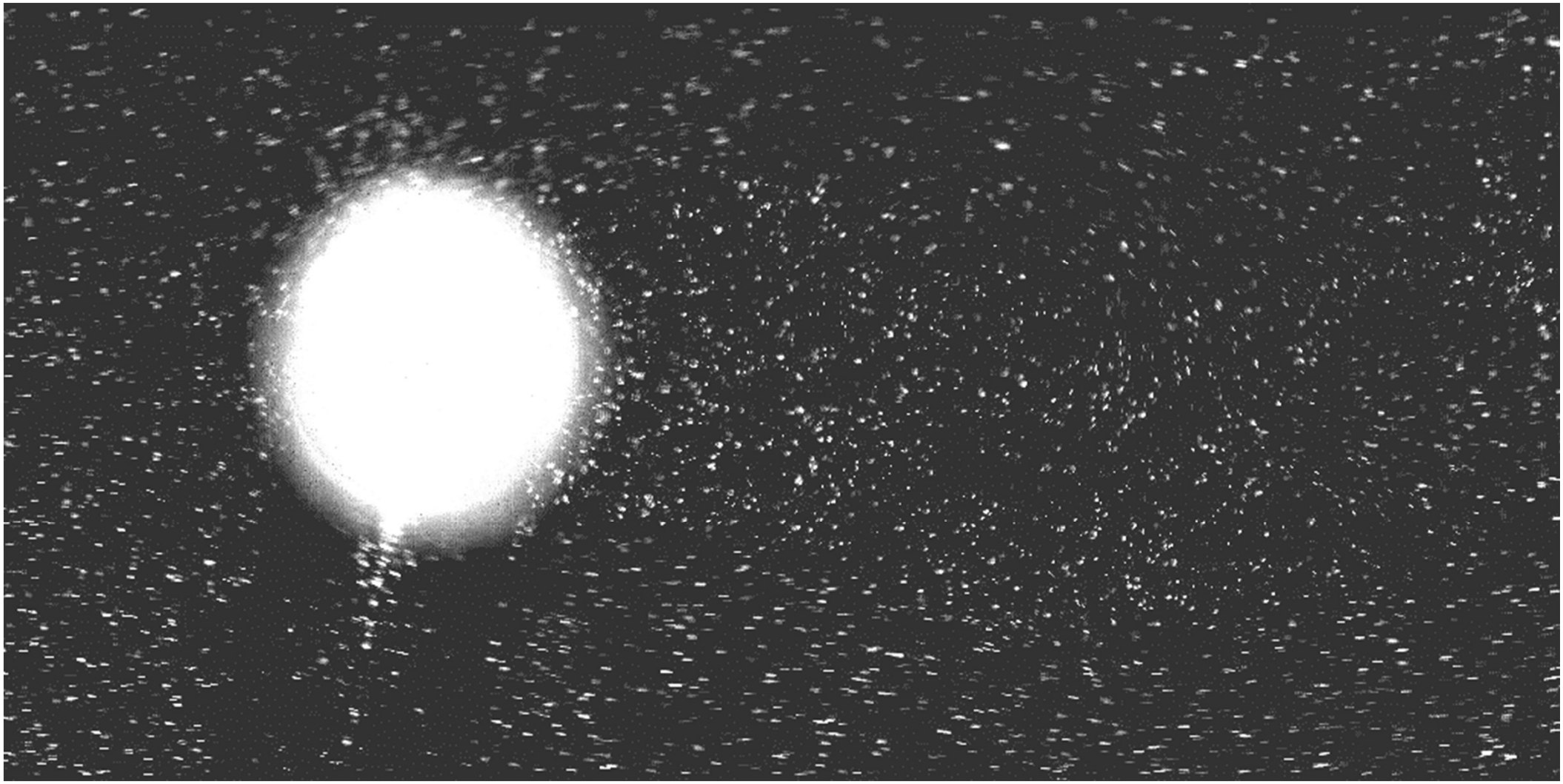
© 2020 Yuki Kato. All Rights Reserved

PIVとは？

- PIV（Particle Image Velocimetry：粒子画像流速測定）は、**流れの速度場を“面”で非接触に測る**光学計測法です。
- しくみ
  - 流体に直径数μm程度のトレーサ粒子を混ぜる（流れに忠実に動く粒子）。
  - レーザーシートで測定平面を照明し、**ごく短い時間差 Δt** で2枚の画像を撮る。
  - 画像を小領域（インタロゲーションウィンドウ）に分け、**相互相関**で粒子模様の平均移動量 Δx を求める。
  - **速度ベクトルは u = Δx / Δt**。これを面内に敷き詰めて2Dの速度場が得られる。
- 特長
  - 非接触・**全視野（フィールド）**のベクトル場を一度に取得
  - 乱流やせん断層など**複雑な流れ構造**の可視化に強い
  - サブピクセル補間で変位精度はおおむね**0.1ピクセル程度**まで追い込める
- 注意点／誤差要因
  - 粒子の**追随性**（ストークス数が十分小さい粒子を選定）
  - 光学アクセス（レーザー安全、反射・屈折、壁面近傍の測定誤差）
  - 粒子濃度、露光時間、Δt、ウィンドウサイズなどの**パラメータ最適化**
  - 高速流ではモーションブラー、低速流では相関ピークが弱くなる

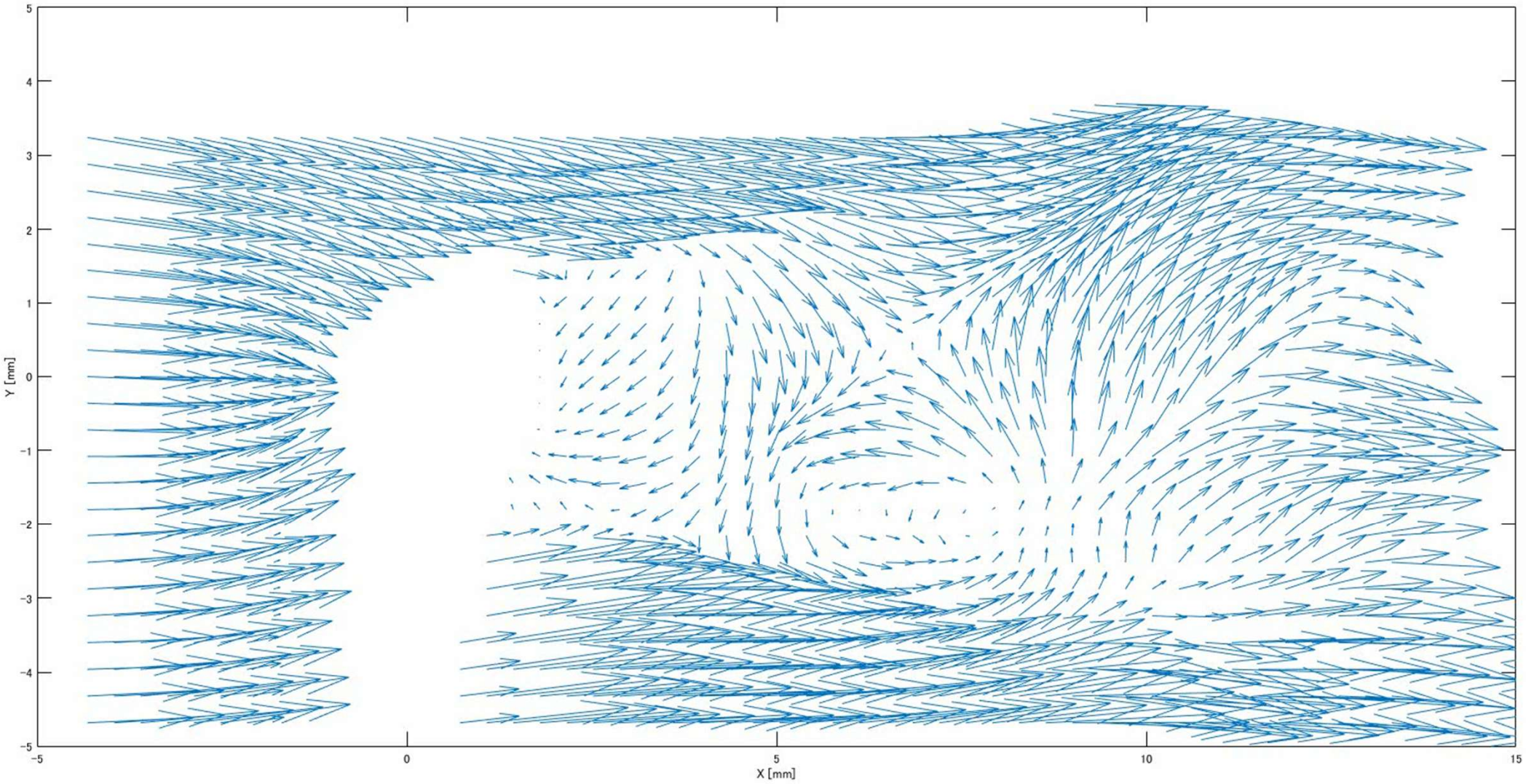
© 2020 Yuki Kato. All Rights Reserved

PIV試験の例



© 2020 Yuki Kato. All Rights Reserved

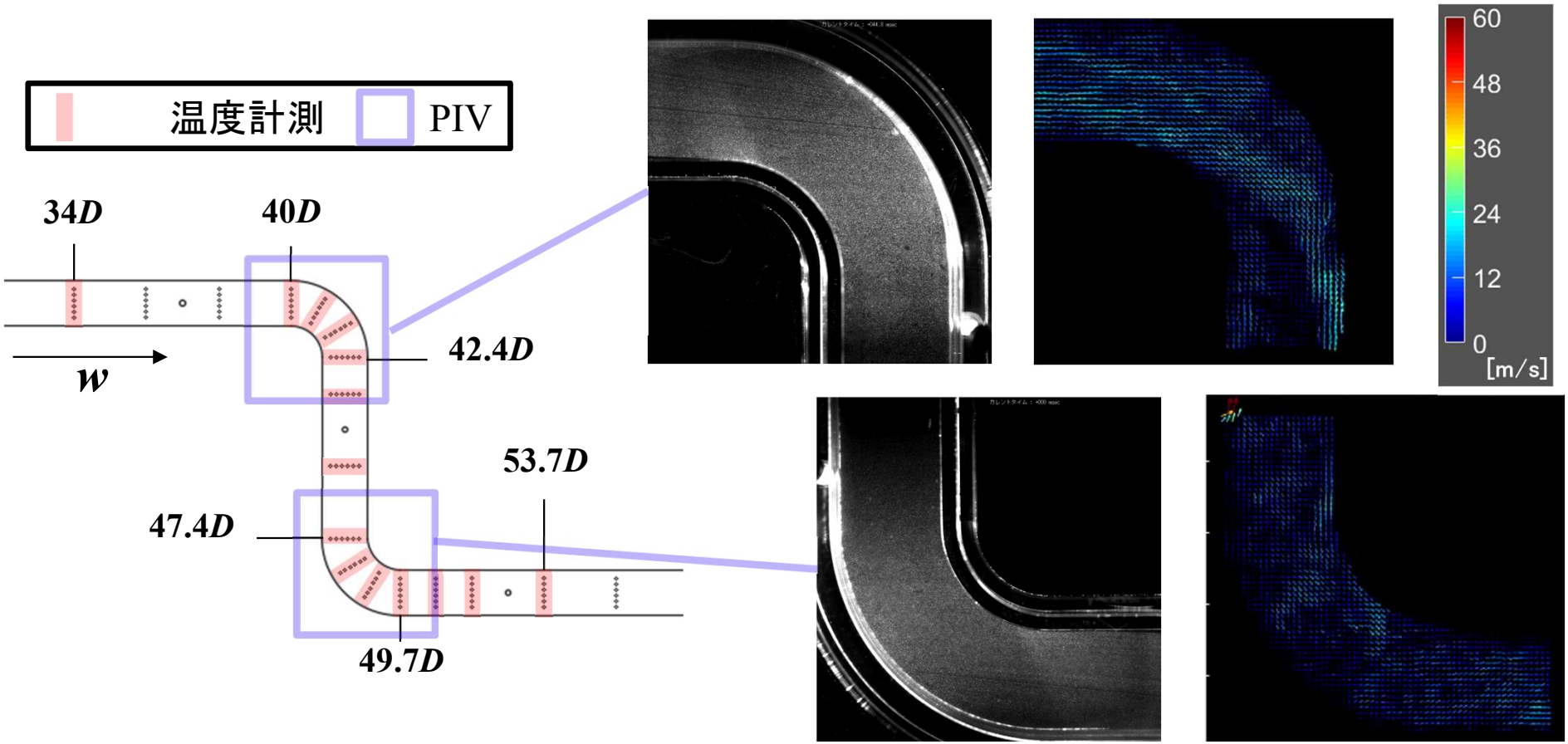
PIV試験の例



© 2020 Yuki Kato. All Rights Reserved

PIVの例（自動車配管）

15Hz



乱れの強さや旋回渦の大きさ，脈動流れの位相などを評価

＜PIV＞

- 中心平面にレーザーを照射　Re：60000
- 34D 温度80℃で計測　f：定常~90Hz

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第九回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布
- 第５回　大数の法則と中心極限定理
- 第６回　時系列データ解析とフィルタ処理1
- 第７回　時系列データ解析とフィルタ処理2，画像解析　✓課題①，②出題

10/28 ■ 第８回　画像解析

10/31 ■ 第９回　最小二乗法，回帰分析　数理モデル，サロゲートモデル構築へ

11/4 ■ 第１０回　最尤推定，ベイズ推定

11/14 ■ 第１１回　変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ

11/18 ■ 第１２回　スパースモデリング

11/21 ■ 第１３回　多変量データ解析の基礎と主成分分析　✓11/14: 課題①②〆，両方とも，講義開始時に集めます

11/25 ■ 第１４回　固有値分解，特異値分解　答案の配布と，課題の解説をどこかで実施予定

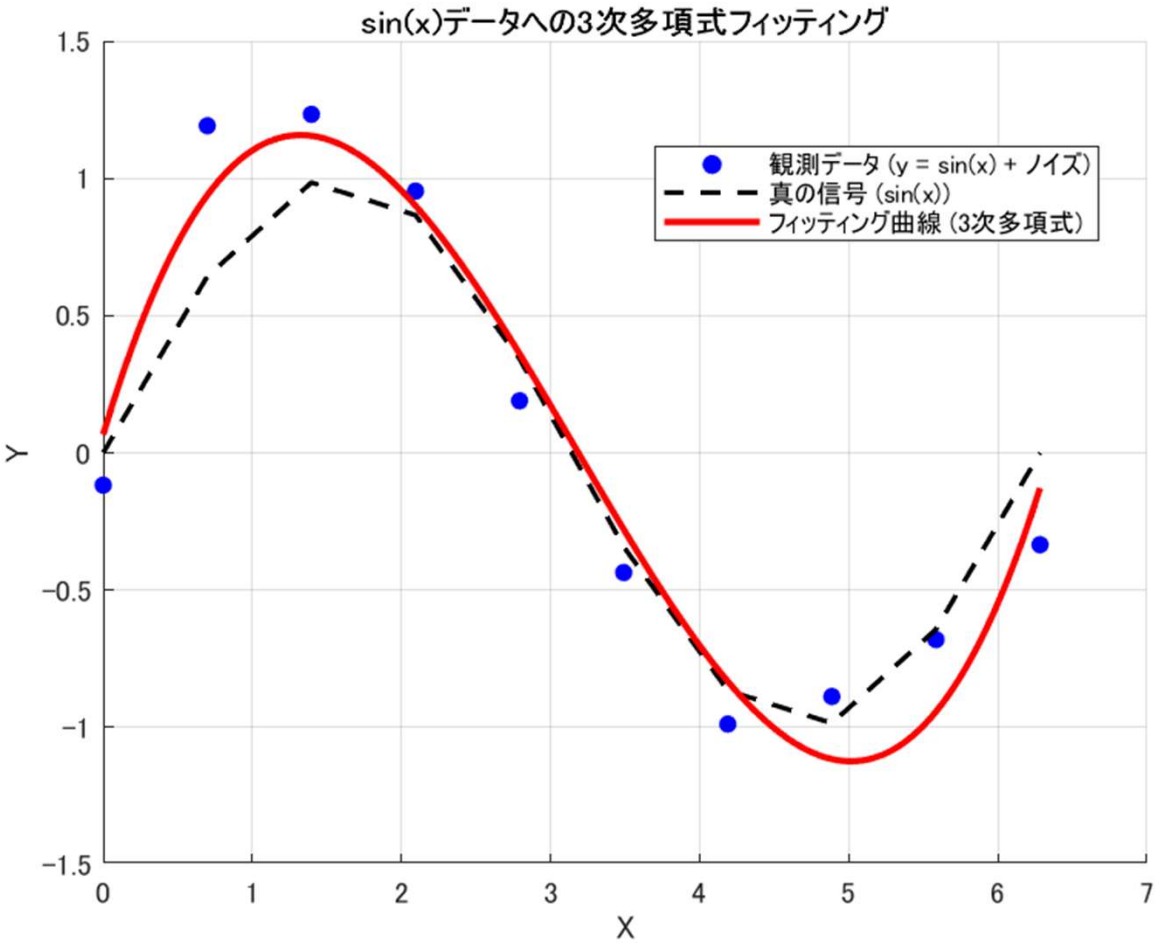
11/28 ■ 第１５回　期末テスト

12/2 ■ 第１６回　返却　⇒返却のみ，補講扱いですが，全員出席とするので任意参加

© 2020 Yuki Kato. All Rights Reserved

最小二乗法とは？

■ 観測データとモデルの差（残差）の二乗和が最小になるように、最適なモデルのパラメータを決定する手法



観測データ

$$\boldsymbol{y}_{obs} = [y_1, y_2 \cdots y_n]$$

$$\boldsymbol{x}_{obs} = [x_1, x_2 \cdots x_n]$$

モデル

$$y = f(x) = ax^3 + bx^2 + cx + d$$

二乗誤差

$$S = \{y_1 - f(x_1)\}^2 + \cdots \{y_n - f(x_n)\}^2$$

$$= \sum_{i=1}^{n} \{y_i - f(x_i)\}^2$$

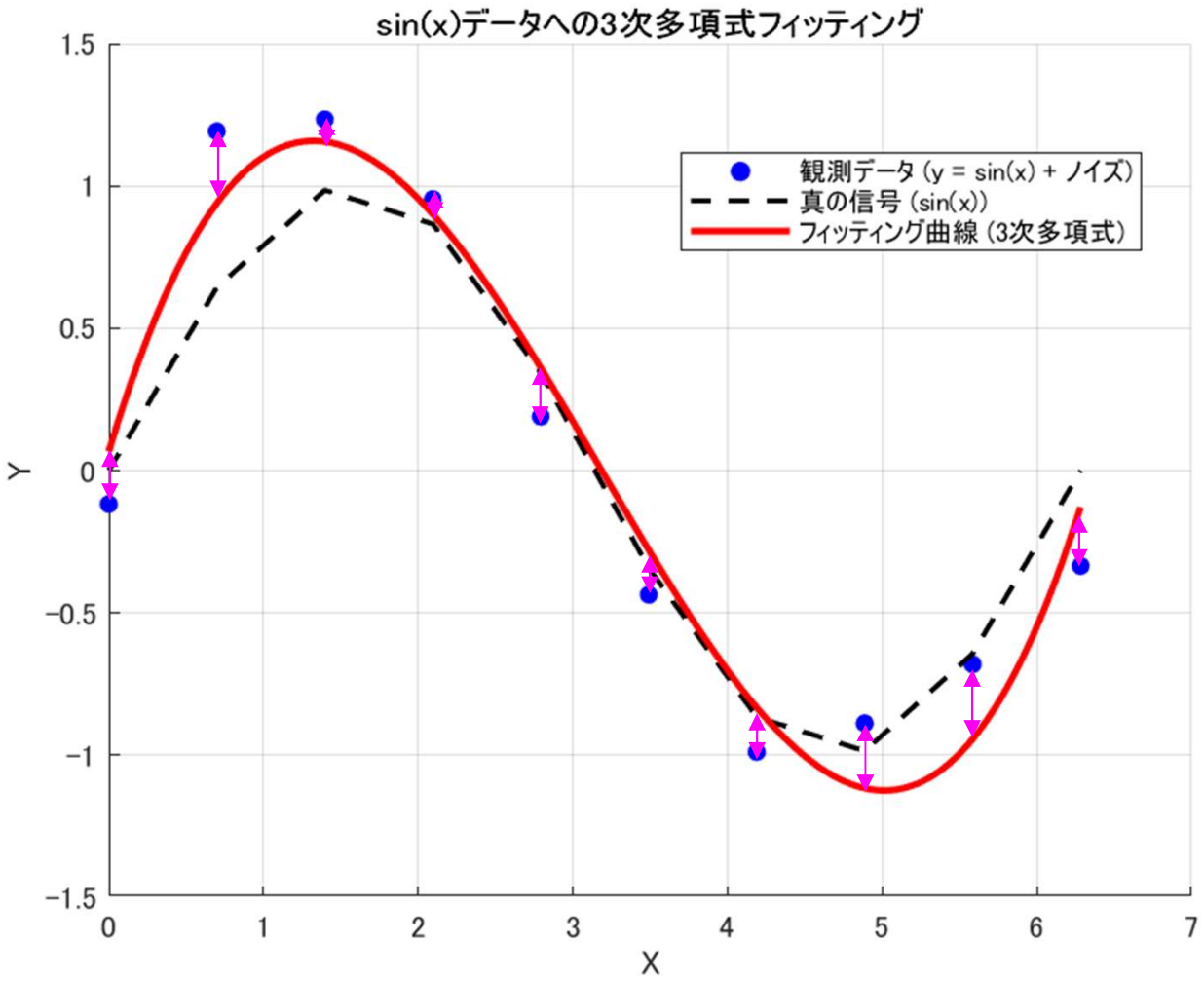
回帰分析

目的変数と説明変数の間の関係を明らかにし，関数（回帰式）で表す統計手法

© 2020 Yuki Kato. All Rights Reserved

最小二乗法とは？

■ 観測データとモデルの差（残差）の二乗和が最小になるように、最適なモデルのパラメータを決定する手法



観測データ

$$\boldsymbol{y}_{obs} = [y_1, y_2 \cdots y_n]$$

$$\boldsymbol{x}_{obs} = [x_1, x_2 \cdots x_n]$$

モデル

$$y = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

二乗誤差

$$S = \{y_1 - f(x_1)\}^2 + \cdots \{y_n - f(x_n)\}^2$$

$$= \sum_{i=1}^{n} \{y_i - f(x_i)\}^2$$

⇒Sが最小になるβ_0~β_3を計算

© 2020 Yuki Kato. All Rights Reserved

最小二乗法：一般化，線形モデル

■ 一般化すると，下記の通り

モデル $f(x, \boldsymbol{\beta})$ は、パラメータ $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ の**線形結合**で表されます。

$$y = f(x, \boldsymbol{\beta}) = \beta_0 g_0(x) + \beta_1 g_1(x) + \cdots + \beta_p g_p(x)$$

- β_k: 推定したいパラメータ（線形）
- $g_k(x)$: 既知の基底関数（x の関数であれば何でもよい。例: $1, x, x^2, \sin(x)$ など）

○ $a\sin(2x)$

× $\sin(ax)$

⇒非線形

⇒解くには非線形最適化が必要

残差二乗和 $S(\boldsymbol{\beta})$ を最小化します。

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{\beta})]^2$$

観測値　モデル

残差二乗和を $\boldsymbol{\beta}$ で偏微分してゼロとする条件（正規方程式）から、解析的に解くことができます（行列計算）。

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

線形モデルは解析解があり，一意に定まる．

嬉しい

いわゆる線形最小二乗法
（Linear Least Squares, LLS）
⇒導出してみよう

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：導出1

1. 問題設定とモデル

n 個のデータ点 $(\mathbf{x}_i, y_i)$ があり、これらのデータに最もよく適合する線形モデル

$$y = f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 g_0(\mathbf{x}) + \beta_1 g_1(\mathbf{x}) + \cdots + \beta_p g_p(\mathbf{x})$$

のパラメータベクトル $\boldsymbol{\beta}$ を求めたいとします。

ここで、

- $\mathbf{y}$：観測値のベクトル（サイズ $n \times 1$）
- $\boldsymbol{\beta}$：推定したいパラメータのベクトル（サイズ $(p+1) \times 1$）
- $\mathbf{X}$：計画行列（Design Matrix）（サイズ $n \times (p+1)$）

観測データ全体の関係は、行列形式で以下のように表されます。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

観測値と解析モデルの差

- $\boldsymbol{\epsilon}$ は誤差（残差）のベクトルです。

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：導出2

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

2. 計画行列 $\mathbf{X}$ の構成

計画行列 $\mathbf{X}$ の各行は、各データ点 $\mathbf{x}_i$ における基底関数 $g_k(\mathbf{x})$ の値で構成されます。

$$\mathbf{X} = \begin{pmatrix} g_0(\mathbf{x}_1) & g_1(\mathbf{x}_1) & \dots & g_p(\mathbf{x}_1) \\ g_0(\mathbf{x}_2) & g_1(\mathbf{x}_2) & \dots & g_p(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(\mathbf{x}_n) & g_1(\mathbf{x}_n) & \dots & g_p(\mathbf{x}_n) \end{pmatrix}$$

最初の例だと・・・　　x = [0, 1, ...n]

$$y = f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0 \qquad \boldsymbol{X} = \begin{pmatrix} 0^3 & 0^2 & 0 & 1 \\ 1^3 & 1^2 & 1 & 1 \\ \dots & \dots & \dots & \dots \\ n^3 & n^2 & n & 1 \end{pmatrix} \begin{pmatrix} \beta_3 \\ \beta_2 \\ \beta_1 \\ \beta_0 \end{pmatrix}$$

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：導出3

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

L2ノルム　式：$||\boldsymbol{v}||_2^2 = v_1^2 + v_2^2 + \ldots + v_n^2 = 2$乗和

3. 残差二乗和の最小化

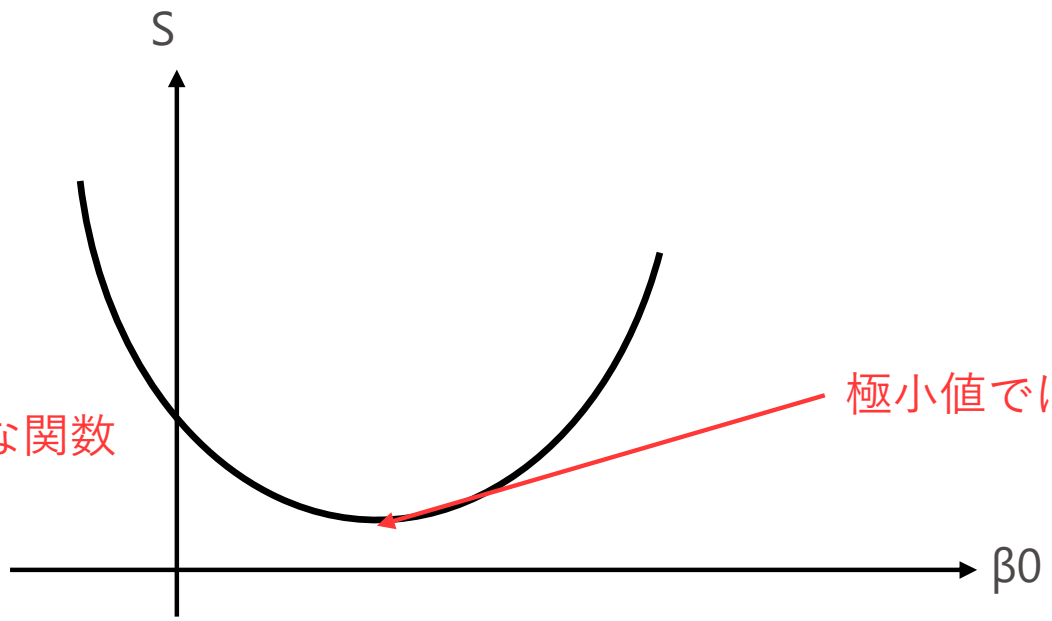
最小二乗法は、**残差二乗和** $S(\beta)$ を最小化するパラメータ β を求めます。

β^2の符号は正

$$S(\beta) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$$

この $S(\beta)$ を最小化するためには、パラメータ β に関する勾配ベクトルをゼロに設定します（$\frac{\partial S}{\partial \beta} = \mathbf{0}$）。

$$\begin{bmatrix} \varepsilon_0 & \varepsilon_1 \ldots \varepsilon_3 \end{bmatrix} \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ .. \\ \varepsilon_3 \end{bmatrix}$$



凸関数：下に凸な関数

極小値では勾配がゼロ

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：導出4

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3. 残差二乗和の最小化

最小二乗法は、**残差二乗和** $S(\boldsymbol{\beta})$ を最小化するパラメータ $\boldsymbol{\beta}$ を求めます。

$$S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

この $S(\boldsymbol{\beta})$ を最小化するためには、パラメータ $\boldsymbol{\beta}$ に関する勾配ベクトルをゼロに設定します（$\frac{\partial S}{\partial \boldsymbol{\beta}} = \mathbf{0}$）。

この式を展開します。行列演算の転置の性質 $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ を使います。

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y}^T - (\mathbf{X}\boldsymbol{\beta})^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}) - (\boldsymbol{\beta}^T\mathbf{X}^T)\mathbf{y} + (\boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

ここで、$\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta})$ と $(\boldsymbol{\beta}^T\mathbf{X}^T)\mathbf{y}$ はどちらも 1×1 のスカラー値（単なる数）になります。スカラーの転置は元のスカラーと等しいため、$(\mathbf{y}^T\mathbf{X}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ より、この2つの項は同じ値です。

$$S(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

この式のβに対する勾配がゼロになれば良い

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：導出5

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$S(\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

この式のβに対する勾配がゼロになれば良い

行列微分に関する公式を使用します。

1. $\mathbf{y}^T\mathbf{y}$ は β を含まない定数項なので、微分すると $\mathbf{0}$ です。

$$\frac{\partial(\mathbf{y}^T\mathbf{y})}{\partial\beta} = \mathbf{0}$$

2. 線形項 $\beta^T\mathbf{A}\mathbf{b}$ の微分公式 $\frac{\partial(\beta^T\mathbf{A}\mathbf{b})}{\partial\beta} = \mathbf{A}\mathbf{b}$ を利用します（$\mathbf{A} = \mathbf{X}^T, \mathbf{b} = \mathbf{y}$ と見なす）。

$$\frac{\partial(2\beta^T\mathbf{X}^T\mathbf{y})}{\partial\beta} = 2\mathbf{X}^T\mathbf{y}$$

3. 二次形式 $\beta^T\mathbf{A}\beta$ の微分公式 $\frac{\partial(\beta^T\mathbf{A}\beta)}{\partial\beta} = 2\mathbf{A}\beta$ を利用します（$\mathbf{A} = \mathbf{X}^T\mathbf{X}$ と見なす。$\mathbf{X}^T\mathbf{X}$ は対称行列です）。

$$\frac{\partial(\beta^T\mathbf{X}^T\mathbf{X}\beta)}{\partial\beta} = 2(\mathbf{X}^T\mathbf{X})\beta$$

$$\frac{\partial S}{\partial\beta} = \mathbf{0} - 2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X})\beta = \mathbf{0}$$

2で割って移項

$\Rightarrow$

正規方程式

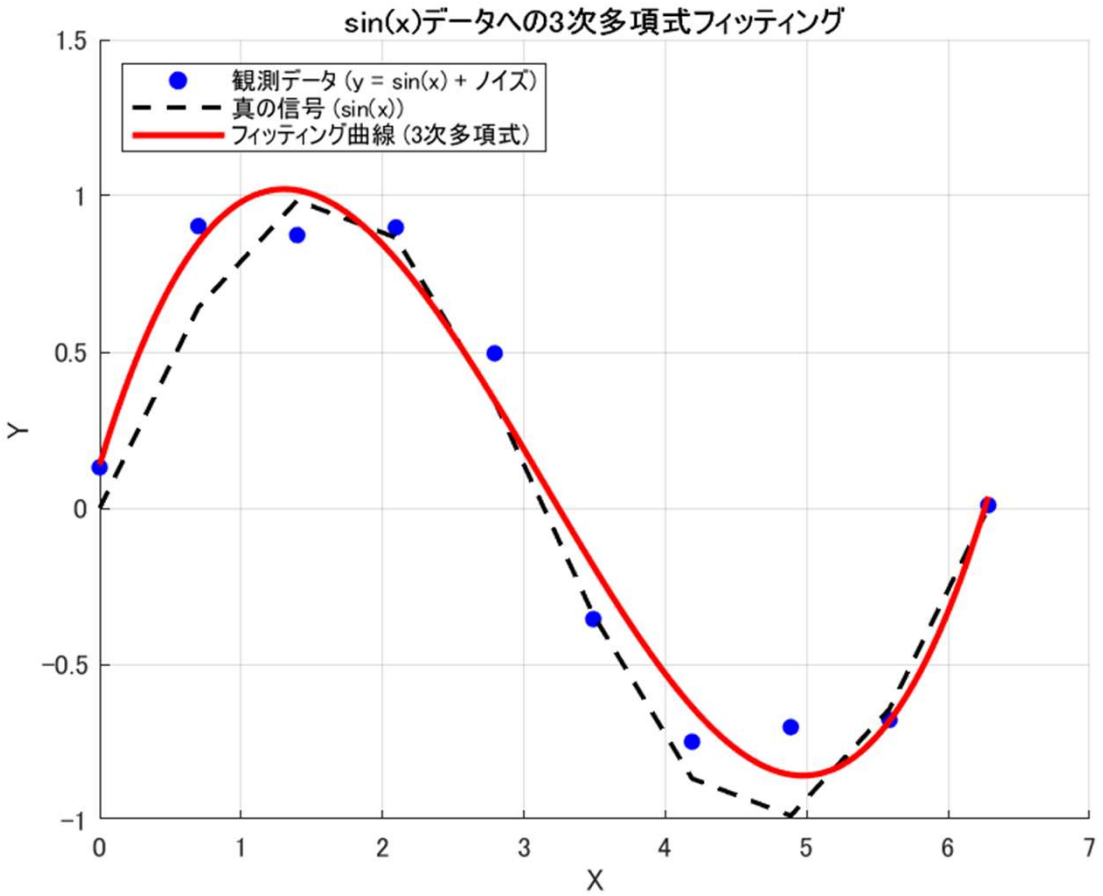
$$(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{y}$$

$(\mathbf{X}^T\mathbf{X})$の逆行列をかければ，$\boldsymbol{\beta}$が求まる

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法：演習

- Matlabで観測値を三次多項式で近似するプログラムを実行
- 1次，5次，11次の多項式ならどうなる？



$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

L2ノルム

$$S(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

2乗和をデータ長Nで
割ったもの⇒MSE
Mean Squared Error

観測値yに対するMSE

MSE): 0.012

真の信号sin(x)に対するMSE

MSE): 0.018

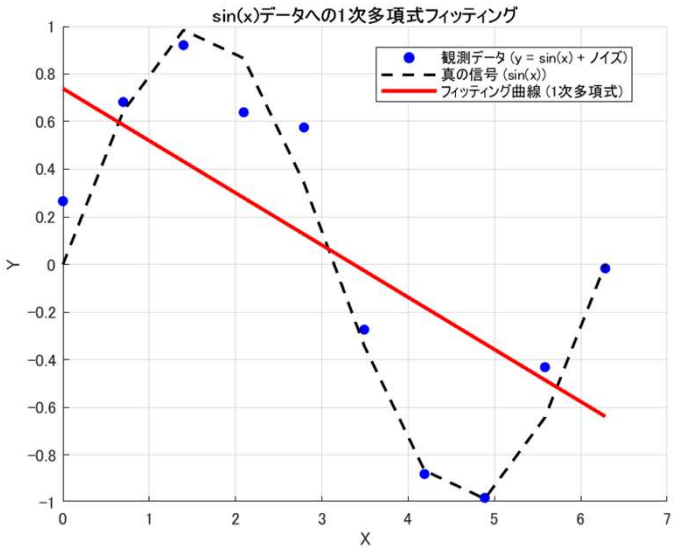
© 2020 Yuki Kato. All Rights Reserved

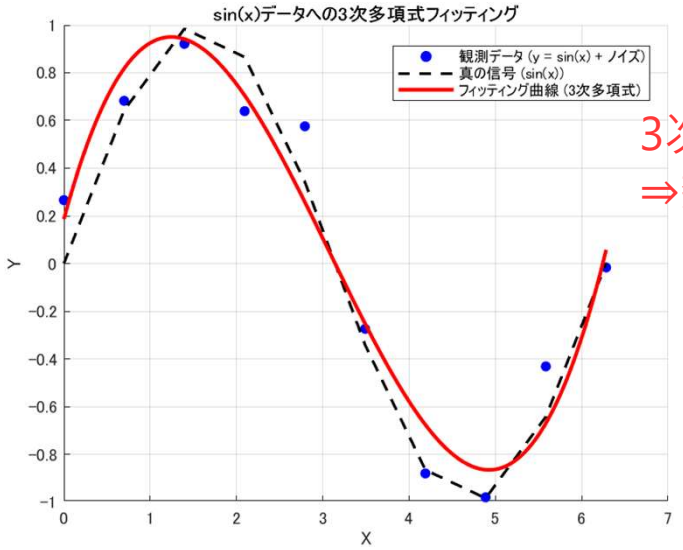
手計算の演習問題

© 2020 Yuki Kato. All Rights Reserved

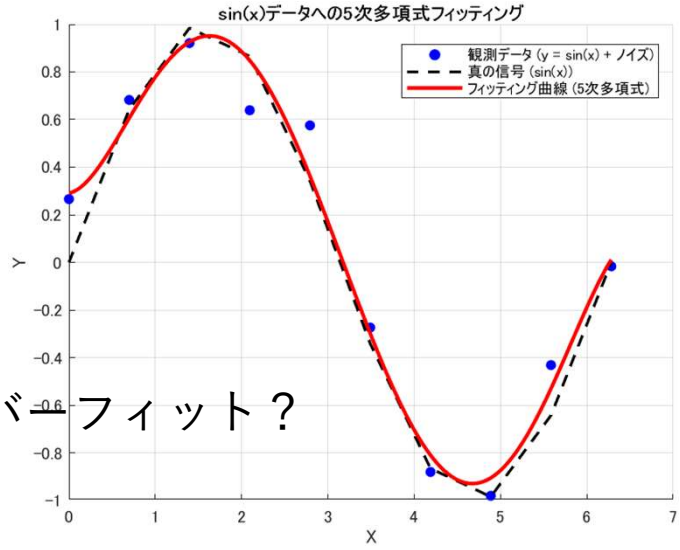
線形最小二乗法：演習

- Matlabで観測値を三次多項式で近似するプログラムを実行
- 1次，5次，11次の多項式ならどうなる？

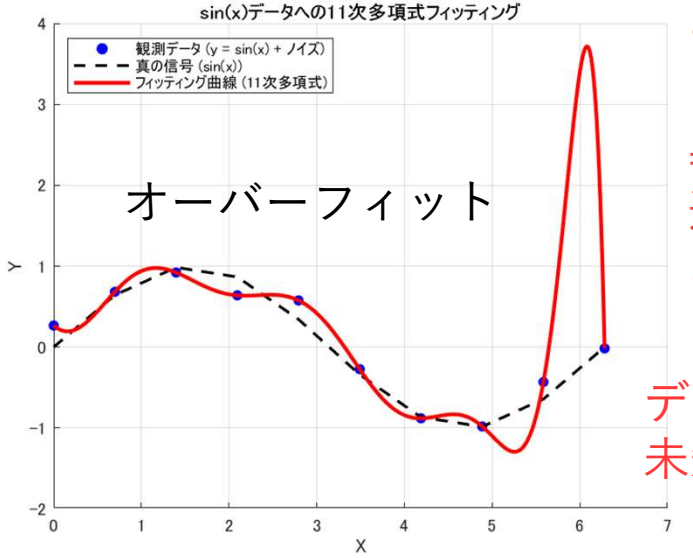




3次が一番いい？
⇒数値実験で検証



オーバーフィット？



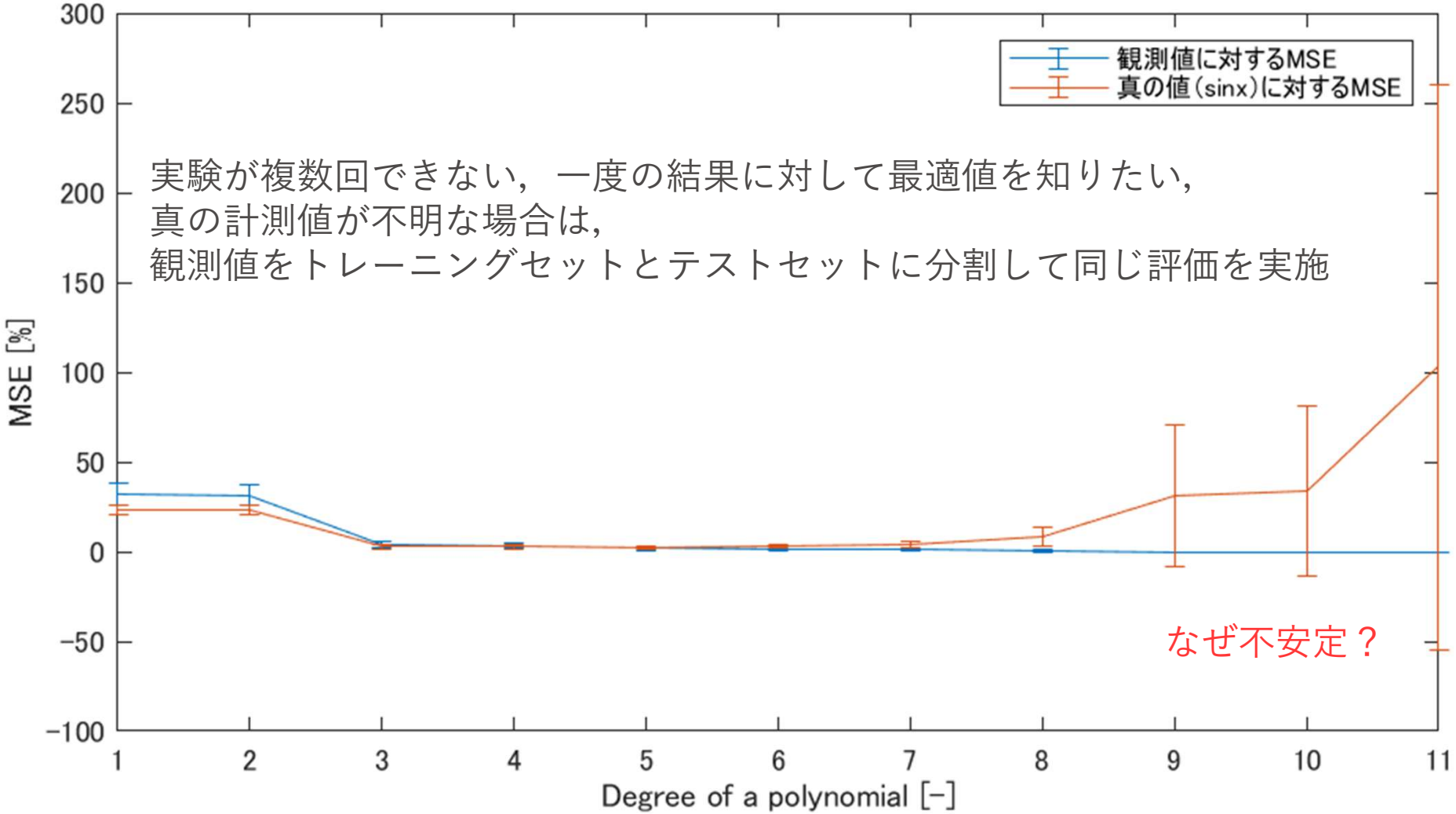
11次は警告が出る
⇒毎回結果が違う可能性がある,不安定

データ：10個
未知変数：12個

© 2020 Yuki Kato. All Rights Reserved

MSEに対する多項式次数の影響（10回実験）

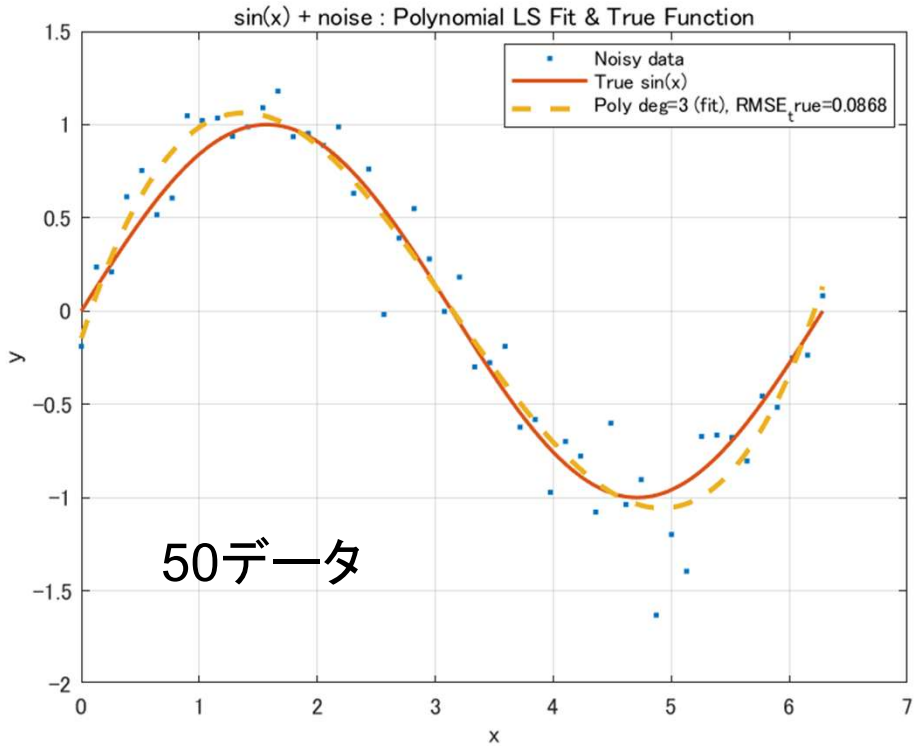
- 次数が3で真の値に対するMSEは極小に（値がサチる）
- 次数を上げていくと，観測値に対するMSEは小さくなるが，真の値に対するMSEが大きく⇒オーバーフィット

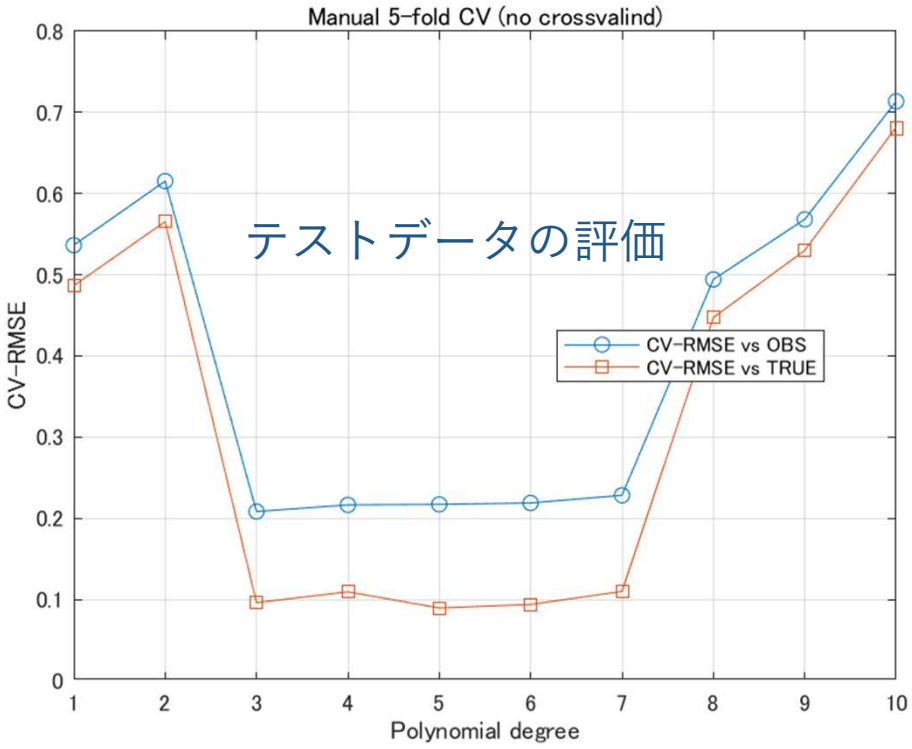


© 2020 Yuki Kato. All Rights Reserved

MSEに対する多項式次数の影響（1回実験）

- 観測値をトレーニングセットとテストセットに分割して同じ評価を実施．下の例は5分割
- 10データ×4：学習用，10データ：テスト用
- 1つのデータから過学習を抑制したモデルが作成できる
- CV：Cross validation





© 2020 Yuki Kato. All Rights Reserved

データ数より次数が多い場合（劣決定）

直感的な理解

1. 自由度が高すぎるモデル 📈

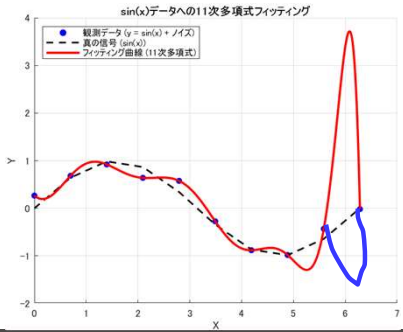
最小二乗法の問題は、観測データ $\mathbf{y}$ とモデル $\mathbf{X}\beta$ の間の誤差を最小化することです。

- **データ点数 (n)**: これは、モデルのパラメータを決定するための**制約条件の数**に相当します。
- **係数（パラメータの数 $p+1$)**: これは、モデルの**自由度**（どれだけ複雑に曲がれるか）に相当します。

$p+1 > n$ **の状況**は、「**データ点の数よりも、モデルが持つ自由度の数の方が多い**」ことを意味します。

劣決定系

例えば、わずか2点（$n=2$）のデータに対して、3次以上の多項式（$p+1 \geq 4$）を当てはめようとするような場合です。この場合、モデルはデータ点を正確に通過するだけでなく、**データを通過しながら無限に多くの異なる形状を取り得る**ため、最適解を一つに絞り込むことができません。



?

© 2020 Yuki Kato. All Rights Reserved

データ数より次数が多い場合（劣決定）2

2. 正規方程式の行列が逆行列を持たない　　p 次多項式を考える

最小二乗解は、正規方程式 $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ を解くことで得られます。

解が一意に決まるためには、係数行列 $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ が**正則**（逆行列を持つ）でなければなりません。

$(\mathbf{X}^T\mathbf{X})$ の逆行列が必要

- $\mathbf{X}$ は $n \times (p+1)$ 行列です。
- $\mathbf{X}^T\mathbf{X}$ は $(p+1) \times (p+1)$ の**正方行列**です。

しかし、係数の数 $p+1$ が行数（データ点数） n を超える（$p+1 > n$）場合、計画行列 $\mathbf{X}$ の最大ランク（列ランク）はデータ点数 n を超えることができません。

$$\text{rank}(\mathbf{X}) \leq n$$

線形独立なベクトルの数: n

このとき、積 $\mathbf{X}^T\mathbf{X}$ のランクも n 以下となります。

$$\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}) \leq n$$

$\mathbf{X}^T\mathbf{X}$ はサイズ $(p+1) \times (p+1)$ の正方行列であるにもかかわらず、そのランクはサイズよりも小さくなってしまいます（$\text{rank}(\mathbf{X}^T\mathbf{X}) < p+1$）。

行列のランクがサイズよりも小さい場合、その行列は**特異行列**（Singular Matrix）となり、**逆行列 $(\mathbf{X}^T\mathbf{X})^{-1}$ が存在しません。**

したがって、正規方程式の解 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ を計算できず、解は**無限に存在する**か、または**不安定**になります。

© 2020 Yuki Kato. All Rights Reserved

データ数より次数が多い場合 （劣決定）3

係数がデータ数より多い場合、最小二乗法は**残差二乗和をゼロにする**（またはゼロに限りなく近づける）ようなパラメータの組み合わせを**無限に見つける**ことができます。これは、解空間の自由度がデータによる制約を上回っているために発生する現象です。

この問題を回避し、一意の解を求めるためには、**リッジ回帰**（L2正則化）や**ラッソ回帰**（L1正則化）といった**正則化手法**を用いて、パラメータの大きさにペナルティを課すことが一般的です。

⇒劣決定系で一意な解を同定するための工夫

- ⇒リッジ回帰では係数が小さい解が，Lasso回帰では係数の数が少ない解が得られる（ 圧縮センシングに関係）
- ⇒逆に過決定系で，スパースな解を求めたい場合にLasso回帰を使う場合もある（スパースモデリング）．

$$S(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

2乗和の最小化

$$\Rightarrow \quad \min_{\beta \in R^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

$$\min_{\beta \in R^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

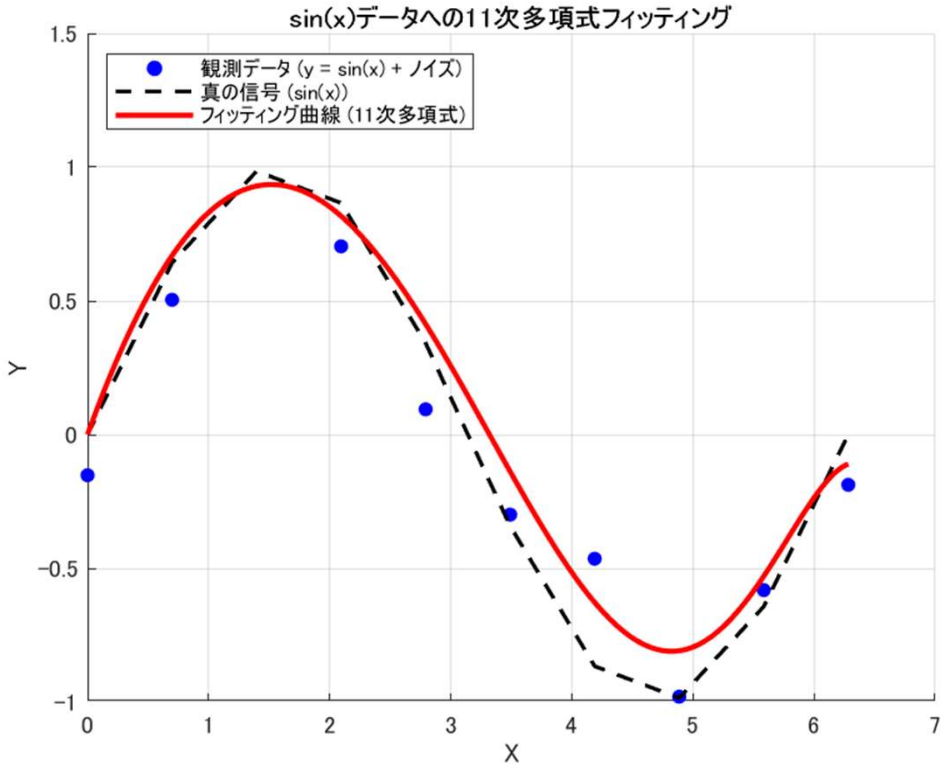
$$\min_{\beta \in R^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2$$

© 2020 Yuki Kato. All Rights Reserved

データ数より次数が多い場合（劣決定）4

- Lasso回帰では係数の数が少ない解が得られる
（ランダム行列を用いてスパース性を担保≃圧縮センシング）
- ⇒逆に過決定系で，スパースな解を求めたい場合にLasso回帰を使う場合もある（スパースモデリング）.
- これらは第14回で説明する

Lasso回帰の例



β	
-0.0000	
0.0000	
0.0017	
-0.0219	
0.0710	
0.3368	
-3.5641	
12.8313	
-23.4642	$\beta_3 x^3$
20.7061	$\beta_2 x^2$
-5.6613	$\beta_1 x$
-0.1521	β_0

劣決定でも
オーバーフィットしない

$$\min_{\beta \in R^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

© 2020 Yuki Kato. All Rights Reserved

線形最小二乗法と凸最適化問題

■ 線形最小二乗法以外でも，解が一意に定まる場合がある

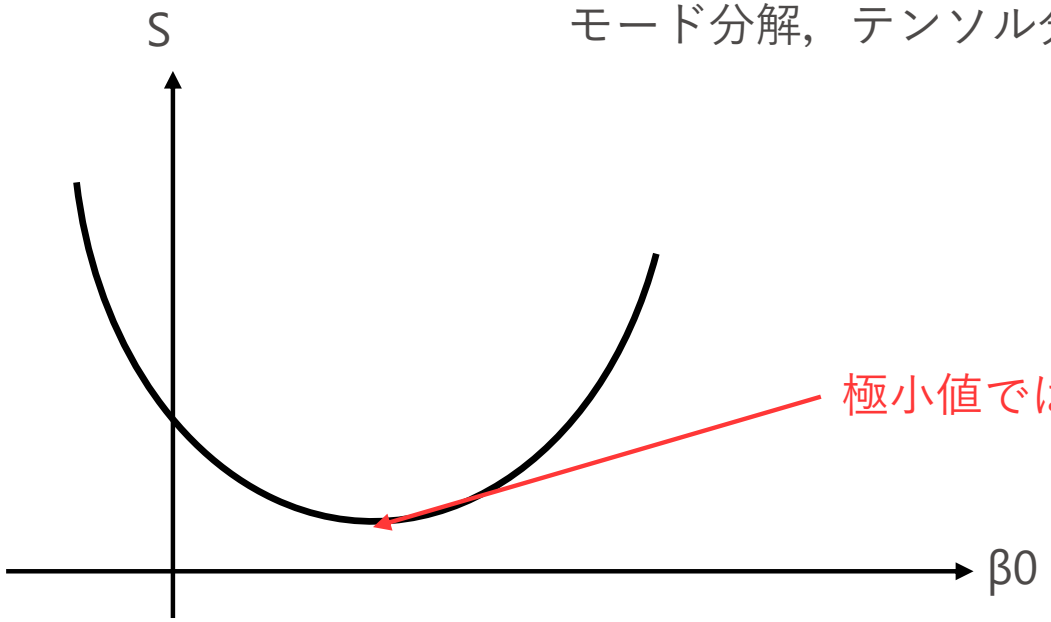
- 目的関数が凸関数である
- 実行可能領域が凸集合である

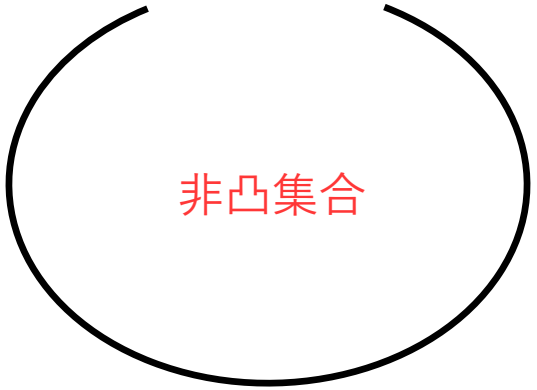
⇒このような問題を凸最適化問題という

■ 解が一意に定まるので厳密性があり，計測等に応用

✓ 数学者が様々な問題を凸最適化問題にするために，研究を推進中

モード分解，テンソル分解，最適化，スパースモデリング. etc...





ある集合 C が凸集合であるとは、集合 C 内の任意の2点 $\mathbf{x}_1$ と $\mathbf{x}_2$ を選んだとき、その2点を結ぶ線分上の任意の点 $\mathbf{x}$ が集合 C に含まれることをいいます。

© 2020 Yuki Kato. All Rights Reserved

非線形最小二乗法（Non-Linear Least Squares, NLLS）

非線形最小二乗法は、モデルが**パラメータ**に対して非線形である場合に適用されます。解析的な解法は存在せず、解を見つけるために**反復的な数値最適化アルゴリズム**が必要です。

○ $a\sin(2x)$

× $\sin(ax)$

モデルの形式

モデル $f(x, \boldsymbol{\beta})$ は、パラメータ $\boldsymbol{\beta}$ に対して非線形な形で表されます。

$$y = f(x, \boldsymbol{\beta})$$

- **例:** 指数関数 $y = \beta_0 e^{\beta_1 x}$ や、振動モデル $y = \beta_0 \sin(\beta_1 x + \beta_2)$ など。

数式と解法

線形最小二乗法と同様に、残差二乗和 $S(\boldsymbol{\beta})$ を最小化しますが、解析的な解はありません。

$$S(\boldsymbol{\beta}) = \sum^{n} [y_i - f(x_i, \boldsymbol{\beta})]^2 \to \min$$

© 2020 Yuki Kato. All Rights Reserved

非線形最小二乗法(NLLS)　解法

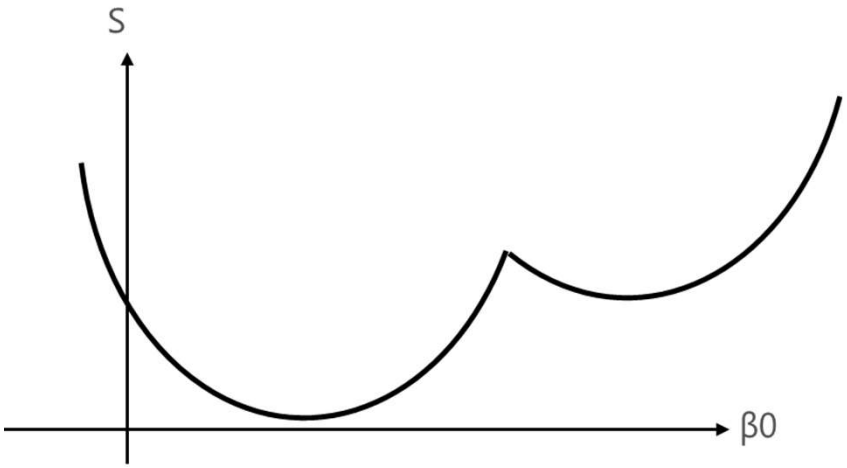
$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{\beta})]^2 \rightarrow \min$$

本講義では割愛
ベータを動かしながら,
Sが小さいところを探す

この最小化は、以下の**反復的アルゴリズム**によって行われます。

1. **初期値の設定:** パラメータ $\boldsymbol{\beta}$ の初期推定値 $\boldsymbol{\beta}^{(0)}$ を設定します。
2. **線形近似:** 現在の推定値の周りでモデルを線形近似します（テーラー展開）。
3. **更新:** 近似された線形問題から、パラメータの修正量 $\Delta\boldsymbol{\beta}$ を計算し、$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \Delta\boldsymbol{\beta}$ でパラメータを更新します。
4. **収束判定:** パラメータの更新量が十分に小さくなるまで、2と3を繰り返します。

- **主なアルゴリズム: ガウス・ニュートン法 (Gauss-Newton)** や **レーベンバーグ・マルカート法 (Levenberg-Marquardt)** が一般的に使用されます。



特徴	線形最小二乗法 (LLS)	非線形最小二乗法 (NLLS)
モデル形式	パラメータに対して**線形**	パラメータに対して**非線形**
解法	**解析的解法**（行列計算）	**反復的数値最適化**（ガウス・ニュートンなど）
計算速度	**高速**	**遅い**（反復計算が必要なため）
最適解	**大域的最適解**が保証される（一意）	**局所的最適解**に収束する可能性がある
初期値	**不要**	**必須**（収束結果に大きく影響する）

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第十回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

11/11が休講になります．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布
- 第５回　大数の法則と中心極限定理
- 第６回　時系列データ解析とフィルタ処理1
- 第７回　時系列データ解析とフィルタ処理2，画像解析　✓課題①，②出題

10/28 ■ 第８回　画像解析
10/31 ■ 第９回　最小二乗法，回帰分析
11/4 ■ 第１０回　最尤推定，ベイズ推定　数理モデル，サロゲートモデル構築へ
11/14 ■ 第１１回　変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ
11/18 ■ 第１２回　スパースモデリング
11/21 ■ 第１３回　多変量データ解析の基礎と主成分分析　✓11/14: 課題①②〆，両方とも，講義開始時に集めます
11/25 ■ 第１４回　固有値分解，特異値分解　答案の配布と，課題の解説をどこかで実施予定
11/28 ■ 第１５回　期末テスト
12/2 ■ 第１６回　返却　⇒返却のみ，補講扱いですが，全員出席とするので任意参加

© 2020 Yuki Kato. All Rights Reserved

回帰分析に置いて，背後にある確率分布を考える

■ モデルに対する誤差が，どのように乗るかを考える

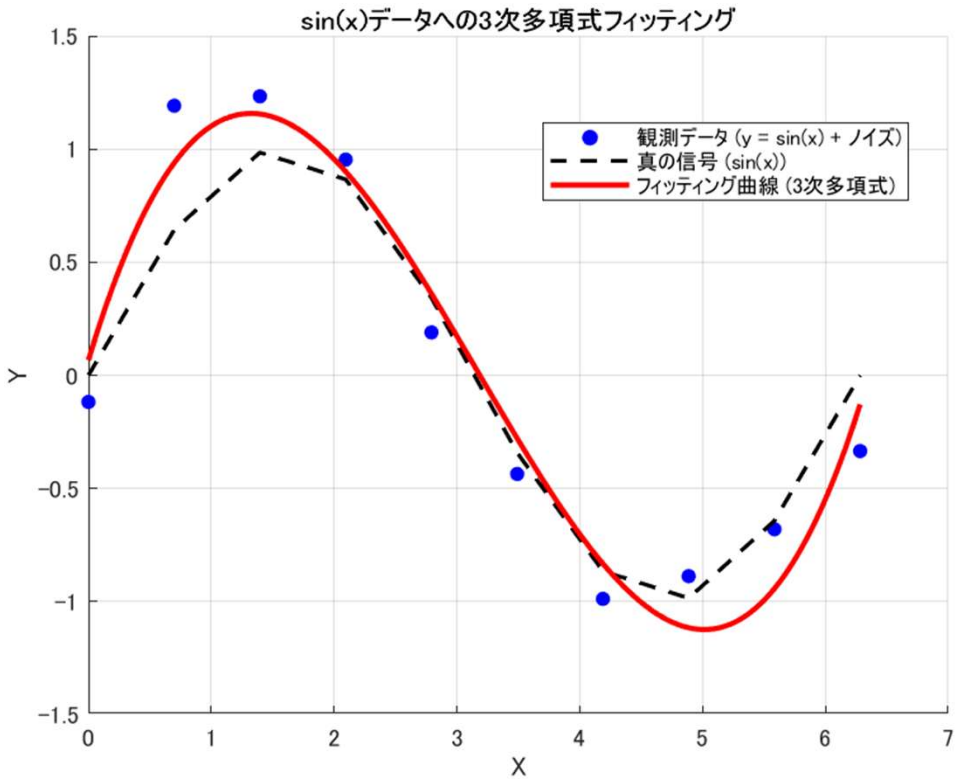
観測データ

$$\boldsymbol{y}_{obs} = [y_1, y_2 \cdots y_n]$$

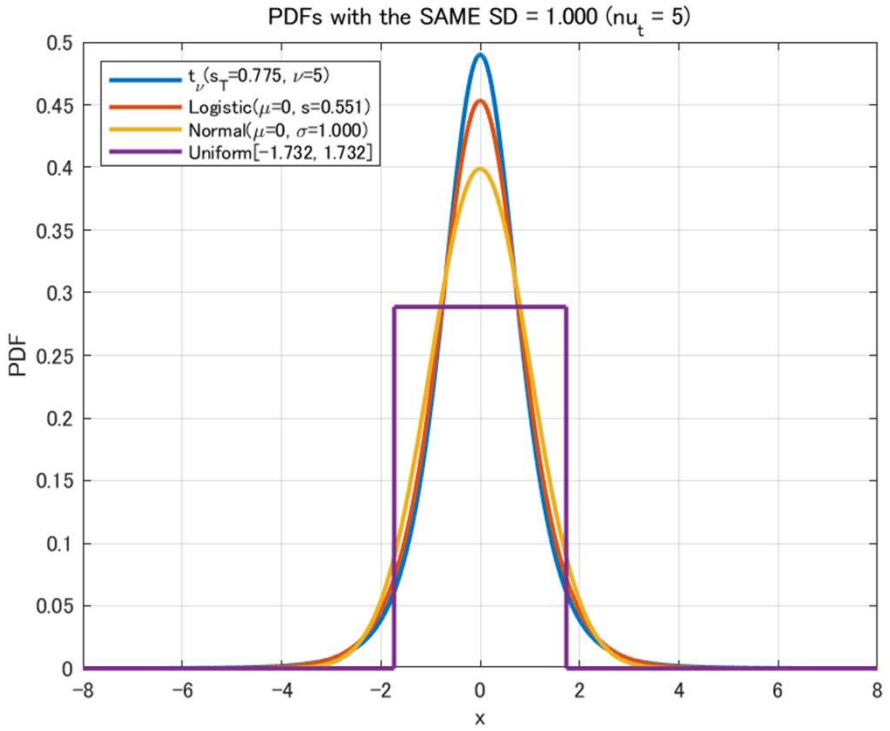
$$\boldsymbol{x}_{obs} = [x_1, x_2 \cdots x_n]$$

モデル

$$y = f(x) = ax^3 + bx^2 + cx + d$$



標準偏差が1の様々な分布



⇒正規分布で仮定するのが自然

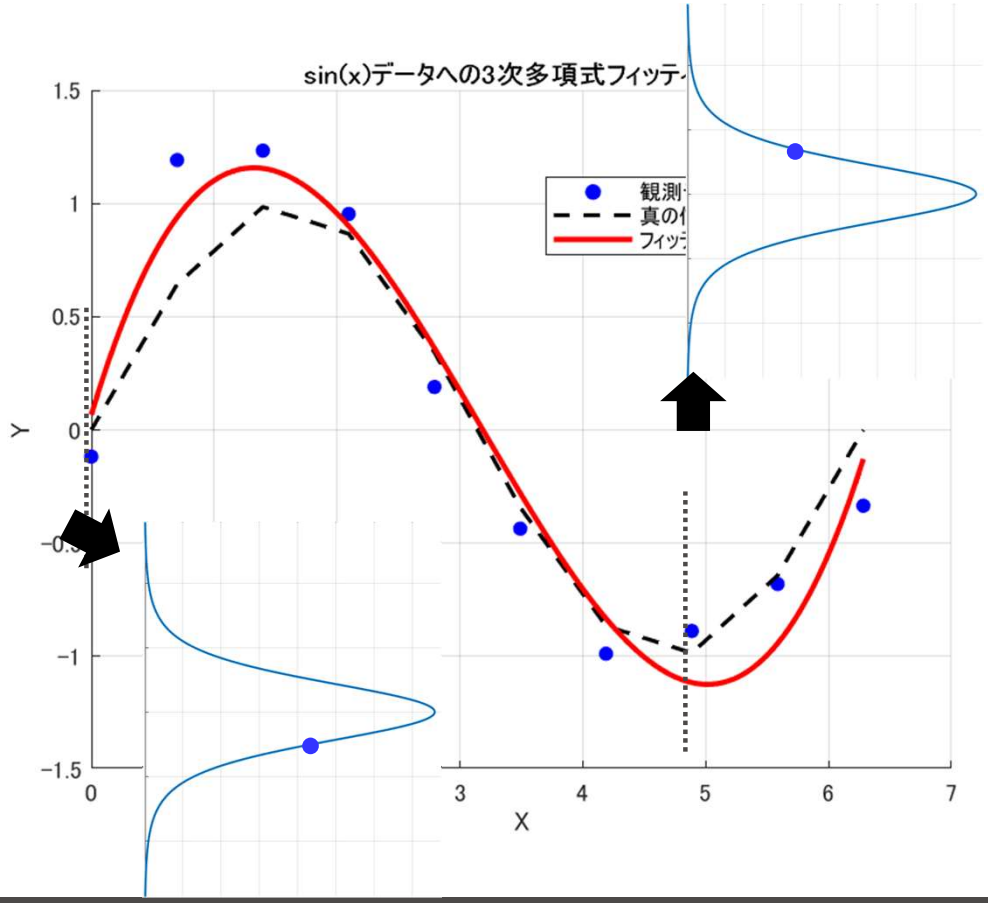
© 2020 Yuki Kato. All Rights Reserved

正規分布が背後にあるとすると

モデルの前提: $y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i$ で、$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$。

このとき、各観測値 y_i は平均 $f(\mathbf{x}_i, \boldsymbol{\beta})$、分散 σ^2 の正規分布に従います。

$$P(y_i|\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2}{2\sigma^2}\right)$$



- ここで，観測値が得られる確率を計算してみる
- ある特定の観測点: x_iについて考えると，そこでy_iが得られる確率は

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

- よって，全ての観測値が得られる確率は

$$\underbrace{\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}_{\text{尤度関数}} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

© 2020 Yuki Kato. All Rights Reserved

尤度関数とは？

- 「観測されたデータが、仮定したモデルパラメータの下で、どれほどありえそうか」を示す関数

$$\underbrace{\mathcal{L}(\beta, \sigma^2)}_{\text{尤度関数}} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2}\right)$$

- 仮説

「観測されたデータは，最も発生確率が高いデータである」

- この仮説の元，上記確率が最大となる様にパラメータを決定する方法を最尤推定法という
- 前頁のグラフから直感的な理解はできる
  - コイントスの表に賭けて，10回連続で裏が出たなら，自分が極度に不運なのではなく，イカサマされて元の確率分布がゆがんでいる

© 2020 Yuki Kato. All Rights Reserved

最尤推定

尤度関数の最大化は、計算の簡便さから、その**対数（自然対数）をとった対数尤度関数** $\ln \mathcal{L}$ の最大化に置き換えられます。

$$\ln \mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2}{2\sigma^2} \right) \right]$$

対数 ln を分解し、整理します。

$$\ln \mathcal{L}(\boldsymbol{\beta}, \sigma^2) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2$$

下線部が最小⇒全体が最大化

パラメータ $\boldsymbol{\beta}$ の推定

対数尤度を最大化するには、$\boldsymbol{\beta}$ を含む最後の項に着目します。この項はマイナスの符号がついているため、この項の絶対値（すなわち**残差二乗和**）を**最小化**すれば、対数尤度全体は最大化されます。

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2 \right]$$

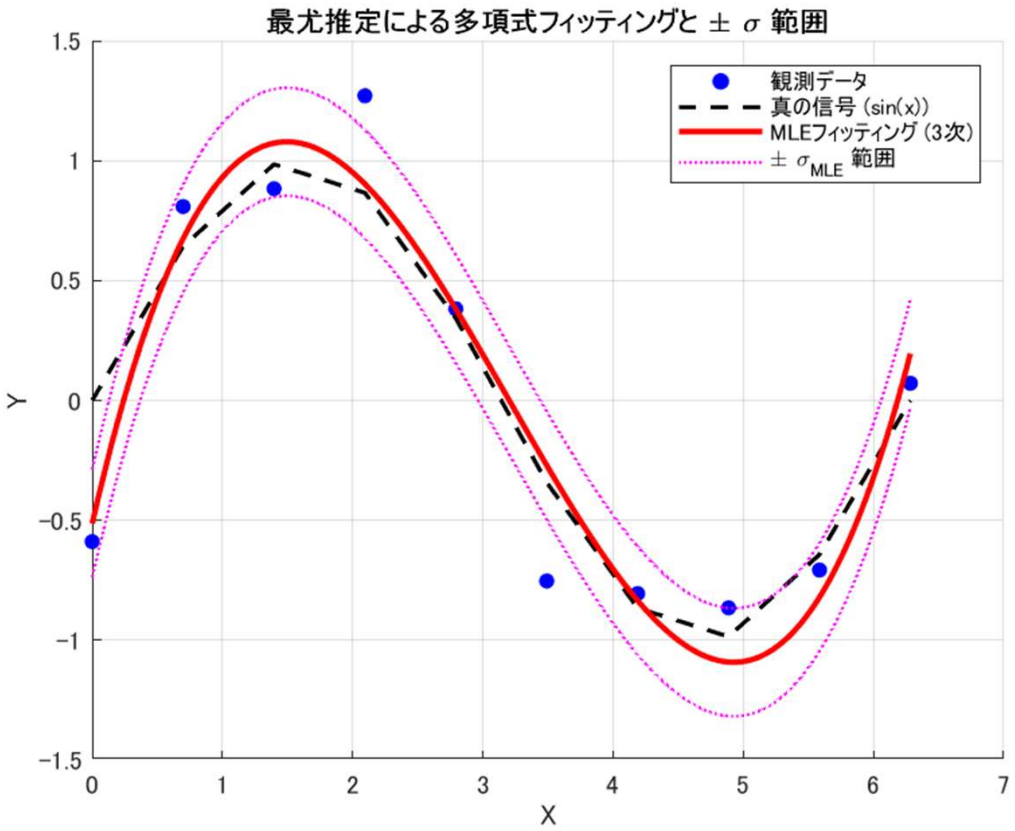
結論: 正規分布を仮定した線形回帰（または多項式フィッティング）の最尤推定量 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ は、**最小二乗解 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ と完全に一致します。**

注意：他の確率分布を仮定すると最小二乗法と一致しない

© 2020 Yuki Kato. All Rights Reserved

最尤推定の例

- 結果は線形最小二乗法と同じ
- 但し，確率分布（標準偏差）が評価できる



モデル

$$y = f(x) = ax^3 + bx^2 + cx + d$$

$$\underbrace{\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}_{\text{尤度関数}} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\underbrace{\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}}_{\text{尤度関数}} \right)$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2 \right]$$

© 2020 Yuki Kato. All Rights Reserved

最尤推定の演習

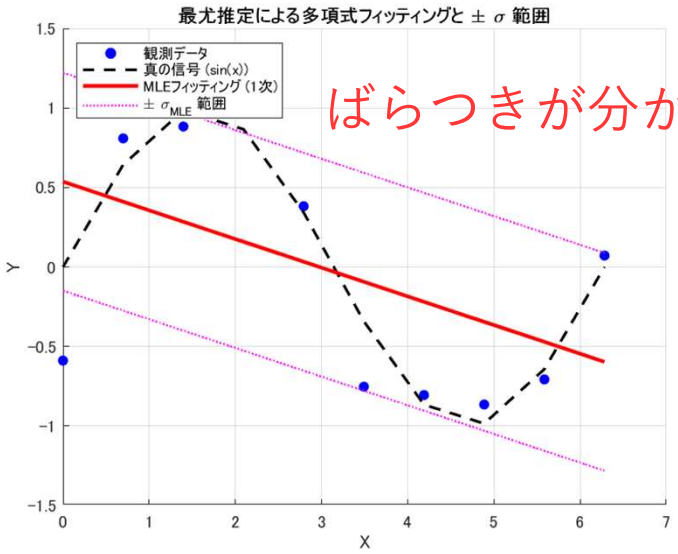
- Matlabで1次，5次，11次の多項式で最尤推定

© 2020 Yuki Kato. All Rights Reserved

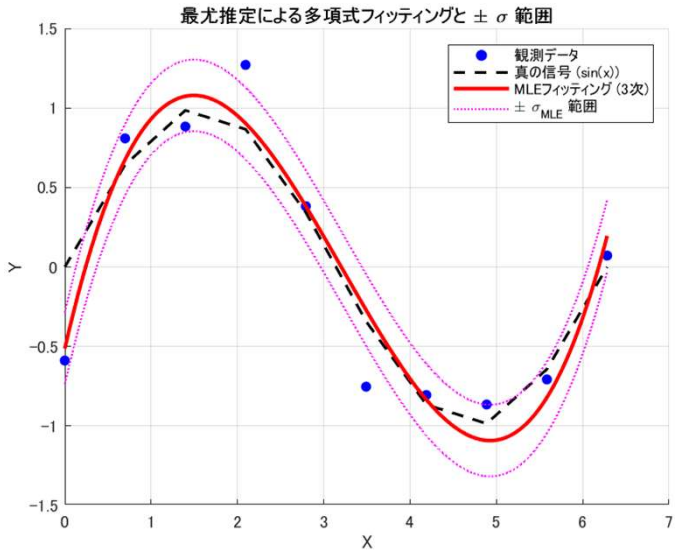
最尤推定の演習

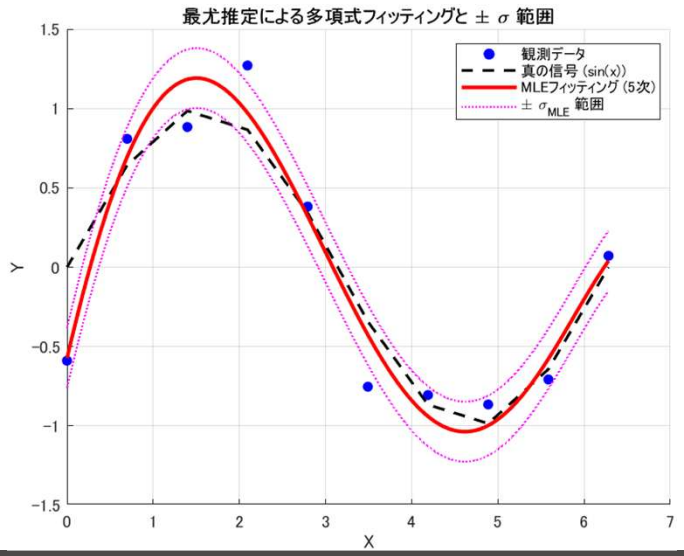
■ Matlabで1次，5次，11次の多項式で最尤推定

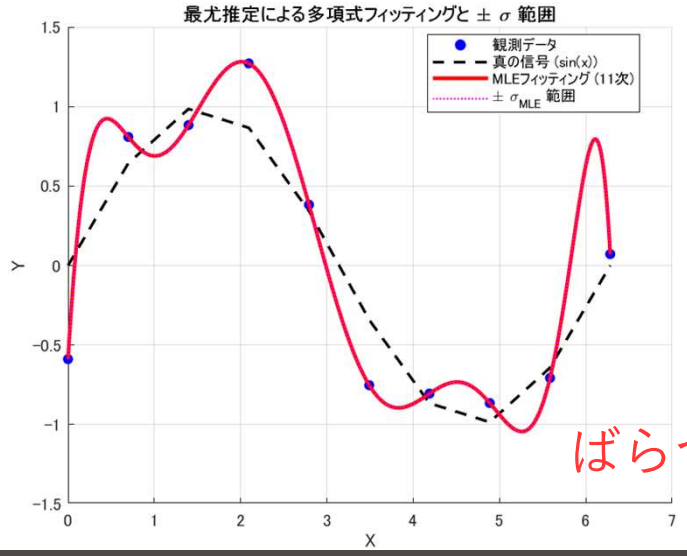
推定値自体はLLSと変わらない…



ばらつきが分かる！







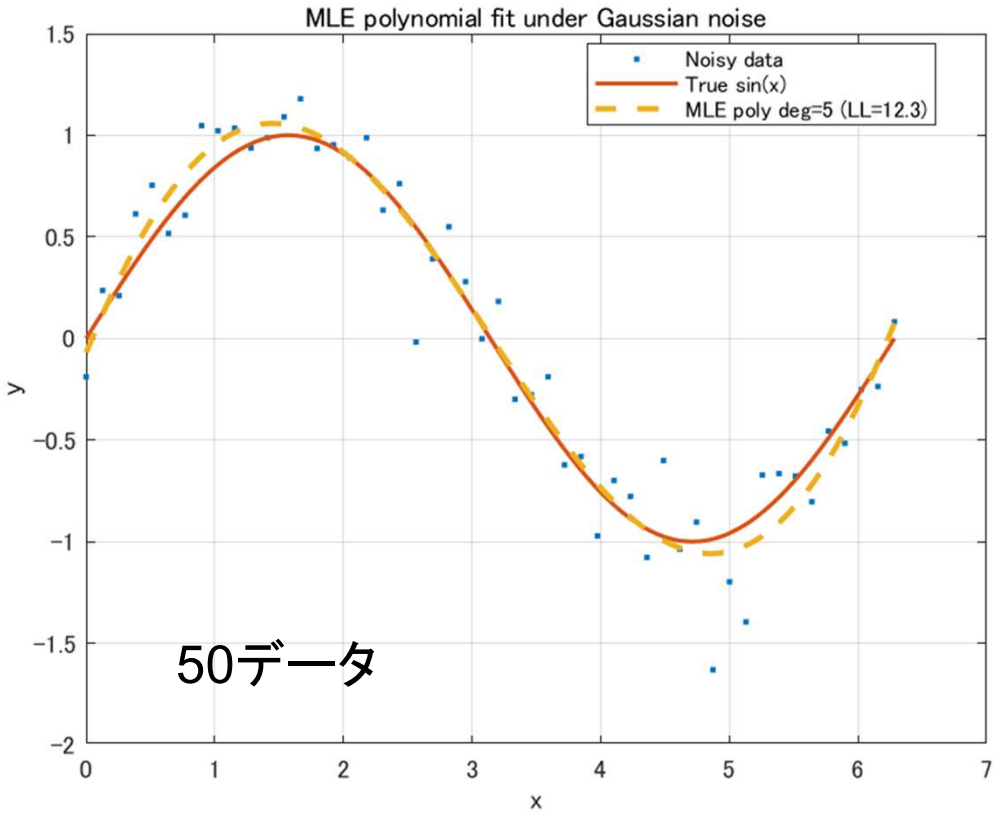
11次だと
オーバー
フィット

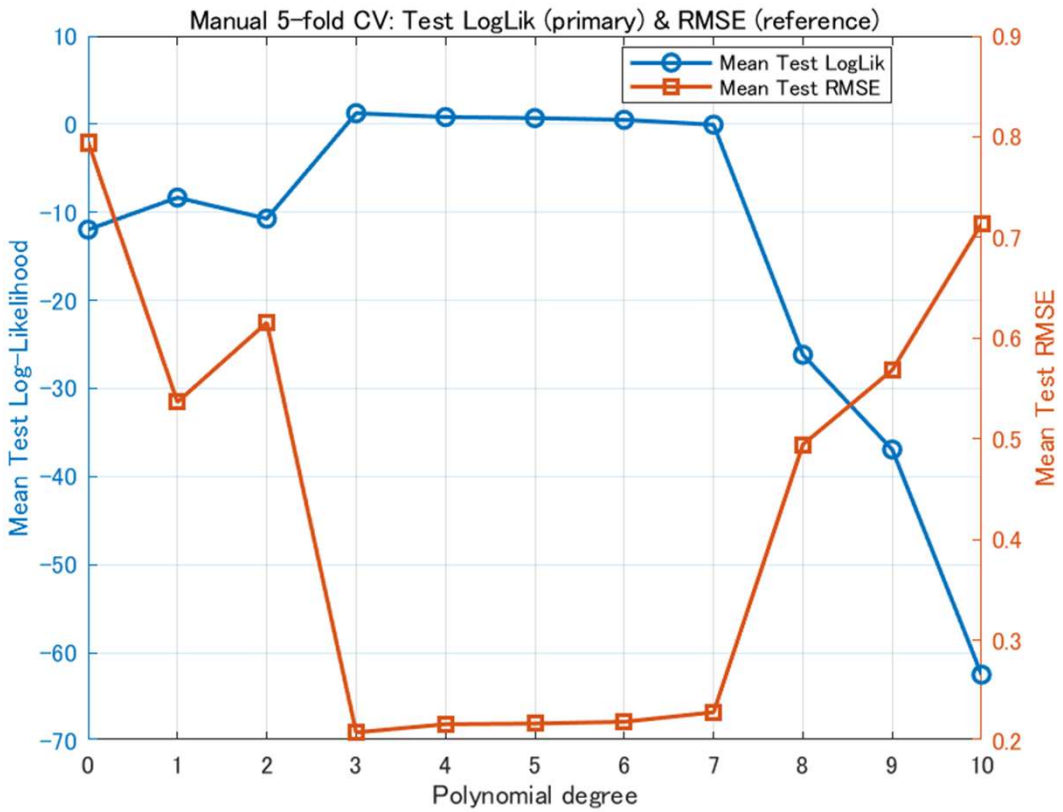
ばらつきが無い・・

© 2020 Yuki Kato. All Rights Reserved

MSEに対する多項式次数の影響（1回実験）

- 観測値をトレーニングセットとテストセットに分割して同じ評価を実施．下の例は5分割
- 10データ×4：学習用，10データ：テスト用
- 対数尤度の最大値からも，最適な次数が選定できる





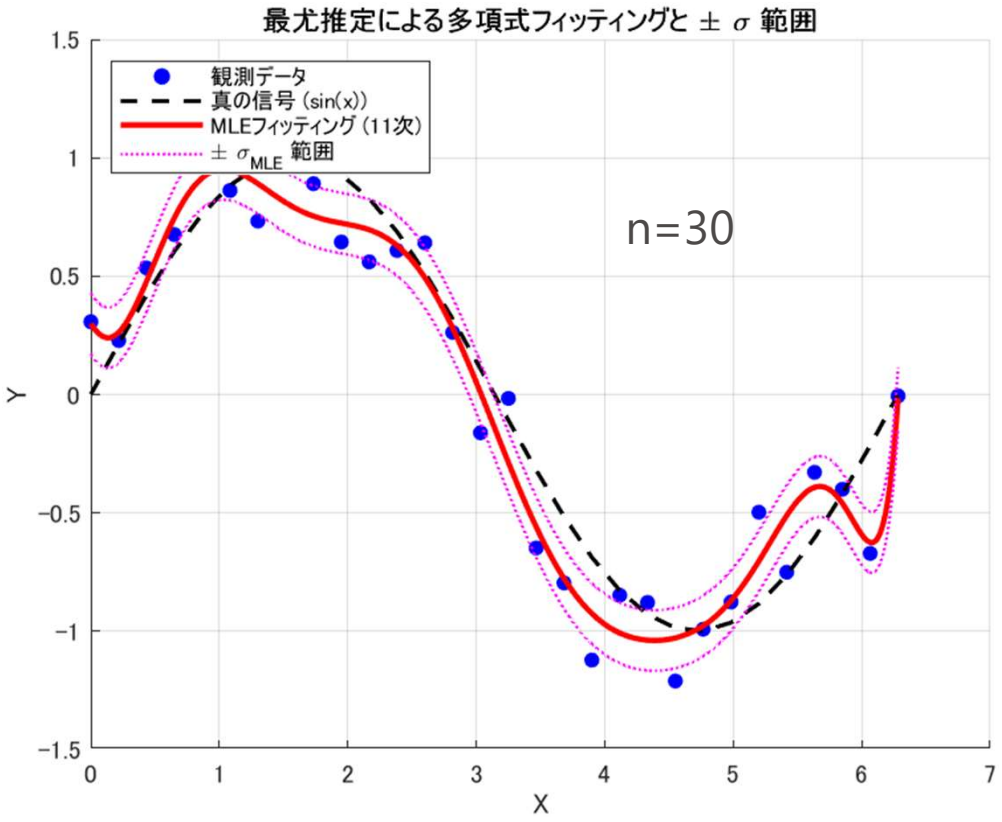
その他：AIC（赤池情報量基準）など

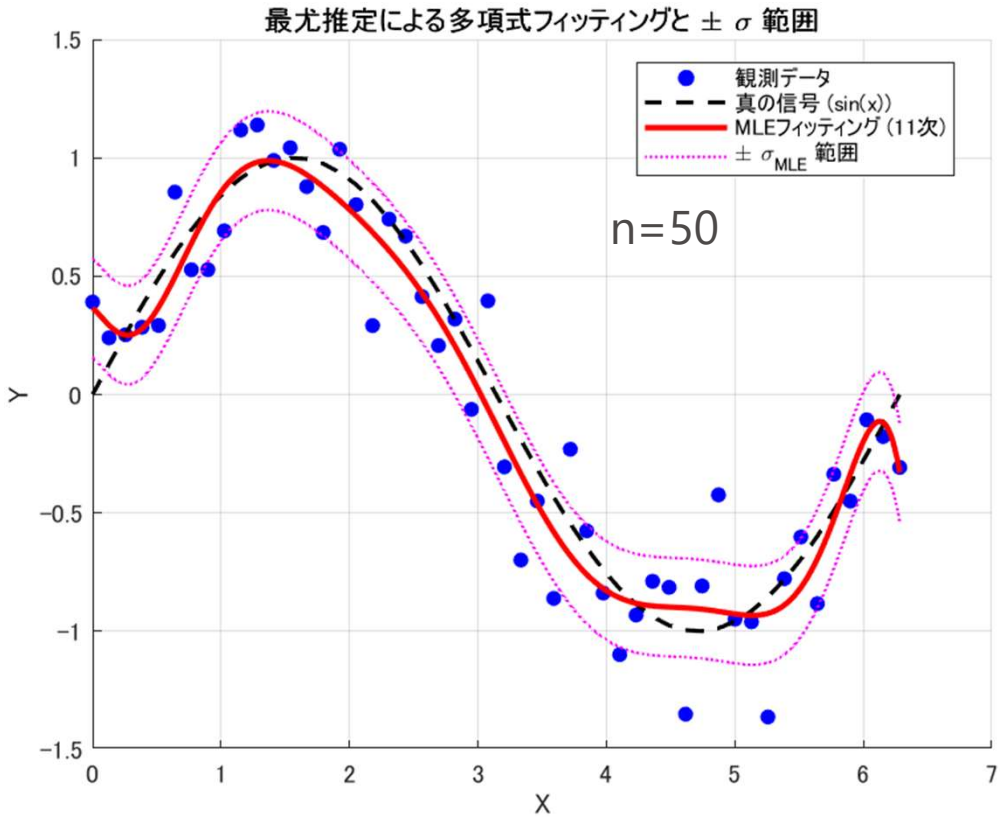
© 2020 Yuki Kato. All Rights Reserved

データ点数を増やすと

- 11次ではオーバーフィットする
- 公差検証により次数を変更して評価することで，最適な次数を決定しても良いが・・・一度に決めたい

⇒ベイズ推定を取り込む





© 2020 Yuki Kato. All Rights Reserved

ベイズの定理（第二回講義より）

- ある条件付き確率が既知の場合，他の条件確率を推定できる
- ベイズ推定では事後分布を求める

事象 B が起こるという条件の下で事象 A が起こる確率を条件付き確率といい、$P(A|B)$ と書く：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{ただし} \quad P(B) \neq 0\,. \tag{1.9}$$

逆に，上式で A と B とを入れ替えると

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

である（乗法公式）．これから

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \text{ただし} \quad P(A) \neq 0 \tag{1.10}$$

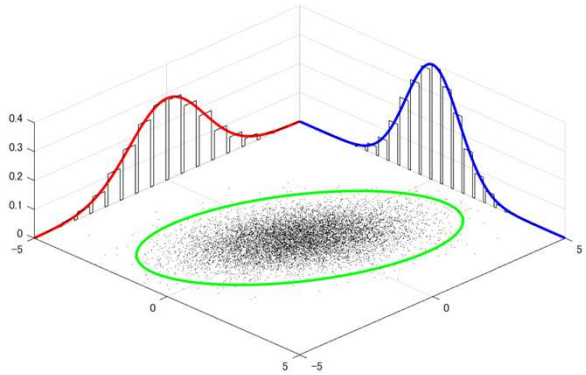
を得る．

© 2020 Yuki Kato. All Rights Reserved

補：多変量正規分布

d 次元ベクトル $\mathbf{x} \in \mathbb{R}^d$ が平均 $\boldsymbol{\mu} \in \mathbb{R}^d$、共分散 $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$（対称・正定値）に従うとき

$$\boxed{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$



の密度は

$$\boxed{p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\!\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$

ここで $|\boldsymbol{\Sigma}|$ は行列式、$\boldsymbol{\Sigma}^{-1}$ は逆行列、二次形式が行列表現です。

対数密度（対数尤度の1点分）は

$$\log p(\mathbf{x}) = -\tfrac{d}{2}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

精度行列（information form）

$$\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^{-1}, \qquad \boldsymbol{\eta} := \boldsymbol{\Lambda}\boldsymbol{\mu}.$$

すると

$$p(\mathbf{x}) \propto \exp\!\left(-\tfrac{1}{2}\mathbf{x}^\top \boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\eta}^\top \mathbf{x}\right) \quad (\text{正規化定数は } (2\pi)^{-d/2}|\boldsymbol{\Lambda}|^{1/2})$$

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析

回帰分析では、推定したいパラメータは**重み（係数）ベクトル $\mathbf{w}$** です。

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

ベイズ回帰では、この $\mathbf{w}$ の事後分布 $P(\mathbf{w}|\mathbf{y})$ を求めます。

パラメータの確率分布も計算する．ここがユニーク

■ 観測値の確率は下記になる

$$\mathcal{N}(y \mid f(x_n), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}\{y - f(x_n)\}^2}$$

MLEと同じ

$$f(x) = \sum_{m=0}^{M} w_m x^m$$

ベイズの慣例に従い，$\beta = \sigma^{-2}$
よって，未知変数がw_mになっている

βは観測におけるばらつき，事前に与える

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析2

- 続いて，未知のパラメータwに対する確率分布を用意
- 既知ならそれを考慮，未知なら正規分布を仮定

MLEでは$\boldsymbol{w}$ が特定の値⇒Bayesでは$\boldsymbol{w}$ が確率分布を持つ

事前分布

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$$

多変数の正規分布

- ここで$\boldsymbol{w}$が決まり$f(x)$が定まると，観測値$\boldsymbol{y}$が得られる確率は

尤度

$$P(\boldsymbol{y} \mid \mathbf{w}) = \mathcal{N}(y_1 \mid f(x_1), \beta^{-1}) \times \cdots \times \mathcal{N}(y_N \mid f(x_N), \beta^{-1})$$

$$= \prod_{n=1}^{N} \mathcal{N}(y_n \mid f(x_n),\ \beta^{-1})$$

$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n) - y_n\}^2\right]$$

MLEと同じ

αはパラメータのばらつき，事前に与える

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析3

■ ベイズの定理を使うと，事後分布が計算できる

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

事後分布

$$P(\mathbf{w}\,|\,\boldsymbol{y}) = \frac{P(\boldsymbol{y}\,|\,\mathbf{w})}{\displaystyle\int_{-\infty}^{\infty} P(\boldsymbol{y}\,|\,\mathbf{w}')P(\mathbf{w}')d\mathbf{w}'} P(\mathbf{w})$$

■ これは，観測データyに基づいて，パラメータwの確率を更新する関係式になる．ただし，分母はwに依存しない定数なので，$\mathrm{P}(w|y)$の正規化定数Zとみなせる

$$P(\mathbf{w}\,|\,\boldsymbol{y}) = \frac{1}{Z} P(\boldsymbol{y}\,|\,\mathbf{w})P(\mathbf{w})$$

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$$

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2\right]$$

よって，

事後分布 $$\mathrm{P}(w|y) = : \mathrm{Const} \times \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n) - y_n\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]$$

αとβが事前に与えられれば，事後分布が求まる

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析4

$$p(\mathbf{x}) \propto \exp\left(-\tfrac{1}{2}\mathbf{x}^\top \mathbf{\Lambda}\mathbf{x} + \boldsymbol{\eta}^\top \mathbf{x}\right)$$
$$\mathbf{\Lambda} := \mathbf{\Sigma}^{-1}, \qquad \boldsymbol{\eta} := \mathbf{\Lambda}\boldsymbol{\mu}.$$

- 下記はwについての二次関数になっているので,

$$\mathrm{P}(w|y) = \mathrm{Const} \times \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n)-y_n\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]$$

- wについて

$$\sum_n \{f(x_n) - y_n\}^2 = \|\Phi\mathbf{w} - \mathbf{y}\|_2^2$$

より指数部は

$$-\tfrac{1}{2}\left[\mathbf{w}^\top(\alpha\mathbf{I} + \beta\Phi^\top\Phi)\mathbf{w} - 2\beta\,\mathbf{w}^\top\Phi^\top\mathbf{y}\right]$$

$$\boxed{\mathbf{b} := \beta\,\Phi^\top\mathbf{y}}$$

$$2\,\mathbf{b}^\top\mathbf{w}$$

です。ここで

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\Phi^\top\Phi$$

と置けば、平方完成から

$$\mathrm{P}(w|y) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \mathbf{S}), \qquad \boldsymbol{\mu} = \mathbf{S}\,\mathbf{b} = \beta\,\mathbf{S}\,\Phi^\top\mathbf{y}$$

平均と分散が求まる

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析5

- 補足：これから，事後分布P($w|y$)が最大になるパラメータwを決定する条件が読み取れる．
- 指数関数の中身が最大となれば良いので，下記誤差関数Eを最小にすればよい

$$E=\frac{\beta}{2}\sum_{n=1}^{N}\{f(y_n)-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- 第一項は最小二乗法や最尤推定と同じ形
- よって，$\alpha=0$ならば同じ解になる
- $\alpha>0$なら，誤差が大きくなる（平均0という事前分布の過程に引きずられて，wが抑制される）

⇒オーバーフィットを防ぐことができる

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析6

ベイズ線形回帰では、新しい観測 y_{new} は、新しい入力 $\mathbf{x}_{\text{new}}$ と、事後分布 $P(\mathbf{w}|\mathbf{D})$ で与えられた重み $\mathbf{w}$ の線形結合から生成されると仮定します。

$$\text{モデル: } y_{\text{new}} = \mathbf{x}_{\text{new}}^T \mathbf{w} + \epsilon_{\text{new}}$$

ここで、

- $\mathbf{w}$：重みベクトル（事後分布 $P(\mathbf{w}|\mathbf{D}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ に従う確率変数）
- ϵ_{new}：ノイズ（正規分布 $\mathcal{N}(0, \beta^{-1})$ に従う確率変数）
- $\mathbf{x}_{\text{new}}$：既知の入力ベクトル（多項式基底関数で変換されたもの）

目標は、**予測 y_{new} の分散** $\text{Var}[y_{\text{new}}]$ を計算することです。

1. 分散の加法性の利用

$\mathbf{w}$ と ϵ_{new} は互いに**独立**であるため、分散の加法性の法則 $\text{Var}[A + B] = \text{Var}[A] + \text{Var}[B]$ を適用できます。

$$\text{Var}[y_{\text{new}}] = \text{Var}[\mathbf{x}_{\text{new}}^T \mathbf{w} + \epsilon_{\text{new}}] = \text{Var}[\mathbf{x}_{\text{new}}^T \mathbf{w}] + \underbrace{\text{Var}[\epsilon_{\text{new}}]}_{=\beta}$$

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析7

■ 分散の計算

$\mathrm{Var}[\mathbf{x}_{\mathrm{new}}^T \mathbf{w}]$ は、**確率変数 $\mathbf{w}$** の分散（共分散行列 $\mathbf{S}_N$）を用いて計算されます。

一般に、確率変数ベクトル $\mathbf{w}$ の共分散行列を $\boldsymbol{\Sigma}$ とするとき、その線形結合 $\mathbf{a}^T\mathbf{w}$ の分散は以下のようになります。

$$\mathrm{Var}[\mathbf{a}^T\mathbf{w}] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$$

この式に、$\mathbf{a} = \mathbf{x}_{\mathrm{new}}$、$\boldsymbol{\Sigma} = \mathbf{S}_N$（$\mathbf{w}$ の事後共分散行列）を代入します。

$$\mathrm{Var}[\mathbf{x}_{\mathrm{new}}^T \mathbf{w}] = \mathbf{x}_{\mathrm{new}}^T \mathbf{S}_N \mathbf{x}_{\mathrm{new}}$$

この項は、重み $\mathbf{w}$ の推定値に含まれる**不確実性**が、予測にどの程度影響するかを示しています。

$$\sigma^2_{\mathrm{pred}}(\mathbf{x}_{\mathrm{new}}) = \mathrm{Var}[y_{\mathrm{new}}] = \underbrace{\mathrm{Var}[\mathbf{x}_{\mathrm{new}}^T \mathbf{w}]}_{\text{パラメータの不確実性}} + \underbrace{\mathrm{Var}[\epsilon_{\mathrm{new}}]}_{\text{ノイズの分散}}$$

$$\sigma^2_{\mathrm{pred}}(\mathbf{x}_{\mathrm{new}}) = \mathbf{x}_{\mathrm{new}}^T \mathbf{S}_N \mathbf{x}_{\mathrm{new}} + \beta^{-1}$$

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析　まとめ

■ 仮定

- 観測値　$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$　　観測値の確率　$\mathcal{N}(y \mid f(x_n), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}\{y - f(x_n)\}^2}$
- 多項式　$f(x) = \sum_{m=0}^{M} w_m x^m$　　パラメータの事前確率　$P(\mathbf{w}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$

■ 結果

- パラメータの事後分布

$$\mathrm{P}(w|y) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \mathbf{S})$$

平均　$\boldsymbol{\mu} = \mathbf{S}\mathbf{b} = \beta \mathbf{S} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{y}$

分散　$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$

- 観測値の分散

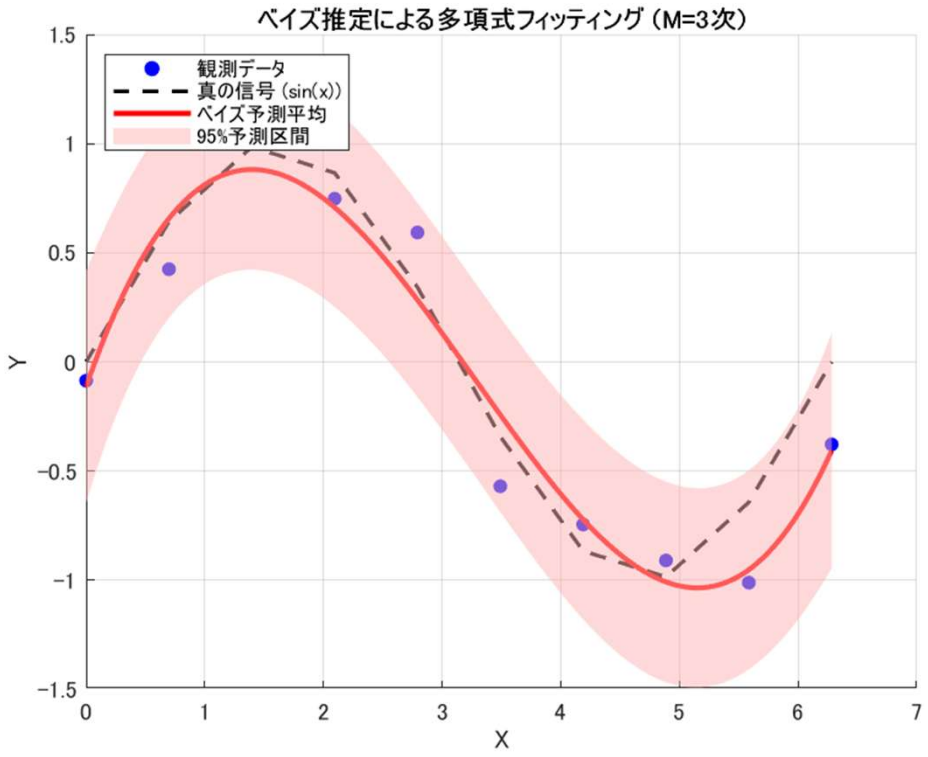
$$\sigma^2_{\mathrm{pred}}(\mathbf{x}_{\mathrm{new}}) = \mathbf{x}^T_{\mathrm{new}} \mathbf{S}_N \mathbf{x}_{\mathrm{new}} + \beta^{-1}$$

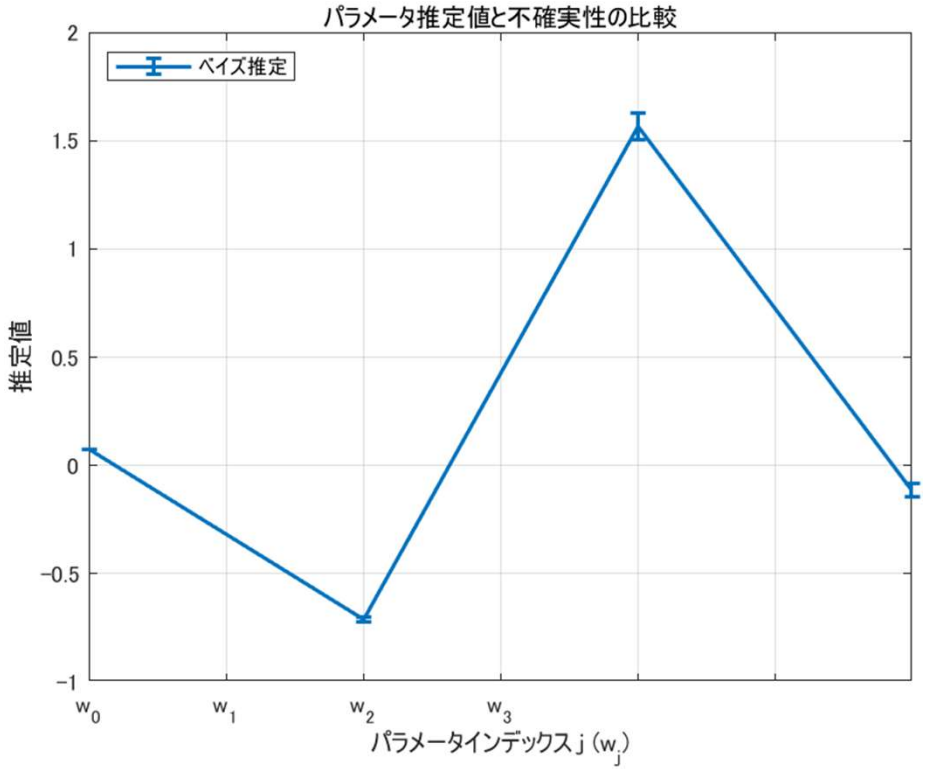
観測値毎に異なる分散になる

© 2020 Yuki Kato. All Rights Reserved

演習：ベイズ推定を用いた回帰分析

- α：1，β：1/0.2^2とした.
- 範囲が分かるのは最尤推定と同じ
- パラメータのばらつきも推定できる





© 2020 Yuki Kato. All Rights Reserved

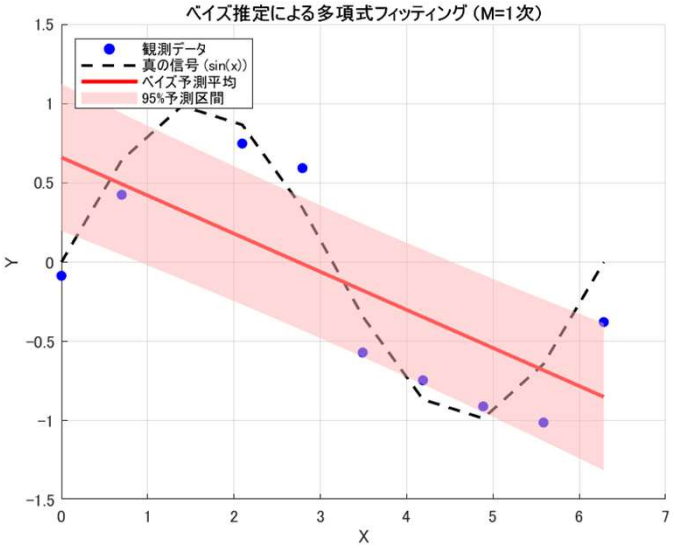
演習：ベイズ推定を用いた回帰分析

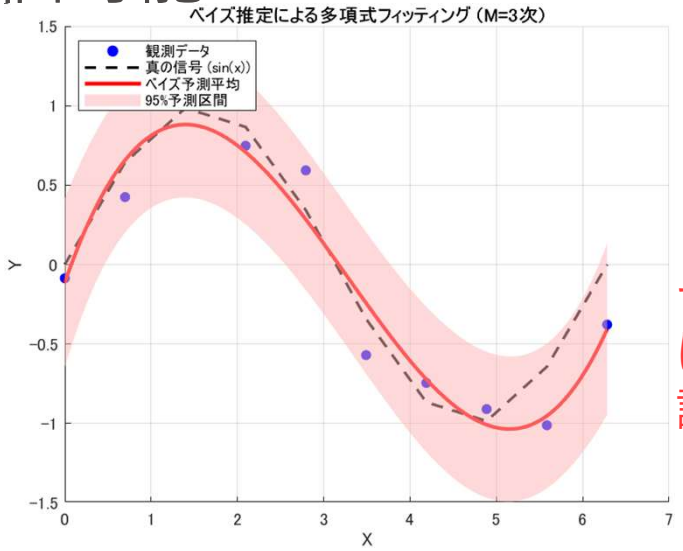
- 1次，5次，11次の多項式で近似しよう

© 2020 Yuki Kato. All Rights Reserved

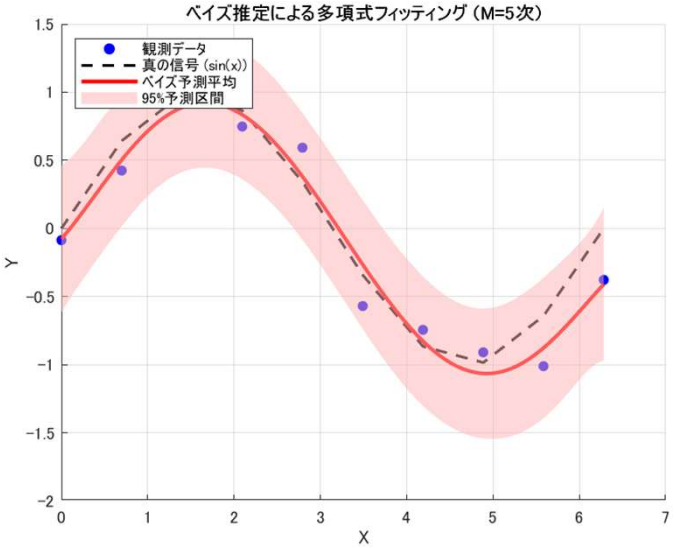
演習：ベイズ推定を用いた回帰分析

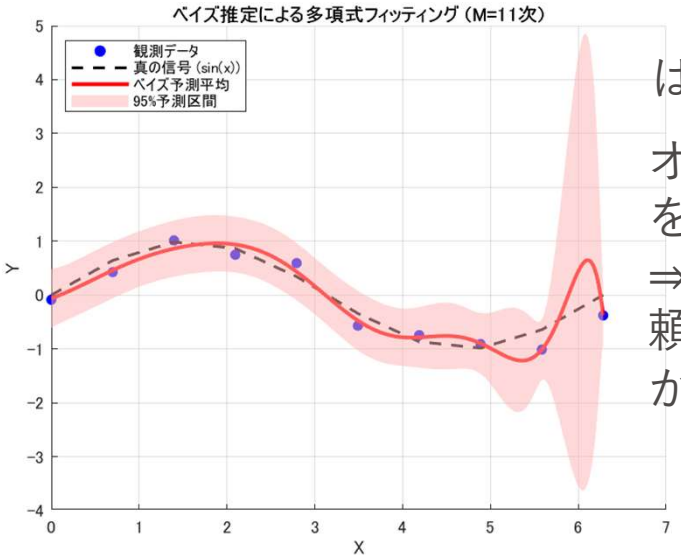
- 1次，5次，11次の多項式で近似しよう
- パラメータのばらつきも評価可能





データごとの
ばらつきが
評価可能





ばらつきを可視化

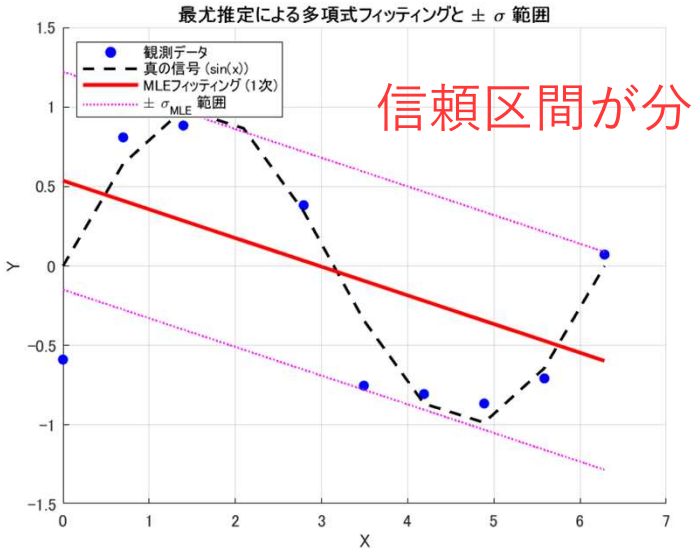
オーバーフィットを抑制
⇒少しするが，信頼区間が大きいのが分かる

© 2020 Yuki Kato. All Rights Reserved

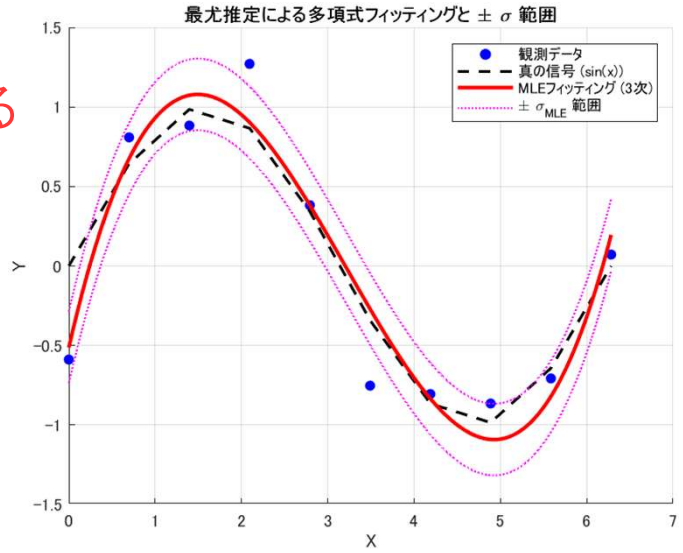
最尤推定の演習

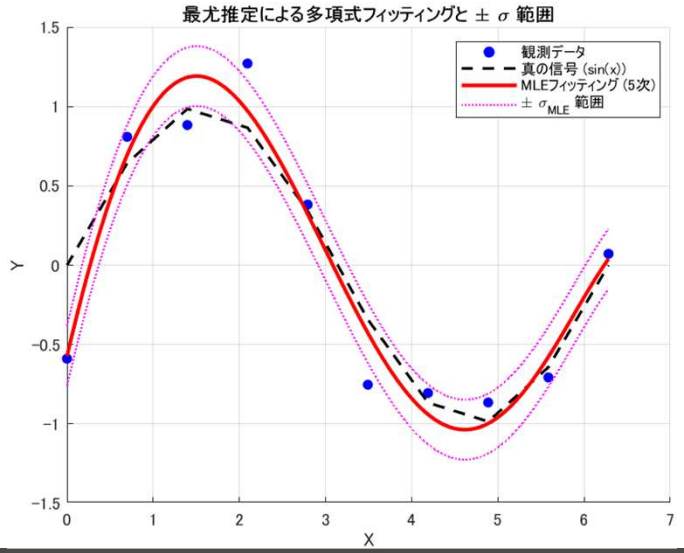
■ Matlabで1次，5次，11次の多項式で最尤推定

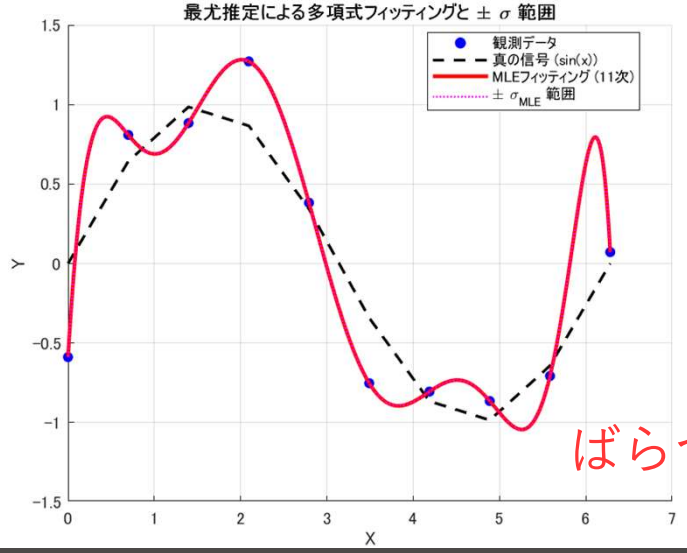
推定値自体はLLSと変わらない…



信頼区間が分かる







11次だと
オーバー
フィット

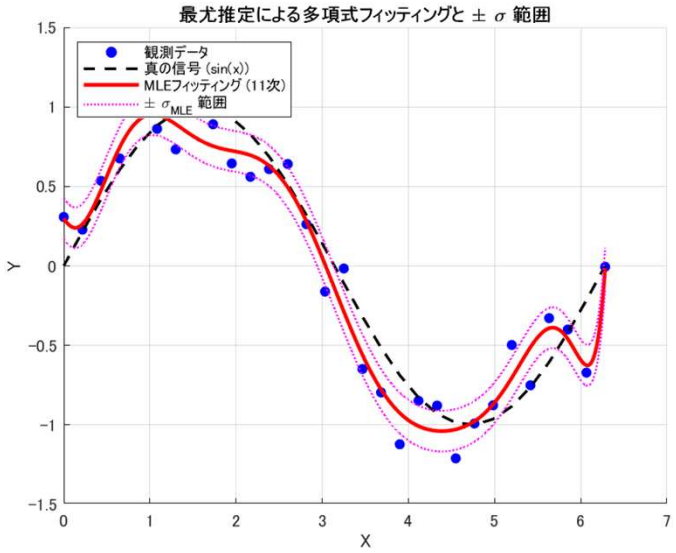
ばらつきが無い・・

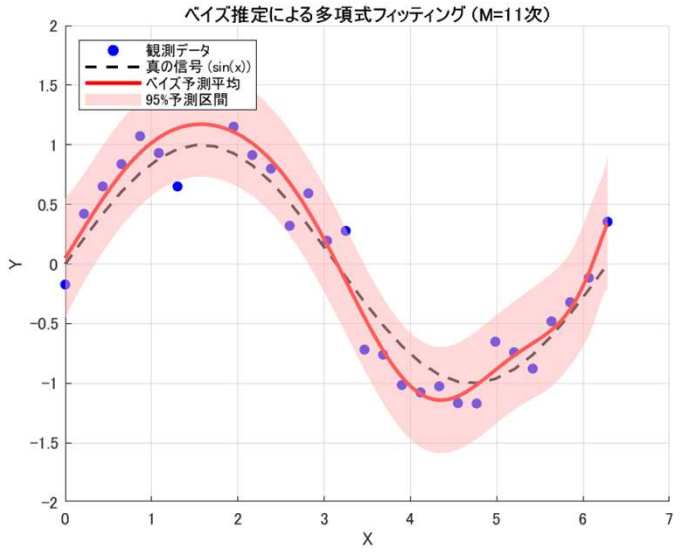
© 2020 Yuki Kato. All Rights Reserved

演習：ベイズ推定を用いた回帰分析

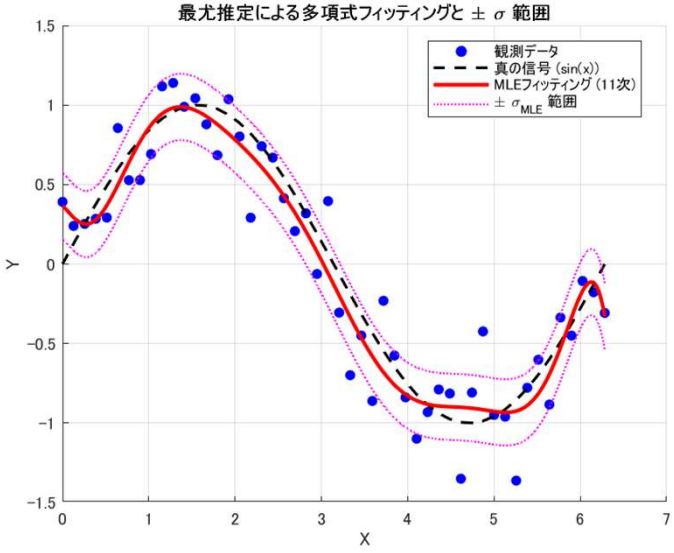
- n数を増やすと，オーバーフィットの改善が分かりやすい
- 但し，事前分布のばらつきを設定する必要がある ⇒変分ベイズ

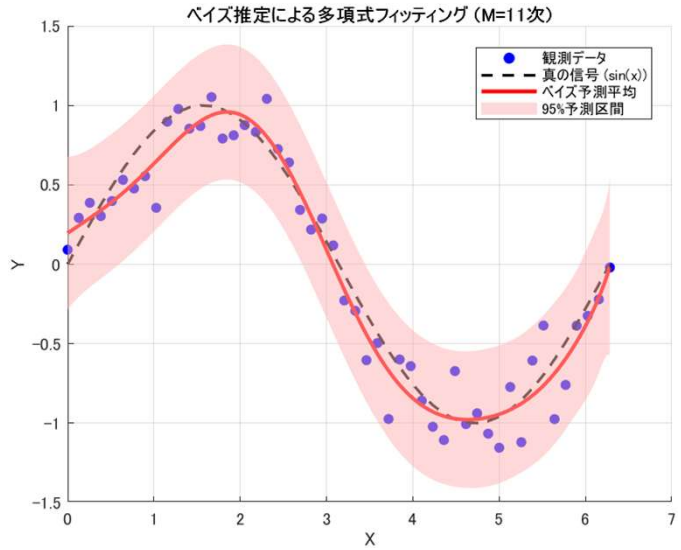
n=30





n=50





© 2020 Yuki Kato. All Rights Reserved

LLS, MLE, Bayesまとめ

手法	基本的な特徴	メリット	デメリット
線形 最小二乗法 (LLS)	観測データとモデルの残差の二乗和を最小化する。 最尤推定の特殊ケース(誤差が正規分布の場合)。	解析解が存在し、計算が高速かつ確実。シンプルで実装が容易。凸最適化問題であり、大域的最適解が保証される。	ノイズの分布に敏感。過学習しやすい(正則化がない場合)。パラメータの不確実性を提供しない。
最尤推定 (MLE)	観測データが出現する尤度(確率)を最大化するパラメータを推定する。	統計モデルの仮定(例:正規分布)の下で、**漸近的に最も効率的**な推定量が得られる. 非常に柔軟で、**任意の確率モデル**に適用可能。	モデルの仮定(例:ノイズ分布)が誤っていると、推定結果も誤る。複雑なモデルでは、解析解がなく、**反復的な数値解法**が必要。
ベイズ推定 (Bayes)	パラメータを確率変数とみなし、事前分布と尤度から事後分布を求める。	不確実性を定量化できる。 **事前知識**(事前分布)を組み込むことで、**過学習を抑制**できる(正則化効果)。	**計算コストが高い**. **事前分布**の設定に主観が入る(適切な設定が難しい場合がある)。⇒変分ベイズ

© 2020 Yuki Kato. All Rights Reserved

© 2020 Yuki Kato. All Rights Reserved

ベイズ推定を組込んだ回帰分析5

- 纏めると，下記はwについての二次関数になっているので，

$$\mathrm{P}(w|y) = \propto \mathrm{Const} \times \exp\left[-\frac{\beta}{2} \sum_{n=1}^{N} \{f(x_n) - y_n\}^2 - \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right]$$

- wについて平方完成すると，下記正規分布になる

$$P(\mathbf{w} \mid \boldsymbol{y}) = \mathcal{N}\left(\mathbf{w} \,\middle|\, \beta \mathrm{S} \sum_{n=1}^{N} y_n \boldsymbol{\phi}(x_n), \mathrm{S} \right)$$

$$\boldsymbol{\mu} = \mathbf{S}\mathbf{b} = \beta \mathbf{S} \Phi^{\mathrm{T}} \mathbf{y}$$

- 多変数の正規分布なので分散Sは共分散行列になり，逆行列が下記で与えられる

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(x) = \begin{pmatrix} x^0 \\ x^1 \\ \vdots \\ x^M \end{pmatrix}$$

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第十一回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布
- 第５回　大数の法則と中心極限定理
- 第６回　時系列データ解析とフィルタ処理1
- 第７回　時系列データ解析とフィルタ処理2，画像解析

10/28 ■ 第８回　画像解析
10/31 ■ 第９回　最小二乗法，回帰分析
11/4 ■ 第１０回　最尤推定，ベイズ推定
11/14 ■ 第１１回　変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ
11/18 ■ 第１２回　スパースモデリング
11/21 ■ 第１３回　多変量データ解析の基礎と主成分分析
11/25 ■ 第１４回　固有値分解，特異値分解
11/28 ■ 第１５回　期末テスト
12/2 ■ 第１６回　返却　⇒返却のみ，補講扱いですが，全員出席とするので任意参加

✓11/14: 課題①②〆，両方とも，講義開始時に集めます
答案の配布と，課題の解説をどこかで実施予定

✓11/21を目途に返却/解答解説を実施予定

© 2020 Yuki Kato. All Rights Reserved

復習：ベイズ推定を組込んだ回帰分析　まとめ

観測ノイズの大きさ等に関連

■ 仮定

- 観測値　$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$　　観測値の確率　$\mathcal{N}(y \mid f(x_n), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}\{y - f(x_n)\}^2}$
- 多項式　$f(x) = \sum_{m=0}^{M} w_m x^m$　　パラメータの事前確率　$P(\mathbf{w}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$

パラメータの不確実性

■ 結果

- パラメータの事後分布

$$\mathrm{P}(w|y) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \mathbf{S})$$

平均　$\boldsymbol{\mu} = \mathbf{S}\,\mathbf{b} = \beta\,\mathbf{S}\,\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}$

分散　$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$

- 観測値の分散

$$\sigma^2_{\mathrm{pred}}(\mathbf{x}_{\mathrm{new}}) = \mathbf{x}^T_{\mathrm{new}}\mathbf{S}_N\mathbf{x}_{\mathrm{new}} + \beta^{-1}$$

観測値毎に異なる分散になる

- コスト関数

$$E = \frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

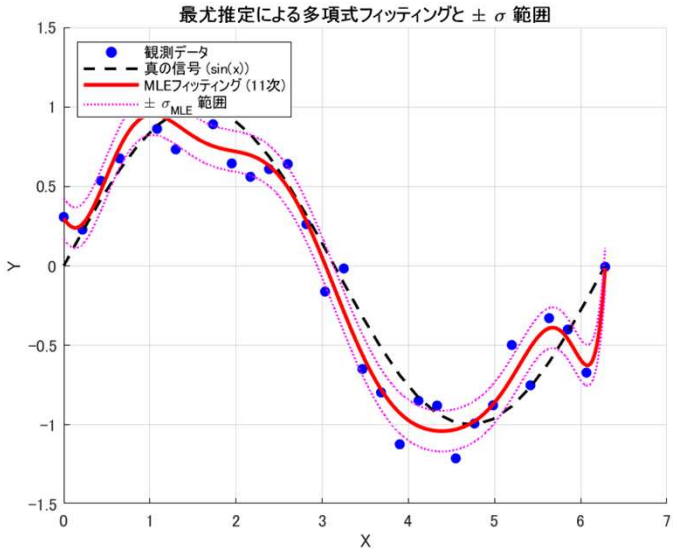
αが大きい程，正則化が効く
⇒オーバーフィットしづらい

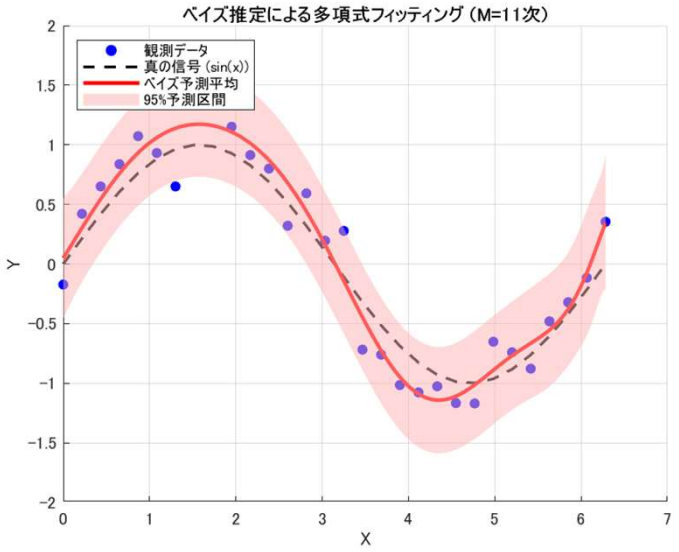
© 2020 Yuki Kato. All Rights Reserved

復習：ベイズ推定を用いた回帰分析

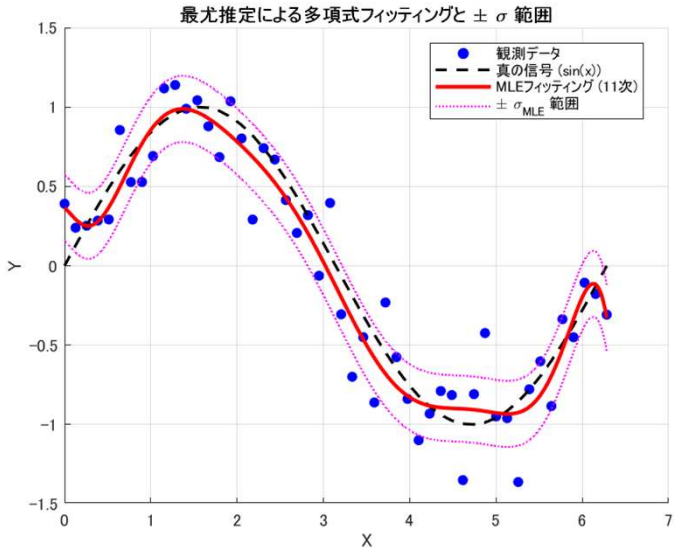
- n数を増やすと，オーバーフィットの改善が分かりやすい
- 但し，事前分布のばらつきを設定する必要がある ⇒変分ベイズ

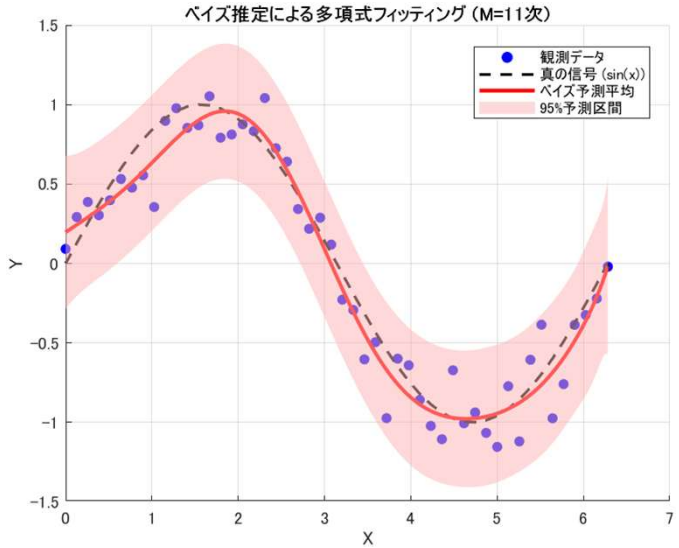
n=30





n=50





© 2020 Yuki Kato. All Rights Reserved

コンテンツ

- 変分ベイズ
- 最尤推定を用いたEMアルゴリズムによる教師無し学習（画像生成器の作成）
- 変分オートエンコーダを用いた教師無し学習（2020~）

	目的関数	変数	上限	推定
EMアルゴリズム	$\ln p(X\|\theta)$ 対数尤度	Z, θ 潜在変数とパラメータは別	$\ln p(X\|\theta)$ 可変値	θ 点推定
変分ベイズ	$\mathrm{KL}[q\|p]$ or $\mathcal{L}[q]$ KLダイバージェンス or 下限	Z 潜在変数にパラメータを含める	$\ln p(X)$ 一定値	$q(Z)$ ベイズ推定
VAE	$\mathrm{KL}[q_\varphi\|p_\theta]$ or $\mathcal{L}[q_\varphi]$ KLダイバージェンス or 下限	Z, θ, φ 潜在変数とパラメータは別	$\ln p_\theta(X)$ 可変値	θ, φ 点推定

Ref: https://academ-aid.com/ml/vae

© 2020 Yuki Kato. All Rights Reserved

変分ベイズ推定とは？

変分ベイズ推定は、先に解説した**解析的なベイズ推定**（共役事前分布に基づく）が適用できない**より複雑なモデル**で、事後分布を近似的に求めるために広く使われる手法です。

数値計算

線形回帰のような比較的シンプルなモデルでは、VBの結果は解析解（共役事前分布に基づく解）とほぼ一致しますが、VBの枠組みは**ハイパーパラメータ（α, β）もデータから自動で推定できる**という点で優れています。

1. 変分ベイズの基本戦略

VBは、真の事後分布 $P(\mathbf{w}|\mathbf{D})$ を、計算が容易な**単純な分布** $Q(\mathbf{w})$ で近似することを目指します。

ベイズでは解析的に出した

解はないので，反復計算を実施
判断基準が必要だが・・・

$$\text{真の事後分布 } P(\mathbf{w}|\mathbf{D}) \approx \text{近似分布 } Q(\mathbf{w})$$

この近似の良し悪しを測るために、**カルバック・ライブラー情報量 (KL Divergence)** $KL(Q||P)$ を最小化します。KLダイバージェンスの最小化は、**ELBO (Evidence Lower Bound)** と呼ばれる量を最大化することと等価です。

© 2020 Yuki Kato. All Rights Reserved

変分ベイズ推定の手順

2. ベイズ線形回帰への適用

変分ベイズを線形回帰に適用する際、通常、重み $\mathbf{w}$ とノイズの精度 β（または分散 σ^2）に加え、重みの事前分布の精度 α も含めた**すべての未知のパラメータ**を推定します。

A. 分布の仮定 (平均場近似)

VBでは、すべてのパラメータ $Z = \{\mathbf{w}, \alpha, \beta\}$ の真の事後分布を、独立な要素の積として近似します（**平均場近似**）。

$$Q(\mathbf{w}, \alpha, \beta) \approx Q_{\mathbf{w}}(\mathbf{w}) Q_{\alpha}(\alpha) Q_{\beta}(\beta)$$

こうなる保証はない
あくまで近似

変分ベイズでは、すべてのパラメータの**期待値**を計算し、その期待値を使って他のパラメータを更新するという手順を繰り返します。この実装は、以下の3つのステップの繰り返しが核心です。

- **Step 1 ($Q_{\mathbf{w}}$ の更新)**: 他のパラメータ（α, β）の現在の推定値を使って、重み $\mathbf{w}$ の事後分布（平均 $\mathbf{m}_N$ と共分散 $\mathbf{S}_N$）を更新します。これは解析的ベイズの計算とほぼ同じ形です。
- **Step 2 (Q_{α} の更新)**: $\mathbf{w}$ の事後平均と分散（$\mathbf{m}_N, \mathbf{S}_N$）を使って、重みの事前精度 α の期待値を更新します。
- **Step 3 (Q_{β} の更新)**: $\mathbf{w}$ の事後平均と分散を使って、ノイズ精度 β の期待値を更新します。

© 2020 Yuki Kato. All Rights Reserved

変分ベイズ推定の手順2

精度パラメータ α（重みの不確実性）と β（ノイズレベル）は、VBの枠組みでは**ガンマ分布** $Q(\cdot) = \mathrm{Gamma}(a, b)$ で近似されます。このステップでは、ガンマ分布のパラメータ a と b を更新し、その期待値 $E[\cdot] = a/b$ を次の反復で使用します。

共役性

A. 重みの事前精度 α の更新

α は**重み $\mathbf{w}$ の大きさ**を制御します。

M:多項式の次元

- **更新式**: α の新しい期待値 $E[\alpha]_{\text{new}}$ は、以下に依存します。

$$a_\alpha = a_{0,\alpha} + \frac{M}{2}$$

$$E[\alpha]_{\text{new}} = \frac{a_\alpha}{b_\alpha} \quad \text{ただし、} b_\alpha \propto E[\mathbf{w}^T\mathbf{w}]$$

$$b_\alpha = b_{0,\alpha} + \frac{1}{2}\langle \mathbf{w}^T\mathbf{w} \rangle$$

  ここで、$E[\mathbf{w}^T\mathbf{w}]$ は $\mathbf{w}$ の二乗和の期待値で、現在の $\mathbf{m}_N$ と $\mathbf{S}_N$ を使って計算されます。

- **意味**: **重み $\mathbf{w}$ の推定値が大きくなる**（$\mathbf{m}_N$ や $\mathbf{S}_N$ が大きくなる）と、分母 b_α が大きくなり、$E[\alpha]$ は**小さく**なります。
  - $E[\alpha]$ が小さい $\Longrightarrow$ 重みの事前分散 $1/E[\alpha]$ が**大きい** $\Longrightarrow$ **正則化が弱くなる**。
  - モデルがデータにフィットするために大きな重みが必要な場合、VBは自動的に正則化を弱める方向に α を調整します。

© 2020 Yuki Kato. All Rights Reserved

変分ベイズ推定の手順3

$$b_\beta = b_{0,\beta} + \frac{1}{2}\langle ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 \rangle$$

$$a_\beta = a_{0,\beta} + \frac{N}{2}$$

N:データ点数

β は**残差の二乗和**（ノイズレベル）を制御します。

- **更新式**: β の新しい期待値 $E[\beta]_{\text{new}}$ は、以下に依存します。

$$E[\beta]_{\text{new}} = \frac{a_\beta}{b_\beta} \quad \text{ただし、} b_\beta \propto E[||\mathbf{y} - \mathbf{X}\mathbf{w}||^2]$$

ここで、$E[||\mathbf{y} - \mathbf{X}\mathbf{w}||^2]$ は**残差二乗和の期待値**であり、$\mathbf{m}_N$ と $\mathbf{S}_N$ を使って計算されます。

- **意味**: **残差が大きく**なる（フィッティングが悪い、またはデータノイズが大きい）と、分母 b_β が大きくなり、$E[\beta]$ は**小さく**なります。
  - $E[\beta]$ が小さい $\Longrightarrow$ ノイズ分散 $1/E[\beta]$ が**大きい**。
  - モデルがデータにフィットしきれない場合、VBは自動的にノイズレベルが大きいと判断し、**ノイズの分散を大きく**する方向に β を調整します。

このように、VBでは、**重みの推定値と不確実性の情報**を用いて、**正則化の強さ**（α）と**ノイズの大きさ**（β）という、モデルの根幹に関わるハイパーパラメータが**自動的かつデータ駆動的**に更新されます。

⇒あとは，反復を止める評価基準があれば良い

© 2020 Yuki Kato. All Rights Reserved

KLダイバージェンスとは？

KLダイバージェンスは、2つの確率分布 P と Q が**どれだけ異なっているか**を示す尺度です。変分ベイズでは、**近似分布 Q が真の事後分布 P にどれだけ近いか**を測るために使用されます。

定義と数式

離散確率分布 $P(\mathbf{z})$ と $Q(\mathbf{z})$ に対するKLダイバージェンス $D_{KL}(Q||P)$ は以下のように定義されます。

$$D_{KL}(Q||P) = \sum_{\mathbf{z}} Q(\mathbf{z}) \log \left(\frac{Q(\mathbf{z})}{P(\mathbf{z})} \right)$$

連続分布の場合は、総和 $\sum$ が積分 $\int$ に変わります。

特徴

- **非負性**: 常に $D_{KL}(Q||P) \geq 0$ です。2つの分布が完全に一致するとき、つまり $Q(\mathbf{z}) = P(\mathbf{z})$ のときに限りゼロになります。
- **非対称性**: $D_{KL}(Q||P) \neq D_{KL}(P||Q)$ です。これは「距離」ではなく「情報の損失」を測る指標であるため、**対称性はありません**。変分ベイズでは、常に**$D_{KL}(Q||P)$ を最小化**します。
- **目的**: $D_{KL}(Q||P)$ を最小化することは、**近似分布 Q を真の事後分布 P に近づける**ことを意味します。

© 2020 Yuki Kato. All Rights Reserved

ELBO（Evidence Lower Bound）との関係

変分ベイズの目標は、真の事後分布 $P(\mathbf{w}|\mathbf{D})$ を、簡単な近似分布 $Q(\mathbf{w})$ で近似すること、すなわち $D_{KL}(Q||P)$ を最小化することです。

しかし、真の事後分布 $P(\mathbf{w}|\mathbf{D})$ は分母に**周辺尤度** $P(\mathbf{D})$ が含まれるため、計算が非常に困難です。

ベイズの定理の対数を取ると、以下の関係式が導かれます。

$$\log P(\mathbf{D}) = \mathcal{L}(Q) + D_{KL}(Q||P)$$

- $\log P(\mathbf{D})$: **対数周辺尤度 (Log Marginal Likelihood / Model Evidence)**。これはデータ $\mathbf{D}$ の確率であり、モデルの良さを示す**定数**です。
- $D_{KL}(Q||P)$: **KLダイバージェンス**。
- $\mathcal{L}(Q)$: **ELBO (Evidence Lower Bound)**。

ELBOの定義と意味

上記の式を変形すると、ELBOの定義が得られます。

$$\mathcal{L}(Q) = \log P(\mathbf{D}) - D_{KL}(Q||P)$$

© 2020 Yuki Kato. All Rights Reserved

ELBO（Evidence Lower Bound）との関係

この式の持つ意味が、ELBOの名の由来です。

1. **Lower Bound（下限）**：$D_{KL} \geq 0$ であるため、$\mathcal{L}(Q) \leq \log P(\mathbf{D})$ が常に成り立ちます。つまり、ELBO $\mathcal{L}(Q)$ は、計算困難なモデルの良さ $\log P(\mathbf{D})$ の**下限**（Lower Bound）を与えます。
2. **Evidence**: $\log P(\mathbf{D})$ はモデルの**エビデンス**（証拠）とも呼ばれます。

変分ベイズの最適化原理

$\log P(\mathbf{D})$ はデータ $\mathbf{D}$ に依存する**定数**なので、ELBO $\mathcal{L}(Q)$ を最大化することは、自動的に**KLダイバージェンス** $D_{KL}(Q||P)$ **を最小化**することにつながります。

$$\max_Q \mathcal{L}(Q) \iff \min_Q D_{KL}(Q||P)$$

したがって、変分ベイズ推定は、「**計算可能なELBOを最大化すること**を通じて、**近似分布** Q **を真の事後分布** P **に最も近づける**」最適化手法であると言えます。

© 2020 Yuki Kato. All Rights Reserved

ELBO（Evidence Lower Bound）の計算

$$D_{KL}(Q||P) = \sum_{\mathbf{z}} Q(\mathbf{z}) \log \left(\frac{Q(\mathbf{z})}{P(\mathbf{z})} \right)$$

1. ELBOの計算 (Step 4)

このセクションで、変分ベイズの目的関数であるELBOを構成する4つの項を計算しています。

$$\mathcal{L}(\mathbf{Q}) = \text{期待対数尤度} - \mathbf{KL}_{\mathbf{w}} - \mathbf{KL}_{\alpha} - \mathbf{KL}_{\beta}$$

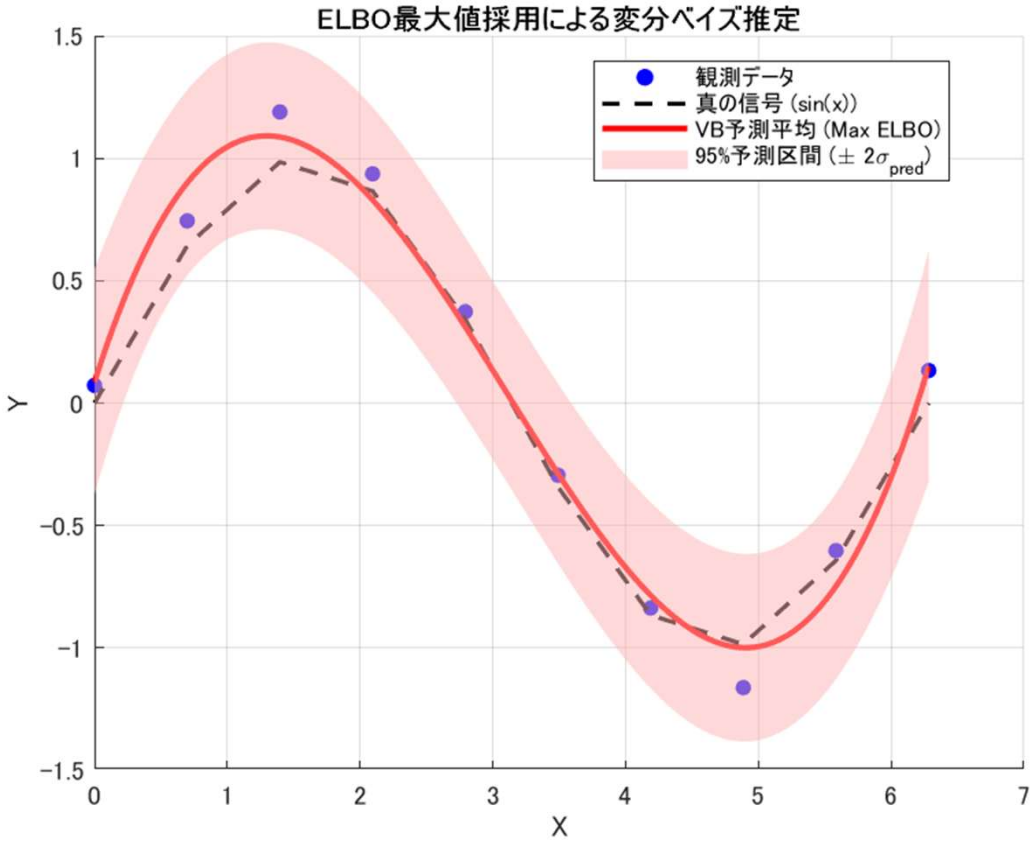
ただし、プログラム内では便宜上、$\text{ELBO} = (\text{期待対数尤度}) + (\mathrm{E}[\ln P(\mathbf{w}, \alpha, \beta)] - \mathrm{E}[\ln Q(\mathbf{w}, \alpha, \beta)])$ の形で計算しています。

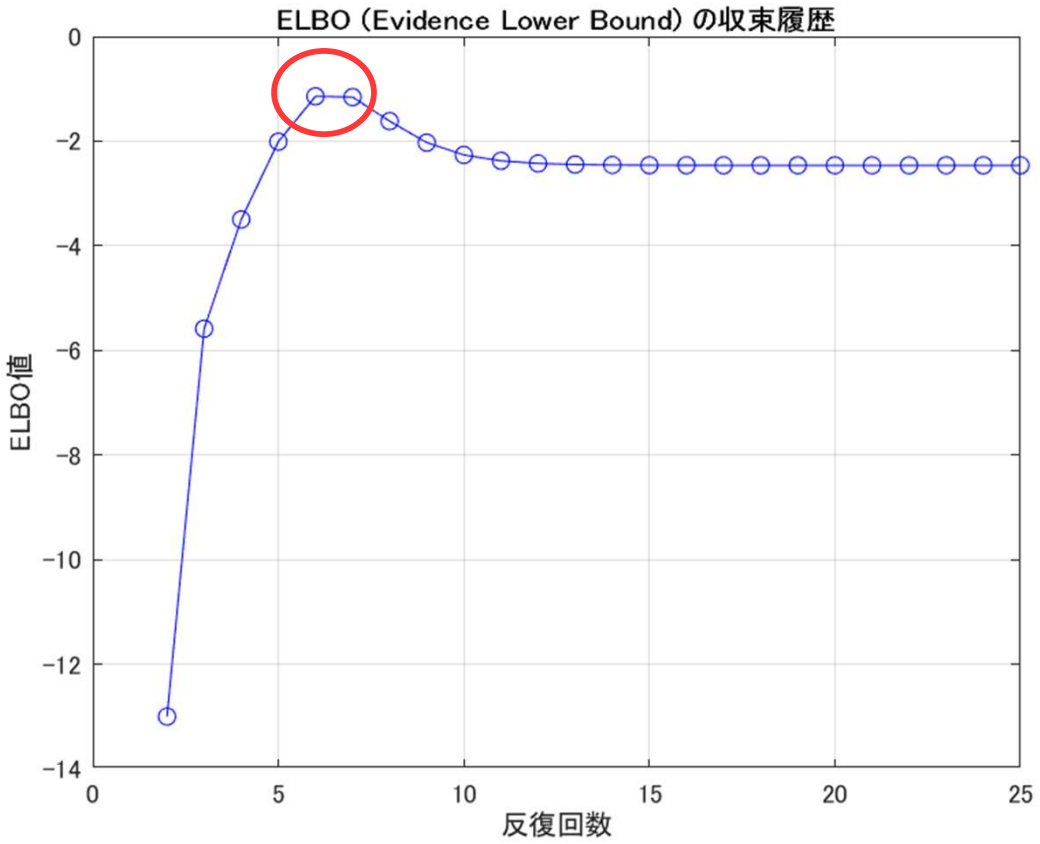
- **期待対数尤度 (`L_lik`)**: データ $\mathbf{y}$ がモデルによって説明される度合いを示します。
- **重み $\mathbf{w}$ の項 (`L_w`)**: 事前分布 $P(\mathbf{w})$ と近似事後分布 $Q_{\mathbf{w}}(\mathbf{w})$ の近さを測ります。この項は、$\mathbf{w}$ のKLダイバージェンスと、$\mathbf{w}$ の事前分布の期待値の差に相当します。
- **精度 α, β の項 (`L_alpha`, `L_beta`)**: ハイパーパラメータ α と β の事前分布と近似事後分布（ガンマ分布）の近さを測ります。ガンマ分布のKLダイバージェンスには、$\log \Gamma(\cdot)$ の計算が必要になります。

© 2020 Yuki Kato. All Rights Reserved

演習：変分ベイズ推定を用いた回帰分析

- パラメータが未知でも計算可能（初期値は必要）
- ELBOによる収束評価が可能





© 2020 Yuki Kato. All Rights Reserved

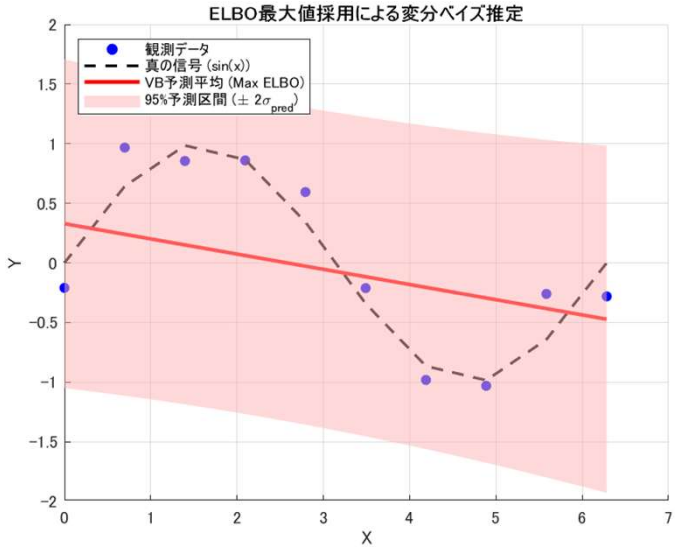
演習：変分ベイズ推定を用いた回帰分析

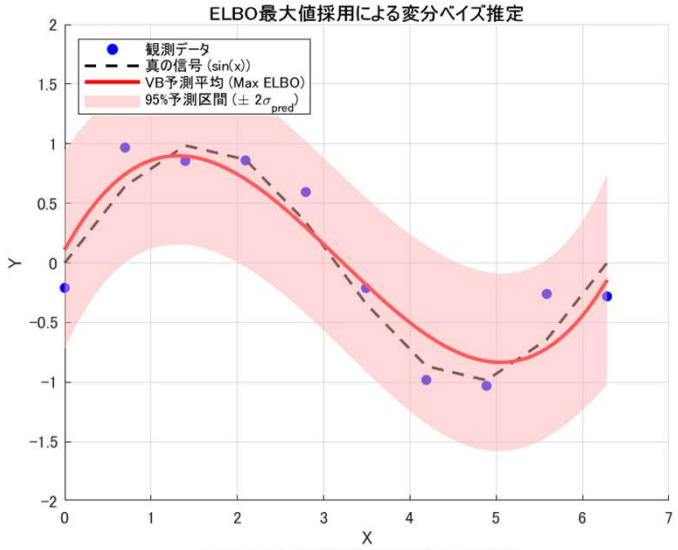
■ 1次，5次，11次の多項式で近似しよう

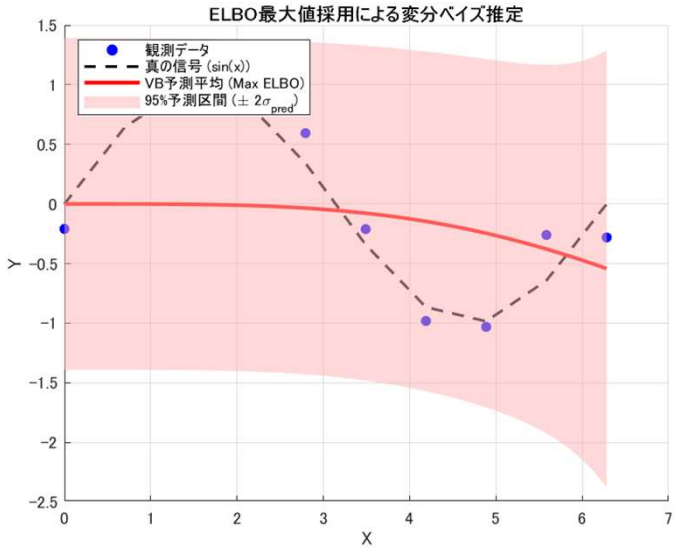
© 2020 Yuki Kato. All Rights Reserved

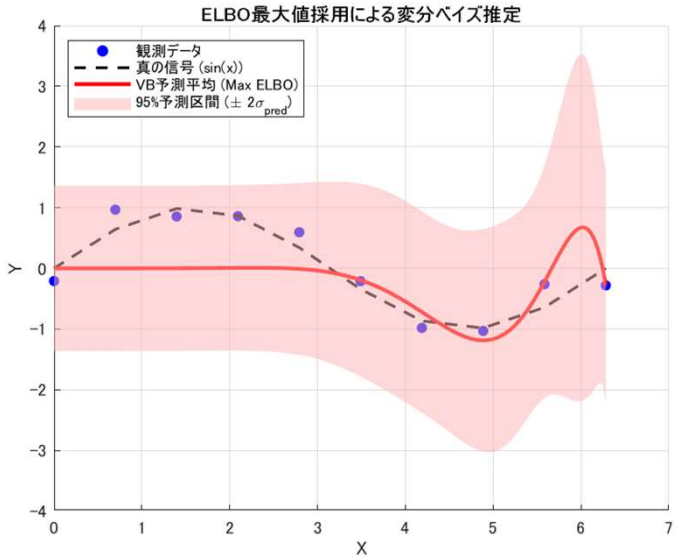
演習：変分ベイズ推定を用いた回帰分析

- 1次，5次，11次の多項式で近似しよう
- データ数が少ないとうまくいかない場合も・・





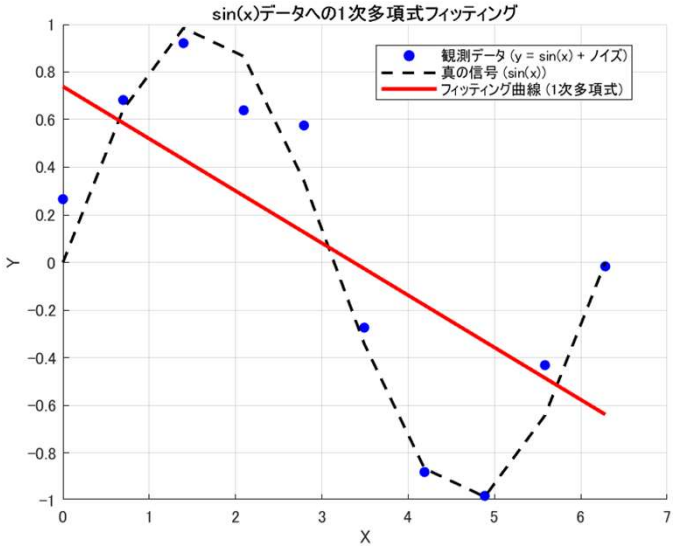


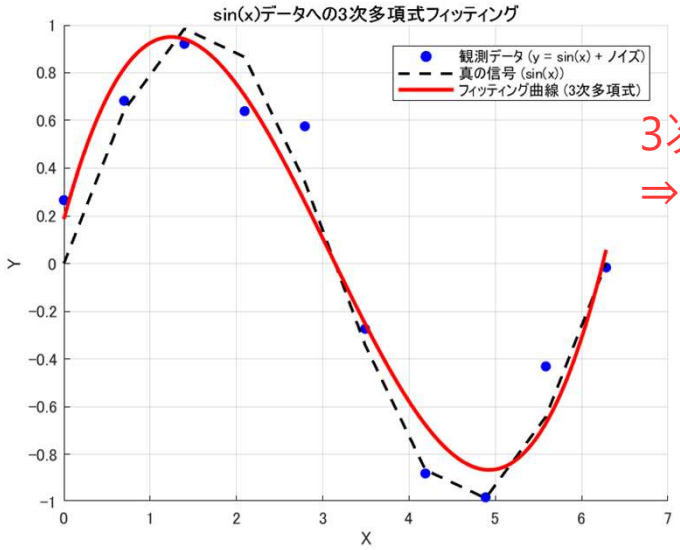


© 2020 Yuki Kato. All Rights Reserved

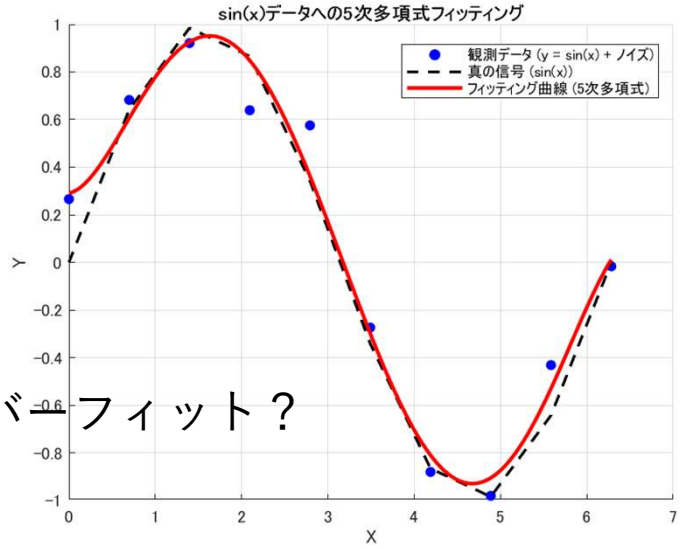
線形最小二乗法：演習

- Matlabで観測値を三次多項式で近似するプログラムを実行
- 1次，5次，11次の多項式ならどうなる？

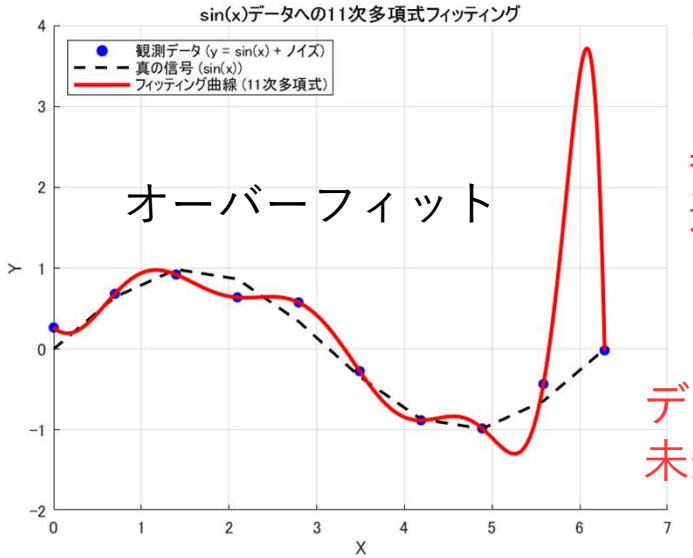




3次が一番いい？
⇒数値実験で検証



オーバーフィット？



11次は警告が出る
⇒毎回結果が違う可能性がある,不安定

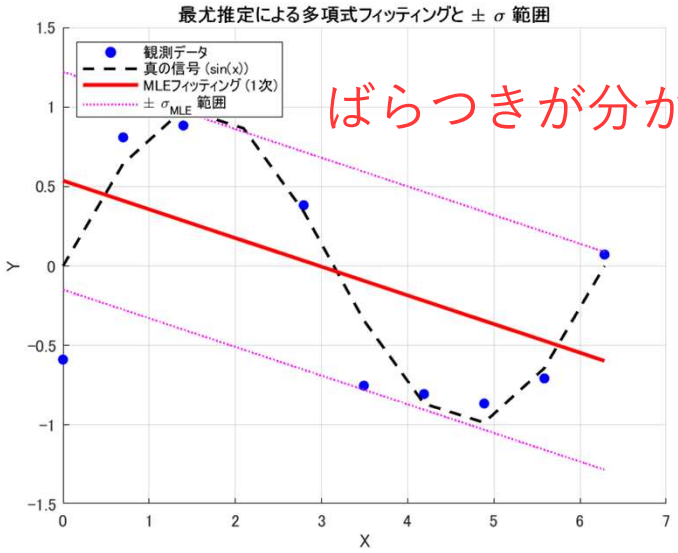
データ：10個
未知変数：12個

© 2020 Yuki Kato. All Rights Reserved

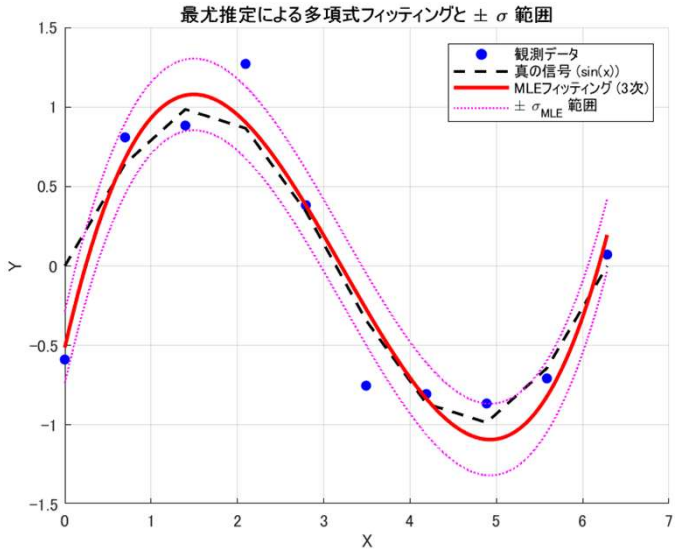
最尤推定の演習

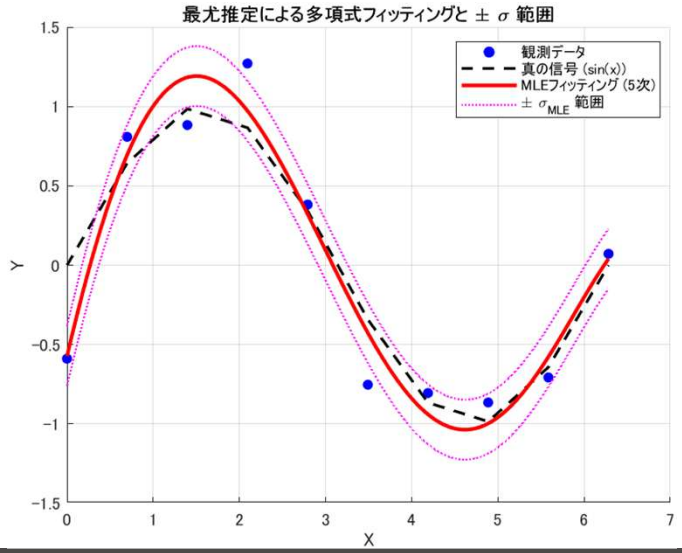
■ Matlabで1次，5次，11次の多項式で最尤推定

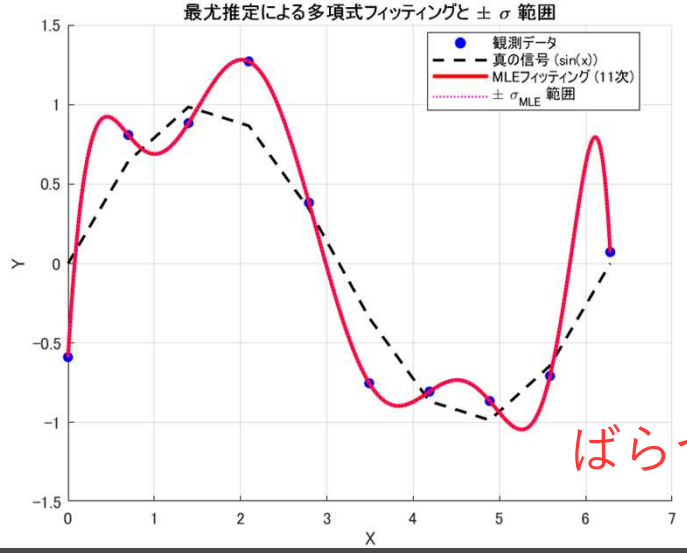
推定値自体はLLSと変わらない…



ばらつきが分かる！







11次だと
オーバー
フィット

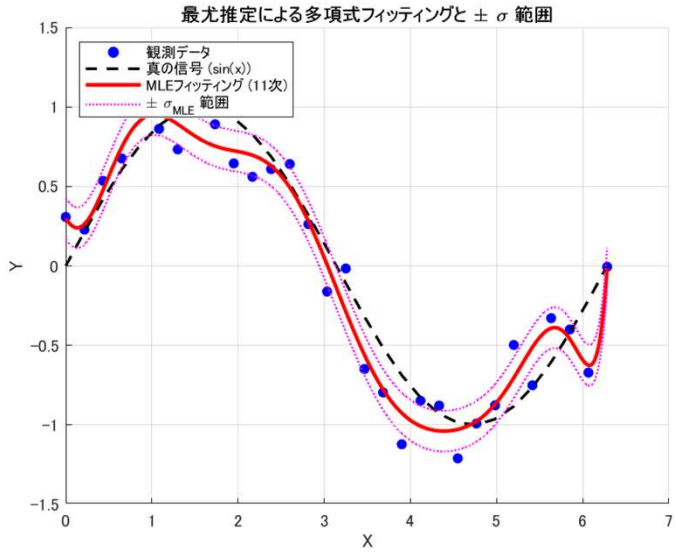
ばらつきが無い・・

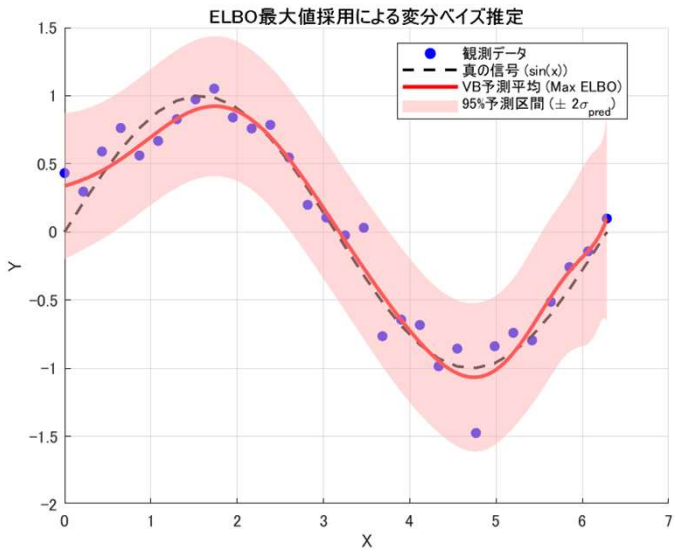
© 2020 Yuki Kato. All Rights Reserved

演習：変分ベイズ推定を用いた回帰分析

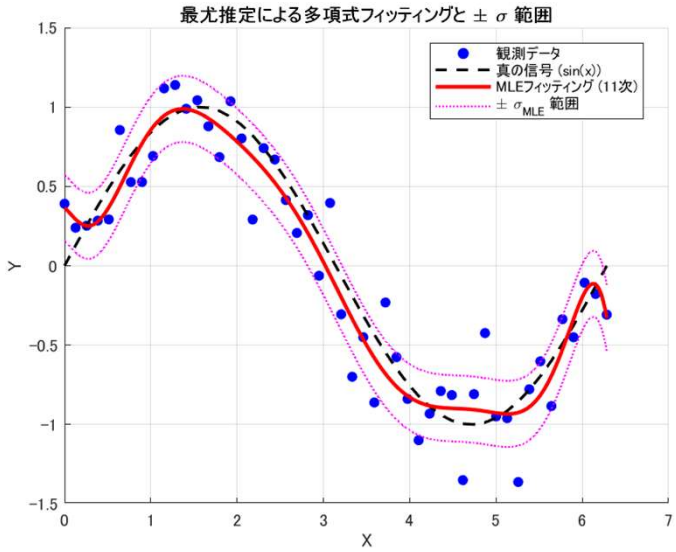
- n数を増やすと良好な推定
- データ数が重要

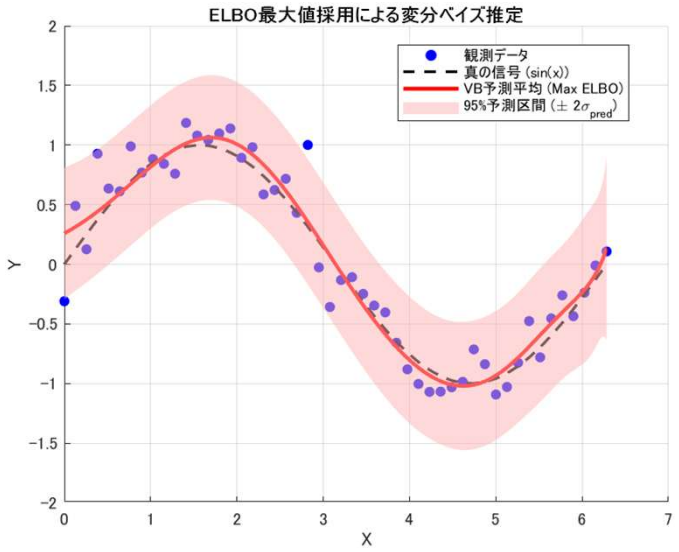
n=30





n=50





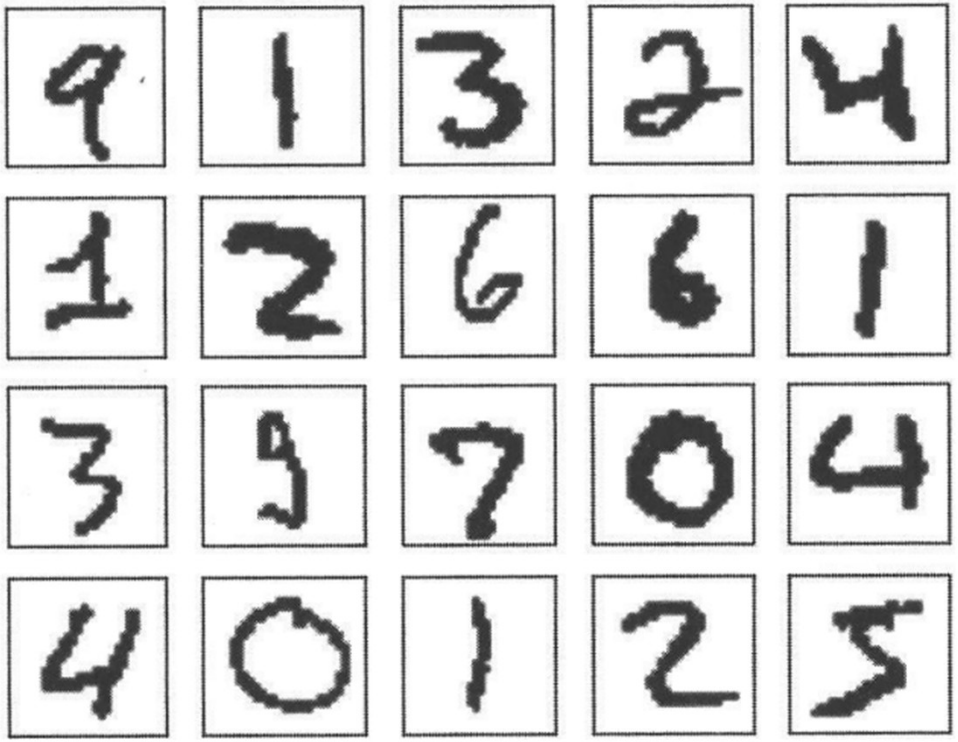
© 2020 Yuki Kato. All Rights Reserved

画像学習

© 2020 Yuki Kato. All Rights Reserved

最尤推定を用いた画像学習（手書き文字）

- 下図のような多数の「手書き数字」（モノクロ２階調）のデータが与えられた際 に、これらを自動的に分類する方法を考える
- まずは，それぞれの手書き数字 を代表する「手書き文字生成器」を用意するとを考える



© 2020 Yuki Kato. All Rights Reserved

手書き文字の平均値学習

- まず，手書き文字の画像に含まれるピクセルを横一列に並べて，各ピクセルの色を0と1の数値で表したベクトル$\boldsymbol{x}_n$を用意する(n番目の画像)
- 第i成分$[\boldsymbol{x}_n]_i$を見ると，i番目のピクセルの色が分かる
- 全ての画像を重ね合わせて平均を取ると，下記の通り

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

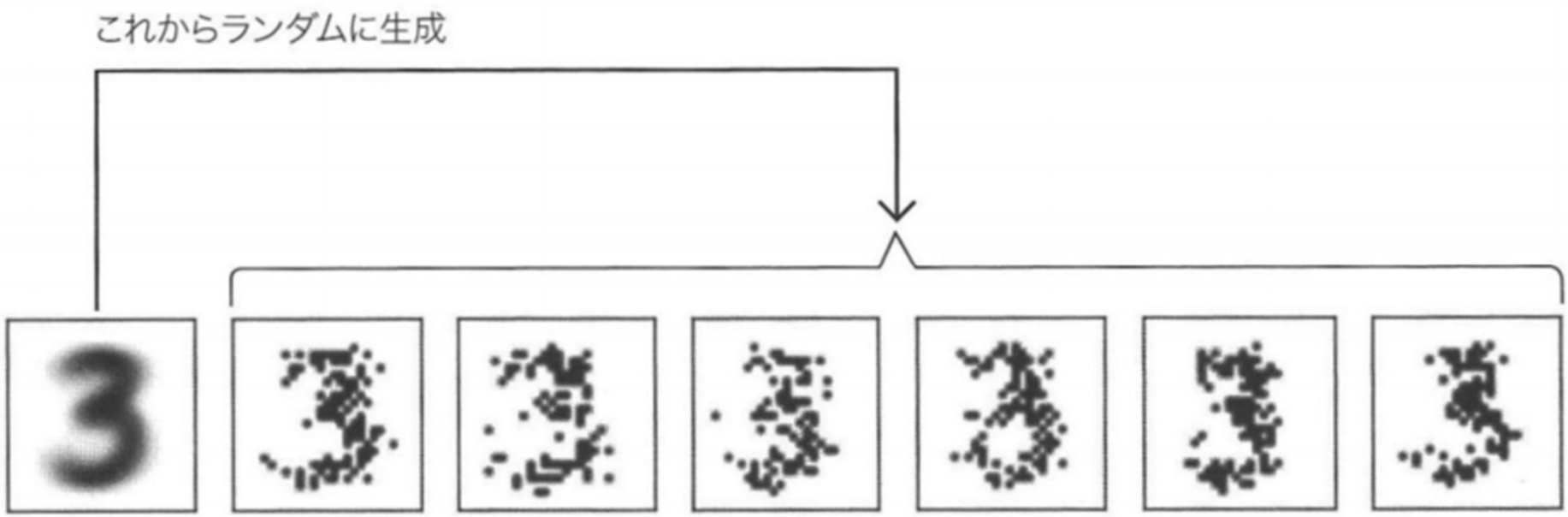
これを画像生成器にしても良いが，
理論的な根拠が欠けている
⇒最尤推定を用いて実施



© 2020 Yuki Kato. All Rights Reserved

「手書き文字生成器」の最尤推定

- 最尤推定には尤度（結果が得られる確率）を導く必要がある
- そこで，手書き文字をランダムに生成する画像生成器を用意
- 各ピクセルの濃淡を0-1の実数値で表したベクトル$\boldsymbol{\mu}$
- この$\boldsymbol{\mu}$の各成分の値を「対応するピクセルが黒になる確率」と考える
- ⇒ $\boldsymbol{\mu}$から下記の様に手書き文字風の画像を生成できる



© 2020 Yuki Kato. All Rights Reserved

「手書き文字生成器」の最尤推定2

- 手書き画像のデータがN個あるとする
- そして，画像生成器を用意して，N個の画像を生成した時，「トレーニングセットとまったく同じデータ群が生成される確率」を考える⇒相当確率が低そうだが，確率は導かれる
- この確率を最大になるような画像生成器を見つければよい

i番目のピクセルについて

$x_i = 1$の場合：$p_i = \mu_i$

$x_i = 0$の場合：$p_i = 1 - \mu_i$

$$p_i = \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

ベルヌーイ分布

$$p(\mathbf{x}) = \prod_{i=1}^{D} p_i = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

一枚の画像において，
全てのピクセルが同じ色になる確率

N枚分

$$P = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N}\prod_{i=1}^{D} \mu_i^{[\mathbf{x}_n]_i}(1-\mu_i)^{1-[\mathbf{x}_n]_i}$$

尤度関数

© 2020 Yuki Kato. All Rights Reserved

「手書き文字生成器」の最尤推定3

■ 最尤推定を実施する

$$P=\prod_{n=1}^{N}p(\mathbf{x}_n)=\prod_{n=1}^{N}\prod_{i=1}^{D}\mu_i^{[\mathbf{x}_n]_i}(1-\mu_i)^{1-[\mathbf{x}_n]_i}$$

尤度関数

■ 対数尤度関数の最大化をしたいので

$$\ln P=\sum_{n=1}^{N}\sum_{i=1}^{D}\{[\mathbf{x}_n]_i\ln\mu_i+(1-[\mathbf{x}_n]_i)\ln(1-\mu_i)\}$$

これをμ_iで偏微分すると、次が得られます。

$$\frac{\partial(\ln P)}{\partial\mu_i}=\sum_{n=1}^{N}\left(\frac{[\mathbf{x}_n]_i}{\mu_i}-\frac{1-[\mathbf{x}_n]_i}{1-\mu_i}\right)$$

対数尤度の極大値

これが0になるという条件から、μ_iは次のように決まります。

$$\mu_i=\frac{1}{N}\sum_{n=1}^{N}[\mathbf{x}_n]_i$$

ベルヌーイ分布
を用いた最尤推定

⇒平均値と一致

教師無学習の初歩

© 2020 Yuki Kato. All Rights Reserved

混合分布を用いた最尤推定

- 次のステップとして，複数の数字の手書き画像データが混在している場合に，画像データを分類する方法を考える
- K種類の数字を含むデータセットがあるとするとし，画像生成器を $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ とする
- 特性の画像生成器$\boldsymbol{\mu}_k$から画像 x が得られる確率は

$$p_{\boldsymbol{\mu}_k}(\mathbf{x}) = \prod_{i=1}^{D} [\boldsymbol{\mu}_k]_i^{x_i} (1 - [\boldsymbol{\mu}_k]_i)^{1-x_i}$$

- どの画像生成器を使うかによって確率が変わるので，これを考慮する．k番目の画像生成器を選ぶ確率を$\boldsymbol{\pi}_k$とすると

$$\sum_{k=1}^{K} \pi_k = 1$$

- よって，特定の画像xが得られる確率は

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_{\boldsymbol{\mu}_k}(\mathbf{x})$$

混合ベルヌーイ分布

© 2020 Yuki Kato. All Rights Reserved

混合分布を用いた最尤推定2

- 最後に，データ数をNとしたとき，N個の画像がトレーニングセットのデータ群と一致する確率は

$$P = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p_{\mu_k}(\mathbf{x}_n)$$

尤度関数

- これを最大化するパラメータを学習すれば良い
- しかし，積（Π）と和（Σ）が混在しているので，対数尤度にしても解析的に解けない
- ⇒EMアルゴリズムを用いて数値計算で解く

© 2020 Yuki Kato. All Rights Reserved

混合分布を用いた最尤推定　EMアルゴリズム

- どれか一つの画像生成器をランダムに選択し，新しい画像を生成するという操作を行うと，画像$\boldsymbol{x}_n$が得られる確率は

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k p_{\mu_k}(\mathbf{x}_n)$$

- 上式はすべての画像生成器が選択される可能性を考慮
- いま，ｋ番目の画像生成器から画像$\boldsymbol{x}_n$が得られる可能性を取り出すと

期待値（expectation, E）ステップ

$$\gamma_{nk} = \frac{\pi_k p_{\mu_k}(\mathbf{x}_n)}{\sum_{k'=1}^{K} \pi_{k'} p_{\mu_{k'}}(\mathbf{x}_n)}$$

Ｎ個の画像を考慮して・・・

- これを基に新たな画像生成器と選択確率を作りなおす

最大化（maximization, M）ステップ

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} \qquad \pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

これを繰り返せば，尤度関数が大きくなり極大値に達する

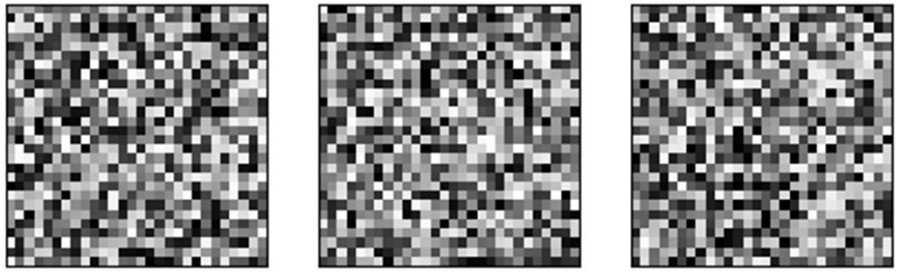
初期値によっては最大値を取らない事に注意が必要

© 2020 Yuki Kato. All Rights Reserved

0，3，6を対象にEMアルゴリズムを可視化

N=600枚

μ_k :



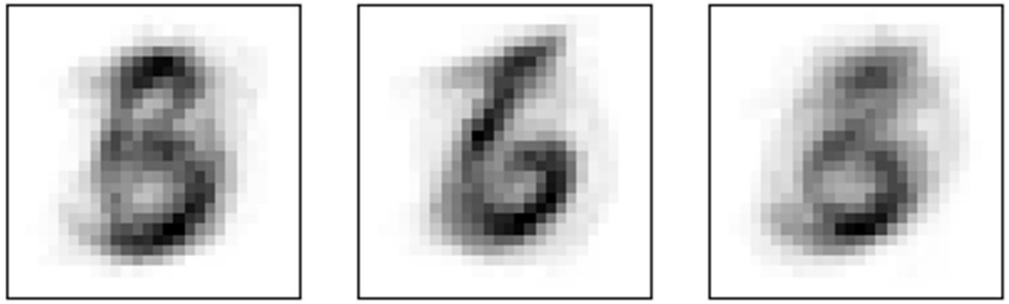
期待値ステップ

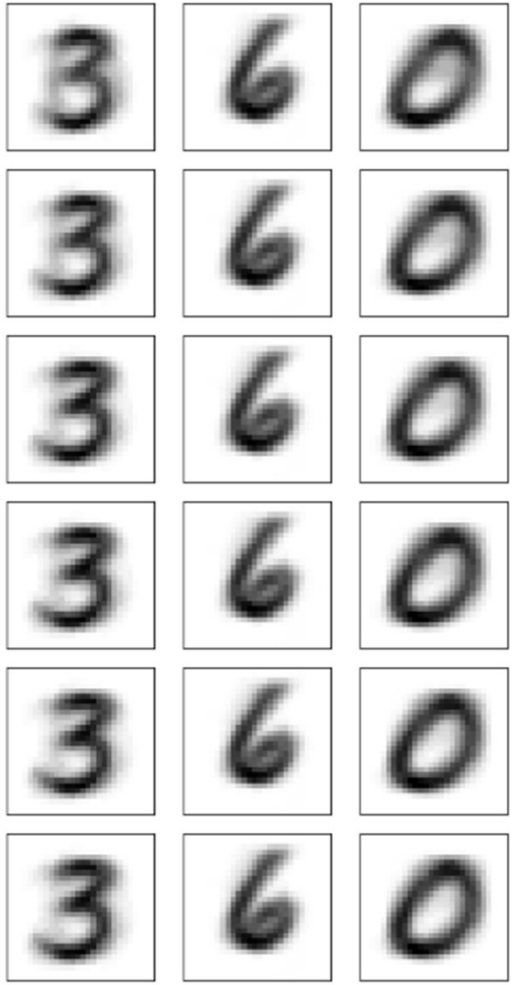
$$\gamma_{nk} = \frac{\pi_k p_{\mu_k}(\mathbf{x}_n)}{\sum_{k'=1}^{K} \pi_{k'} p_{\mu_{k'}}(\mathbf{x}_n)}$$

最大化ステップ

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} \qquad \pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

μ_k^1

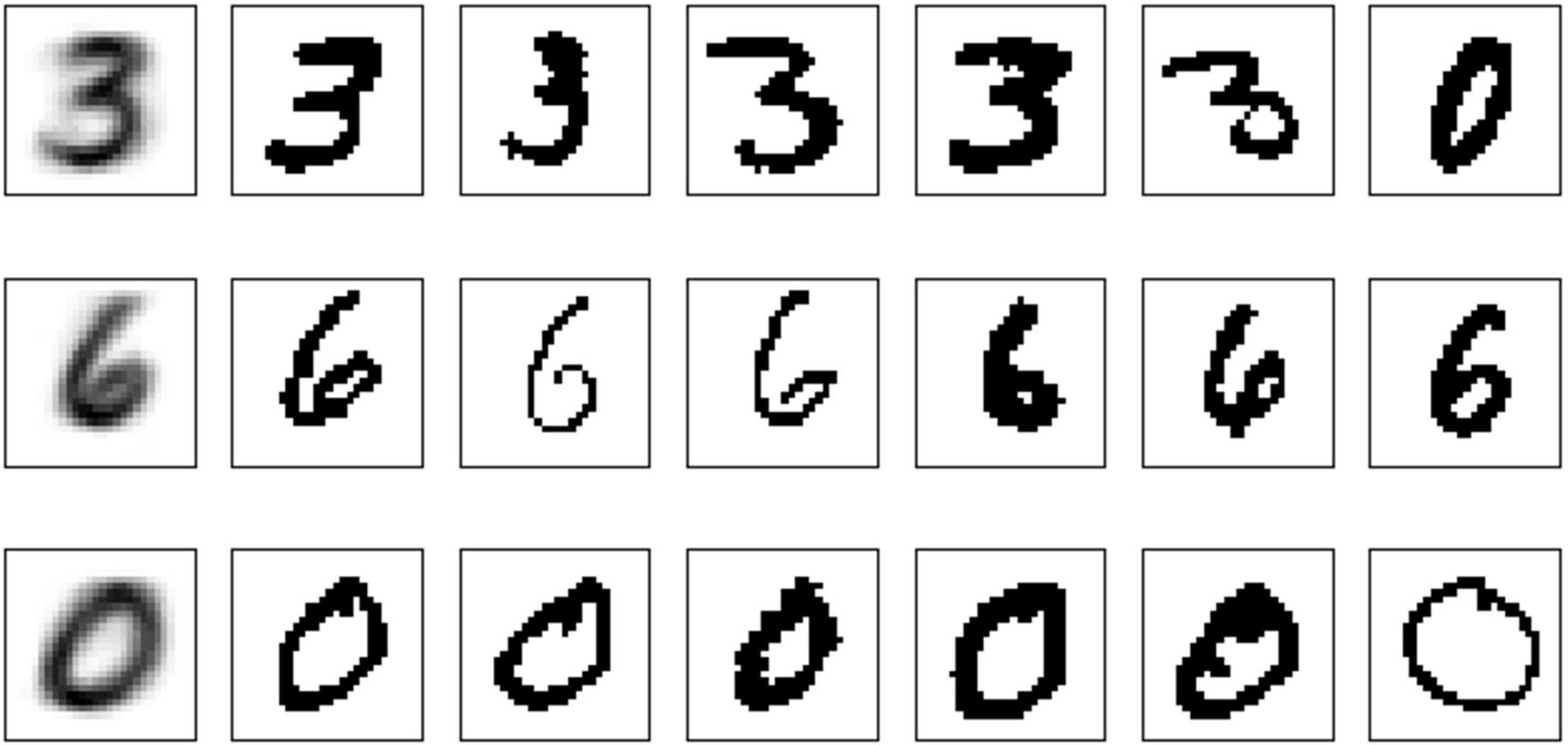




© 2020 Yuki Kato. All Rights Reserved

0，3，6を対象にEMアルゴリズムを可視化2

■ 分類結果を可視化すると



© 2020 Yuki Kato. All Rights Reserved

画像学習

■ 画像生成

- 機械学習
  - （畳み込み）変分オートエンコーダ
- 深層学習
  - オートエンコーダ
  - GAN

ベイズ+最尤推定（EM）+DNN

滑らかな変化を学習⇒ぼやけやすい

一致度合いが高い

■ 画像分類

- 機械学習
  - ロジスティック回帰
  - サポートベクターマシン
- 深層学習
  - DNN
  - Transformer
  - …

© 2020 Yuki Kato. All Rights Reserved

（畳み込み）変分オートエンコーダによる画像生成器

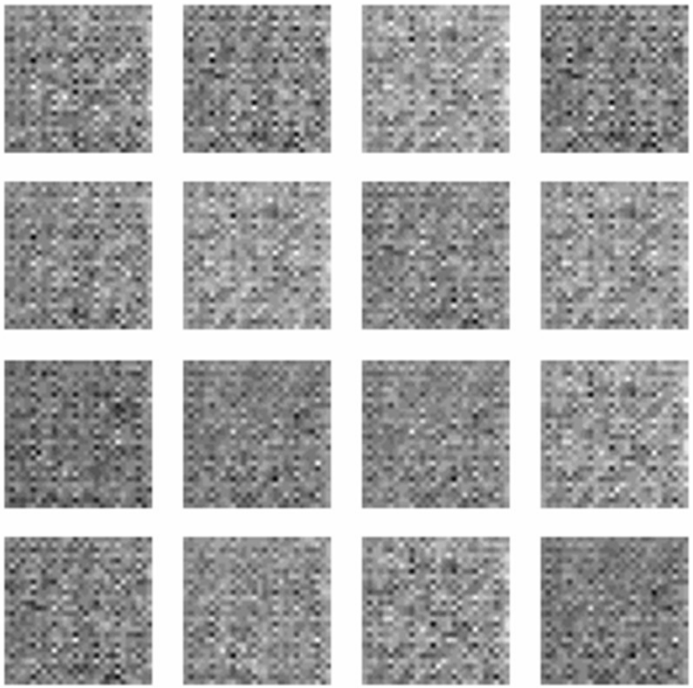
VAE は、限界対数尤度の証拠下限（ELBO）を最大化することでトレーニングします。

$$\log p(x) \geq \text{ELBO} = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right].$$

実際には、この期待値の単一サンプルのモンテカルロ推定を最適化します。

$$\log p(x|z) + \log p(z) - \log q(z|x),$$

とし、z は $q(z|x)$ からサンプリングされます。

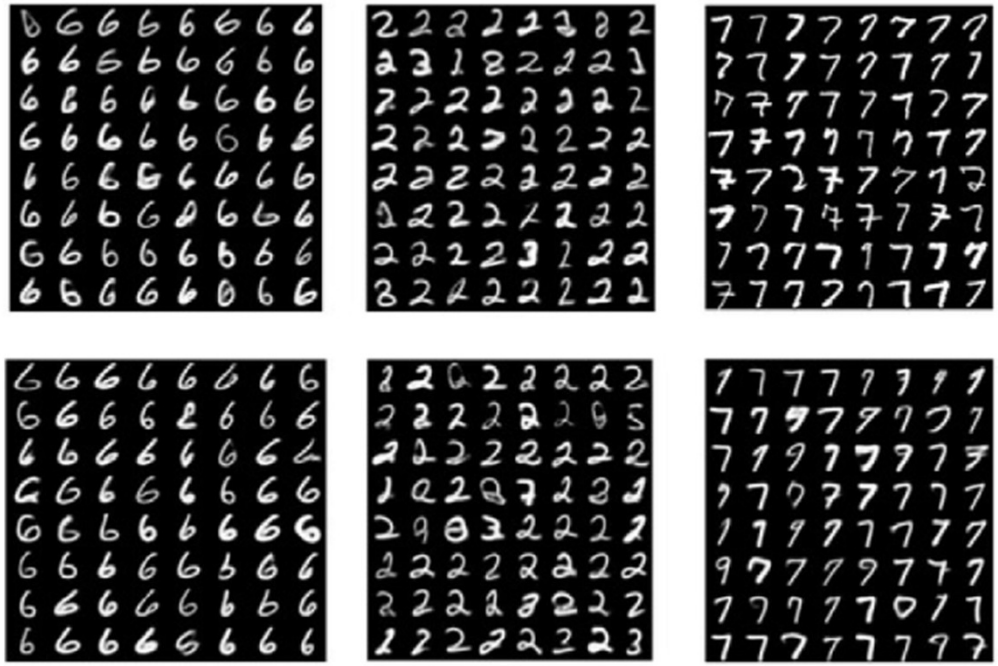


https://www.tensorflow.org/tutorials/generative/cvae?hl=ja

滑らかな変化を学習⇒ぼやけやすい

データ

出力



Ref:Mak、Hugo & Han、Runze & Yin、Hoover

© 2020 Yuki Kato. All Rights Reserved

画像学習

■ 画像生成

- 機械学習
  - （畳み込み）変分オートエンコーダ
- 深層学習
  - オートエンコーダ
  - GAN

■ 画像分類

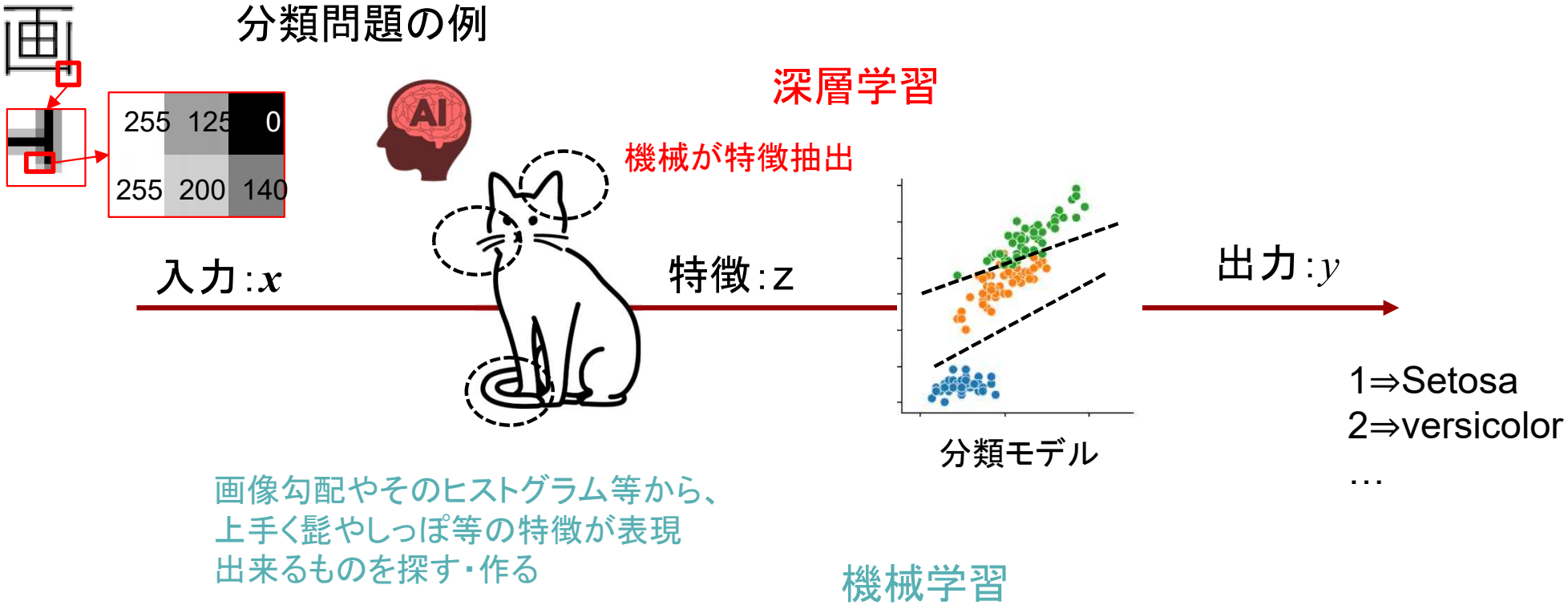
- 機械学習
  - ロジスティック回帰
  - サポートベクターマシン
- 深層学習
  - DNN
  - Transformer
  - …

© 2020 Yuki Kato. All Rights Reserved

機械学習と深層学習

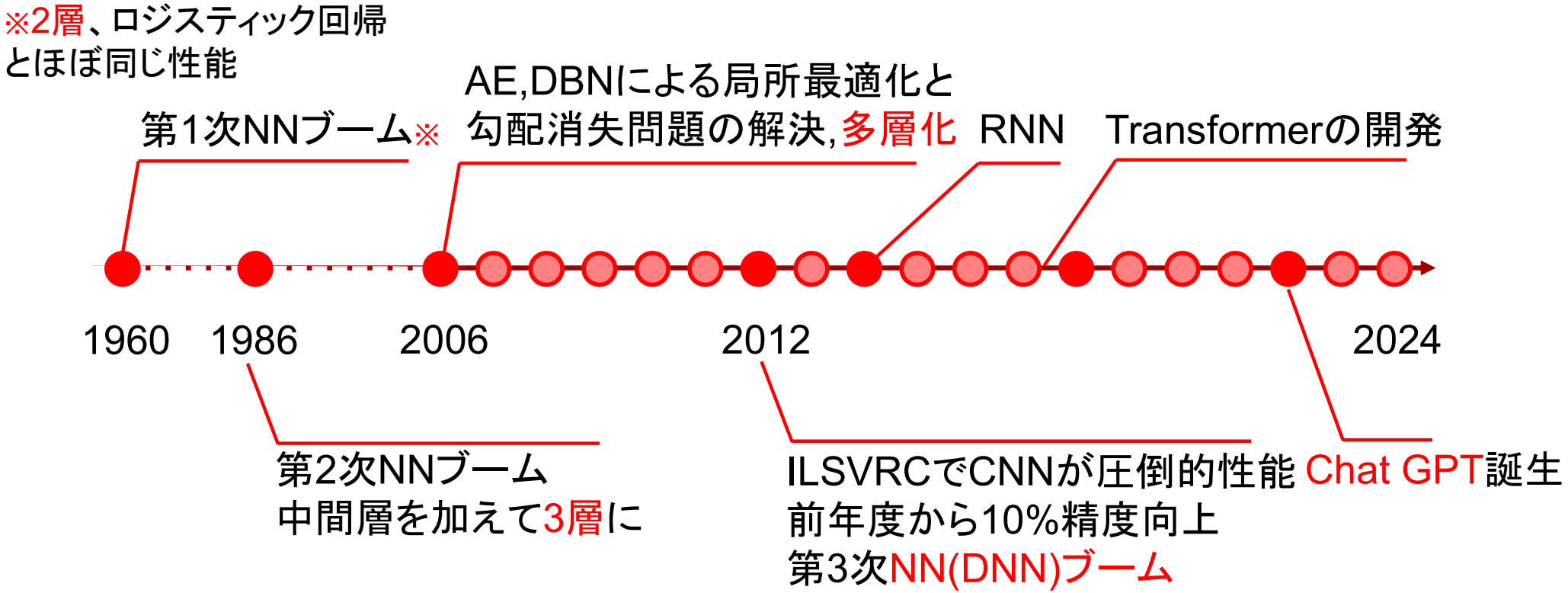
- 機械学習は人間が特徴量を設定（特徴抽出）、回帰モデルや分類モデルで機械に学習させる
- 深層学習は機械自体が特徴抽出とモデル化を行う

⇒人間の目を超える性能/特徴量設定の手間の減少が期待



© 2020 Yuki Kato. All Rights Reserved

深層学習（Deep Neural Network）の歴史



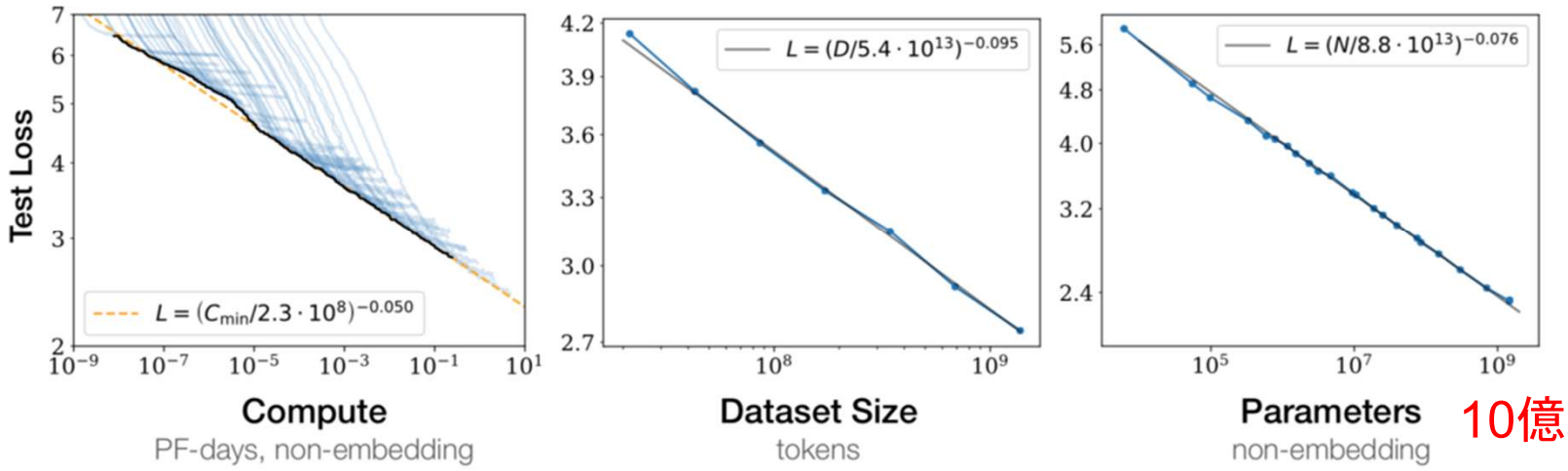
- いまはTransformer(Generative Pretrained Transformer)が主流
- 生成AIはもちろん画像認識も置き換わりつつある
（他のモデルもある、進化が速い）

© 2020 Yuki Kato. All Rights Reserved

生成AIが進化した理由

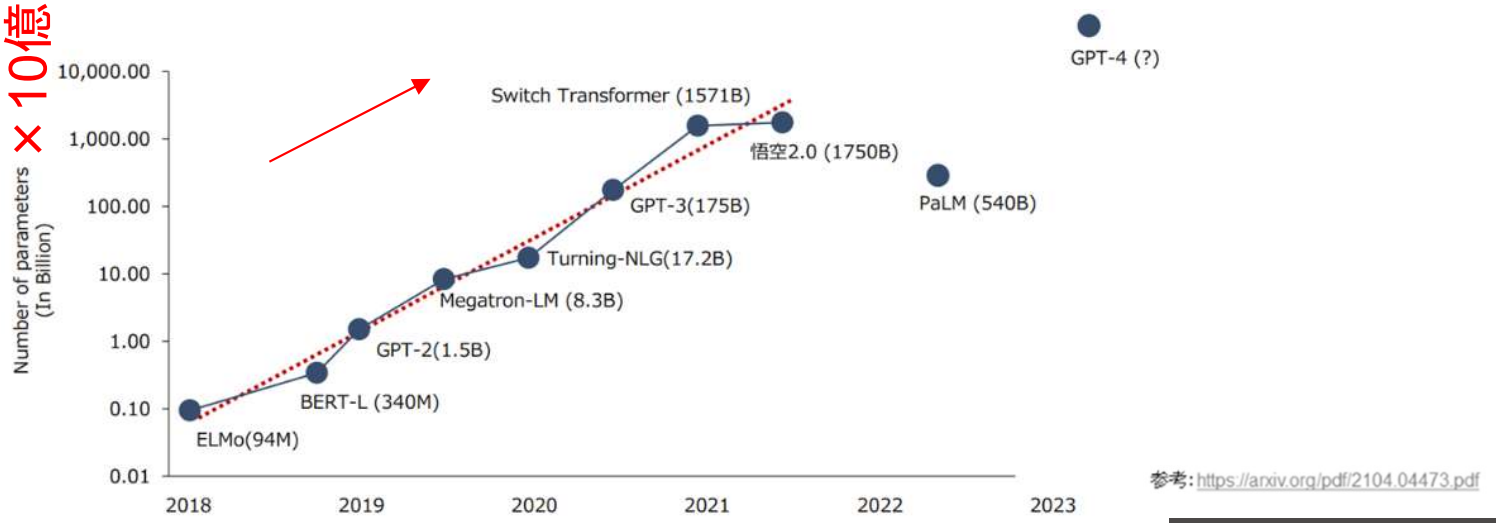
- スケール則が発見
- 計算予算、データセットサイズ、パラメータを上げるほど性能が上がる
- （他のモデルでは過学習等があるのだが…）

膨大な計算資源、膨大なデータ、膨大なパラメータ数があれば性能大
⇒マシン性能　、webデータ　、メモリの増加により可能となった



10億

"Scaling Laws for Neural Language Models"（Jared Kaplan, et al. @ OpenAI, arXiv, 2020.）

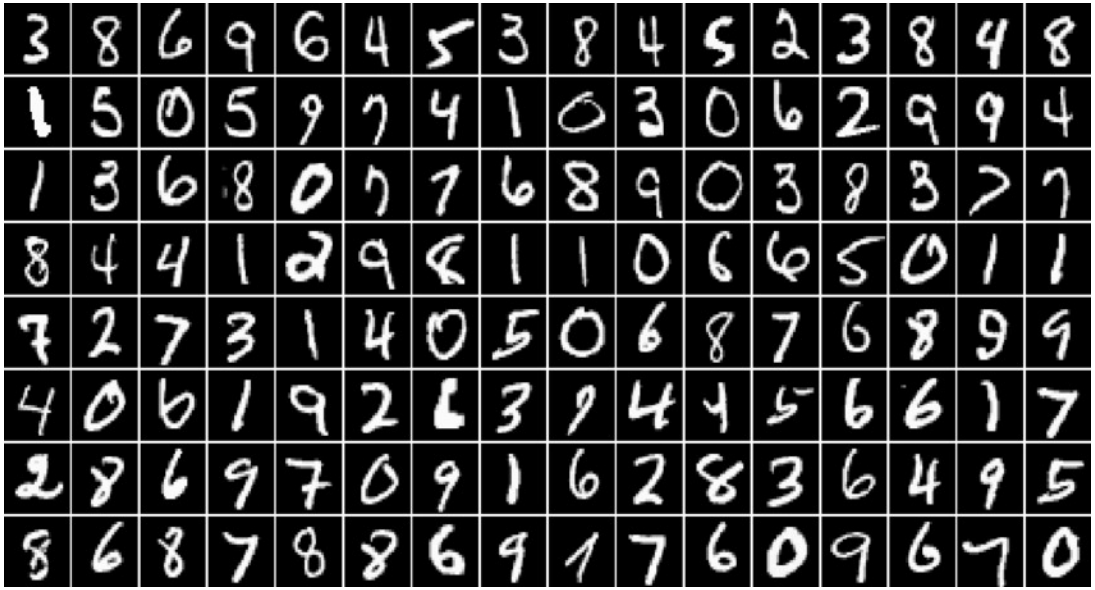


参考: https://arxiv.org/pdf/2104.04473.pdf

© 2020 Yuki Kato. All Rights Reserved

MNISTデータセット

画像認識に用いられる手書き数字画像のデータベース



7万枚の画像とラベルで構成

MNISTは、主に画像認識を目的としたディープラーニング／機械学習の初心者向けチュートリアルでよく使われており、データがきれいに整形さているので「高い精度が出やすい」という特長がある（逆に言うと、MNISTでは精度が高すぎて、機械学習の手法間で精度を比較したい場合には使いづらいという欠点がある[1]

指標：1~9に正しく分類⇒成功

[1] https://atmarkit.itmedia.co.jp/ait/articles/2001/22/news012.html
[2] Building Neural Network From Scratch For Digit Recognizer Using MNIST Dataset, Medium, https://medium.com/@niranjanshah474/building-neural-network-from-scratch-for-digit-recognizer-using-mnist-dataset-30397be28f5e

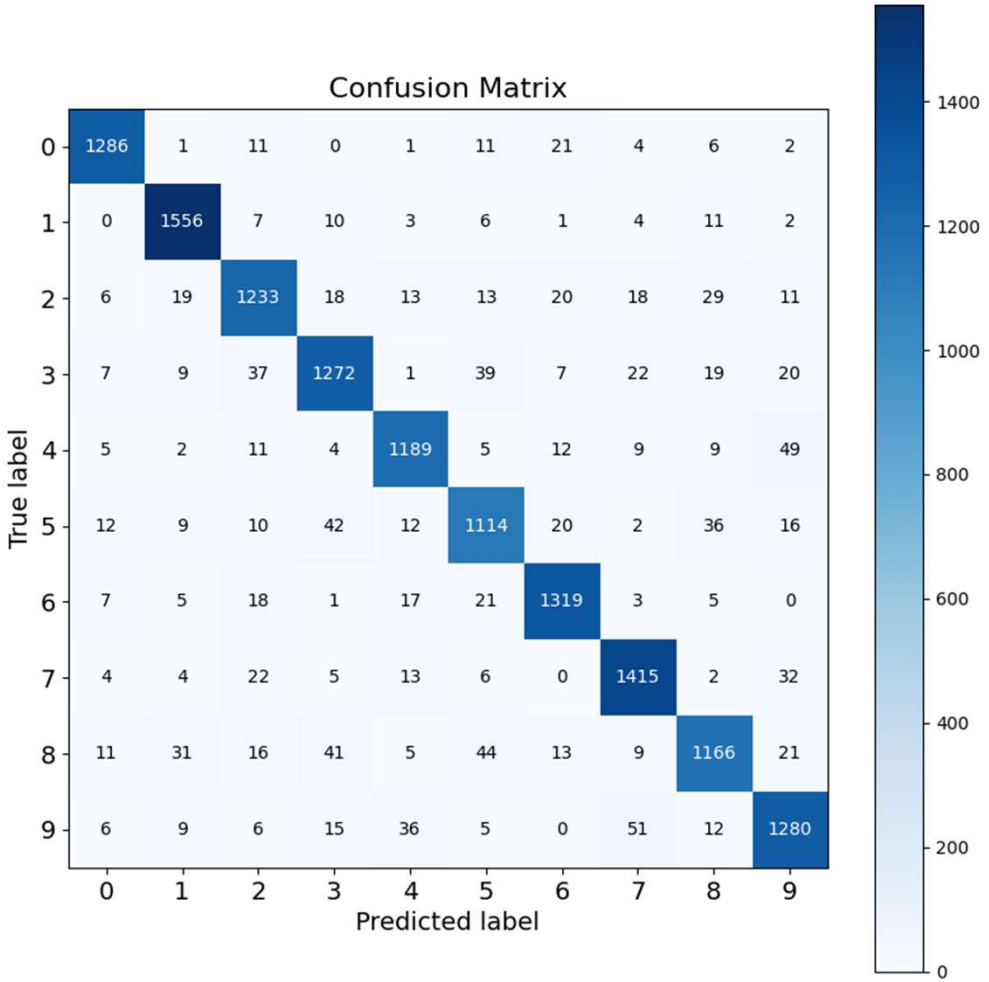
© 2020 Yuki Kato. All Rights Reserved

MNISTデータセット

- 教師あり学習で０～９の他クラス分類を行う
- モデルにはロジスティック回帰モデルを用いる

```
Accuracy: 0.9164285714285715
```

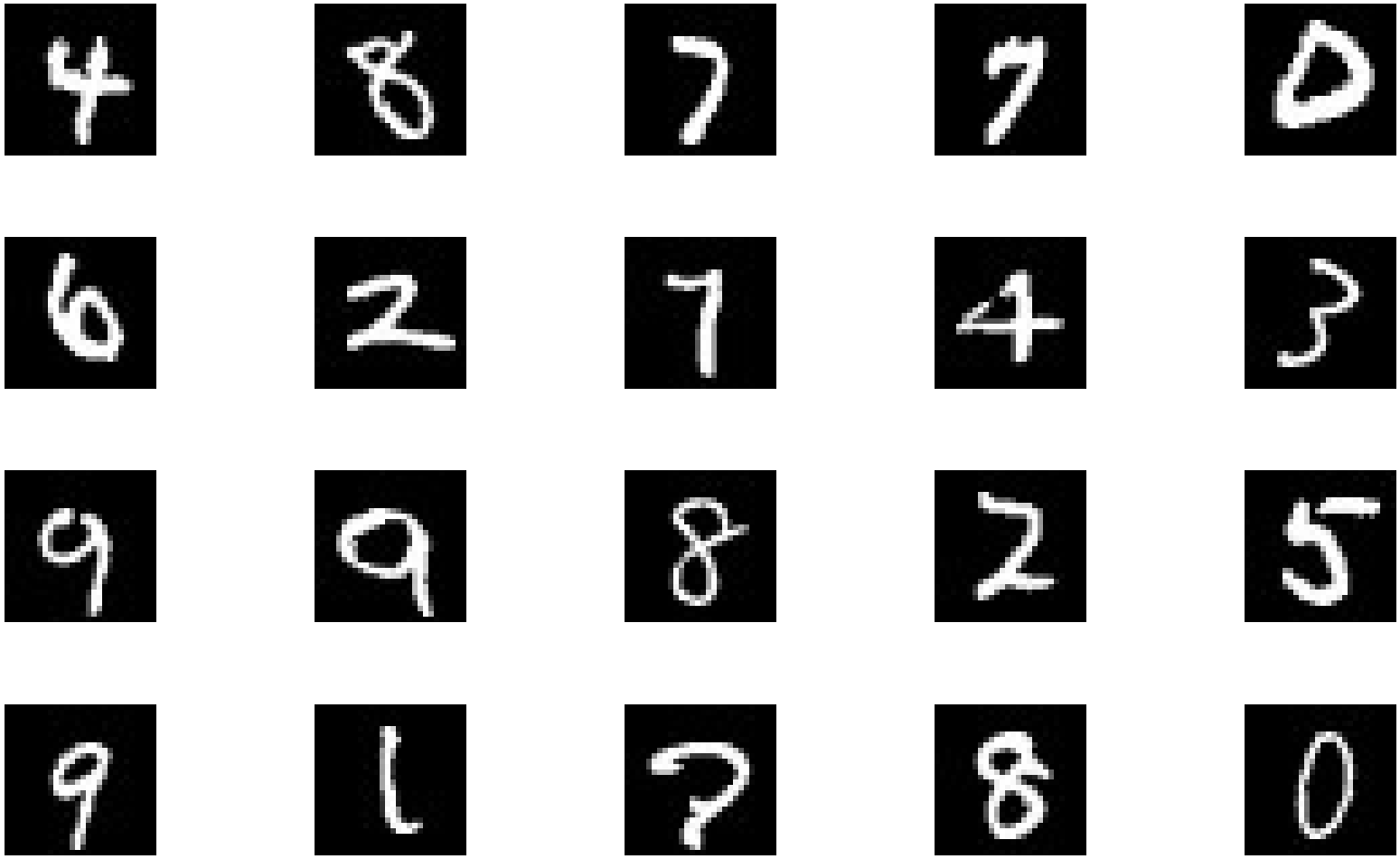
92%の精確性



https://colab.research.google.com/drive/1l0PnE89mGqPWQrvO4xBgIVAqwGa_SvD7?usp=sharing

© 2020 Yuki Kato. All Rights Reserved

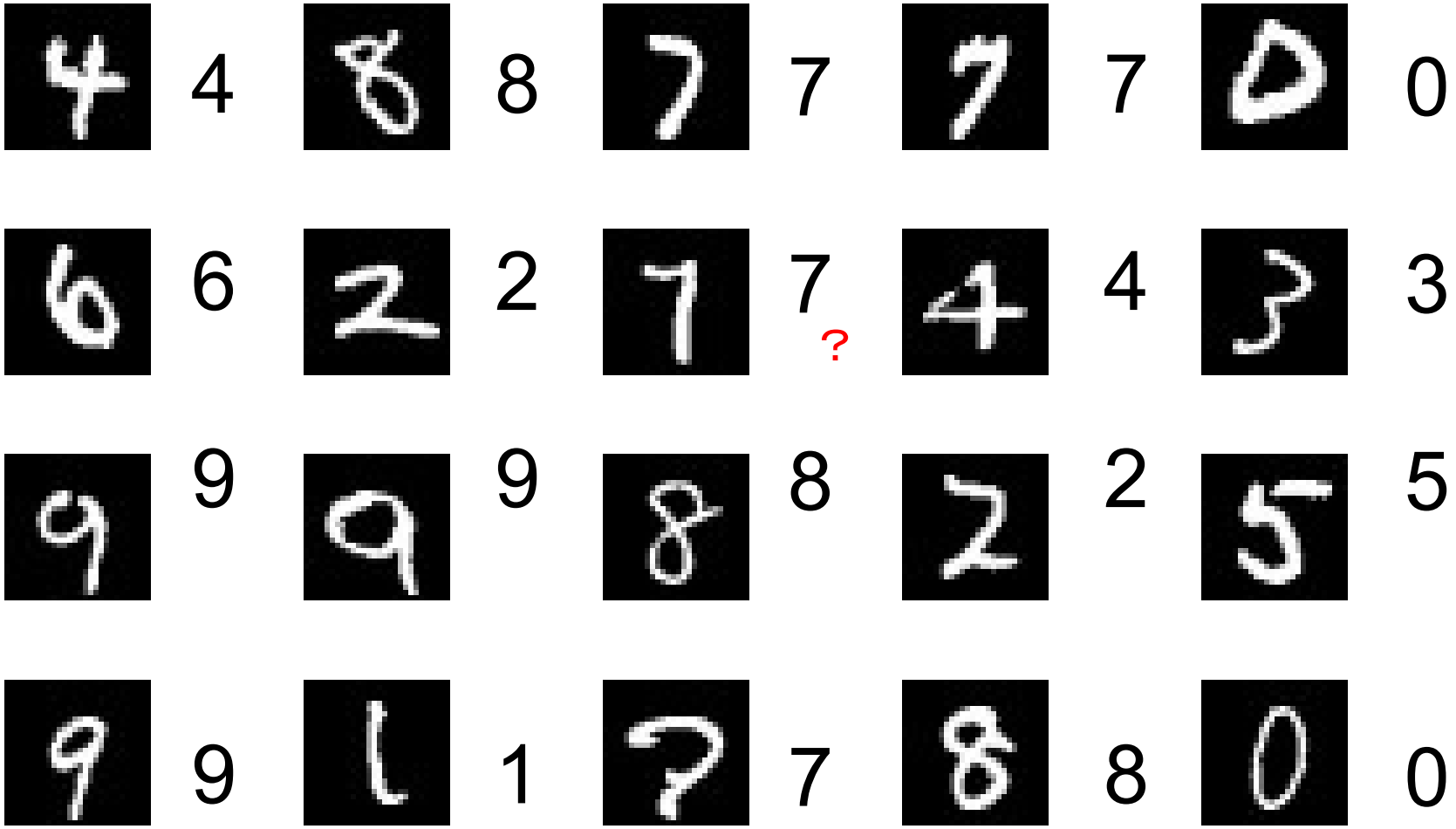
人間の目で分類してみよう



メモやエクセルに自分の回答を記入してください

© 2020 Yuki Kato. All Rights Reserved

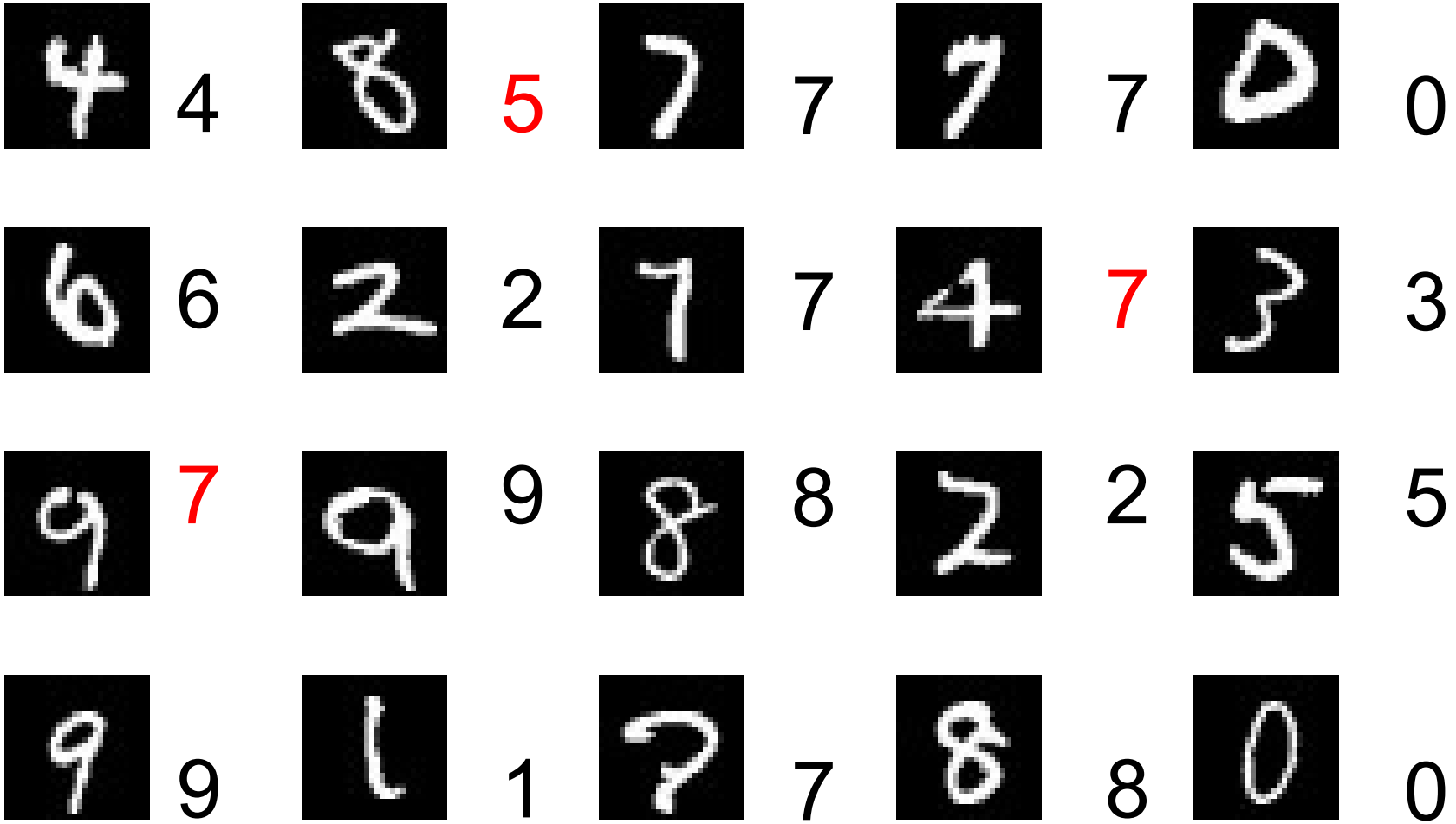
人間の目で分類してみよう



正解率は何%でしたか？

© 2020 Yuki Kato. All Rights Reserved

機械学習（ロジスティック回帰）による分類



17/20⇒85%

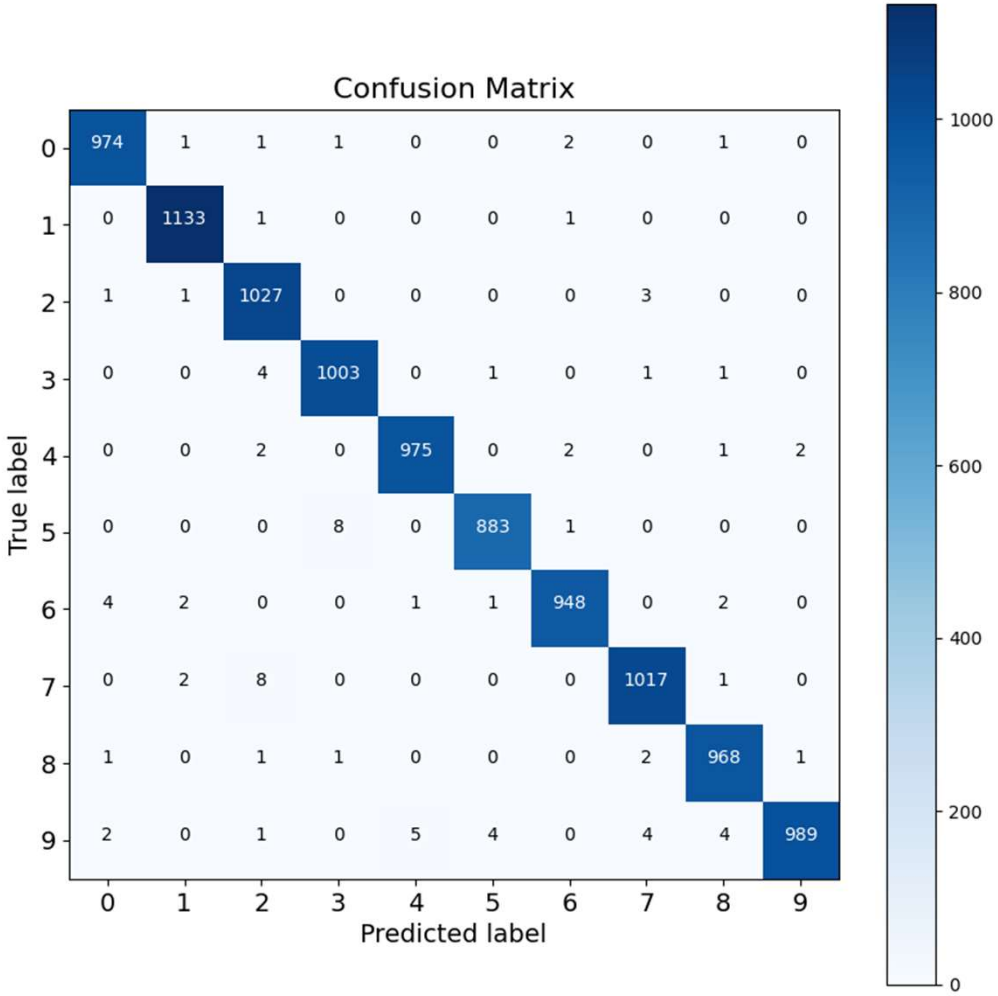
© 2020 Yuki Kato. All Rights Reserved

MNISTデータセット

- 教師あり学習で０～９の他クラス分類を行う
- 深層学習を用いる

```
Test loss: 0.02863382175564766
Test accuracy:
0.9916999936103821
```

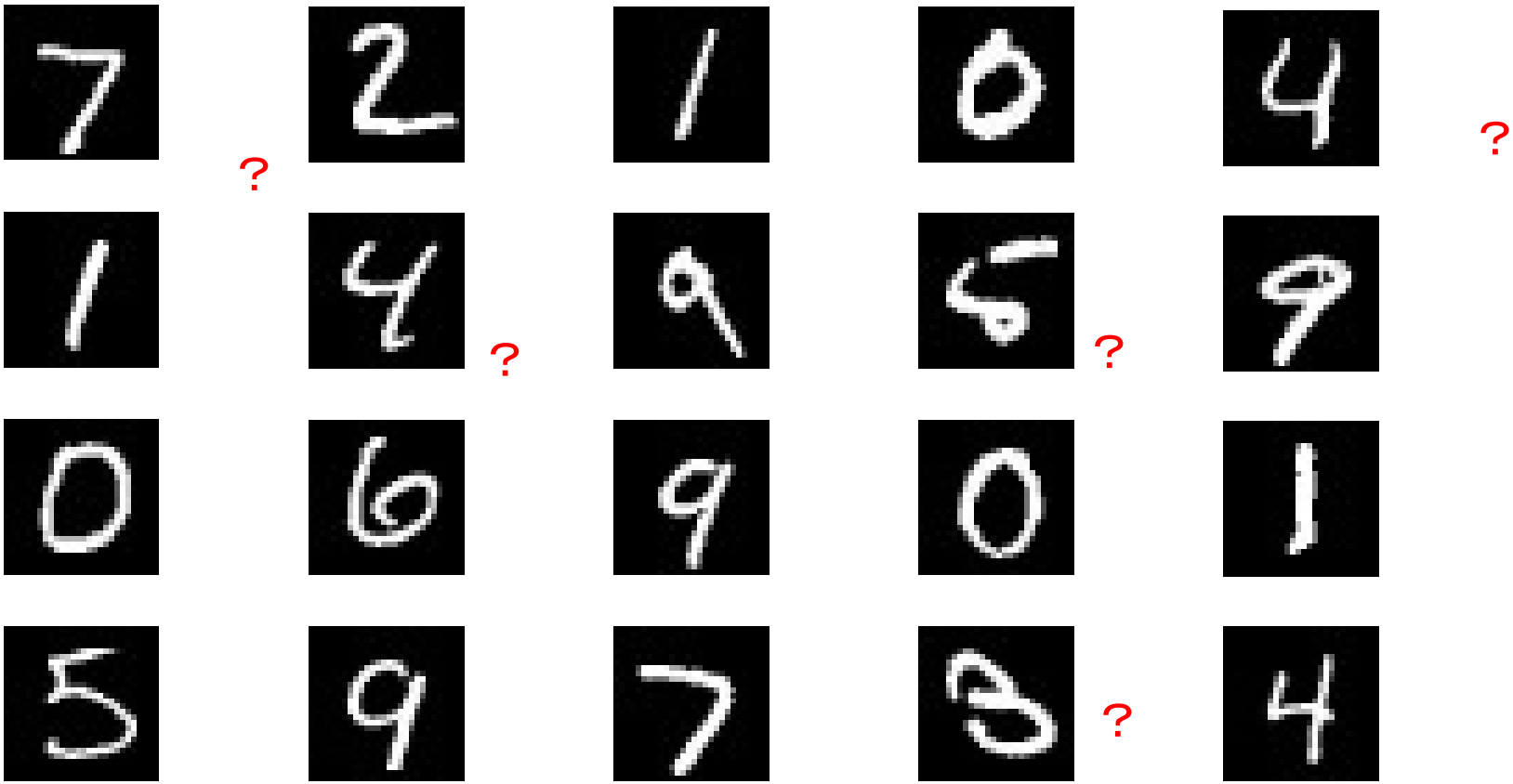
99%の精確性



https://colab.research.google.com/drive/1ngXL1VHS72GujBhGphEufHqiixuE2Y94?usp=sharing

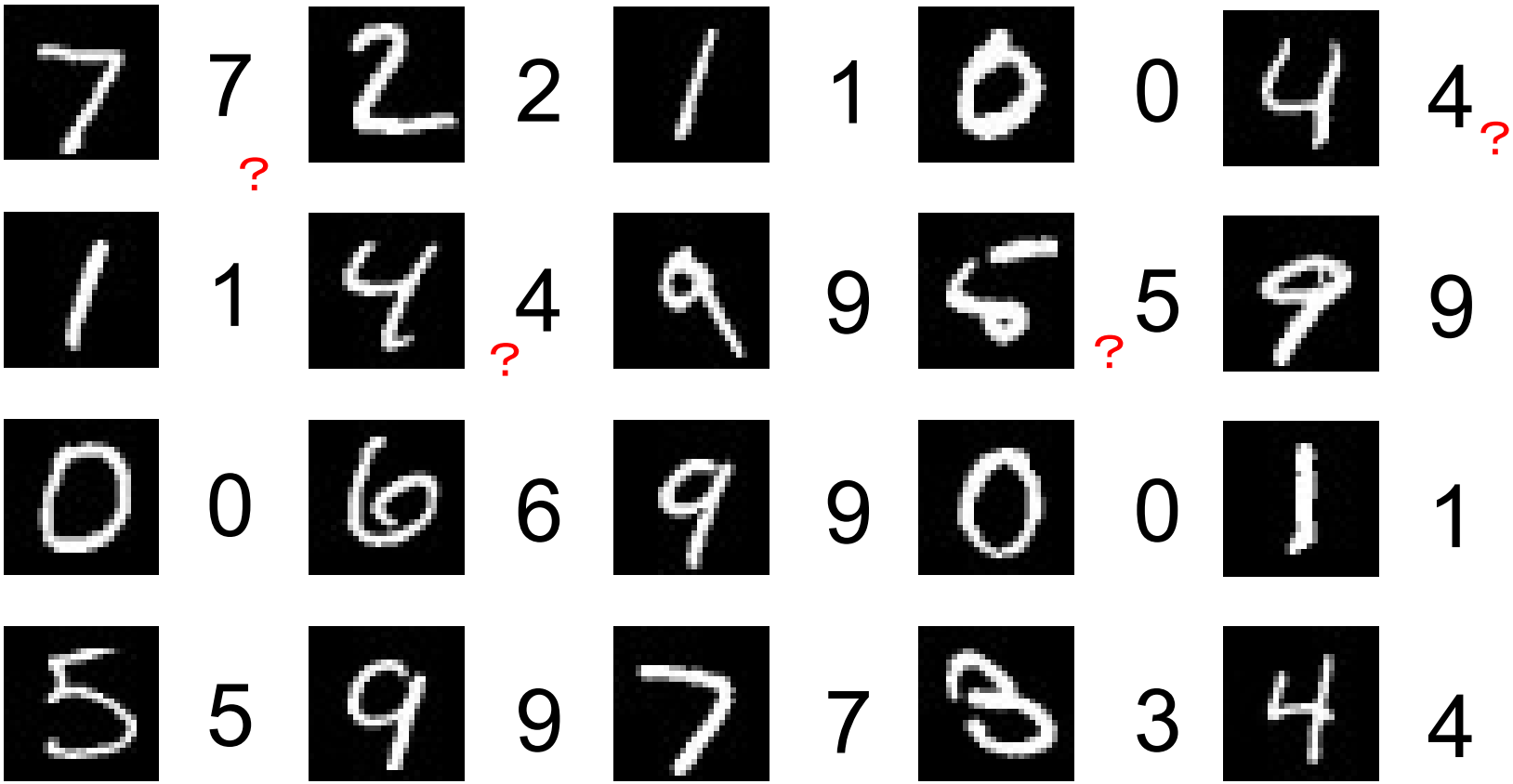
© 2020 Yuki Kato. All Rights Reserved

人間の目で分類してみよう



© 2020 Yuki Kato. All Rights Reserved

答え



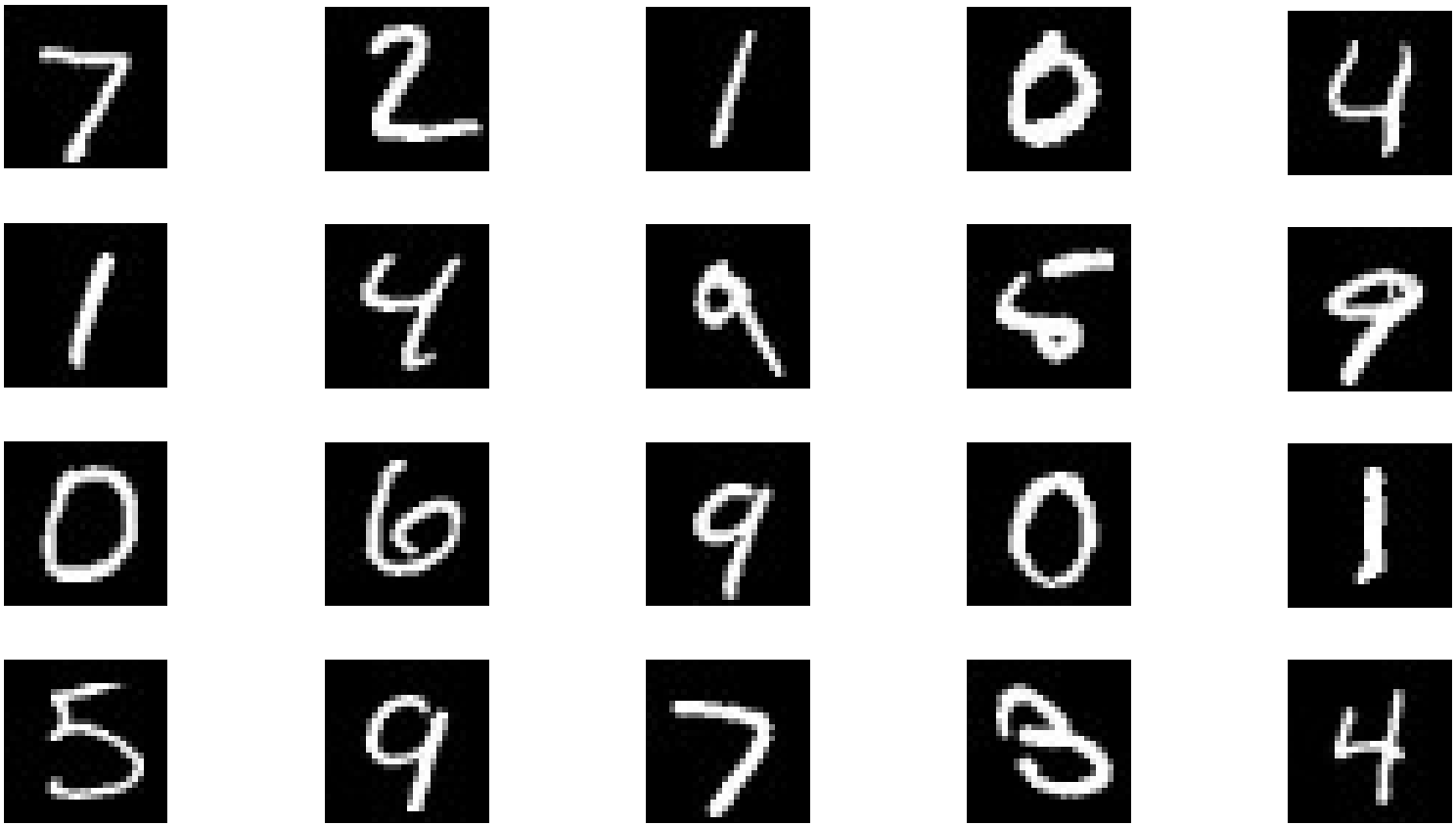
正解率は何%でしたか？⇒DNNは全問正解（20/20）

327

© 2020 Yuki Kato. All Rights Reserved

補：Transformer（Gemini）では．．．

- 画像分類用のTransformerではないが，Geminiに分類させてみよう



© 2020 Yuki Kato. All Rights Reserved

© 2020 Yuki Kato. All Rights Reserved

補：EMアルゴリズム

	定義	具体例
Eステップ	「隠れ変数」の期待値を求める	「π_k（画像生成器を選択する確率）」を求める
	確率モデルのパラメータを固定し、「隠れ変数の事後確率 (＝モデルの尤度の期待値)」を求める	「μ_k（画像生成器）」を固定し、「π_k（画像生成器を選択する確率）」の事後確率を求める
Mステップ	E ステップで求めた「隠れ変数の事後確率 (＝モデルの尤度の期待値)」を踏まえて、これを最大化するパラメータを求める	E ステップで求めた「π_k（画像生成器を選択する確率）」の事後確率を踏まえて、新たな「μ_k（画像生成器）」を求める
反復	M ステップで求めたパラメータを用いて次の E ステップで使う「隠れ変数」の分布を再計算する	M ステップで求めた「μ_k（画像生成器）」を用いて次の E ステップで使う「π_k（画像生成器を選択する確率）」を再計算する

© 2020 Yuki Kato. All Rights Reserved

© 2020 Yuki Kato. All Rights Reserved

© 2020 Yuki Kato. All Rights Reserved

統計データ解析
第十二回

加藤由幹



高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

© 2020 Yuki Kato. All Rights Reserved

講義内容（シラバスより抜粋）

- 講義の前半では，確率分布や中心極限定理を学び，実験データの誤差をどのように評価/グラフ化するかについて学ぶ．また，回帰分析を用いてデータの特徴を分析する手法や，時系列データ解析や画像解析で重要となるフィルタリング処理について学ぶ．
- 講義の後半では，主成分分析や固有値分解，特異値分解を用いた多変量データ解析手法を習得する．

- 第１回　統計データの性質とMatlabを用いたデータ解析の基礎
- 第２回　確率分布の性質
- 第３回　グラフと確率分布を用いた実験データの可視化
- 第４回　さまざまな分布
- 第５回　大数の法則と中心極限定理
- 第６回　時系列データ解析とフィルタ処理1
- 第７回　時系列データ解析とフィルタ処理2，画像解析

10/28 ■ 第８回　画像解析
10/31 ■ 第９回　最小二乗法，回帰分析
11/4 ■ 第１０回　最尤推定，ベイズ推定
11/14 ■ 第１１回　変分ベイズ推定，EMアルゴリズム，変分オートエンコーダ
11/18 ■ 第１２回　スパースモデリング
11/21 ■ 第１３回　多変量データ解析の基礎と主成分分析　課題返却，振り返り
11/25 ■ 第１４回　固有値分解，特異値分解
11/28 ■ 第１５回　期末テスト
12/2 ■ 第１６回　返却　⇒返却のみ，補講扱いですが，全員出席とするので任意参加

© 2020 Yuki Kato. All Rights Reserved

スパースとは？

■ データを行列やベクトルで表現したときに，殆どの要素が0でごくわずかな要素だけが値を持つ状態

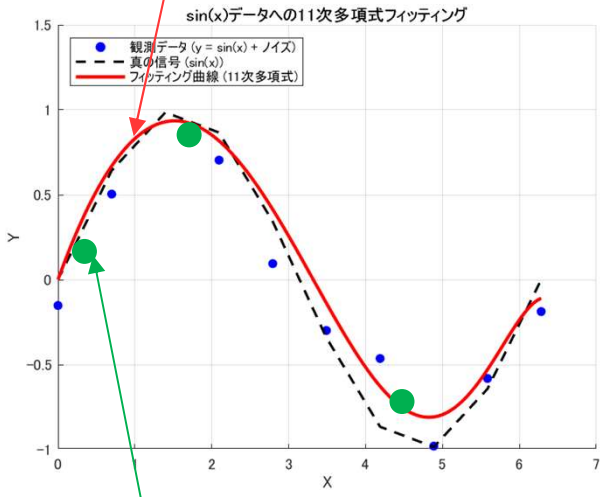
■ スパースモデリング~モデル化~

- スパース性を利用して、本当に重要な情報（非ゼロの要素）だけを抽出・選択し、データ構造などをシンプルかつ効率的に表現（モデル化）する

■ 圧縮センシング~計測~

- データのスパース性を利用して，従来のサンプリング理論の限界を超え、信号を劇的に少ないデータ（サンプル）で取得・復元するための画期的な計測・信号処理技術

オーバーフィットを防ぎシンプルなモデルを作成したい



エイリアシングが起きないようにランダムに計測

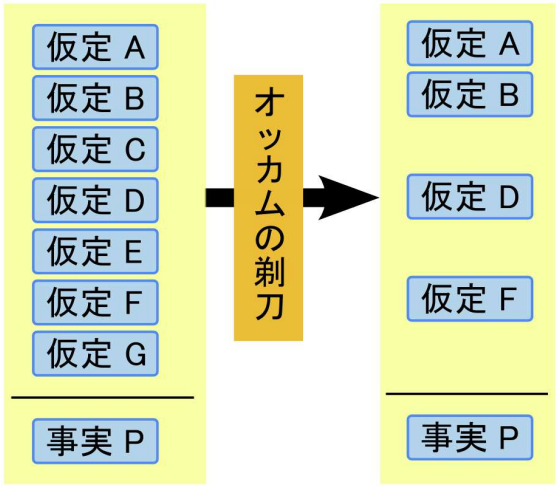
係数：β_i
0
0
0
0
0
0
0
12.83
-23.5
20.7
-5.6
0

© 2020 Yuki Kato. All Rights Reserved

スパースモデリングの歴史と考え

■ オッカムの剃刀

- ある事柄を説明するためには、必要以上に多くを仮定するべきでない
- 14世紀の哲学者・神学者であるオッカム村のウィリアムが多用したことで有名に





■ アイザック・ニュートン

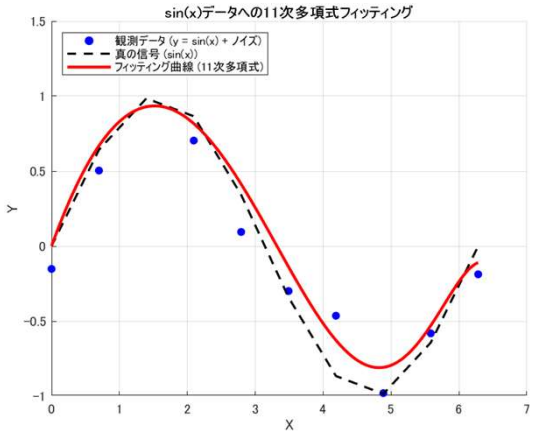
- 「自然物に関しては、事実で かつ充分な原因だけを認めるべきだ。同じ自然的影響については、できるかぎり、同じ原因を用いて説明すべきなのだ。」

Ref: 三浦俊彦『論理学が分かる事典』ISBN 4534037104 pp.204-205「5.14　説明はスリム化すべきである？ オッカムの剃刀」

© 2020 Yuki Kato. All Rights Reserved

L0ノルム最小化

- もっとも単純なスパースモデリングの実現方法は $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ を満たす中で最も係数（β）の数が小さくなるものを選択すればよい。



n 個のデータ点 $(\mathbf{x}_i, y_i)$ があり、これらのデータに最もよく適合する線形モデル

$$y = f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 g_0(\mathbf{x}) + \beta_1 g_1(\mathbf{x}) + \cdots + \beta_p g_p(\mathbf{x})$$

のパラメータベクトル $\boldsymbol{\beta}$ を求めたいとします。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} +$$

複雑になるので，一旦
誤差は考えない
劣決定（p>n）とする

$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{\beta}\|_0 \text{ subject to } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

$\{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_p\}$

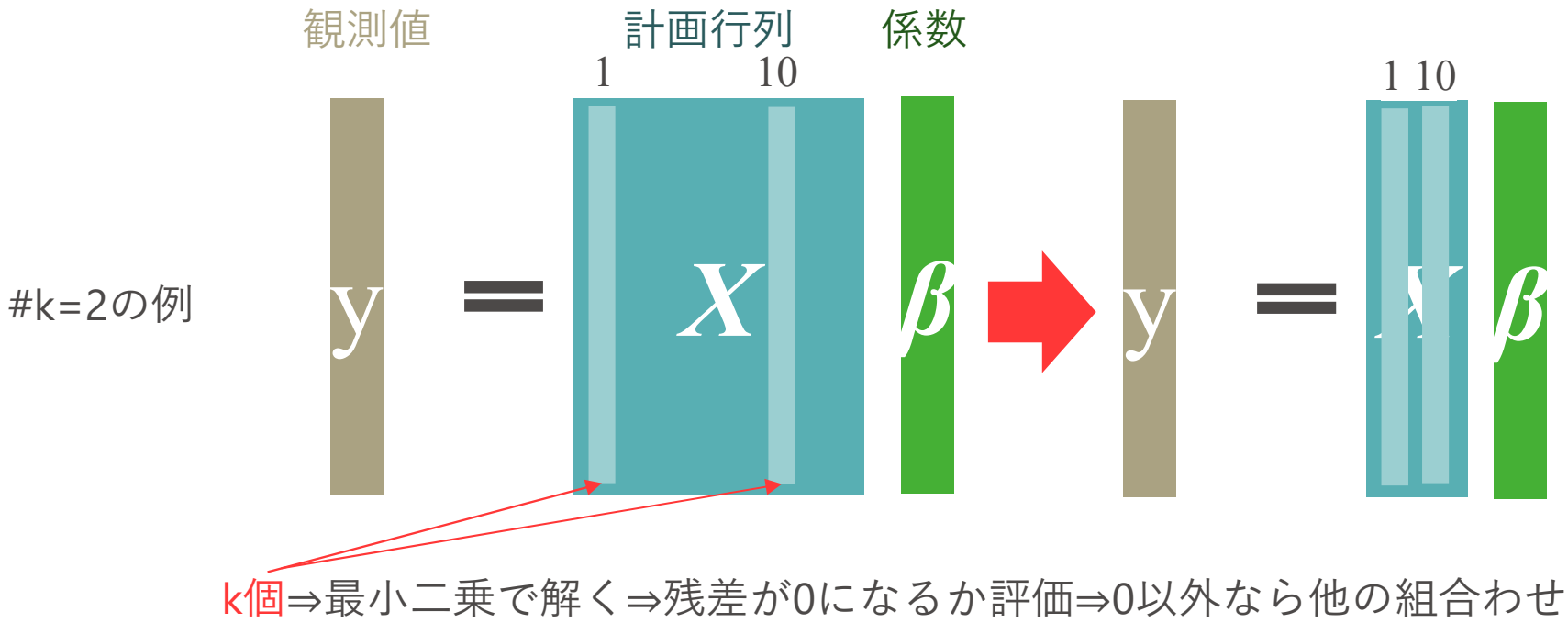
- $k = 1$から$k = n$まで，すべての組み合わせを探していく
- $\Sigma_n C_k$回の計算が必要（$O(n^k)$）⇒組み合わせ爆発

© 2020 Yuki Kato. All Rights Reserved

L0ノルム最小化2

- もっとも単純なスパースモデリングの実現方法は $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ を満たす中で最も係数（β）の数が小さくなるものを選択すればよい．

$$\hat{\beta} = \arg\min \|\beta\|_0 \text{ subject to } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$



- $k = 1$から$k = n$まで，すべての組み合わせを探していく
- $\Sigma_n C_k$回の計算が必要（$O(n^k)$）⇒組み合わせ爆発

© 2020 Yuki Kato. All Rights Reserved

OMPによる近似解法（L0ノルム最小化）

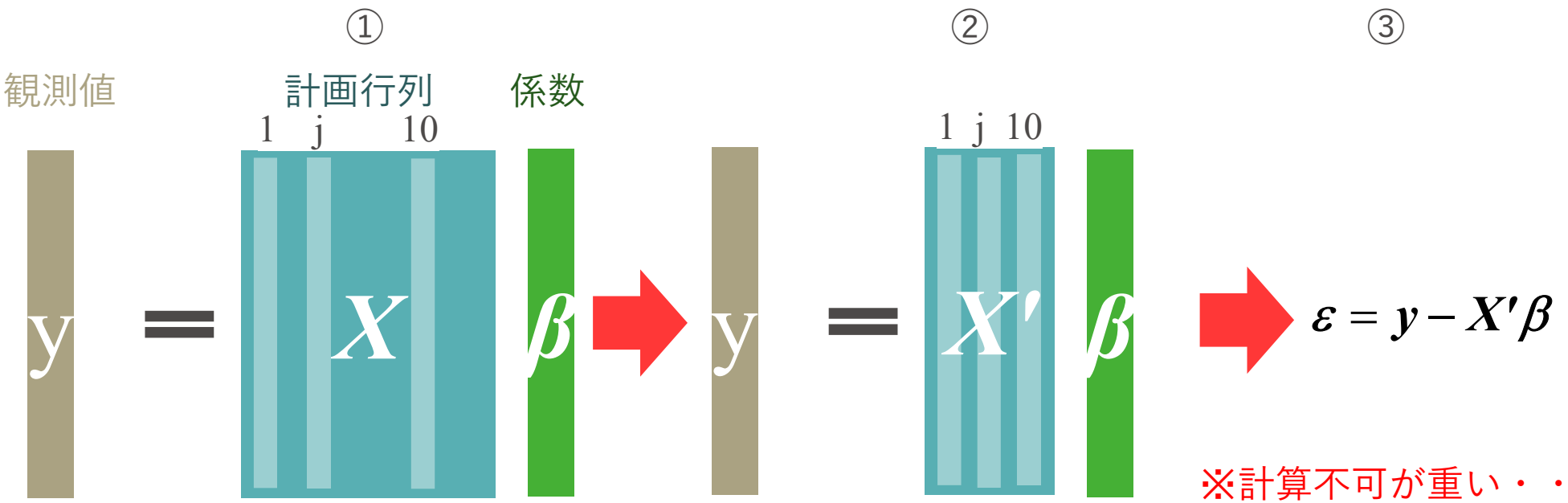
- OMP（Orthogonal Matching Pursuit）は、信号を表現するための「基底」から、信号を最もよく説明できる要素を一つずつ、貪欲に（その時点での最適解を選びながら）選択していく手法．貪欲法の一種

① 残差$\boldsymbol{\varepsilon}$に相関が高い列を選択（最大内積） $j = \arg\max \left| \langle \boldsymbol{x}_j, \boldsymbol{\varepsilon} \rangle \right|$

② この列とそれまでに選んだ列集合上で最小2乗を解く

③ 新しい残差を作る（残差は常に選ばれた列空間に直交）

疑似逆行列を使っても可



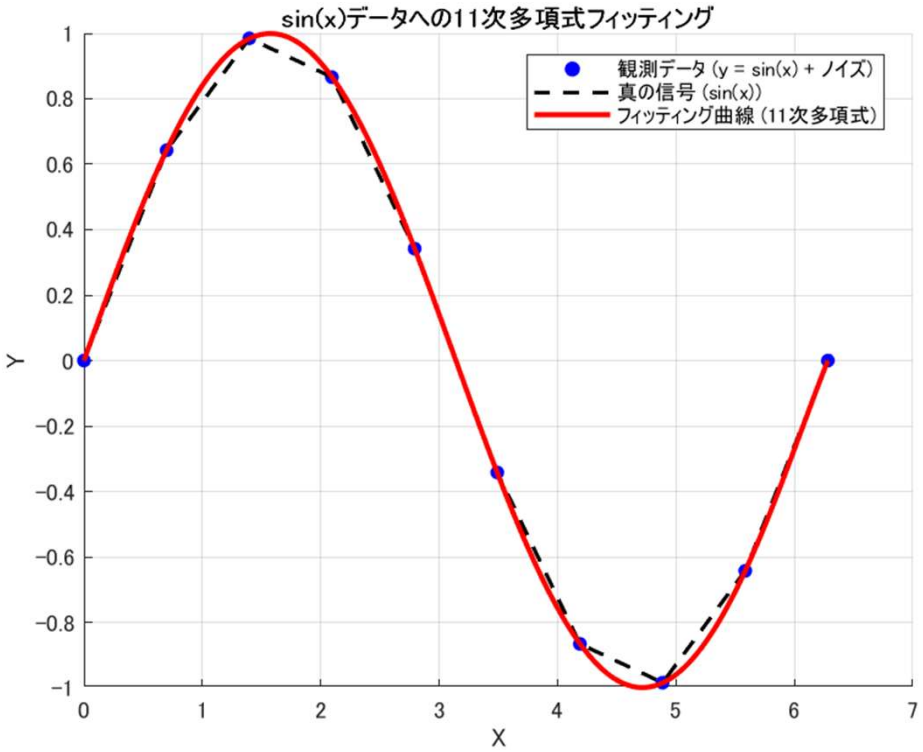
※計算不可が重い・・

© 2020 Yuki Kato. All Rights Reserved

例題：回帰分析におけるL0ノルム最小化

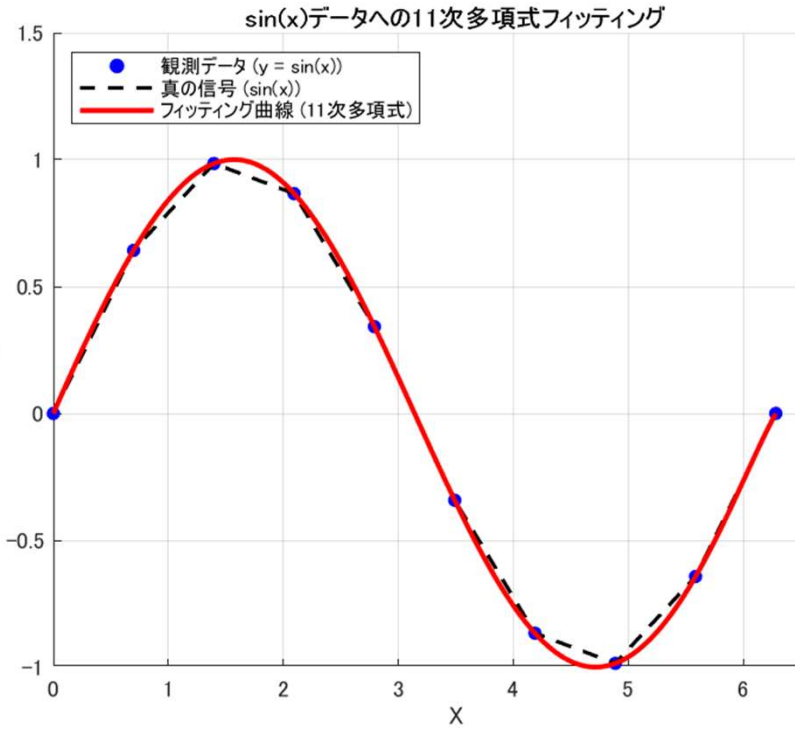
- OMPによるスパースモデリングを実施
- 解がスパースであれば，スパースな解が求まる
- 最小二乗法ではスパースになりづらい

OMP



β
0
0
0
0
0
0
0.0067
0
-0.1526
-0.0246
1.0116
0

LLS（線形最小二乗）



β
0
0
-0.00056
0.003786
-0.0146
0.028095
0
-0.08027
0
-0.13434
1.038585
0

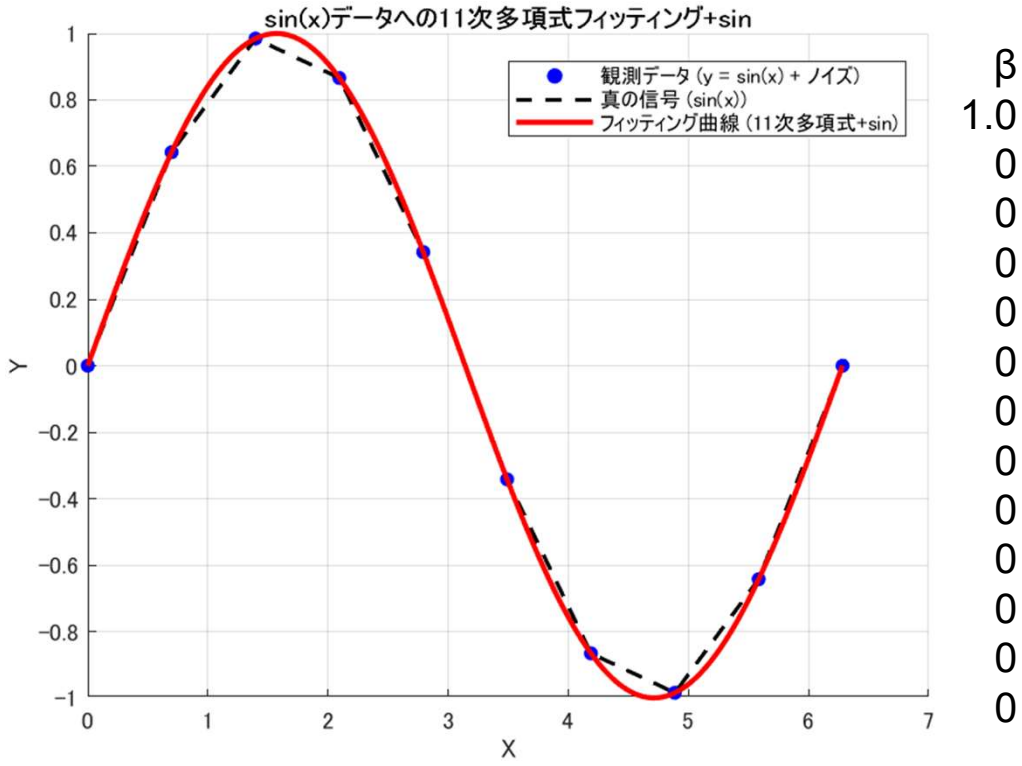
© 2020 Yuki Kato. All Rights Reserved

演習：モデルにsin(x)を追加しよう

© 2020 Yuki Kato. All Rights Reserved

演習：モデルにsin(x)を追加しよう

■ 基底として良い関数を用意しなければ，スパースモデリングは失敗する（正しい仮説が必要）



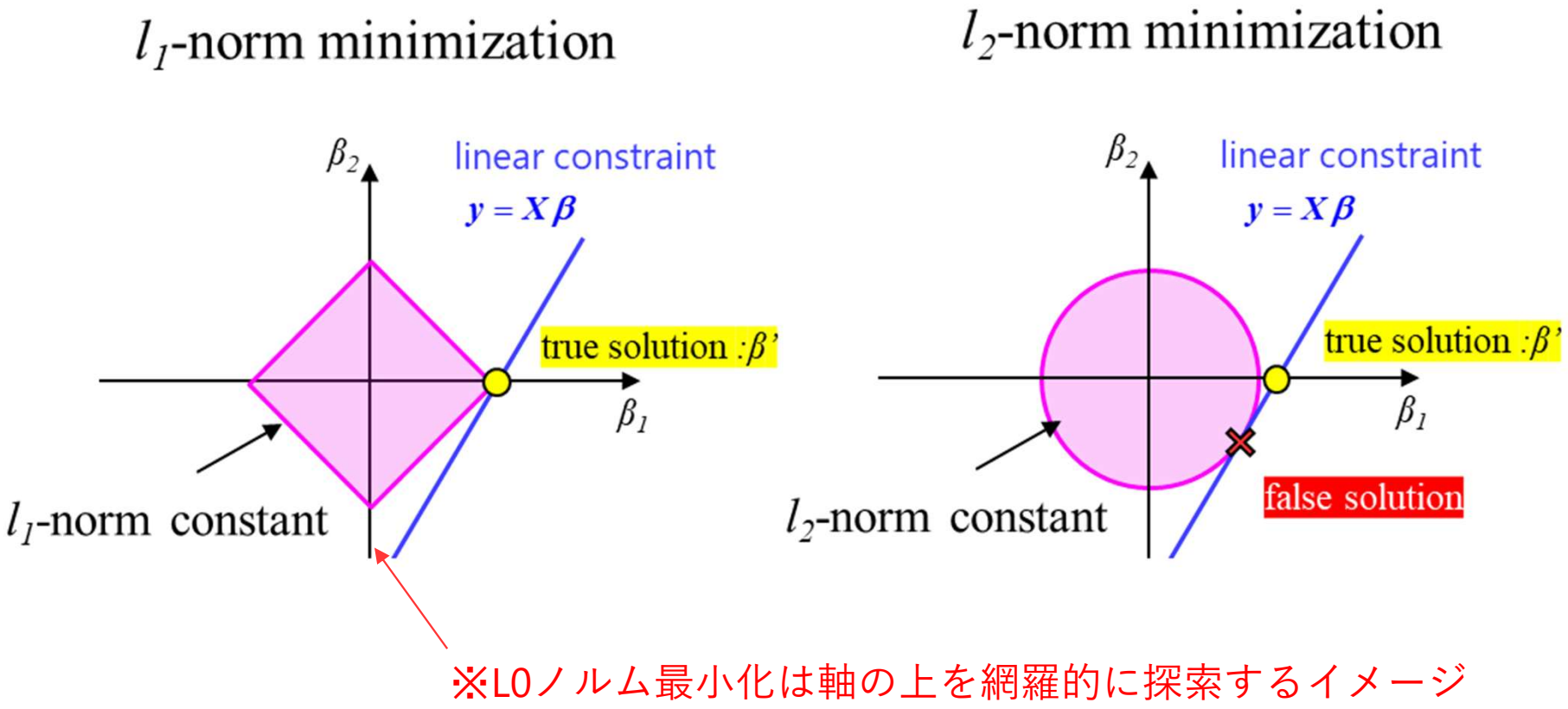
⇒OMPは計算不可が重い・・，別の方法で解けないか？

© 2020 Yuki Kato. All Rights Reserved

L1ノルム最小化

- L1ノルム（係数の絶対値）の最小化を実施すれば，解空間の尖りを利用してスパースな解が求まる

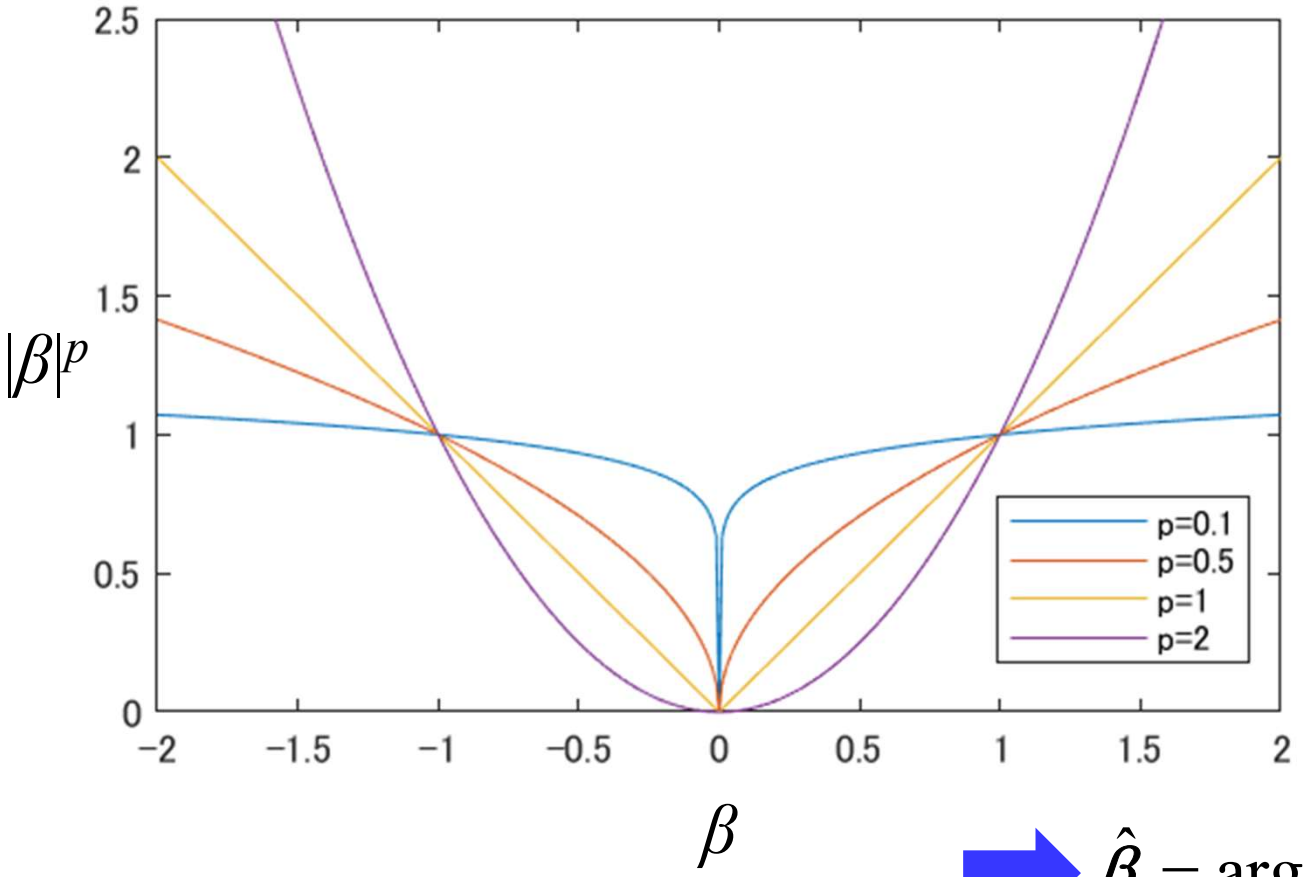
$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$$



© 2020 Yuki Kato. All Rights Reserved

L1ノルム最小化と凸関数

- L1ノルムは下記の通り凸関数
- ⇒最小二乗法と同じく，これの最小化は凸緩和法であり大域的最適解が求まる
- 加えて，L1ノルムは解空間で尖りを持つので優秀



$p \geqq 1$

⇒凸関数

$p \leqq 1$

⇒解空間で尖りを持つ
⇒スパース同定が可能

$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$$

© 2020 Yuki Kato. All Rights Reserved

L1ノルム最小化を線形計画法に置き換える

■ $\beta = u - v$

u : βの正成分はそのまま，負成分を0
v : βの正成分は0，負成分はそのまま

■ $z = [u^{\mathrm{T}}, v^{\mathrm{T}}]^{\mathrm{T}}$

■ $\|\beta\|_1 = 1^{\mathrm{T}}(\mathrm{u}+\mathrm{v}) = 1^{\mathrm{T}}z$ ■ $y = X\beta = X(\mathrm{u}-\mathrm{v}) = [X, -X]\, z$

どちらも正

$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$$



$$\arg\min 1^{\mathrm{T}}z \text{ subject to } y = [X, -X]\, z \text{ and } z > 0$$

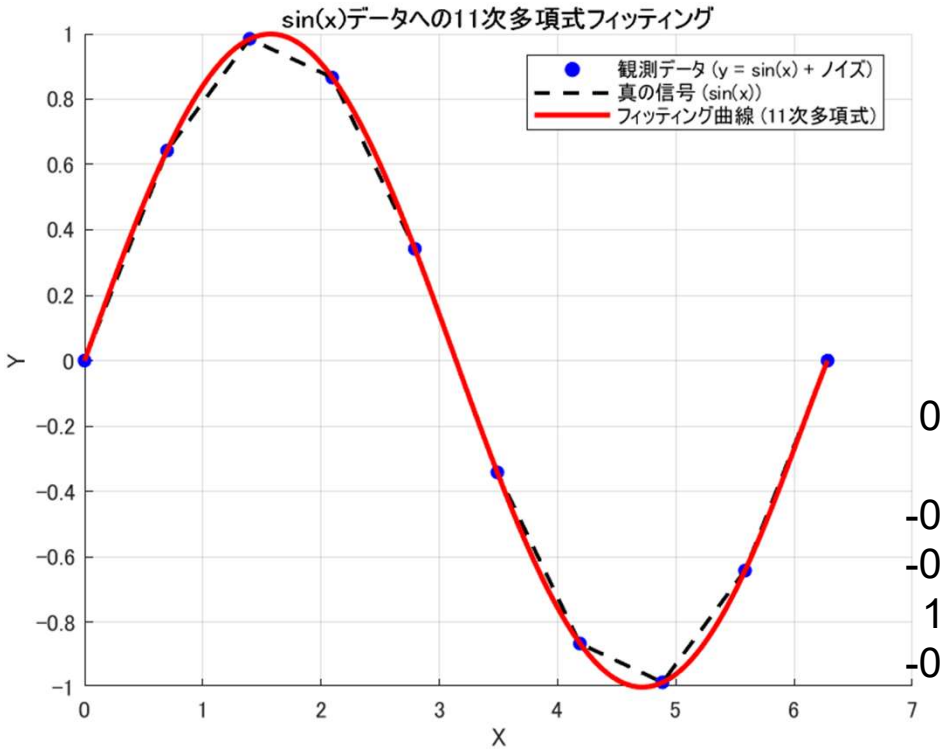
⇒線形計画問題に置き換えて解ける

© 2020 Yuki Kato. All Rights Reserved

例題：回帰問題のL1ノルム最小化

■ OMPと同様にスパース同定が可能

OMP



β

0

0

0

0

0

0.0067

0

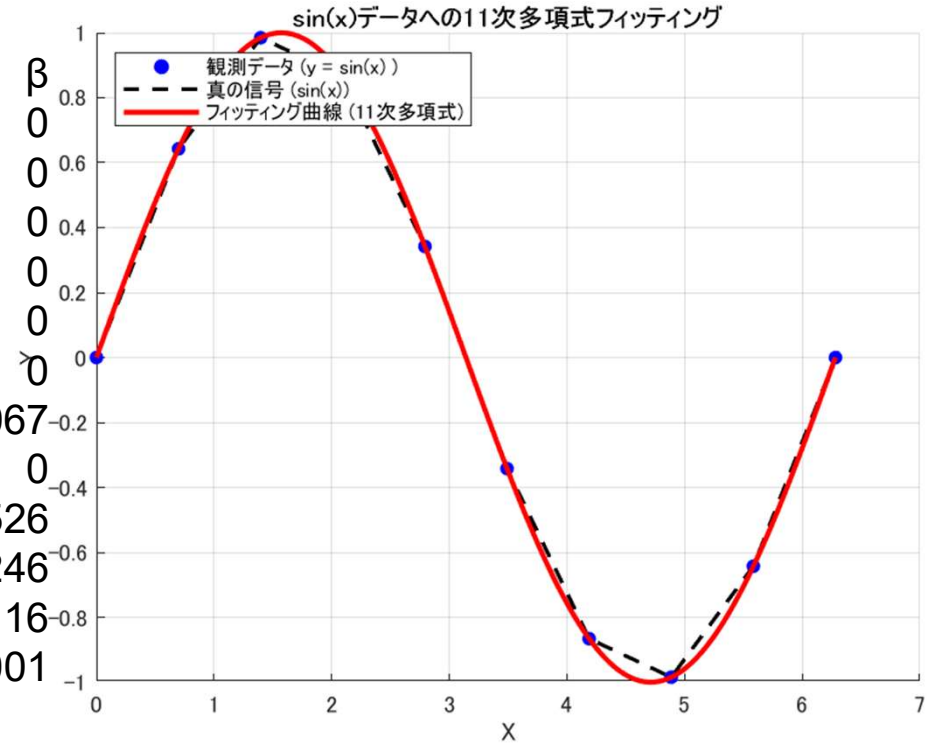
-0.1526

-0.0246

1.0116

-0.0001

L1 min



β

0

0

0

0

-0.00018

0

0.00830

0

-0.1665

-0.0001

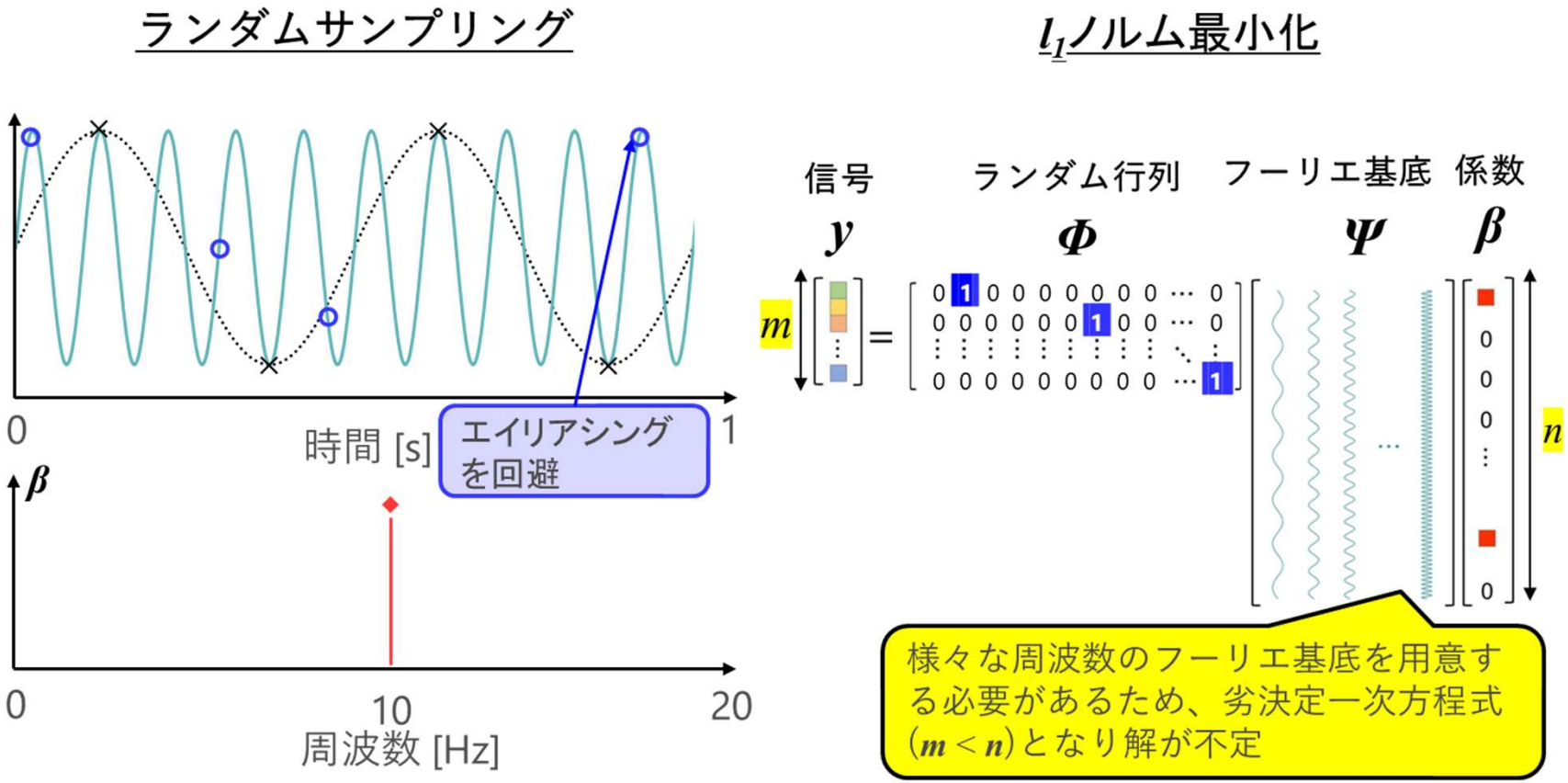
1.0000

0

© 2020 Yuki Kato. All Rights Reserved

補：圧縮センシングとは？

- スパース同定を計測に応用した手法
- 10 Hzの振動を4 Hzのサンプリングレートで計測すると，**標本化定理により振動の計測が不可能(×)**
- しかし，時間方向にランダムサンプリング(〇)し，スパース同定による信号復元を実施すれば計測できる

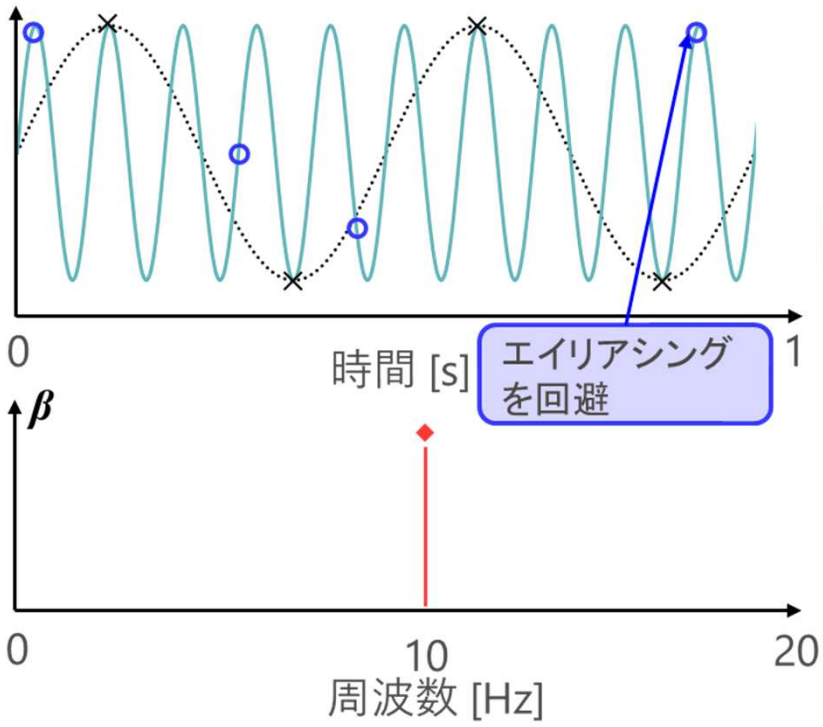


© 2020 Yuki Kato. All Rights Reserved

補：圧縮センシングとは？

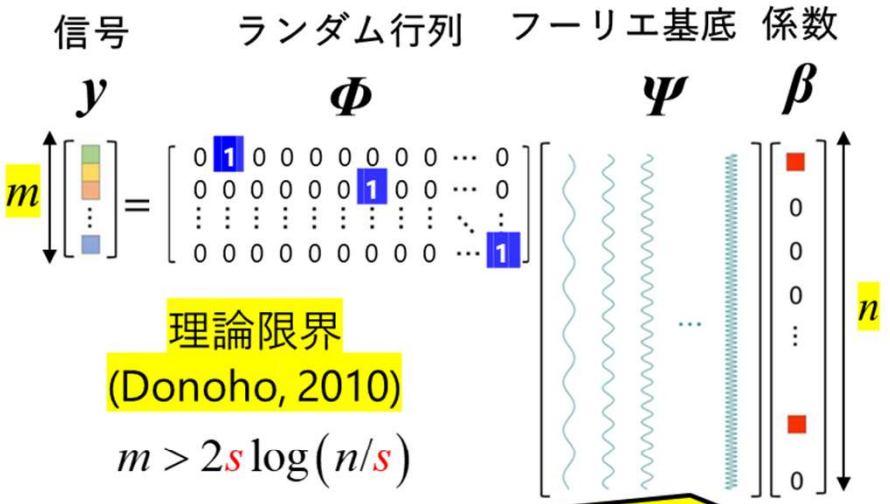
- スパース同定を計測に応用した手法
- 10 Hzの振動を2 Hzのサンプリングレートで計測すると，**標本化定理により振動の計測が不可能(×)**
- しかし，時間方向にランダムサンプリング(〇)し，スパース同定による信号復元を実施すれば計測できる

ランダムサンプリング



l_1ノルム最小化

$\Phi\Psi$がRIPを満たす必要有



理論限界
(Donoho, 2010)

$m > 2s\log(n/s)$

$m >> s$（スパース）なら解が定まる

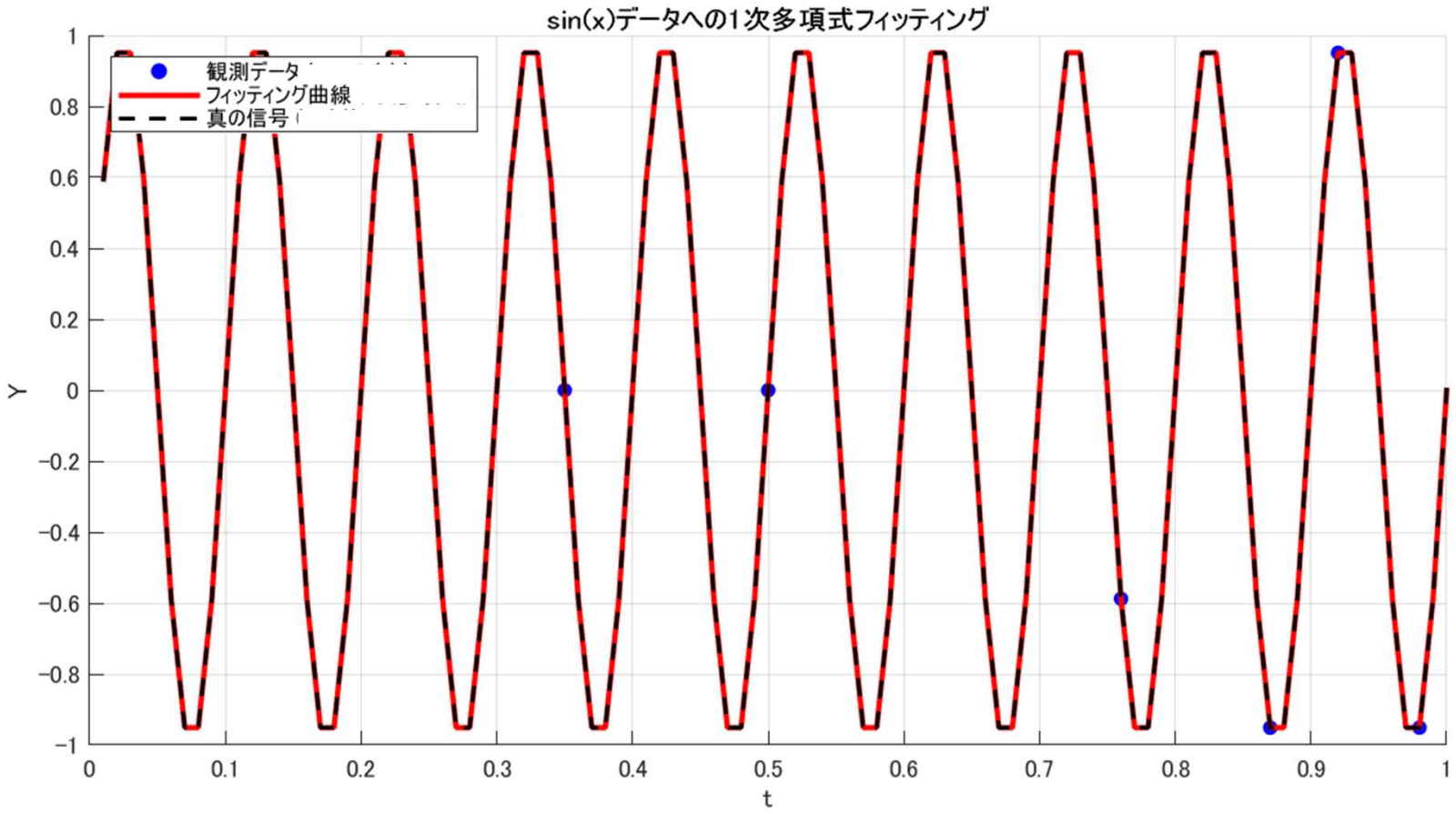
$\Rightarrow \hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$

によりスパースな解が求まる*

© 2020 Yuki Kato. All Rights Reserved
*D. L. Donoho and J. Tanner, in IEEE Transactions on Information Theory, vol. 56, no. 4, pp. 2002-2016, April 2010,

演習：基底を変更する-振動/波長計測

- 観測値を時系列方向にランダムサンプリングし，0-20 Hz の基底を用いて圧縮センシングにより信号復元せよ
- 計測時間: 10s, 計測点数(m): 40点,
- 基底は0.1 Hz刻み⇒ n = 20*2*10 = 400



© 2020 Yuki Kato. All Rights Reserved

演習：振動波形を変更する-振動計測

① 振動成分が1, 5.5, 10 Hzの場合

② 振動成分が1,2,3,4,5,6,7,8,9,10 Hzの場合

■ それぞれ計測可能か，求めよ

© 2020 Yuki Kato. All Rights Reserved

演習：振動波形を変更する-振動計測

① 振動成分が1, 5.5, 10 Hzの場合

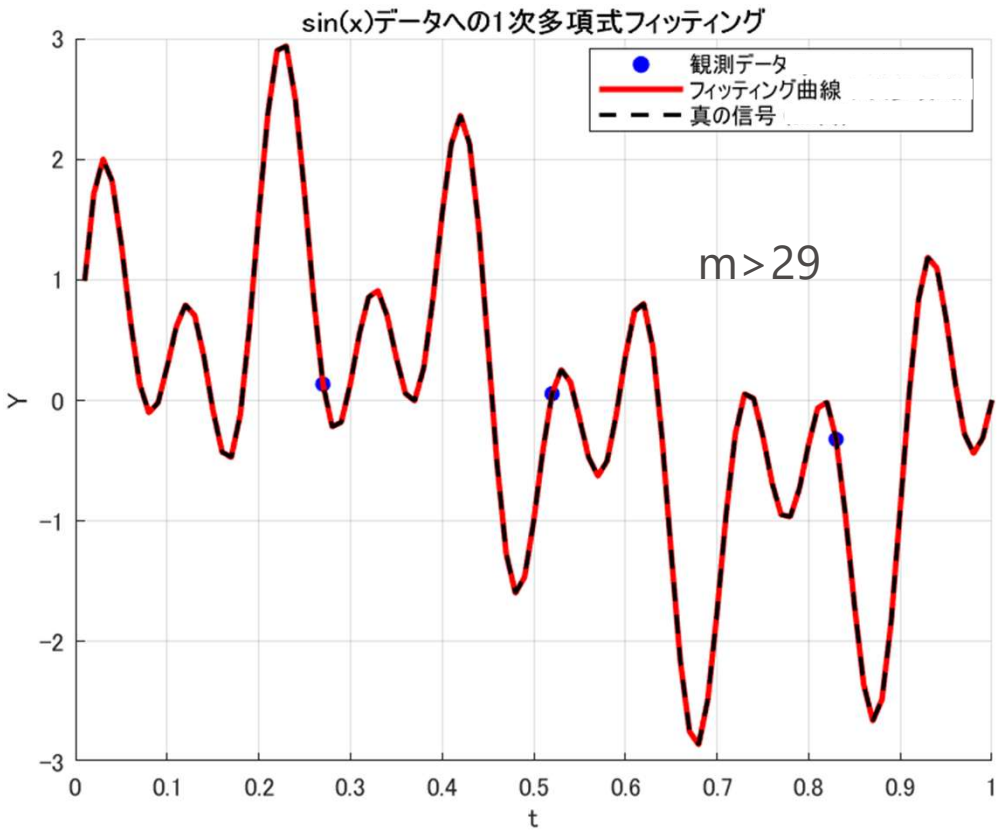
② 振動成分が12,3,4,5,6,7,8,9,10 Hzの場合

■ それぞれ計測可能か，求めよ

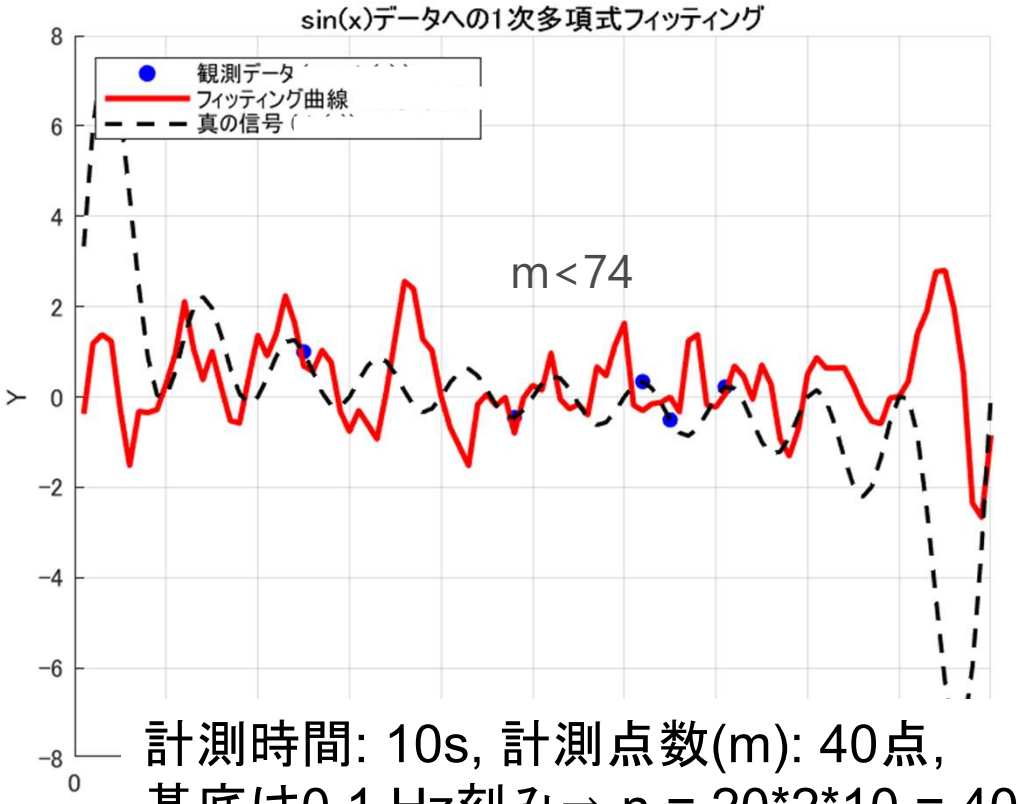
理論限界
(Donoho, 2010)

$$m > 2s\log(n/s)$$

①



②



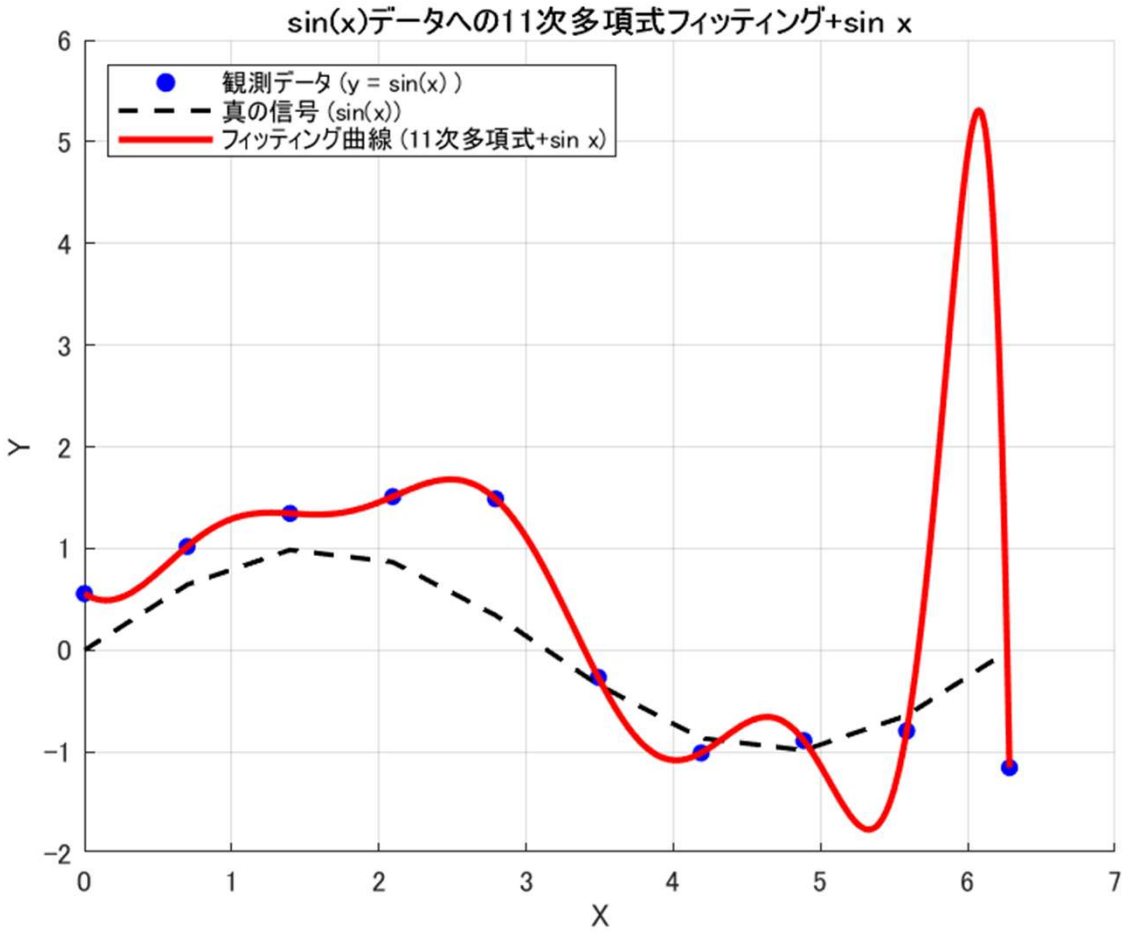
計測時間: 10s, 計測点数(m): 40点,
基底は0.1 Hz刻み⇒ n = 20*2*10 = 400

© 2020 Yuki Kato. All Rights Reserved

ノイズ/モデル化誤差があると・・・？

- 11次多項式+sin xの問題をL1ノルム最小化で解く
- 解が求まらない or 非スパース解が出る

⇒解を縛り過ぎると良くない⇒Lasso(正則化)を用いる



$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$$

コレ

β
0.101848634147003
-0.0003691312367236 56
0.009812229029434 31
-0.106242653250789
0.594496209217458
-1.76863401533222
2.38378964664618
0
-2.64466571351826
0
2.43089720616320
0
-0.057975447477347 3

$$\min_{\beta \in R^n} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

© 2020 Yuki Kato. All Rights Reserved

Lasso：Least Absolute Shrinkage and Selection Operator

■ 最小二乗法にL1ノルム最小化を組込んだ方法

最小二乗法 $\min_{\beta \in R^n} \frac{1}{2} \|y - X\beta\|_2^2$

L1ノルム最小化 $\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } y = X\beta$

Lasso

$$\min_{\beta \in R^n} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

■ モデルや計測の誤差を許容しつつ，スパースな解を求めることができる

■ こちらも凸緩和問題になるので，解が一意に定まり嬉しい

■ 解き方はADMMなど，詳しくは割愛

■ 正則化

- 過学習を防ぐために、モデルの複雑さに対するペナルティ（罰則項）を目的関数（損失関数）に追加すること

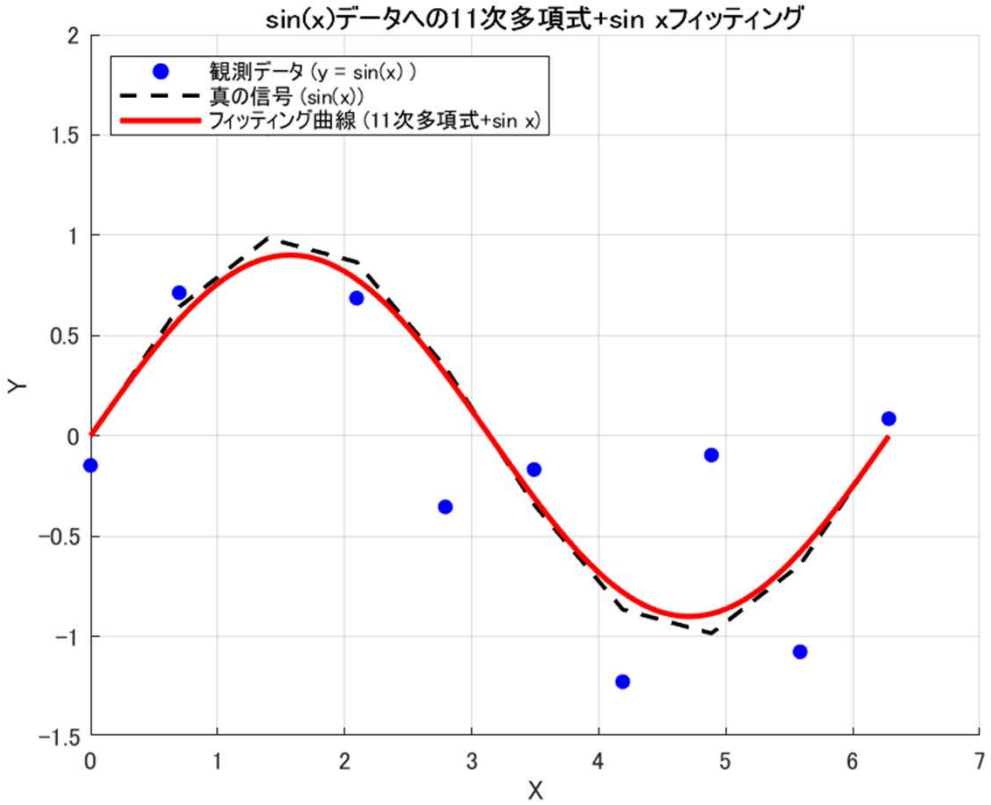
$$\text{目的関数} = \text{損失関数} + \lambda \times \text{正則化項}$$

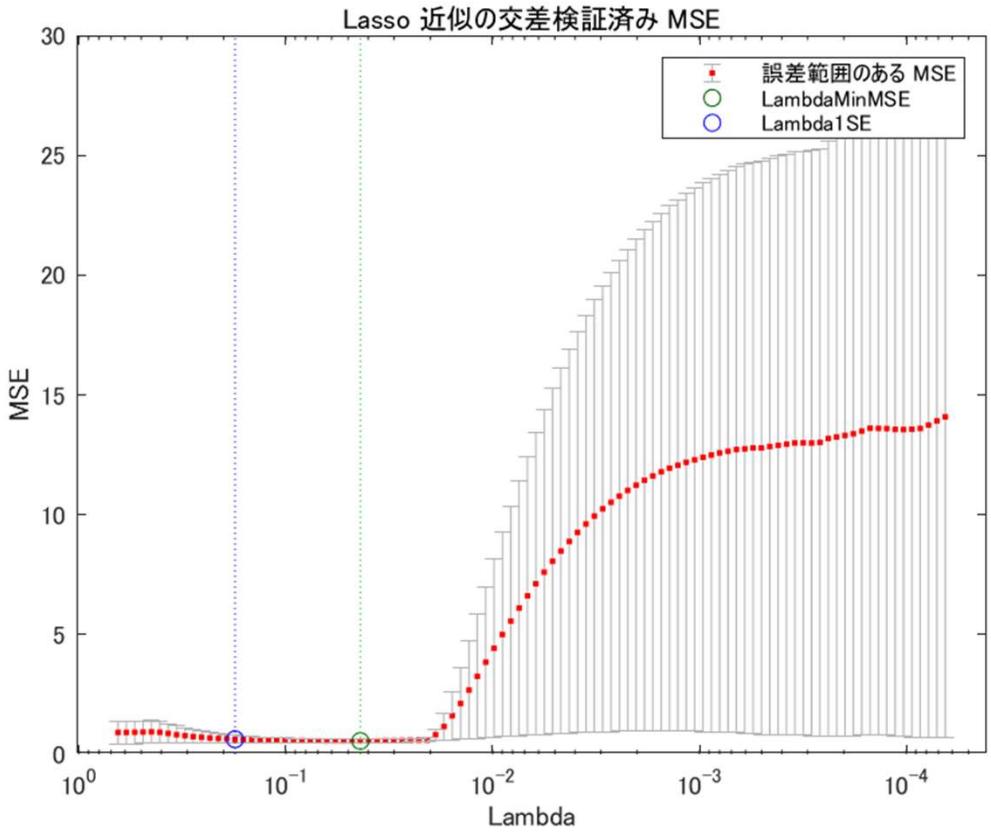
© 2020 Yuki Kato. All Rights Reserved

例題：Lassoによるスパースモデリング

- ノイズを含むデータでも，スパースな解が求まる
- λは公差検証により最適な値を求める必要がある

$$\min_{\beta \in R^n} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1$$





© 2020 Yuki Kato. All Rights Reserved

回帰問題まとめ

- 回帰問題を例に様々な解法を学んだ
  - 線形最小二乗法
  - 非線形最小二乗法
  - 最尤推定法
  - ベイズ推定
  - 変分ベイズ推定
  - スパース同定
    - L0ノルム最小化
    - L1ノルム最小化
    - Lasso
- 解く問題やモデルによって使い分けるのが重要

© 2020 Yuki Kato. All Rights Reserved